PF$^3$ fit: Hierarchical Flexible Fitting in 3D EM

Radhakrishna Bettadapura, Antje Vollrath and Chandrajit Bajaj

September 3, 2012

Abstract

We develop a hierarchical solution (PF$^3$ fit) for fitting an atomic structure to a density map. We use scoring functions that account for steric clashes while maximizing the degree of fit between the protein and the density map, a non-uniform rotationally exhaustive Fourier-based search scheme, and a flexibility model that parametrizes shear and hinge bending motions available to each protein domain. This framework rests on a hierarchical domain decomposition of the input protein, whereby domains at one level in the hierarchy are split into subdomains at the next level. We show that our flexible fitting technique can quickly and efficiently search over the space of motions available to the domains of a protein at each level in the hierarchy, and yields flexible fits that compare favorably with existing pre-MD as well as all-atomistic MD-based flexible fitting schemes. More importantly, our flexible fitting technique is capable of annotating the kind and degree of motion that each domain undergoes, making it a useful tool to analyze protein conformational change in cases where only a protein and density map are available.

Availability. PF$^3$ fit is available from our website:

1 Introduction

In flexible fitting, a protein crystal structure is deformed to maximize its correlation with a cryo- or tomo-EM density map. The resulting flexibly deformed crystal structure combines the high-resolution information in the protein with the functional state information in the density map, providing biologists with an all-atomistic model of the in vitro state of the protein. Approaches to the flexible fitting problem begin by defining a scoring function that reflects the goodness of fit between the protein and the density map, and a flexibility model that parametrizes the space of motions available to the protein. With an appropriate search scheme, the scoring function is optimized over the space of motions of the protein, thus flexibly fitting it to the density map.

At the minimum, a flexible fitting scoring function $S$ measures the fit between the protein $\mathcal{P}$ and the density map $\mathcal{M}$. Additionally, $S$ can account for the energetic interactions between the atoms of $\mathcal{P}$ as it is deformed. Scoring functions that use the former approach predominantly rely on the correlation between a Gaussian-blurred version of $\mathcal{P}$ and a suitably preprocessed version of $\mathcal{M}$ [47, 39]. The addition of an energetics-based term to the scoring function has lately also become computationally feasible; this term can either be an MD-based force-field, reflecting the molecular mechanical energy of $\mathcal{P}$ [49, 48, 32, 19], or an elastic- or Gaussian-network-based approximation [45, 55, 44].

The protein flexibility model is the most important aspect of a flexible fitting approach. A flexibility model must be reasonably accurate, i.e., it must not result in unrealistic deformations of $\mathcal{P}$. It must also be reasonably efficient, reducing the degrees of freedom available to $\mathcal{P}$ so that the search can be conducted relatively quickly. Accuracy can be enforced with an MD-based, all-atomistic approach, whereby the deformation of $\mathcal{P}$ at each time
Figure 1: Fitting 4AKEa into a 10 Å simulated map of 1AKEa at multiple hierarchical levels. (A) Initial rigid-body fit. Domains detected at level 1 by HDD are colored red and green respectively. The initial Cα RMSD is 7.8 Å, and the initial correlation with respect to the Gauss-CCS is 0.49. Hinge-bending motions are assigned between the domains; the primary hinge axis points out of the plane of the page, and the secondary hinge axis points right to left. (B) Conformation obtained after search over hinge bending motions at hierarchical level 1. Rotations about the primary and secondary axes are respectively 68 and 30 °. The Cα RMSD is 4.3 Å. (C) Second hierarchical level, side view. The yellow domain, which protrudes out of the density map, is assigned hinge bending motions. (D) Conformation obtained after search over hinge bending motions at level 2. Rotations about the primary and secondary axes are respectively 12 and 20 °. The Cα RMSD is 3.3 Å. (E) Final MD-based flexible fit. Final Cα RMSD is 0.6 Å.

step is carried out by exerting minute forces on each atom, maneuvering it into the density map while ensuring that it is not driven too far from equilibrium [49, 32]. On the other hand, if efficiency is a concern, P can be partitioned into rigid domains, each of which is then fit independently into M. Domain-based approaches can choose to enforce inter-domain constraints during the search process [48, 19, 33] or afterwards [51]. Flexible search schemes are either local, global, or a combination of the two. Local search relies on the assumption, often retrospectively justified, that the conformation of P that fits optimally into M is close to an initial guess. All-atomistic MD-based flexible fitting [49, 32], while in theory capable of capturing global conformational change, can be practically used only for local searches. Exhaustive global searches, on the other hand, can only be performed with an initial domain decomposition of the protein. Often this decomposition consists of only a few domains, and either a generic search scheme such as gradient descent, conjugate gradients, Monte-Carlo [47, 48], simulated annealing [46], or a targeted MD simulation [49, 19, 33] is used.

In addition to the scoring, flexibility parametrization, and search problems, all flexible fitting approaches require an initial guess of the fit of P into M, i.e., they require an initial rigid-body fit. Popular exhaustive [52, 16] or local [34] rigid-body fitting implementations can be used to provide this fit. Exhaustive algorithms are typically Fourier-based, taking advantage of translational or rotational FFTs to compute a set of correlations over the space of rigid-body transformations of P. Local search algorithms are based on variants of gradient descent, and refine an initial placement of the protein into the density map.

This work on flexible fitting is motivated by two observations. The first of these relates to the scoring and search phases. Can a domain-based flexible fitting search scheme aspire to conduct flexible correlations on a chosen sample set of motions, while simultaneously ensuring that inter-domain energetic interactions are favorable according to a chosen score? At first glance, this question seems to be answered by the MD and GNM-
based search schemes discussed above; however, without exception, all such schemes are iterative and have the potential to oscillate about local extrema. One way to address the performance issues is through domain-based MD, which however retains the limitation common to all iterative search schemes.

The second observation motivating this work is biological. It has long been known [18] that a large variety of protein domains undergo two complementary kinds of motion: shearing across a plane, and hinge bending about a single axis. Whereas a variety of approaches have sought to characterize hinges in proteins, starting either from two conformations [53, 41] or a single one [13, 23], shearing motion, either by virtue of its complex nature, or its relatively low effect on overall conformational change, remains not as well understood. A recent paper [6] outlines the major factors governing shearing motion—the existence of a continuously-maintained interface; the presence of packing constraints—but concludes that it is still difficult to predict whether, and to what degree, a given protein in a single conformation undergoes shear. Current MD- and GNM-based flexible fitting approaches are capable of accounting for shearing motions in proteins, but at the expense of significantly altering the conformations of the domains participating in shear; these alterations make it difficult to isolate that portion of domain motion explained wholly by shearing across a plane.

![Figure 2: Control flow of the flexible fitting algorithm developed in this work.](image-url)
Our hierarchical polar Fourier flexible fitting (PF^3 fit) approach takes steps towards addressing the above drawbacks. It combines a backbone flexibility model with a non-uniform rotational Fourier-based search scheme [3, 5]. The flexibility model assumes that backbone flexibility in proteins is governed by shear and hinge bending motions, that these motions are mutually exclusive, and that the ranges of motions in shear are small relative to those of hinge bending; using these assumptions, we explicitly parametrize shear and hinge bending motions to discover if an improvement in the flexible fit of the protein with the density map can be found. The rotational Fourier-based search scheme is capable of (Bajaj et. al [3]) searching over arbitrary subsets of the space $SE(3)$ of rigid-body motions; we use this property to flexibly fit each domain of $\mathcal{P}$ under its range of motion.

PF^3 fit requires a hierarchical domain decomposition of the input protein. Domains can be detected from crystal structures using techniques based on elastic-network [21] or Gaussian network [24] normal mode analysis, graph-theory and the pebble game [22], dynamic programming [1], or a combination of two or more of the above methods [23]. Most of these techniques include a sensitivity parameter that controls the size of the detected domains; for instance, the works in Abyzov et. al [1] and Jacobs et. al [22] respectively define a distance difference cutoff and a flexibility threshold. This parameter can be varied to obtain a composite picture of the motion of the protein at several different motion scales. In this work, we present HDD, an alternative to existing domain decomposition techniques that uses an analogous sensitivity parameter to come up with not just a single domain decomposition but a hierarchy of decompositions, each progressively reflecting smaller scales of motion. HDD detects domains from an initial rigid-body fit into the density map. We emphasize that PF^3 fit is largely independent of HDD, and is general enough to be used with most other domain decomposition techniques.

Our primary goal in this work is to show that, given an input crystal structure $\mathcal{P}$ and density map $\mathcal{M}$, our methods can be used to infer, via flexible fitting to $\mathcal{M}$, the degree of shear and hinge bending that each domain in $\mathcal{P}$ undergoes. Our secondary goal is to show that our scoring functions, accounting as they do for steric clashes, provide a basic pairwise energetic model that resolves a large percentage of stereochemical errors, mitigating the need for a final energy minimization. At the very least, PF^3 fit can be used as a plausible alternative to existing pre-MD or pre-ENM search phases that involve finding the best fit of each domain in $\mathcal{P}$ (See, for instance, [48]). In certain cases, we also show that PF^3 fit performs nearly as well as rival all-atomistic techniques while surveying a smaller but more meaningful search space.

2 Overview of methods

Our flexible fitting algorithm (PF^3 fit) makes use of several well-known properties of proteins, and comprises five stages. See also Figure 2 for an overview of how each of these stages fit in the overall framework.

We establish some notation. Let the inputs to the fitting algorithm be a crystal structure and a density map. Throughout, $\mathcal{P}$ refers to the set of atoms that belong to the crystal structure, while $\mathcal{M}: \mathbb{R}^3 \rightarrow \mathbb{R}$ is the scalar-valued function corresponding to the density map. Further, $\mathcal{M}_\mathcal{P}: \mathbb{R}^3 \rightarrow \mathbb{R}$ is a scalar-valued representation of $\mathcal{P}$.

1. **Rigid-body fitting.** The first stage is rigid-body fitting [3], where a rigid-body score $S(\mathcal{P}, \mathcal{M})$ is maximized under rigid-body transformations of $\mathcal{P}$ (Section 3).

2. **Flexibility model and search.** Our protein flexibility model/search scheme consists of the following broad steps (Section 4).
   
   • **Hierarchical Domain Decomposition.** We introduce HDD (Section 5), a technique to decompose $\mathcal{P}$ into a hierarchy $\mathcal{HD} = \{\mathcal{DD}_1, \ldots, \mathcal{DD}_n\}$ of domain decompositions $\mathcal{DD}_i$ based on the geometric properties of $\mathcal{P}$ as well as a score $S(\mathcal{P}, \mathcal{M})$. 

<table>
<thead>
<tr>
<th>Name</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA Gauss C</td>
<td>Gaussian represents $P$</td>
</tr>
<tr>
<td>SC Scattering potential C</td>
<td>Scattering potential represents $P$ [5]</td>
</tr>
<tr>
<td>NUExtnon-uniform exterior penalty C</td>
<td>Non-uniform-grid based version of ETR [5]</td>
</tr>
<tr>
<td>PP Pocket penalty C</td>
<td>Penalizes pocket-target overlaps [5]</td>
</tr>
<tr>
<td>ETR External-total ratio</td>
<td>Num. atoms outside chosen isocontour of $M$ [34]</td>
</tr>
<tr>
<td>MIS Mutual information</td>
<td>Information shared by $P$ and $M$ [40, 50]</td>
</tr>
<tr>
<td>SSS Skeleton-secondary structure</td>
<td>Correlates skeletal features on $M$ with secondary structures on $P$ [5]</td>
</tr>
</tbody>
</table>

Table 1: Scoring functions implemented in PF³ fit. The boldface letter C signifies that the scoring function is an FFT-amenable correlation, and the trailing letter ‘S’ on the MIS and the SSS stand for ‘score’. We do not include a citation for the GA score as its use is widespread. The SC, NUExt, PP, and SS scores have been described in a recent paper on scoring functions and rigid-body fitting by the authors.

- **Motion assignment and sampling.** At each stage in the hierarchy, we make use of a pairwise shear/hinge bending based flexibility model to assign motions to domains, generating motion samples of each domain on the basis of these assignments (Section 5.2).
- **Flexible search.** We maximize a correlation-amenable score for each pair of domains at each step in the hierarchy, thereby bringing about flexible fitting (Section 7).
- **Biased molecular dynamics.** As a final, optional step, we use a biased molecular dynamics step that includes a term $S(P, M)$ that accounts for the score between $P$ and $M$. We emphasize that the MD-based step is not the thrust of this work, and, as we shall show, it plays a relatively minor role in improving the flexible fit.

## 3 Scoring

PF³ fit uses two different kinds of scoring functions to effect a flexible fit of the crystal structure into the density map. The first kind are the fitting functions, which measure the goodness of fit of the atoms of $P$ in $M$; we denote each of these scores $S_{fit}^{score}$, where the superscript is replaced by the name of the measure used. The second kind are interaction functions $S_{inter}^{score}$, which measure the degree to which a domain of $P$ intrudes into another.

### 3.1 Scoring functions for fitting

In Bettadapura et. al [5], we introduce several new scoring functions for density map fitting; we describe these as well as some other well-known scoring functions, all of which are available in PF³ fit, briefly below. See also Table 1.

**FFT-amenable scoring functions.** All FFT-amenable scoring functions take on the form

$$CCS(A, B) = \int_{\mathbb{R}^3} A(Rx + t)B(x)dx,$$

(1)

where $A$ and $B$ are scalar-valued functions derived from $P$ and $M$ respectively, and $(R, t) \in SE(3)$ is a rigid-body transformation. For the functions we introduce in Bettadapura et. al [5], $A$ are $B$ comprise two
components, the target score $\text{target}$, derived from the volumes occupied by $\mathcal{P}$ and $\mathcal{M}$, and the complementary score $\text{comp}$, derived from the volumes complementary to $\mathcal{P}$ and $\mathcal{M}$ respectively. The target score is given by one of the Gaussian, scattering potential, or non-uniform exterior penalty scores; the complementary score is given by the pocket penalty.

The target scores.

1. **Gaussian.** In the Gaussian score,
   \[
   A_{\text{target}}(x) := A_{gc}(x) = \sum_{i \in \text{atoms of } \mathcal{P}} G^i(x),
   \]
   where
   \[
   G^i(x) = \exp\left(-\beta \|x - x_i\|^2\right) = \exp\left(-\frac{\pi^2 \ln 2}{R^2} \|x - x_i\|^2\right)
   \]
   is the Gaussian centered at atom center $i$. $B_{\text{target}}$ is just $\mathcal{M}$ or a suitably filtered version of it.

2. **Scattering potential.** In the scattering potential score,
   \[
   A_{\text{target}}(x) := A_{sc}(x) = \sum_{i \in \text{atoms of } \mathcal{P}} V^i_{sc}(x)
   \]
   where
   \[
   V^i_{sc}(x) = \frac{16\pi^2 \hbar^2}{m_0 e^2} \sum_{j=1}^{5} \frac{a_j}{b_j^2} \exp\left(-\frac{4\pi^2 \|x - x_i\|^2}{b_j + R^2}\right)
   \]
   is the scattering potential due to atom $i$. $B_{\text{target}}$ is just $\mathcal{M}$ or a suitably filtered version of it.

3. **Non-uniform exterior penalty.** Let $\mathcal{P}_s$ be a chosen subset of atoms of $\mathcal{P}$, and let $X_s$ be the union of spheres of the atoms of $\mathcal{P}_s$. Then, in the non-uniform exterior penalty,
   \[
   A_{\text{target}}(x) := A_{nu}(x) = \begin{cases} 1, & x \in X_s \\ 0, & \text{otherwise.} \end{cases} \]
   Similarly, let $m \in \mathbb{R}$ be a chosen scalar intensity value. Then
   \[
   B_{\text{target}}(x) := B_{nu}(x) = \begin{cases} 1, & \mathcal{M}(x) \geq m \\ 0, & \text{otherwise.} \end{cases} \]

4. **The complementary score. Pocket Penalty.** Let $V_\mathcal{P}$ be the volume occupied by the Gaussian molecular surface of $\mathcal{P}$, and $V_\mathcal{M} \subset \mathbb{R}^3$ is the volume occupied by a suitably chosen isosurface of $\mathcal{M}$. The complementary volume $\overline{V}_\mathcal{P} \subset \mathbb{R}^3 = \text{conv}(V_\mathcal{P}) - V_\mathcal{P}$, and similarly for $\overline{V}_\mathcal{M}$. Smooth, scalar-valued representations of $\overline{V}_\mathcal{P}$ and $\overline{V}_\mathcal{M}$ can be extracted from, respectively, the molecular surface of $\mathcal{P}$ and a suitable isosurface of $\mathcal{M}$ [54]. Then, in the pocket penalty,
   \[
   A_{\text{comp}}(x) := \begin{cases} 1 - \sqrt{-1}, & x \in \overline{V}_\mathcal{P} \\ 0, & \text{otherwise.} \end{cases}
   \]
\[ B_{\text{comp}}(x) := -A_{\text{comp}}(x) \quad (9) \]

**Non-FFT-amenable scoring functions.** We also use a number of scoring functions whose general form does not make them amenable to a Fourier-based speedup.

5. **External-total Ratio (ETR).** The ETR is a ratio of the total number of atoms of \( P \) outside a chosen isocontour of \( M \) to the total number of atoms in \( P \). Lower ETR scores correspond to better fits.

6. **Mutual Information Score (MIS).** For a pair of scalar-valued functions \( A \) and \( B \) representing \( P \) and \( M \), the MIS is given by

\[ MIS = \sum_{x \in B} \sum_{y \in A} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right), \quad (10) \]

where \( p(x) \) and \( p(y) \) are the percentage of voxels in \( B \) and \( A \) that take on intensities equal to \( x \) and \( y \) respectively and \( p(x, y) \) is the percentage of voxels in \( B \) with intensity \( x \) that are aligned with voxels in \( A \) with intensity \( y \). Since the range of intensities in a density map is typically very high, a binning scheme with about 20 bins is used. Higher mutual scores correspond to better fits. The MIS measures the amount of information that one random variable contains about another [40]. Additionally, it is seen, in Vasishtan and Topf [50], to possess two features that make it attractive as an alternative scoring function for cryoEM fitting: a) It correlates with the Gaussian-based cross-correlation score in Equation 1 and b) correlates well with the RMSD in experiments involving synthetic density maps. It is thus an independent measure of the goodness of fit.

7. **Skeleton-secondary structure score (SSS).** Let \( H_M \) and \( H_P \) respectively be the set of helices detected from \( M \) and \( P \). Each helix consists of an axis \( r \), with \( \|r\|_{\ell_2} = 1 \), and a mid point \( p \). Let \( h_i^M \) be a helix in \( H_M \), and let \( h_j^P \) be a helix in \( H_P \). Let \( d(., .) \) be the Euclidean distance function, \( \langle ., . \rangle \) be the dot product, and \( w_1 \in \mathbb{R}^-, w_2 \in \mathbb{R}^+ \) be respectively negative and positive weights. Then the per-helix score and the secondary structural score are respectively given by Equations 11 and 12 below.

\[ SSS_{h_j^P} = \max_i w_1 d(p_i^M, p_j^P) + w_2 |\langle r_i^M, r_j^P \rangle| \quad (11) \]

\[ SSS = \sum_j SSS_{h_j^P} \quad (12) \]

where \( w_1 = -1, w_2 = 1 \). The theoretical range of the per-helix score \( SSS_{h_j^P} \) is \((-\infty, 1]\).

### 3.2 Scoring functions for steric clashes

To measure the intrusion of the atoms of a moving domain of \( P \) with respect to a stationary domain, we use the following non-uniform-grid-based steric clash function. Let \( P_s \subset P \) be a domain. Define the \( i^{th} \) layer \( L_i, \forall i \geq 0 \) of \( P_s \) as the set of solvent-exposed atoms obtained when the \( i-1^{th} \) layer is removed from \( P_s \). Let the scalar-valued function

\[ A^s_{\text{steric}}(x) := \begin{cases} (1.1)^i \sqrt{1 - 1}, & x \in L_i, \forall i \geq 0 \vspace{1mm} \\ 0, & \text{otherwise}. \end{cases} \quad (13) \]

represent \( P_s \). Then maximizing the real part of Equation 1, where \( A = A^\text{moving}_{\text{steric}} \) represents the moving domain \( P^\text{moving} \), and \( B = A^\text{stat}_{\text{steric}} \) represents the stationary domain \( P^\text{stat} \), penalizes steric clashes between \( P^\text{moving} \) and \( P^\text{stat} \). In particular, the maximum value of the following interaction score
\[ S_{\text{inter}} = CCS(\text{moving}_{\text{steric}}, \text{stat}_{\text{steric}}), \]  

is 0, while the minimum value is proportional to \(-M \times N\), where \(M\) and \(N\) are the number of atoms in the stationary and moving domains.

### 3.3 Total score

The total scoring function is a weighted sum of each of the scoring functions

\[ S_{\text{tot}}(A, B) = w_{\text{fit}}S_{\text{fit}}(A, B) + w_{\text{inter}}S_{\text{inter}}(A, B), \]  

where \(S_{\text{fit}}\) is one of the scoring functions for fitting in Section 3.1, and \(S_{\text{inter}}\) is the clash score in Equation 15. Maximizing the real part of this scoring function yields an optimal fit.

### 3.4 Choice of scoring functions

In Bettadapura et. al [5], we provide compelling evidence that the scattering potential results in better quality fits than the Gaussian. We thus use the scattering potential score for real density map fitting, and, to provide a fair comparison with other works, we use the Gaussian blur score for synthetic density map fitting. Further, we set \(w_{\text{fit}} = 1\), while we discuss the appropriate value of \(w_{\text{inter}}\) in Section 9.2.

### 4 Flexibility model and search

We introduce a hierarchical domain-based flexibility model that is based on assigning motions to domains of \(\mathcal{P}\). Let a domain \(D_i\) be a set of secondary structural elements together with a set of connectors that lead from \(D_i\) to any adjacent domain \(D_j\); the set of connectors may have one or more members. Let a domain decomposition \(\mathcal{D} = \{D_1, \ldots, D_n\}\) of \(\mathcal{P}\) be a collection of domains. We require that domains \(D_i\) and \(D_j\) for any \(i, j\) be disjoint, and that \(\bigcup_j D_j = \mathcal{P}\). A hierarchical domain decomposition \(\mathcal{H} = \mathcal{D} \ldots \mathcal{D}_{n-1}\) is a collection of domain decompositions such that each domain in \(\mathcal{D}_i\) belongs either wholly or partially to some domain in \(\mathcal{D}_{i-1}\).

There are three broad steps comprising the flexibility model and search scheme in PF3.fit.

1. **Hierarchical domain decomposition and pairwise motion assignment.** The input is the protein \(\mathcal{P}\) and the density map \(\mathcal{M}\), and the output is a hierarchical domain decomposition \(\mathcal{H}\) along with pairwise domain motion assignments at each hierarchical level (Section 5).

2. **Flexibility tree.** The input is the hierarchical domain decomposition \(\mathcal{H}\) and the density map \(\mathcal{M}\), and the output is the flexibility tree \(\mathcal{F}\) (Section 6).

3. **Flexibility tree traversal/search.** The input is the flexibility tree \(\mathcal{F}\) and the density map \(\mathcal{M}\), and the output is the protein \(\mathcal{P}\) flexibly fit to \(\mathcal{M}\) (Section 7).

There are also a set of parameters implicit to each step that are controllable by the user. We discuss these parameters, as well as each of the above steps in turn, in the sections to follow.
Figure 3: Schematic of hierarchical motion domain decomposition and assignment at the secondary structural level. (A) Secondary structural level. (B) Motion graph, secondary structural level. The dot-dash line connects a pair of domains undergoing relative motion. The angular and rectangular glyphs denote hinge bending and shear respectively.

Figure 4: Schematic of hierarchical motion domain decomposition and assignment at the penultimate level \( i = 2 \). (A) Penultimate level. Solid lines indicate containment; for instance, \( S_1 \) is contained in \( D_{20} \). Dotted lines indicate partial containment; \( S_2 \) is distributed between the disjoint domains \( D_{21} \) and \( D_{2m-1} \). In practice, partial containment is rarely encountered. (B) Penultimate level, motion graph. See also Figure 3.

5 Hierarchical domain decomposition and pairwise motion assignment

Hierarchical Domain Decomposition (HDD) detects domains in a crystal structure based on an initial rigid-body fit into the density map. The process of domain detection involves three stages. See also Figures 3, 4, and 6.

1. **Detection of secondary structural elements (SSEs).** Secondary structural elements \( SS = \{ S_j \} \) are assigned to \( P \) using the publicly available software Stride [15]. The set \( SS \) is allowed to contain helices (either \( \alpha \) or \( 3_{10} \)) and \( \beta \) sheets only, i.e., we cluster adjacent \( \beta \) strands into a single \( \beta \) sheet.
Let $S_i$ be associated with layer $L_j$ (See Section 3.2) if one of its backbone atoms belongs to $L_j$. Denote $L_i$ as the set of layers with which the SSE $S_i$ is associated. An SSE $S_i$ is uni-layer if its associated $L_i$ has size equal to one; otherwise, it is multi-layer.

2. **Segmental SSE Subdivision.** Long/broad SSEs $S_j \in SS$ are replaced by $m$ segmental SSEs $S_{j1}, S_{j2} \ldots S_{jm}$ using a layered subdivision technique (Section 5.1).

3. **Motion assignment.** Motions are assigned between each segmental SSE as well as normal SSEs using the pairwise motion assignment procedure, resulting in a motion graph that corresponds to $SS$. The process of motion assignment is shared by SSEs as well as domains (Section 5.2).

4. **Hierarchical motion graph creation.** With a cluster criterion, the elements of $SS$ are clustered into a set $DD$ of domains $\{D_i\}$. Between each pair of domains, a motion is assigned using the pairwise domain motion assignment procedure. Varying the cluster criterion then results in a set $HD$ of domain decompositions $\{DD_j\}$, i.e., results in a hierarchical domain decomposition (Section 5.3). The user has control over this stage of the algorithm, i.e., the user is able to choose the discrete cluster criteria at which he desires a domain decomposition. See Section 9.1.

We discuss items 2, 3 and 4 in further detail below.

5.1 **Segmental SSE subdivision**

The input to this stage of the algorithm is a helix length threshold $l_H$ as well as a beta sheet length $l_S$ and width $w_S$ threshold, each expressed in Å. If either a helix or a beta sheet $S_j$ belonging to $SS$ is a multi-layer SSE, it is subdivided into $|L_j|$ single-layer SSEs, after which the following single-layer subdivision is carried out.

2.1 **Helix subdivision.** Let the length of a helix $H_j \in SS$ be $l_j$. If $l_j > l_H$, $H_j$ is replaced by $m$ segmental helices $H_{j1} \ldots H_{jm}$ such that the length of the first $m - 1$ helices is $l_H$ and the sum of the lengths of the $m$ helices equals the length of the original helix.

2.2 **$\beta$-sheet subdivision.** Assume that a $\beta$ sheet $S_j \in SS$ is an open surface with four extreme points. Let the length and width of the smallest rectangle enclosing these four points be $l_j$ and $w_j$ respectively. Then, in a procedure identical to helix subdivision, $S_j$ is replaced by $m$ segmental sheets $S_{j1} \ldots S_{jm}$ in the length (respectively width) direction if $l_j$ (respectively $w_j$) is greater than $l_S$ (respectively $w_S$).

5.2 **Motion assignment to domains**

To assign motion to domains/SSEs, we begin with the findings in Gerstein et al. [18] and Gerstein and Krebs [17], in which the authors conclude that, for a vast spectrum of proteins, conformational change is brought about by two kinds of domain motion: shear and hinge bending [17].

Shear (Figure 5) is lateral motion between a pair of domain interfaces. Shear-based conformational change occurs when several domains simultaneously undergo small ranges ($< 2\AA$) of lateral motion relative to each other, and is indicated whenever there is a continuously maintained inter-domain interface comprising buried side-chain atoms. Hydrogen bonding and electrostatic interactions at the interface cause incremental changes to several backbone torsional angles, leading to relative motion along the plane of shear.

Hinge bending, the counterpart of shear, occurs when a few domains undergo large ranges of motion. It occurs when there are few or no packing constraints, and when inter-domain distances are typically higher. The mechanism underlying hinge bending is similar to that of shear, i.e., the variation of backbone torsion angles. However, it has been empirically observed that fewer torsional angles participate in hinge bending motions, and
that these torsional angles vary by large amounts. Hinge bending also results in the variation of side chain angles by small amounts.

We describe shear and hinge bending by a pairwise parametrization of motions. Motion assignment consists of two stages. The first, or detection stage, takes as input a pair of domains and a desired motion, and produces as output a True/False value that indicates whether or not the desired motion occurs between the input domains. If the first stage returns True, the second, or parametrization stage, takes the pair of domains, the desired motion, and a sampling parameter as input, and returns as output a set of samples of the second domain relative to the first under the desired motion. Both the first and second stages are specific to the kind of motion. We set them down in Algorithms ShearDetect, ShearParam, HingeDetect, and HingeParam.

Let \( D_1 \) and \( D_2 \) be a pair of domains at any level in the hierarchy, and let \( MS_1 \) and \( MS_2 \) be triangulations representing the molecular surfaces of \( D_1 \) and \( D_2 \) respectively. Let an atom on \( D_1 \) (resp. \( D_2 \)) be an interface atom of \( D_1 \) if it is within a certain distance, termed the interface width, from any atom in \( D_2 \). Further, let the interface area between \( D_1 \) and \( D_2 \) be the sum of the areas of the triangles in \( MS_1 \) and \( MS_2 \) whose vertices are separated by at most the interface width (We discuss the choice of the interface width in Section 9.1).

**Algorithm ShearDetect.** We detect shear between \( D_1 \) and \( D_2 \) if one of the interface criteria 1 and 2 is met.

1. **Interface size.** If the number of interface atoms shared between \( D_1 \) and \( D_2 \) is greater than a certain percentage, termed the threshold interface ratio, of the total number of side-chain atoms on both \( D_1 \) and \( D_2 \). (See )

2. **Interface area.** If the area of the interface between \( D_1 \) and \( D_2 \) is greater than a certain percentage, termed the threshold interface area ratio, of the total surface area of the molecular surfaces defined by \( D_1 \) and \( D_2 \).

Additionally, if the domains \( D_1 \) and \( D_2 \) are secondary structures, we also require that they satisfy the following packing criterion.

1. **Packing criterion.** If \( D_1 \) (respectively \( D_2 \)) shares an interface with a domain other than \( D_2 \) (respectively \( D_1 \)), i.e., the number of interface atoms between \( D_1 \) (respectively \( D_2 \)) and any \( D_i \) for \( i \) not equal to 1 (respectively 2) is greater than the threshold interface ratio.

**Algorithm ShearParam.** We describe a shearing motion of \( D_2 \) relative to \( D_1 \) by a triple \((t_1, t_2, n)\), where \( t_1 \) and \( t_2 \) are vectors in the plane of shear, and \( n \) is the normal to that plane. Let \( c_1 \) and \( c_2 \) be the centroid of the shared interfaces on \( D_1 \) and \( D_2 \) respectively. The normal \( n \) is the line joining the centroids \( c_1 - c_2 \), and \( t_1 \) and \( t_2 \) are computed by assuming that \( c_2 \) lies in the plane of shear.

**Algorithm HingeDetect.** We detect hinge bending between \( D_1 \) and \( D_2 \) if one of the following criteria are met.

1. **Interface size.** If the number of interface atoms between \( D_1 \) and \( D_2 \) is zero for any interface width below the interface width cutoff (see Section 9).

2. **Connector length.** If the size of the longest connector between \( D_1 \) and \( D_2 \) is greater than six residues, or if an unpaired domain between \( D_1 \) and \( D_2 \) is greater than six residues in size.

Additionally, if \( D_1 \) and \( D_2 \) are segmental SSEs belonging to the same parent SSE, then hinge bending motions are assigned between them. No shearing motions are assigned between segmental SSEs.

**Algorithm HingeParam.** Hinge bending motions are parametrized by a triple \((h, p, s)\), where \( h \) is the hinge point and \( p \) and \( s \) are the primary and secondary bending axes about which rotation occurs. Let the centroids of the domains be \( c_1 \) and \( c_2 \), and the centroid of the connector joining the two domains be \( c_{12} \). Then
the hinge point \( h := c_{12} \), the primary hinge bending axis \( p \) is the normal to the plane containing \( c_1, c_2, \) and \( c_{12} \), and the secondary hinge bending axis is the vector \( s = c_1 - c_2 \).  

5.3 Hierarchical motion graph creation  

1. **Rigid-body fitting.** Our rigid-body search algorithm is based on exhaustively sampling a chosen subset of \( SE(3) \), the space of rigid-body motions. In a previous work [5], we show how our non-uniform search and scoring functions can be combined into a rigid-body fitting tool that is fast, resolution-robust, and accurate. The scoring functions have been explained in Section 3; we explain our rigid-body search scheme in Sections 7.1, 7.2, as it is also part of our flexible search scheme.  

2. **Measuring local fit of each secondary structure.** For a given scoring function \( S \) (see Section 3), HDD collects all the secondary structures that score poorly with respect to \( S \), forming \( n \) clusters of secondary structures \( D_1, D_2 \ldots D_n \). The scoring function used here is typically the ETR, which returns a list of all the secondary structures with atoms excluded outside a given isocontour of \( M \). This step is skipped if the ETR is zero for all secondary structures in the protein. See also Section 9.1.  

3. **Clustering secondary structures.** An interface ratio and area threshold is set (See Section 9.1 for a discussion of how to set these parameters). To each cluster \( D_i \), a candidate secondary structure is added if it satisfies the shear criterion in Algorithm ShearParam, i.e., if it shares an interface with any of the secondary structural members of \( D_i \). The process is repeated for all \( D_i, i \in 1 \ldots n \) until no more secondary structures can be added to any of the clusters. Each cluster is now a domain; all unassigned secondary structures are also designated as domains.  

\(^1\)All rotation vectors are normalized to have length = 1.
Figure 6: Schematic of hierarchical motion domain decomposition and assignment at all levels, and resulting flexibility tree. (A) Hierarchical domain decomposition/motion graph of entire protein. (B) Flexibility tree. See also Figure 3 and 4.

To form the hierarchy, the process above is repeated with a higher interface ratio and area threshold, resulting in a larger number of domains. As the interface ratio threshold approaches one, the number of domains approaches the number of secondary structures. The secondary structures are the lowest level in the hierarchy.

Note that in the hierarchy we form, a domain \( D \in \mathcal{DD}_i \) at any level \( i \) is not necessarily a strict subset of a single domain at the previous level \( i - 1 \), and in fact may be distributed among several different, necessarily disjoint domains at \( i - 1 \) (Figure 4). However, we have yet to encounter this partial containment (Section 9) in practice, i.e., each domain at level \( i \) is usually a strict subset of a single domain at level \( i - 1 \).

**Motion graph.** Given a domain decomposition \( \mathcal{DD} \) at any level in the hierarchy, we construct the graph of possible motions, or motion graph, of \( \mathcal{P} \) under rigid-body motions of \( D_i \in \mathcal{DD} \). Let \( D_i, D_j \) be two domains between which either shear or hinge bending motions are assigned. Then the set of linkers between \( D_i \) and \( D_j \) forms a flexor \( F_{ij} \). The motion graph \( MG = (D_i, F_{ij}), i, j \in \{1 \ldots n\} \) is the set of vertices \( D_i \) and the set of edges \( F_{ij} \). Each flexor \( F_{ij} \) is associated with the type of motion that \( D_i \) and \( D_j \) undergo relative to each other. The process is repeated for all hierarchical levels to produce the hierarchical motion graph. See also Figures 3, 4, and 6.

### 6 Flexibility tree

The motion graph associated with each \( \mathcal{DD}_i \in \mathcal{HD} \) enumerates the space of motions available to the protein, but this space is too vast to sample exhaustively, and in any case consists of relative orientations of \( \mathcal{DD}_i \) that score poorly with respect to either \( S_{fit} \) or \( S_{inter} \) (Section 3). In order to flexibly fit \( \mathcal{P} \) to \( \mathcal{M} \), we reduce the motion graph specified by \( \mathcal{HD} \) to a tree \( \mathcal{F} \) that we term the flexibility tree. See also Figure 6.
In the following algorithms, the letter $i$ indexes the distance from level 0, i.e., the level, and the letter $j$ indexes the node at a given level $i$.

Let the number of levels in $\mathcal{HD}$ be $n_{\mathcal{HD}} + 1$, with the zeroth level denoting the entire protein. We begin at the first level $\mathcal{DD}_1$ of the hierarchy specified by $\mathcal{HD}$. The score $S(D_i, M)$ of each domain $D_i$ in $\mathcal{DD}_1$ with respect to $M$ is computed. The most poorly scoring domain $D_j$ is chosen as the left node of $F_1$, while each of the other domains in $\mathcal{DD}_1$ are collapsed into a single domain, the right node of $F_1$.

For levels $2 - n_{\mathcal{HD}} - 1$, we use the following scheme to populate $F$.

**Single-child tree.** In this scheme, each node at level $1 \leq i \leq n_{\mathcal{HD}} - 1$ has exactly one child. The most ill-fitting domain at each hierarchical level of $\mathcal{HD}$ is chosen as the child of the left node, whereas the rest of the protein is collapsed into a single domain, chosen as the right node. Note that the number of nodes at any level $i$ is exactly 2, i.e., $0 \leq j \leq 1$.

Finally, for the secondary structural level $n_{\mathcal{HD}}$, we group all secondary structures belonging to a particular domain at level $n_{\mathcal{HD}} - 1$ into pairs, between which we then assign motions using the scheme in Section 5.2. If there are an odd number of secondary structures in any domain, we choose that secondary structure that has the highest packing constraint (See Section 5.2) as stationary. Note that, if the secondary structural level is not chosen by the user for motion, then $F$ is a single-child binary tree.

### 7 Flexibility tree traversal/search

At each level $i$, we search recursively over the space of relative rigid-body motions of $F_{i,j}$, thus flexibly fitting the protein $\mathcal{P}$ into the density map $\mathcal{M}$. Define the local score of a node of the flexibility tree $F_{i,j}$ as

$$S_{local} = S_{tot}(A_{i,j}, B)$$

where $A_{i,j}$ is the representation of $F_{i,j}$, $B$ is the representation of $\mathcal{M}$, and $S_{tot}$ is given by Equation 15. Define also the global score

$$S_{global} = S_{tot}(A_c, B)$$

where $A_c$ is the representation of the current conformation of the crystal structure $\mathcal{P}$, and is a union of the representation of all the nodes $j$ at the current hierarchical level $i$, i.e.,

$$A_c = \bigcup_j A_{i,j}$$

For each level $i$ in the tree $F$, we perform the following operations.

1. **Motion assignment.** The motions $T_i \in SE(3)$ of $F_{i,j}$ relative to $F_{i,k}$ are computed using the pairwise criterion in Section 5.2. Here $j$ and $k$ are left and right nodes.

2. **Rigid-body search, left node.** The position of $F_{i,j}$ relative to $F_{i,k}$ that maximizes Equation 16 is then found using the either the hinge bending or the shear search algorithms (Section 7.1, 7.2), depending on the motion assigned.

3. **Rigid-body search, right node.** Step 2 is repeated with the roles of $j$ and $k$ interchanged.

4. **Rigid-body search, repeat.** Steps 2 and 3 are repeated $n_{repeat}$ times.
For levels \( i \geq 1 \), the following local search procedure is adopted: the space of motions available to any child of \( F_i \) is restricted to the volume occupied by \( F_i \). Our technique to compute correlations on arbitrary subsets of \( SE(3) = \mathbb{R}^3 \times SO(3) \) \([3, 5]\), makes this local search procedure efficient and practical; note that by contrast a standard Fourier-based search scheme is not capable of such local refinement.

7.1 Search over hinge bending motions

A hinge bending motion is a special case of a pure rotational motion. Let \((h, p, s)\) specify the hinge bending motion, and let \(\theta\) and \(\psi\) be the angle \(p\) makes with the positive z-axis in the y-z plane and x-z planes respectively, and \(\phi\) the angle \(s\) makes with the positive y-axis in the x-y plane. Let \(R = R_z(\phi)R_x(\theta)R_y(\psi)\), where \(R_x, R_y, \) and \(R_z\) denote rotations about the x-, y-, and z-axes respectively. Then the rigid-body transformation

\[
T = \begin{bmatrix}
R & Rh
\end{bmatrix}
\]

translates the hinge-point \(h\) to the origin, and aligns the primary and secondary axes \(p, s\) with the z- and y-axes respectively. Following this transformation, hinge bending motions about \(p\) and \(s\) can be expressed as the z-y-z Euler angle triple \((\alpha, \beta, \gamma)\), where \(\beta\) is a rotation about the secondary axis and \(\alpha, \gamma\) are rotations about the primary axis. A search over hinge bending motions thus entails a search over the Euler angular parameters \(\alpha, \beta\) and \(\gamma\). In a previous work \([3]\) , we explain how to conduct a fast search over a chosen set of samples from the space of rotations \(SO(3)\). We briefly recount that scheme below.

Our search scheme is based on a framework in which scalar-valued functions \(A: \mathbb{R}^3 \rightarrow \mathbb{C}\) are expressed in terms of basis coefficients \(\hat{a} \in \mathbb{C}\) where the bases are either pure spherical or mixed radial/spherical eigenfunctions.\(^2\)

The mixed representation is computed, for instance, as follows. Let \(u = (\theta, \phi), \quad \theta \in [0, \pi], \quad \phi \in [0, 2\pi]\), and \(r \in \mathbb{R}^+\). A scalar valued function \(A(r, u) : \mathbb{R}^+ \times S^2 \rightarrow \mathbb{C}\) can be expanded as

\[
A(r, u) = \sum_{k=1}^{L} \sum_{l=0}^{k-1} \sum_{m=-l}^{l} \hat{a}_{k,l,m}(r) R_k(l, m) Y_l^m(u)
\]  \hspace{1cm} (19)

where \(R_k(l, m)\) and \(Y_l^m(u)\) are the radial and spherical basis functions respectively, and \(L\) is a finite expansion degree. Following Ritchie et.al \([38]\), we choose generalized Laguerre-polygonal-scaled Gaussians for \(R_k(r)\), whereas \(Y_l^m(u)\) are the spherical harmonic functions.

The coefficients \(\hat{a}_{k,l,m}\) represent the original function \(A\) in the new basis. Given the representations \(\hat{a}_{k,l,m}\) and \(\hat{b}_{k,l,m}\) for two scalar-valued functions \(A\) and \(B\), the correlation between \(A\) and \(B\) under a rigid-body rotation \(R \in SO(3)\) of \(B\) is given by

\[
C(R) = \int_{\mathbb{R}^3} A(Rx)B(x)dx
\]

\[
= \sum_{k,l,m,m'} (-1)^m \hat{b}_{k,l,m}(-1)^{m'} \overline{\hat{a}_{k,l,m}} D_{k,l,m}^{m,m'}(R),
\]  \hspace{1cm} (20)

where the Wigner-D functions \(D_{k,l,m}^{m,m'}\) are a basis for functions in \(L^2(SO(3))\).

In Bajaj et. al and Bettadapura et. al \([3, 5]\), we introduce a technique to compute Equation 20 in \(O(L^4 + N_{\text{rot}})\) steps, where \(N_{\text{rot}}\) is the number of rotations of \(B\).

\(^2\)The choice of the basis function depends on the application. For pure rotational fitting, we prefer to use pure spherical bases, whereas the mixed bases are more appropriate for exhaustive rigid-body motions. See Bajaj et. al \([3]\) for a detailed consideration of basis functions.
7.2 Search over shearing motions

A shearing motion can be expressed by three parameters. Two of these, \( c_1 \) and \( c_2 \), describe translations along the in-plane axes \( t_1 \) and \( t_2 \), and the last describes a rotation \( \theta \) about the normal axis \( n = [n_x, n_y, n_z]^T \) passing through the point \( c \). Together, these three parameters specify the rigid-body motion \((R, t) \in SE(3)\). Let

\[
\hat{R} = \begin{pmatrix} 0 & -n_z & n_y \\ n_z & 0 & -n_x \\ -n_y & n_x & 0 \end{pmatrix}
\]

be the antisymmetric hat operator for rotation about an axis \( [4]\).

\[
R = I + \hat{R} \sin \theta + \hat{R}^2 (1 - \cos \theta),
\]

and

\[
t = c_1 t_1 + c_2 t_2 + c - Rc,
\]

where \( I \) is the identity matrix. A set of samples describing rigid-body motion along a plane is thus a subset of \( SE(3) \). In a previous work \([3]\), we explain how to conduct a fast search over a chosen set of samples from \( SE(3) \). We briefly recount that scheme below.

Given the representations \( \hat{a}_{klm} \) and \( \hat{b}_{klm} \) (Equation 19) for two scalar-valued functions \( A \) and \( B \), the correlation between \( A \) and \( B \) under a rigid-body transformation \((R, t) \in SE(3)\) of \( B \) is given by

\[
C(R, t) = \int_{\mathbb{R}^3} A(Rx + t)B(x)dx
\]

\[
= \sum_{klmn} \hat{b}_{klm} D^B_{n,m}(R) \sum_{k'm'} (-1)^n \hat{a}_{k'l'm'} D^{-n,m'}_{k'l'}(R^A) T^{[n]}_{k'k'l'}(z),
\]

where \( T^{[n]}_{k'k'l'}(z) \) are the translation matrix entries for the basis functions in Equation 19 \([11, 37]\), and \((R^A, R^B, z)\) expresses \((R, t)\) as a single translation about the z-axis and a set of rotations about the z- and y-axes.

In Bajaj et. al and Bettadapura et. al \([3, 5]\), we introduce a technique to compute Equation 20 in \( O((L^6 + L^4 N_{Ra} + N_{Ra} N_{Rb}) T) \) steps, where \( T \) is the number of z-axis translations.

7.3 Biased all-atomistic molecular dynamics

After the initial domain-based flexible search, we conduct an additional biased molecular dynamics search to improve the flexible fit. This fitting tool is implemented in GROMACS and uses the CHARMM27 \([28]\) force field in vacuo at a temperature of 300K and a pressure of 1 atm, maintained using a Langevin thermostat. Following several recent works (see the introduction for a summary), the biasing function is additive and is proportional to unity minus the correlation between the Gaussian blurred representation of \( \mathcal{P} \) and \( \mathcal{M} \). We have also implemented an additional harmonic term to restrain backbone torsions \( \phi \) and \( \psi \); following Trabuco et. al\([49]\), the spring constant is set to 200 kcal mol\(^{-1}\) rad\(^{-2}\), thus maintaining the identity of secondary structural elements.
<table>
<thead>
<tr>
<th>Name</th>
<th>PDB</th>
<th>Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemagglutinin</td>
<td>1IBN, 1IBO [20]</td>
<td></td>
</tr>
<tr>
<td>Adenylate kinase</td>
<td>1AKE [31], 4AKE [30]</td>
<td></td>
</tr>
<tr>
<td>Guanylate kinase</td>
<td>1EX6, 1EX7 [9]</td>
<td></td>
</tr>
<tr>
<td>Ribose binding protein</td>
<td>1URP [8], 2DRI [7]</td>
<td></td>
</tr>
<tr>
<td>GroEL</td>
<td>1OEL [10], 1WE3 [42]</td>
<td>EMD 5001 [27]</td>
</tr>
<tr>
<td>Citrate synthase</td>
<td>1CTS, 2CTS [36]</td>
<td></td>
</tr>
<tr>
<td>Hexokinase</td>
<td>1HKG [43], 2YHX [2]</td>
<td></td>
</tr>
<tr>
<td>Aspartyl-T-RNA-Synthetase</td>
<td>1EQR [35], 1C0A [12]</td>
<td></td>
</tr>
<tr>
<td>SIV</td>
<td>3DNO [26]</td>
<td>EMD 5020 [26]</td>
</tr>
</tbody>
</table>

Table 2: Datasets used in this study.

8 Validation
We choose a standard technique to validate PF$^3$fit. Let a protein $P$ available in two different conformations $P^1$ and $P^2$. We generate a synthetic density map $M^2$ from $P^2$ at a chosen resolution $R$ that varies between 5 and 15 Å. We then flexibly fit $P^1$ to $M^2$ and measure the final RMSD between the flexibly fit version of $P^1$ and $P^2$. See Table 2 for the list of chosen datasets.

We note that all the results in the following section are obtained with the following parameters.

- **Steric clash.** The steric clash score is used, i.e. $w_{\text{inter}} > 0$. We discuss the role of this score and the value of $w_{\text{inter}}$ in Section 9.2.
- **SSE subdivision.** We do not subdivide SSEs, i.e, $l_H$ and $l_S$ are both set to infinity (Section 5.1).
- **Number of repetitions.** The number of repetitions $n_{\text{repeat}}$ (Section 6) is set to 0.

9 Results and discussion

**Hemagglutinin.** Hemagglutinin is a small protein available in the pH-7 (1IBO) and ph-5 (1IBN) conformations. At an interface width of 6.5 Å, the ratio of the number of shared interface atoms to total atoms on the domains is 0.4, i.e., 40% of the atoms in Hemagglutinin occur at the interface. However, since both domains are secondary structural elements, and since there are no packing considerations (See Algorithm ShearDetect and Figure 9 (A)), hinge bending motions are assigned, yielding the flexible fit in Figure 7. Note that since the $3_{10}$ helix in the pH 5 conformation doesn’t exist in the pH 7 conformation, no significant improvement in RMSD is obtained by subjecting the result in Figure 7 (C) to all-atomistic molecular dynamics-based flexible fitting.

**Guanylate kinase.** For Guanylate Kinase, the two domains detected by HDD do not share a significant interface (the interface ratio = 0.1 at an interface width = 6.5 Å); the domains are thus assigned hinge bending motions. The resulting flexible fit yields a conformation that is very close to the target (Figure 8). MD-based flexible fitting is further used to improve the fit; however, PF$^3$fit captures a majority of the motion of this kinase (Table 3) in a fraction of the time required by the MD-based flexible fitting tool.

**Ribose-binding protein.** The two domains detected from 1URP, one of the open forms of the ribose binding protein, share a small interface (interface ratio = 0.06 at interface width = 6.5 Å), and hinge bending motions are thus assigned. In addition, the initial rigid-body fit places each of the domains poorly into the
Figure 7: Fitting the pH 5 structure of Hemagglutinin (1IBN) into a 5 Å synthetic density map of the pH 7 (1IBO) structure. (A) Initial rigid-body fit. The domains detected by HDD are colored red and green. The initial Cα RMSD is 4.23 Å and the initial correlation coefficient with respect to the Gauss CCS is 0.62. Hinge bending motions are assigned between the domains; the primary hinge axis points out of the plane of the page, and the secondary hinge axis points right to left. (B) Conformation obtained after search over hinge bending motions. Rotations about the primary and secondary axes are respectively 2° and 45°. The final correlation coefficient is 0.88. (C) Final fit structure superimposed on 1IBO. The final Cα RMSD is 2.5 Å.

Figure 8: Fitting the closed structure (1EX7) of Guanylate kinase into a 8 Å synthetic density map of its open structure (1EX6). (A) Initial rigid-body fit. The domains detected by HDD are colored red and green. The initial Cα RMSD is 4.76 Å, and the initial correlation coefficient with respect to the Gauss CCS is 0.72. Hinge bending motions are assigned between the domains; the primary hinge axis points out of the plane of the page, and the secondary hinge axis points right to left. (B) Conformation obtained after search over hinge bending motions. Rotations about the primary and secondary axes are respectively 33° and 37°. The correlation coefficient is 0.93, and the Cα RMSD is 1.43 Å. (C) All-atomistic molecular dynamical flexible fitting with GROMACS, using the structure in B. The final correlation coefficient is 0.99. (D) The final fit structure superimposed on 1EX6. The final Cα RMSD is 0.7 Å.

density map (ETR > 0.1 for both domains), and thus the same hinge axes and hinge point are used to move both domains. The resulting flexible fit captures a great deal of the overall motion of the protein (Table 3), and
Figure 9: Flexibility trees for a few proteins on which PF$^3 fit$ is applied. (A) Hemagglutinin. (B) Aspartyl-T-RNA-Synthetase. See also Figure 10.

Figure 10: Flexibility trees for a few proteins on which PF$^3 fit$ is applied. (A) Adenylate-kinase. (B) Citrate synthase; the final level is the SSE level. See also Figure 9.

...the MD-based flexible fitting tool improves the RMSD to 1.16 Å.

**Aspartase RNA-Synthetase.** HDD detects two spatially distinct domains in the unbound structure of Aspartase RNA-Synthetase at the first hierarchical level, three at the second, and four at the third hierarchical level (Figure 12, Figure 9 (B)). The two domains (yellow and red) at the first level share an interface ratio of 0.18 with each other, and the two at the second level share an interface ratio of 0.11. Like in the above discussions, hinge bending motions are assigned between each of the domains at both levels 1 and 2. However, the search over hinge bending motions yields a conformation, in the case of each of the moving domains, that does not significantly improve the fit. When shearing motions are assigned instead, the correlation increases by a larger amount (0.24 vs 0.13 in the hinge bending case), suggesting that the true motion is closer to shear than to hinge bending. At the third level in the hierarchy, the purple domain in Figure 12 is assigned hinge bending motions due to the 20 residue connector that connects it to the rest of the protein, resulting in a further small improvement in flexible fit. We did not conduct MD-based flexible fitting on this protein.

**Hexokinase.** HDD detects two domains from Hexokinase(Figure 13). In relative terms, these domains
<table>
<thead>
<tr>
<th>Name</th>
<th>Init./1/2/3/MD*</th>
<th>% PF\textsuperscript{3} fit/% MD\textsuperscript{†}</th>
<th>t\textsubscript{h}/t\textsubscript{m}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemagglutinin</td>
<td>4.23/2.5\textsuperscript{H}/−/−/2.3</td>
<td>90/10</td>
<td>40/−</td>
</tr>
<tr>
<td>Guanylate kinase</td>
<td>4.76/1.43\textsuperscript{H}/−/−/0.7</td>
<td>82/18</td>
<td>35/200</td>
</tr>
<tr>
<td>Ribose binding protein</td>
<td>4.5/2.7\textsuperscript{H}/−/−/1.16</td>
<td>54/46</td>
<td>45/600</td>
</tr>
<tr>
<td>Aspartyl-T-RNA-Synthetase</td>
<td>1.99/1.4\textsuperscript{S}/1.2\textsuperscript{S}/1.1\textsuperscript{H}/−</td>
<td>100/−</td>
<td>600/−</td>
</tr>
<tr>
<td>Hexokinase</td>
<td>2.8/1.9\textsuperscript{H}/−/−/−</td>
<td>100/−</td>
<td>300/−</td>
</tr>
<tr>
<td>Adenylate kinase</td>
<td>7.8/4.3\textsuperscript{H}/3.3\textsuperscript{H}/−/0.4</td>
<td>61/39</td>
<td>20/600</td>
</tr>
<tr>
<td>Citrate synthase</td>
<td>3.0/2.5\textsuperscript{H}/1.9\textsuperscript{S}/−/0.6</td>
<td>42/58</td>
<td>230/1100</td>
</tr>
<tr>
<td>GroEL</td>
<td>17/3.8\textsuperscript{H}/3.3\textsuperscript{S}/2.9\textsuperscript{S}/0.9</td>
<td>87.5/12.5</td>
<td>400/2500</td>
</tr>
</tbody>
</table>

Table 3: Report of motions involved in flexible fitting each protein available in two conformations to its density map. \(^{\dagger}\): Cα RMSD. Init. = Initial. 1/2/3 = RMSD at first /second/third level in the hierarchy using PF\textsuperscript{3} fit. MD = RMSD after MD-based fitting. The superscripts in this column stand for hinge (H) or shear (S) respectively. \(^{\dagger}\): Percentage of overall motion captured. \(^{‡}\): Time taken in seconds as measured on a single-threaded dual-core Macbook Pro with 8 GB main memory and a processor speed of 2.16 GhZ. t\textsubscript{h} = Time taken by PF\textsuperscript{3} fit. t\textsubscript{m} = Time taken by MD.

...share a small interface with each other (Interface ratio = 0.15); however, the interface area between them is high (720 Å\(^2\)) due to the four helices and two beta sheets that exist at the interface. Shearing motions are thus assigned between the two domains, resulting in a final conformation that improves the Cα RMSD by 0.8 Å. MD-based flexible fitting failed on this protein due to the presence, in both conformations, of many instances of amino-butanoic acid, a residue not parametrized by the CHARMM27 force field.

**Adenylate kinase.** HDD detects two domains from Adenylate kinase at the first hierarchical level, and further splits the bottom domain (Figure 1, 10(A)) into two at the second hierarchical level. Similar to the case of Guanylate kinase, hinge bending motions are assigned at both hierarchical levels (Figure 10), resulting in an overall improvement in Cα RMSD of about 4Å. MD-based flexible fitting further lowers the RMSD to < 1Å.

**Citrate synthase.** HDD detects two domains from citrate synthase at the first hierarchical level. We further choose each secondary structural element as a domain for the second level of the hierarchy (Figure 10). At the first level, a clear hinge motion is detected between the two domains, as they share no interface (Interface ratio = 0.01). At the second hierarchical level, shearing motions are assigned due both to the high interface ratios as well as the packing density (see algorithm ShearDetect). The improvement in RMSD after both stages of hierarchical fitting is 1.1Å. A further improvement of 1.4 Å is obtained with MD-based flexible fitting.

**GroEL.** GroEL presents a case of very large conformational change. The initial Cα RMSD between the open and closed conformations is 17 Å, and is almost entirely due to the hinge bending motions of a single domain. HDD detects two domains at the first hierarchical level, and splits the bottom domain into two at the second hierarchical level. At the third hierarchical level, we choose each domain as a secondary structural element. Hinge-bending motions at the first and second hierarchical levels, and shear at the third hierarchical level between a single helix and its neighbors results in an overall improvement in Cα RMSD of 14Å, after which domain-based MD is able to reduce the RMSD to < 1Å.
Figure 11: Fitting 1URP into a 15 Å simulated density map of 2DRI. (A) Initial rigid-body fit. Domains detected by HDD are colored red and green respectively. The initial Cα RMSD is 4.5 Å, and the initial correlation with respect to the Gauss-CCS is 0.85. Hinge bending motions are assigned between the domains; the primary hinge axis points out of the plane of the page, and the secondary hinge axis points right to left. (B) Conformation obtained after search over hinge bending motions. Rotations about the primary and secondary axes are respectively 98° and 20° for the red domain, and 70 and 13° for the green domain. The correlation coefficient is 0.92. The Cα RMSD is 2.7 Å. (C) All-atomic molecular dynamical flexible fitting with GROMACS, using the structure in B. The final correlation coefficient is 0.989. (D) Final fit structure superimposed on 2DRI. The final Cα RMSD is 1.16 Å.

9.1 Domain decomposition

HDD, the hierarchical domain decomposition technique introduced in this study, clusters domains based on two parameters. The first of these, the interface width, measures the distance within which two neighboring domains can be said to share an interface. Following Miyazawa and Jernigan [29], we set this to the residue-residue contact distance, 6.5Å. This resembles the decision in Pandurangan and Topf [33], with the distinction that while that work uses the contact distance to measure the distance between neighboring secondary structures, we use it to measure the influence of the interface.

The second set of parameters is the interface ratio and area thresholds, both of which are related to each other. Varying this threshold has the effect of changing the number of domains detected, and creates a hierarchy of domain decompositions. In all the experiments above, varying the interface ratio threshold between 0.3 and 0.5 and the interface area threshold between 700 and 1000Å² yields domains that either (A) compare well with domains detected in the same proteins in the literature or (B) yields a good flexible fit, or (C) both of the above. The degree of user intervention required to obtain a good domain decomposition, i.e., one that satisfies either A or B in the previous sentence, depends on the initial quality of the rigid-body fit. If the initial fit is relatively poor, with several domains possessing an ETR of > 0.1, HDD uses the protruding atoms in an effective manner to come up with a good guess of the domain decomposition. This is the case in Hemagglutinin, Guanylate Kinase, Ribose binding protein, Adenylate Kinase (levels 1 and 2), Citrate Synthase (level 1), and GroEL. However, where the rigid-body fit already places the protein in an approximately correct position with
Figure 12: Fitting the unbound structure of Aspartase RNA-Synthetase (1EQRa) to a synthetic density map of its RNA-bound conformation (1C0Aa) at 8 Å. (A) Domains at multiple hierarchical levels. The domains at the first hierarchical level are colored red, and yellow; the one colored green belongs to the red domain at hierarchical level 1, and the ones colored purple and cyan belong to the red colored domain at level 2. The initial correlation with respect to the Gauss CCS is 0.58, and the initial Cα RMSD is 1.99 Å. Shearing motions are assigned between the yellow/red and the green/red domains respectively. (B) Improvement in flexible fit after search over shearing motions. The yellow domain moves 3.2 and 2.2 Å in the plane of the page, with a slight rotation of 2.8° about the normal axis pointing out of the page. The green domain also undergoes a shearing motion with in-plane motions of 2.8 and 3.7 Å respectively, and 5.3° rotation about the normal axis. (C) Hinge-bending motions are assigned to the purple domain at hierarchical level 3, resulting in a final RMSD of 1.1 Å.

low RMSD, information about protruding domains is not available, and the domain decompositions then become purely a function of the interface ratio and area thresholds. In these situations a certain user discretion is required if unrealistic fits are not to be obtained. This is currently a limitation with all domain detection techniques that depend on a single conformation.

9.2 Steric clashes, overfitting and protein quality

The steric interaction score (Equation 14) proves most successful in multi-domain situations where the range of hinge or shear motions result in intrusions with other domains. Let $P^1$ be the initial conformation, and let $P_{fit}^1$ be the conformation obtained by PF$^3 fit$ following the flexible search over shear and/or hinge bending motions. The effectiveness of the steric interaction score can be said to be directly proportional to the number of energy minimization steps required to bring the molecular mechanical energy of $P_{fit}^1$, as measured by the CHARMM27 force field, to the order of the molecular mechanical energy of the initial conformation $P^1$. Along with the number of steric collisions with and without the steric interaction score, this information is presented in Table 4. We see that the steric interaction score is successful in eliminating a large number of the clashes that inevitably occur in a flexibility parametrization such as ours, and results in conformations that are closer to an energy minimum, as expected. However, there is one important proviso, i.e., the choice of the weight $w_{inter} \in \mathbb{R}^+$. The magnitude of this weight should result in a steric interaction score of the same order as that of the fitting score $S_{fit}$. To enable this, we set $w_{inter} = (i_2 - i_1)^2$, where $i_1 \leq i \leq i_2$ is the range of intensity values in the target density map $M$. 

22
Figure 13: Fitting 1HKG into a 7 Å simulated density map of 2YHX. (A) Initial rigid-body fit, top view. Domains detected by HDD are colored red and green respectively. The initial Cα RMSD is 2.8 Å, and the initial correlation with respect to the Gauss-CCS is 0.72. Shearing motions are assigned between the domains; the normal to the shearing plane points out of the plane of the page. (B) Conformation obtained after search over shearing motions. Translations along the shearing plane are respectively 6 and 2 Å. The correlation coefficient is 0.89. (C) Side view of fit. (D) Fit structure superimposed on 2YHX. The Cα RMSD is 1.9 Å.

Apart from steric clashes, motions assigned by our flexibility model tend to impair the quality of the input protein in another significant way, i.e. by producing long bonds. This is seen typically to be more of an issue with shear rather than hinge bending, a fact also reflected by the number of energy minimization steps required to bring the protein back to a sterically feasible state (Table 4). However, due to our limiting the range of shearing motions to < 5 Å, there occur no situations in which bond lengths cannot be reinstated to their correct values by an energy minimization routine. In addition, since the typical number of steps of energy minimization is < 50, the energy minimized configuration is very close to the input flexibly fit conformation $P^1_{fit}$, with all-atom RMSDs that vary between 0.05 and 0.2 Å.

9.3 Inferring shearing motions: the SIV spike protein

We consider the SIV (Simian Immunodeficiency Virus) spike protein complex in Liu et. al [26], which consists of three proteins. The first is the trimeric gp120 spike protein, each of whose monomers forms a complex with two other proteins: CD4, which mediates immune system responses in the viral host, and 17b, a neutralizing antibody. The gp120 trimer is available in the PDB as 3DNO [26], and the gp120-CD4-17b trimeric complex is available in the PDB as 1GC1.pdb [25]. Additionally, a density map is also available in the EMDB (Electron Microscopy Database) as EMD 5020; this density map contains density information for the gp120-CD4-17b complex as well as other regions on the spike [26].

Liu et. al report a rigid-body fitting of 1GC1 into the density map EMD 5020. For this study, we ask the question: can we improve the fit between the gp120 complex and EMD 5020? Specifically, since gp120 has a significant interface between both CD4 and 17b, can the assignment of shearing motions lead to a better fit between the complex and 5020?

The results in Figure 14 shows an improved fit resulting from assigning shearing motions. The improvement is arguably low in magnitude, but dramatizes the potential of the shear parametrization. Since we can assume
Table 4: Information about the steric interaction score (Equation 14). *: Maximum number of steric clashes in any result belonging to the top ten flexible fitting results. N1 = without steric interaction score. N2 = with steric interaction score. †: Number of energy minimization steps required to bring the molecular mechanical energy of the flexibly fit conformation to the order of the molecular mechanical energy of the pre-fit conformation. E1 = without steric interaction score. E2 = with steric interaction score. The letters DNC stand for "does not converge". We could not run energy minimization on Hexokinase due to its possessing a great many residues unparametrized by the CHARMM27 force field.

<table>
<thead>
<tr>
<th>Name</th>
<th>N1/N2 *</th>
<th>E1/E2†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemagglutinin</td>
<td>0/0</td>
<td>3/3</td>
</tr>
<tr>
<td>Guanylate kinase</td>
<td>3/0</td>
<td>10/4</td>
</tr>
<tr>
<td>Ribose binding protein</td>
<td>0/0</td>
<td>7/7</td>
</tr>
<tr>
<td>Aspartyl-T-RNA-Synthetase</td>
<td>0/0</td>
<td>30/30</td>
</tr>
<tr>
<td>Hexokinase</td>
<td>20/0</td>
<td>–</td>
</tr>
<tr>
<td>Adenylate kinase</td>
<td>313/25</td>
<td>357/28</td>
</tr>
<tr>
<td>Citrate synthase</td>
<td>846/328</td>
<td>DNC/35</td>
</tr>
<tr>
<td>GroEL</td>
<td>791/115</td>
<td>DNC/35</td>
</tr>
</tbody>
</table>

Figure 14: The SIV spike protein complex. (A) The 17b and gp120 domains are seen to the bottom in green, while the CD4 domain is in red. The large interface between CD4 and gp120 indicates shearing motions between them. (B) CD4 sheared by in-plane translations of 0.5 and 1 Å respectively, superimposed on its original copy. The normal vector to the shearing plane approximately points from the bottom to the top of the page. (C) Top and (D) side views of final flexible fit of all three monomeric complexes using shearing motions. For each of the complexes, CD4 is sheared by in-plane translations of 1Å and 1Å, and 17b is sheared by in-plane translations of −0.9Å and 1Å. There are no in-plane rotations. The fit results in an improved correlation coefficient with respect to the one by Liu et. al (0.85 to 0.787). The resulting flexible fit excludes 36 fewer atoms inside the zero isocontour of EMD 5020 than the fit in Liu et. al.
that EMD5020 represents the native state of the spike complex, the conformation corresponding to the improved fit can be interpreted as being closer to the native state than the initial complex in Liu et. al. Whether or not shearing is applicable in this case can only be verified through experiment: the role of this flexible fitting exercise is to surmise that such a motion is possible by demonstrating an improvement in fit.

10 Conclusions

We have shown that, across a range of resolutions, PF$^3 fit$, our flexible fitting routine, is capable of generating conformations that fit well into the density map (Section 9) while exhibiting steric feasibility (Section 9.2). Additionally, we have shown that most of the motion of the proteins in Table 2 can be captured by either of our parametrizations (Table 3), with recourse to an MD-based approach only at the final stage. PF$^3 fit$ also contributes to existing work on flexible fitting in the following ways.

- **Search scheme.** To our knowledge, ours is the first scheme to use a Fourier-based approach, rotational or translational, to optimize the fit of $P$ by searching the space of motions available to each of its domains. Additionally, the non-uniformity inherent to our search scheme enables searching over restricted subsets of the rigid-body motion space $SE(3)$, where the restriction is determined by the type of motion (shear or hinge bending). By contrast, uniform Fourier search schemes, limited as they are to uniform grids, cannot be used for the same purpose. See also Bajaj et. al and Bettadapura et. al [3, 5].

- **Search scheme, part 2: fast rotations.** Our fast rotational search scheme (Equation 20) enables a quick scan of the rotational degrees of freedom available to the domain. This means that a hinge bending search can be conducted very efficiently by either (A) allowing PF$^3 fit$ to assign the hinge point and axes, or (B) requiring that such a motion occurs about a given point and axis, and examining the resulting improvement in fit. Such searches would find widespread application in flexible fitting routines wherein a quick estimate of the hinge point and amount of rotations about the primary and secondary axes are required. It would also find application in docking routines where the flexibility in either or both the docked proteins is known to be explained wholly or mostly by hinge bending. See Flores and Gerstein [14] for several examples of docking in the special case of hinge bending flexibility.

- **Motion detection.** PF$^3 fit$ detects motions from input domains based on pairwise interface statistics. This detection has proven successful with most proteins surveyed in this work, with one notable ambiguity occurring in the case of Aspartase RNA-Synthetase. To our knowledge, this work is the first to explicitly parametrize, and search over, the space of shearing motions available to the protein without user intervention. It is also the first flexible fitting tool to parametrize and search over hinge bending motions, also without user intervention.

- **Motion detection, part 2.** PF$^3 fit$ results in flexible fits that take a protein in a particular conformation $P^1$ most of the way to a target conformation $P^2$. For instance, in the case of Hemagglutinin, Ribose binding protein, Hexokinase, Guanylate/Adenylate Kinase and GroEL, it is able to explain more than 50% and in most cases more than 80% of the motion of the protein. The MD-based flexible fitting is then used only in the final stage. Additionally, results from Ribose binding protein and GroEL by PF$^3 fit$ are better than previously reported fits using all-atomistic approaches (see Topf et. al and Pandurangan and Topf [48, 33]), and the use of the MD-based tool improves the fit to amounts that have not been so far obtained. For proteins like Hexokinase, which contain many residues unparametrized by any molecular dynamics force field, PF$^3 fit$ provides an alternative flexible fitting procedure.

Given the above observations, we see PF$^3 fit$ as fulfilling the following role in flexible fitting: it can be used as a tool to annotate and quantify the motions available to the domains of a protein at several hierarchical levels.
Thus, if the approximate hinge axis and rotational angles, or the approximate shearing plane and translational amounts are desired, the user can turn to PF$^3$ fit to determine this information, using the MD-based flexible fitting tool only at the very end, and only if he needs to. The time-efficiency of our approach relative to normal MD-based approaches (Table 3, column 4) reinforces this role.

Our scheme to detect shearing motions between domains currently admits of only two parameters, the interface ratio/area thresholds and, in the case of secondary structural elements, the degree of packing. As a mechanism, domain-based or secondary structural-based shear is complicated, involving many small backbone and side-chain torsional rotations that result in the canonical “sliding across an interface”. We emphasize that we do not claim, in this work, to have solved the problem of predicting shear in proteins. However, it is relatively easy to introduce a new parameter, say the angle between participating helices in secondary structural shear, into PF$^3$ fit, with the goal of training it against a set of proteins known to undergo shear. This may be prove to be a powerful way to understand, and thus predict, the as yet unsolved puzzle of shearing motion in proteins.

References


27


