1   ***Exposing hidden alternative backbone conformations***
2   ***in X-ray crystallography using qFit***
3

4   Daniel A. Keedy[1], James S. Fraser[1], Henry van den Bedem[2*]
5
6   [1] - Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco,
7   San Francisco, California, United States of America
8   [2] - Division of Biosciences, SLAC National Accelerator Laboratory, Stanford University, California,
9   United States of America
10   [*] corresponding author
11   Email: vdbedem@slac.stanford.edu (HvdB)
12

13   **Abstract**

14   Proteins must move between different conformations of their native ensemble to perform their
15   functions. Crystal structures obtained from high-resolution X-ray diffraction data reflect this
16   heterogeneity as a spatial and temporal conformational average. Although movement between natively
17   populated alternative conformations can be critical for characterizing molecular mechanisms, it is
18   challenging to identify these conformations within electron density maps. Alternative side chain
19   conformations are generally well separated into distinct rotameric conformations, but alternative
20   backbone conformations can overlap at several atomic positions. Our model building program qFit uses
21   mixed integer quadratic programming (MIQP) to evaluate an extremely large number of combinations of
22   sidechain conformers and backbone fragments to locally explain the electron density. Here, we
23   describe two major modeling enhancements to qFit: peptide flips and alternative glycine conformations.
24   We find that peptide flips fall into four stereotypical clusters and are enriched in glycine residues at the
25   $n+1$ position. The potential for insights uncovered by new peptide flips and glycine conformations is
26   exemplified by HIV protease, where different inhibitors are associated with peptide flips in the "flap"
27   regions adjacent to the inhibitor binding site. Our results paint a picture of peptide flips as
28   conformational switches, often enabled by glycine flexibility, that result in dramatic local
29   rearrangements. Our results furthermore demonstrate the power of large-scale computational analysis
30   to provide new insights into conformational heterogeneity. Overall, improved modeling of backbone
31   heterogeneity with high-resolution X-ray data will connect dynamics to the structure-function
32   relationship and help drive new design strategies for inhibitors of biomedically important systems.
33

34   **Author Summary**

35   Describing the multiple conformations of proteins is important for understanding the relationship
36   between molecular flexibility and function. However, most methods for interpreting data from X-ray
37   crystallography focus on building a single structure of the protein, which limits the potential for biological
38   insights. Here we introduce an improved algorithm for using crystallographic data to model these
39   multiple conformations that addresses two previously overlooked types of protein backbone flexibility:
40   peptide flips and glycine movements. The method successfully models known examples of these types
41   of multiple conformations, and also identifies new cases that were previously unrecognized but are well
42   supported by the experimental data. For example, we discover glycine-driven peptide flips in the
43   inhibitor-gating "flaps" of the drug target HIV protease that were not modeled in the original structures.
44   Automatically modeling "hidden" multiple conformations of proteins using our algorithm may help drive
45   biomedically relevant insights in structural biology pertaining to, e.g., drug discovery for HIV-1 protease
46   and other therapeutic targets.
47

48 **Introduction**

49 Even well-folded globular proteins exhibit significant flexibility in their native state [1]. However, despite

50 advances in nuclear magnetic resonance dynamics experiments and computational simulations,

51 accurately characterizing the nature and extent of biomolecular flexibility remains a formidable

52 challenge [2]. While traditionally X-ray crystallography is associated with characterizing the ground

53 state of a biomolecule, the ensemble nature of diffraction experiments means that precise details of

54 alternative conformations can be accessed when the electron density maps are of sufficient quality and

55 resolution [3]. These maps represent spatiotemporal averaged electron density from conformational

56 heterogeneity across the millions of unit cells within a crystal [4, 5].

57

58 Computational methods have made strides toward uncovering and modeling conformational

59 heterogeneity in protein structures from crystallographic data [3]. However, there is currently no

60 automated approach to recognize the features of extensive backbone flexibility in electron density

61 maps, model the constituent alternative conformations, and validate that the incorporation of

62 heterogeneity improves the model. B-factors theoretically model harmonic displacements from the

63 mean position of each atom, but in practice are often convolved with occupancies of discrete alternative

64 positions when multiple backbone conformations partially overlap [5]. Statistical analyses of electron

65 density using Ringer has revealed evidence for a surprising number of "hidden" alternative

66 conformations in electron density maps [6, 7]. The phenix.ensemble_refinement method [8] uses

67 electron density to bias molecular dynamics simulations, then assembles snapshots from this trajectory

68 into a multi-copy ensemble model. However, energy barriers of the simulation may prevent sampling of

69 well separated backbone conformations. Accurately modeling protein conformational heterogeneity, in

70 particular when the mainchain adopts distinct conformations for one or a number of contiguous

71 residues, remains a difficult task. The spatial overlap of electron density of multiple conformations and

72 the relatively similar profiles of branching mainchain and sidechains blur structural features that can

73 guide the human eye to reduce the large number of possible interpretations [9].

74

75 We have previously developed qFit [10], a method for automatically disentangling and modeling

76 alternative conformations and their associated occupancies, which are represented by the variable q

77 (for "oc*cu*pancy") in standard structure factor equations. The qFit algorithm examines a vast number of

78 alternative interpretations of the electron density map simultaneously. To propitiously explore a high-

79 dimensional search space, conformational sampling is guided by the anisotropy of electron density at

80 the Cβ atom position, the nexus of backbone and sidechain in polypeptides [11]. For each slightly

81 shifted Cβ atom position, qFit samples sidechain conformations with a rotamer library [12] and uses

82 inverse kinematics to maintain backbone closure [9]. Finally, it selects a set of one to four

83   conformations for each residue that, collectively, optimally explain the local electron density in real

84   space.

85

86   However, the anisotropy of the Cβ atom limits the exploration radius of qFit to model backbone

87   conformational heterogeneity. While protein backbone motions are often associated with large-

88   amplitude conformational flexibility of surface loop regions, subtle motions can have important ripple

89   effects in closely packed areas via sidechain-backbone coupling. For example, fast (ps-ns) backbone

90   NH and sidechain methyl order parameters from spin relaxation experiments are highly correlated with

91   each other in flexible regions [13], suggesting that mainchain and sidechain motions collectively sample

92   conformational substates. For example, a backbone backrub motion [14] repositions the Cα-Cβ bond

93   vector in a plane perpendicular to the chain direction, enabling the sidechain to access alternative,

94   often sparsely populated rotamers that otherwise would be energetically unfavorable. We previously

95   linked coupled transitions between alternative sidechain conformations, like "falling dominos", to

96   enzymatic turnover and allostery [15, 16].

97

98   Additionally, qFit cannot model discrete conformational substates such as peptide flips, which are >90°

99   rotations of a peptide group while minimally perturbing the flanking residues. Some structure validation

100   methods highlight incorrect peptide orientations [17] and even automate subsequent model rebuilding

101   [18]. However, rebuilding fits a correct, unique conformation rather than multiple well-populated

102   alternative peptide conformations. Peptide flips can have important functional roles in proteins. For

103   example, flavodoxin undergoes peptide rotations between functional states as part of the catalytic cycle

104   [19], and peptide flips that convert β-sheet to α-sheet have been linked to amyloid formation [20].

105   Furthermore, high-resolution crystal structures have shown that alternative conformations related by a

106   peptide flip may be populated in the same crystal [14].

107

108   Modeling alternative conformations of glycine residues, which lack a Cβ atom, is also a current

109   limitation of qFit. The lack of a Cβ atom allows glycine residues to access otherwise forbidden regions

110   of conformational space [11] and thereby fill special structural roles such as capping helix C-termini

111   [21]. In addition, the flexibility of glycines may contribute directly to function at flexible inter-domain

112   linkers or conformationally dynamic enzyme active sites [22]. Automatically modeling such cases as

113   alternative conformations with qFit paves the way toward understanding their contributions to protein

114   function. Increasingly, new experiments are being proposed which, combined with computational

115   analysis, can extract the spatiotemporal ensemble from electron density maps [15, 23, 24].

116

117   Adding the capability to model peptide flips and alternative conformations for glycines will increase our

118   power to uncover conformational heterogeneity. While the number of sampled conformations for

119   glycines is modest owing to a missing side-chain, including peptide flips for all amino acids adds

120   significant computational complexity to the qFit algorithm. A powerful quadratic programming algorithm

121   lies at the core of qFit and is necessary to determine non-zero occupancies for up to four conformations

122   from among hundreds or even thousands of candidate conformations for each residue. Even for

123   modest sample sizes, around 500, the number of combinations of candidate conformations is

124   enormous, exceeding 10^9. As more backbone motion is incorporated into qFit, the computational

125   complexity increases, demanding a parallelized approach to refinement on a residue by residue basis.

126   Although this moves rebuilding away from a single node towards a larger compute cluster, the

127   combination of data-driven sampling and selection has enabled qFit to automatically build

128   multiconformer models that have illuminated intramolecular networks of coupled conformational

129   substates [16] and the effects of cryocooling crystals [25, 26]. Similar hybrid approaches using robotics

130   sampling and selection based on experimental NMR data are also being extended to nucleotide

131   systems such as the excited state of HIV-1 TAR RNA [27].

132

133   Here we introduce qFit 2.0, an updated version of the qFit algorithm with new capabilities for modeling

134   near-native backbone conformational heterogeneity in crystal structures. We first describe the

135   quadratic programming procedure that allows selection of a small set of conformations per residue that

136   collectively account for the local electron density, and discuss its extension to fitting backbone atoms in

137   addition to sidechain atoms. We then describe new conformational sampling features of qFit 2.0, in

138   particular glycine shifts and peptide flips. Finally, we validate the updated algorithm with both synthetic

139   and experimental X-ray data. qFit 2.0 is freely available by webserver and source code is available for

140   download at https://simtk.org/home/qfit.

141

142   **Results**

143   Improved backbone sampling and selection in qFit

144   To automatically identify alternative backbone conformations, including peptide flips, we augmented the

145   sample-and-select protocol in qFit (see **Figure 1** and Methods). Previously, conformations were

146   sampled based on anisotropy of the Cβ atom and were selected based on the fit between observed

147   and calculated electron density for the sidechain (Cβ atom and beyond) only. Alternative

148   conformations for mainchain atoms were ultimately included in the multiconformer model only because

149   they accommodated the best sidechain fits. In qFit 2.0, we now select conformations based on the fit

150   between observed and calculated electron density for the sidechain atoms and also the backbone O

151   atom. The O atom is an excellent yardstick for identifying backbone conformational heterogeneity for

152   two reasons. First, it is furthest from the Cα-Cα axis so its density profile is somewhat isolated and is

153   displaced most by rotations around that axis [14]. Second, it has more electrons than other backbone

154   heavy atoms, so is most evident in electron density maps. This change allows us to select peptide flips

155   outside of α-helices and β-sheets, where flips are prevented by steric and hydrogen-bonding

156   constraints, then directly select flipped conformations. This procedure is effective because the large

157   movement of the backbone O during a peptide flip leaves a major signature in the electron density.

158

159   Glycine modeling

160   Incorporating the backbone O atom also enhances the detection of less discrete backbone

161   conformational changes. In particular, we now sample alternative glycine conformations based on

162   anisotropy of the electron density for the O atom, by analogy to the Cβ-driven sampling for all other

163   amino acids. This results in alternative glycine conformations that are dictated by their own local

164   electron density. After sampling, we select combinations of conformers from a pool of candidates

165   based on both sidechain and backbone O atoms for all amino acids, including glycines. This addition

166   results in greater potential to discover alternative conformations throughout the protein and include

167   additional conformational heterogeneity in the final multiconformer model.

168

169   Characterizing peptide flip geometry

170   The nullspace inverse kinematics procedure of qFit [9] naturally encodes backrub [14], crankshaft [28],

171   and shear [29, 30] motions (**Figure S1**) where they are dictated by the anisotropy of the electron

172   density for the Cβ atom. However, this anisotropy cannot identify more discrete substates of the

173   backbone, such as peptide flips. Peptide flips are large, ~180° rotations of a peptide plane in protein

174   backbone with minimal disturbance of adjacent peptide conformations. Enumerating many peptide flip

175   candidate conformations with the nullspace inverse kinematics procedure would quickly lead to

176   prohibitively large sample sizes. We therefore examined common geometries of discrete peptide flips to

177   expedite sampling of discrete backbone substates in qFit 2.0.

178
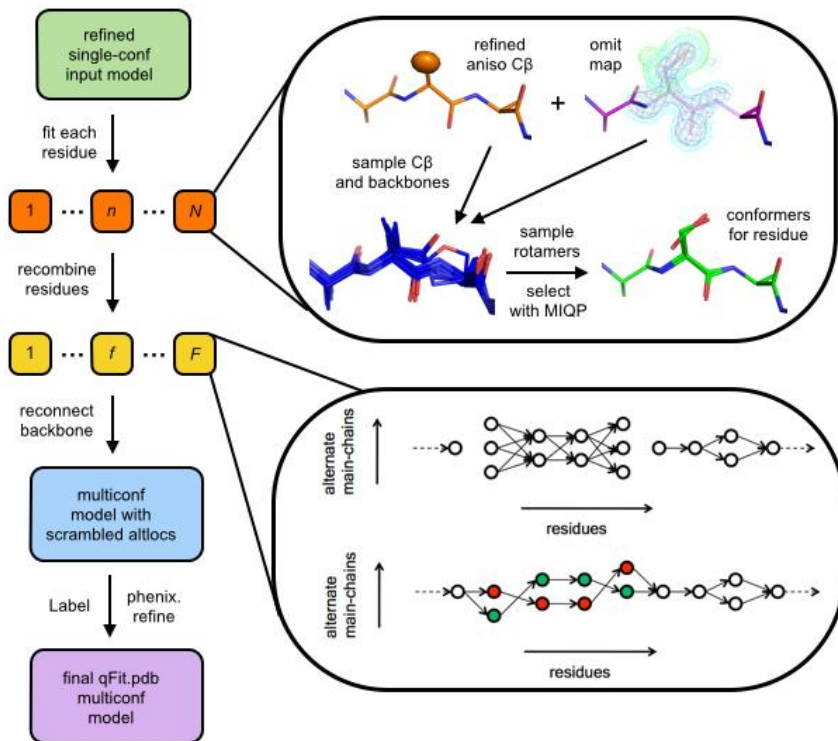
179

180 **Figure 1: Flowchart of the qFit 2.0 algorithm.** qFit can operate on each residue in the protein (orange boxes) in

181 parallel (1 ≤ n ≤ N indices are for residues in the protein). Anisotropic refinement gives a thermal ellipsoid for the

182 Cβ (orange model), and refinement with occupancies set to 0 gives an omit map (purple model). These inputs

183 are combined, backbone translations and peptide flips are sampled (blue models), each backbone is decorated

184 with sidechain rotamers, and an MIQP is used to select 1-4 conformations for the residue. Residues with

185 consecutive multiple backbone conformations, called fragments (yellow boxes), are then subjected to a second

186 MIQP to trace compatible alternative backbone conformations across residues. Residues and fragments are

187 combined into an intermediate model. Finally, a Monte Carlo procedure is used to adjust alternative conformation

188 labels ("altloc" identifiers) to minimize steric overlaps, and the final model is refined.

189

190 Steric interactions prevent arbitrary rotations of the peptide plane, much like sidechains adopt preferred

191 rotamer conformations. To identify plausible geometries for peptides relative to a single input peptide,

192 we examined cases where the peptide rotates by 90-180° around the Cα-Cα axis. We identified 147

193 peptide flips modeled as alternative conformations in high-quality structures. After filtering this set of

194 peptide flips with structure validation criteria and reserving some examples for a test set, we retained

195 79 examples that clustered around four geometries (**Table S1**). We observed that peptide flips often

196 included rotation and translation within the peptide plane such that the first Cα moves "below" the Cα-

197 Cα axis and the second Cα moves "above" it (from the view in **Figure 2A,C**). These in-plane

198 movements justify sampling geometries found in natural peptide flips in qFit 2.0 rather than, e.g., simply

199 rotating the peptide 180° around the Cα-Cα axis. The first two clusters, "simple down" (**Figure 2A,C**,

200 blue) and "tweaked down" (**Figure 2A,C**, red), feature a very nearly 180° rotation around the Cα-Cα

201   axis, but with different in-plane adjustments. By contrast, the second two clusters, "left" (**Figure 2B,D**,

202   green) and "right" (**Figure 2B,D**, brown), feature rotations closer to 120°, but in opposite directions.

203   Our dataset here is sufficient to propose plausible, well-validated peptide flip geometries for sampling in

204   qFit 2.0, and suggests that the four clusters could also be used to inspire moves in protein design.

205

Cluster	# of examples	# (%) in tight turn	# (%) with Gly as first residue	# (%) with Gly as second residue
"tweaked down" (red)	26	13 (50%)	0 (0%)	16 (62%)
"simple down" (blue)	29	12 (41%)	4 (14%)	19 (66%)
"left" (green)	10	0 (0%)	1 (10%)	2 (20%)
"right" (brown)	14	2 (14%)	1 (7%)	3 (21%)

206   **Table 1: Peptide flip geometries aggregate into distinct clusters.** Colors refer to **Figure 2**.

207

208   <u>Structural context of flips</u>

209   We found that the two "down" clusters were more common in tight turns between β-strands: 41-50% of

210   flips in these clusters were found in turns, as compared to 0-14% for the other two flip clusters (with a

211   conservative definition of a turn; see Methods) (**Table 1**). The flip is nearly always associated with a

212   transition between Type I/I' and II/II' turns. The "left"/"right" clusters were dispersed among many

213   irregular structural contexts, but not α-helices or β-sheets. Across the four clusters, the first residue of

214   the peptide was a glycine 7.5% of the time, in line with the general abundance of glycines in proteins

215   (7-8%). However, the second residue of the peptide was a glycine significantly more frequently (50%, p

216    $< 10^{-22}$). This was true for the "left"/"right" clusters (21%, $p < 0.05$) and especially the two "down"

217   clusters (**Figure 2C**) (64%, $p < 10^{-24}$). This may be in part because a glycine as the second residue of

218   a peptide can lower the flip transition energy [31]. These results generally agree with reports of flip-like

219   conformational differences between the same tight turn in separate homologous structures [32].

220
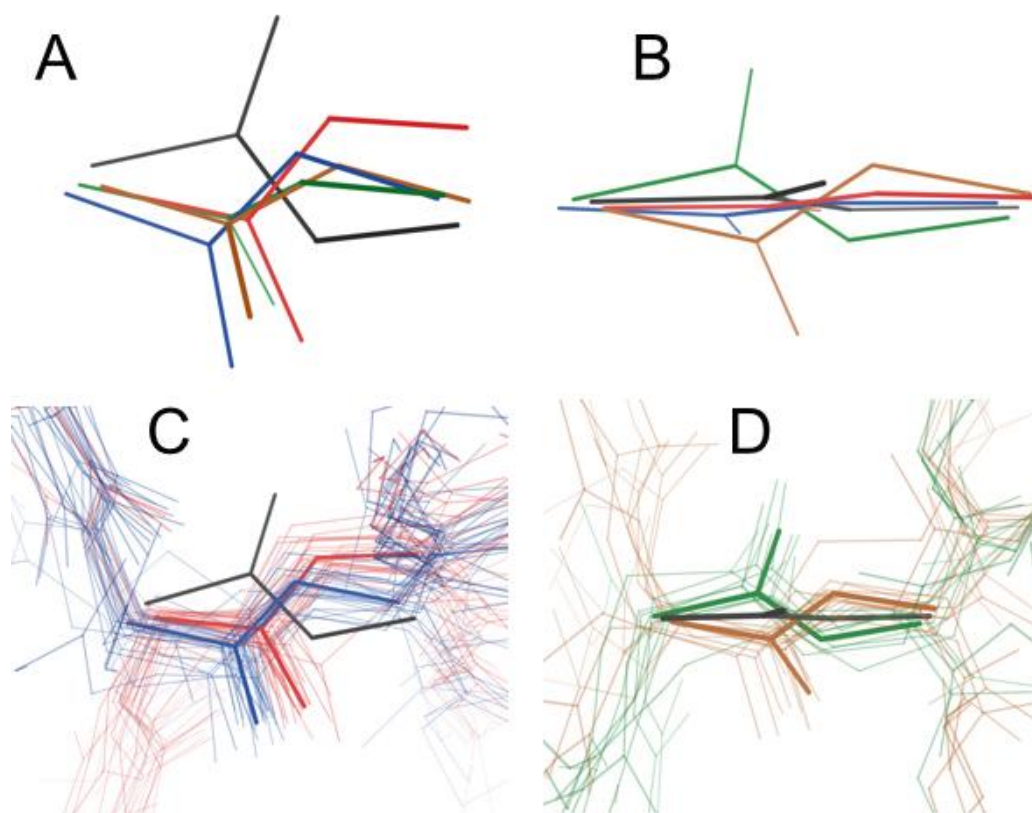
Figure 2: Geometry and distribution of peptide flips in training set. **(A,B)** Reference primary conformation peptide (black) and four cluster centroids for secondary peptide conformations (colors), from the side **(A)** or "top-down" **(B)**. **(C)** Members from the training set segregate into two ~180° rotated clusters with different translations in the peptide plan (blue vs. red). View from roughly the same angle as **(A)**. **(D)** Other members from the training set segregate into +120° and -120° rotated clusters (green vs. brown). View from roughly the same angle as **(B)**.
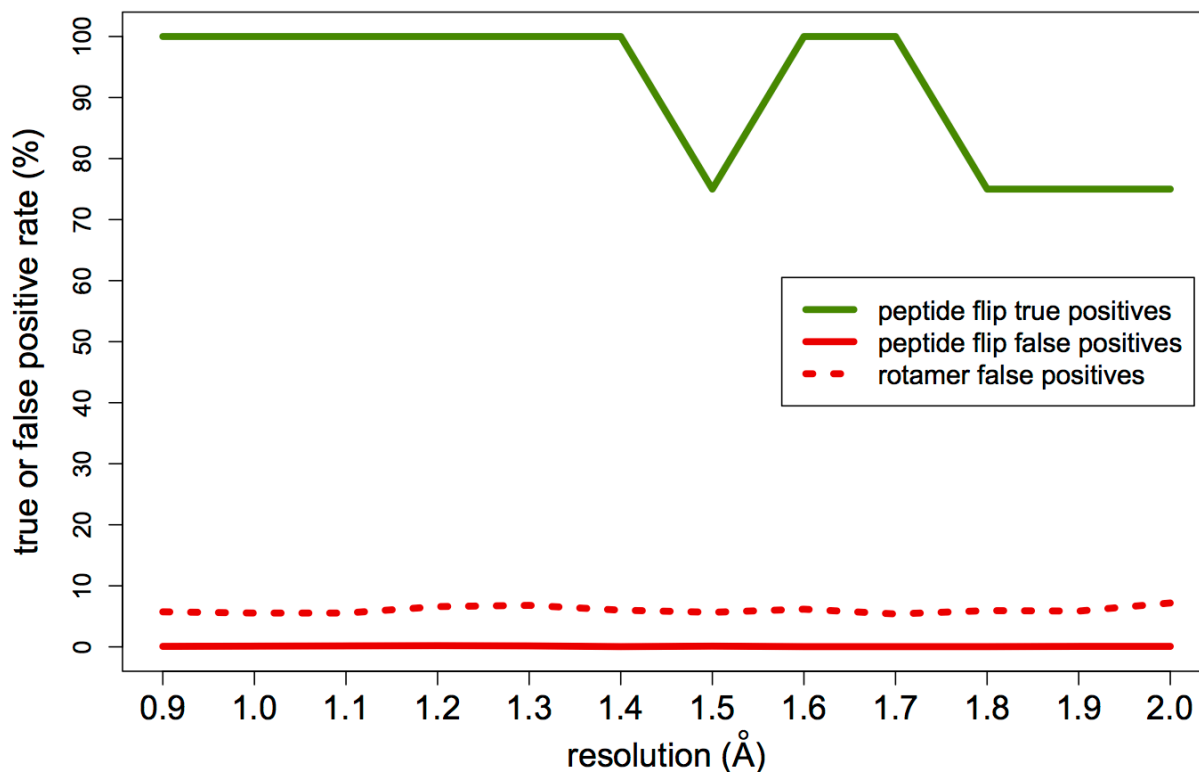
Tests with synthetic datasets

To test these advances, we first explored synthetic datasets spanning resolutions from 0.9 to 2.0 Å with increasing B-factors as a function of resolution and Gaussian noise added to structure factors (see Methods). We used the Top8000 peptide flip geometry cluster centroids, with the alternative conformations at 70/30 occupancies for the "tweaked down" cluster and 50/50 occupancies for the other three clusters. Because qFit uses these geometries to sample peptide flips, we expected it would be able to successfully identify each flipped alternative conformation starting from the primary (labeled "A") conformation at high-to-medium simulated resolution, but less well at lower simulated resolution. Indeed, qFit 2.0 successfully finds the flipped conformations for most peptide flip geometry clusters across resolutions with a 92% success rate overall; this rate drops only slightly with resolution from 0.9 to 2.0 Å (**Figure 3**). Since we rebuilt the entire protein chain, we also assessed the performance on other residues. By contrast to the true positive peptide flip results, the peptide flip and rotamer false positive rates remain quite low across clusters and resolutions (**Figure 3**). These results indicate that

242   qFit 2.0 is effective at identifying peptide flip alternative conformations across a wide range of

243   crystallographic resolutions without introducing spurious conformations.

244

245

246   **Figure 3: True vs. false positives with synthetic data.** Peptide flip true positives = percent of peptide flips in

247   the actual synthetic model that are present in the qFit 2.0 model. Peptide flip false positives = percent of residues

248   with a peptide flip in the qFit 2.0 model that are not in the actual synthetic model. Rotamer false positives =

249   percent of sidechain rotamers (as defined by MolProbity [12, 33]) in the qFit 2.0 model that are not in the actual

250   synthetic model. True positives in green; false positives in red. Peptide flips in solid lines; rotamers in dotted line.

251   Data is averaged over all four synthetic datasets (corresponding to the four peptide flip geometry clusters in

252   **Figure 2**) and all three mainchain amplitudes are considered; see Methods.

253

254

255   Tests with experimental datasets

256   Although tests with synthetic datasets offer insight into resolution dependence, a more direct test of the

257   usefulness of qFit 2.0 involves crystal structures with real data. We combined structures left out of the

258   training set from the Top8000 peptide flip examples with a few more manually curated examples for a

259   total of 15 test cases (**Table 2**). When comparing qFit 2.0 models to rerefined original structures, R_{free}

260   is better for 7/15 cases and R_{work} is better for 8/15 cases (**Figure S2**). However, after rerefinement with

261   automated removal and addition of water molecules to allow the ordered solvent to respond to the new

262   protein alternative conformations modeled by qFit (see Methods), R_{free} is better for the qFit 2.0 model

263   for 10/15 cases and R_{work} is better for 13/15 cases (**Figure 4**). The differences generally are small: the

264   average ΔR_{free} is ~0.1%. Overall, these results suggest that qFit 2.0 models explain experimental

265   crystallographic data as well as or better than traditional refinement protocols at a global structural

266   level.

267

PDB ID	Resolution (Å)	Chain	Peptide	Found flip?
1w0n	0.80	A	42-43	Y
1nki	0.95	A	53-54	Y
1nki	0.95	B	53-54	Y
1x9i	1.16	A	65-66	Y
1c9o	1.17	A	36-37	Y
1c9o	1.17	B	36-37	n
2c6z	1.20	A	227-228	n
2qd6	1.28	A	50-51	n
2qd6	1.28	B	150-151	Y
3n6z	1.30	A	177-178	Y
2c0c	1.45	A	44-45	n
3h0u	1.50	C	115-116	Y
3el5	1.60	A	50-51	Y
3el5	1.60	B	50-51	Y
3ii7	1.63	A	539-540	Y
3ie5	1.69	A	61-62	Y
2ozv	1.70	A	54-55	Y
3f7w	1.85	A	48-49	Y

268   **Table 2:** List of positive-control peptide flip test cases. Last column indicates whether or not qFit 2.0 found the

269   peptide flip alternative conformations for at least one of the three backbone amplitude parameters. Overall, 14/18

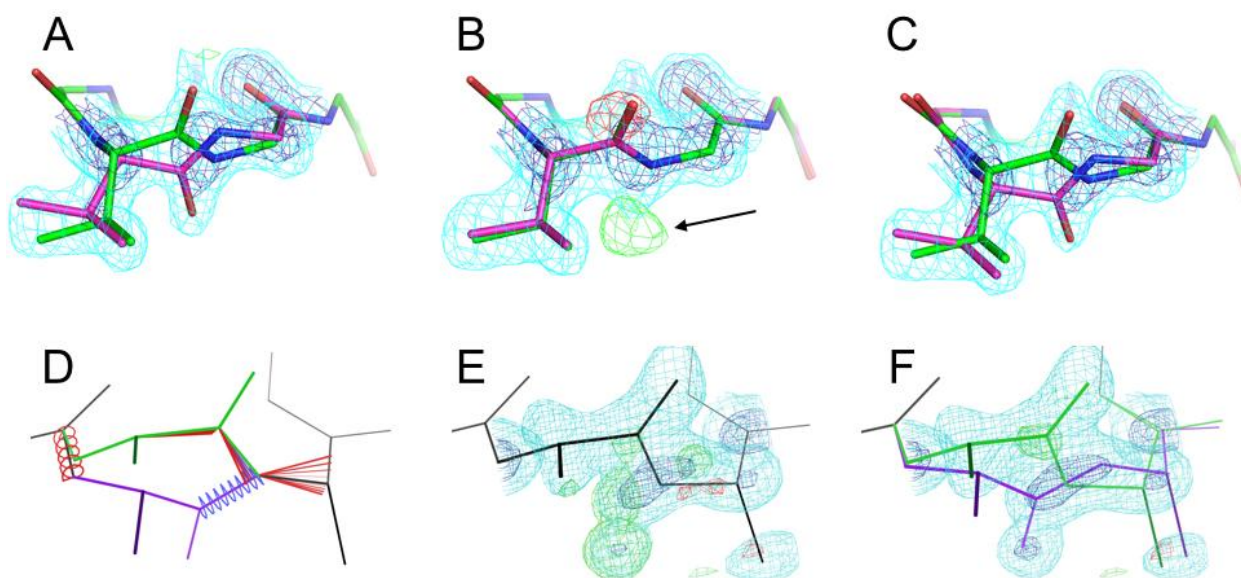270   (78%) peptide flips were successfully identified.

271

272
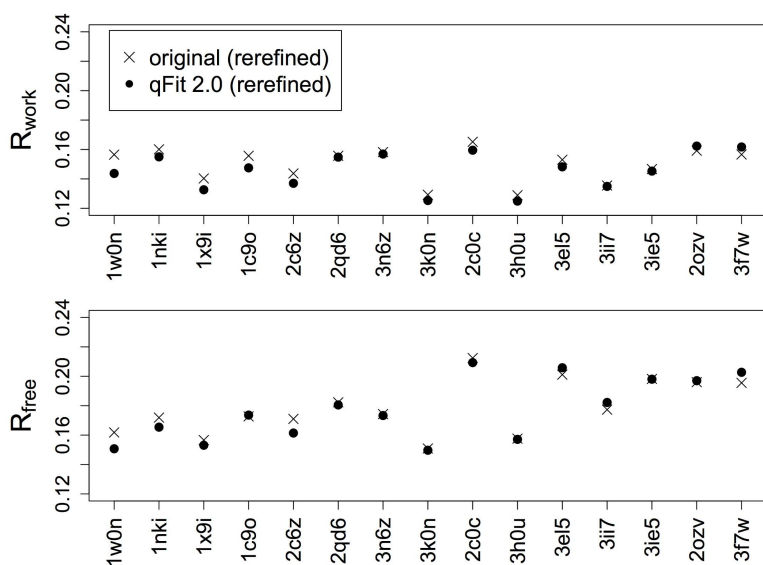
Figure 4: Multiconformer modeling with qFit results in similar or better crystallographic R-factors. R_{work} and R_{free} are plotted vs. PDB ID sorted from high to low resolution. X's indicate rerefined original structures and filled circles indicate qFit 2.0 models; both are after refinement with water picking.

276

277   While global metrics are important, a major focus of the current work is correctly identifying local

278   alternative backbone conformations. To explore this aspect, we compared results from qFit 2.0 to those

279   from qFit 1.0 and original deposited structures for our test set (**Table 2**). qFit 2.0 successfully models

280   both flipped conformations in 14/18 (78%) cases. For example, Val539-Gly540 in the Kelch domain of

281   human KLHL7 is modeled with two alternative conformations related by a peptide flip (1.63 Å, PDB ID

282   3ii7) (**Figure 5A**). qFit 1.0 fails to discover the flip, resulting in significant difference electron density

283   peaks (**Figure 5B**). By contrast, qFit 2.0 beautifully recovers both alternative conformations (**Figure**

284   **5C**). In another example, Asn42-Gly43 in carbohydrate binding domain 36 at high resolution (0.8 Å,

285   PDB 1w0n) adopts flipped peptide conformations -- yet MolProbity flags geometry errors in the

286   deposited structure that indicate it re-converges too quickly, with alternative conformations for only the

287   Asn42 and not also Gly43 (**Figure 5D**). qFit 1.0 fails to capture the flip (**Figure 5E**). However, qFit 2.0

288   not only identifies both peptide flip conformations for Asn42, but also includes split conformations for

289   Gly43, thereby repairing the covalent backbone geometry (**Figure 5F**). In both cases, the peptide flip

290   and glycine sampling enhancements in qFit 2.0 combine to model discrete backbone heterogeneity as

291   accurately as or even better than the original structure.

292

293

294 **Figure 5: qFit 2.0 successfully identifies known peptide flips. (A-C)** Val539-Gly540 in the Kelch domain of

295 human KLHL7 at 1.63 Å (PDB ID 3ii7). 2mFo-DFc electron density is contoured at 1.2 σ (cyan) and 2.5 σ (blue);

296 mFo-DFc electron density is contoured at +3.0 σ (green) and -3.0 σ (red). **(A)** The deposited model includes

297 alternative conformations for this peptide, which are well justified by the electron density. **(B)** qFit 1.0 starting

298 from single-conformer input fails to find the second conformation, resulting in peaks in the difference density map

299 (arrow). **(C)** qFit 2.0 finds both conformations, resulting in the disappearance of the difference peaks. **(D-E)**

300 Asn42-Gly43 in carbohydrate binding domain 36 at 0.8 Å resolution (PDB ID 1w0n). The Asn42 sidechain (left,

301 darker green/purple) points up out of the image so is visually truncated. In **(E-F)**, 2mFo-DFc electron density is

302 contoured at 1.5 σ (cyan) and 2.5 σ (blue); mFo-DFc electron density is contoured at +3.0 σ (green) and -3.0 σ

303 (red). **(D)** The deposited structure includes alternative conformations (green and purple) related by a peptide flip,

304 but re-converges too early at the Gly43 backbone N atom, resulting in >4 σ bond length (red and blue fans) and

305 bond angle (red and blue springs) outliers [33]. **(E)** qFit 1.0 fails to identify the flip, leaving significant difference

306 density map features. **(F)** qFit 2.0 identifies the flip at Asn43 and also correctly splits Gly43 into separate

307 conformations, thereby flattening the difference map relative to qFit 1.0 and eliminating the covalent geometry

308 errors in the original structure.

309

310 <u>Discovering new conformational heterogeneity from experimental datasets</u>

311 In addition to retrospective positive-control tests, we also looked prospectively for "hidden" peptide flip

312 alternative conformations that are unmodeled in existing structures. One such example is Met519-

313 Thr520 in RNA binding protein 39. In chain A of the room-temperature structure (PDB ID 4j5o), the

314 mFo-DFc difference electron density map around this peptide has significant positive and negative

315 peaks, indicating it is mismodeled as a single conformation (**Figure 6A**). Other instances of this

316 peptide -- including in chain B of the room-temperature structure and both chains of the cryogenic

317 structure -- feature conformational diversity, much of which may be related to crystal contacts; however,

318 these conformations fail to account for the room-temperature chain A mFo-DFc peaks (**Figure 6B**).

319 However, using the room-temperature data, qFit 2.0 identifies a peptide flip in this region, which

320   repositions Met519 and flattens the local difference density (**Figure 6C**). By contrast, it does not

321   identify a peptide flip for this region in either chain using the cryogenic data, which is in accord with

322   previous reports that cryocooling crystals can conceal or otherwise perturb conformational

323   heterogeneity that is present at room temperature [25, 26].

324
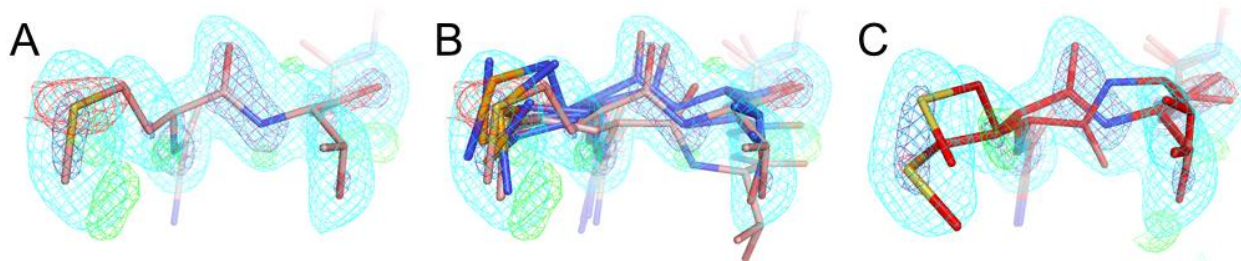
325

326   **Figure 6: qFit 2.0 finds a hidden peptide flip at room temperature. (A)** Met519-Thr520 in mouse RNA binding

327   protein 39 (RBM39) is modeled with just a single conformation in chain A of the 1.11 Å room-temperature

328   structure (PDB ID 4j5o, pink). There appear to missing unmodeled conformations based on mFo-DFc difference

329   electron density contoured at +3.0 σ (green) and -3.0 σ (red). 2mFo-DFc electron density is shown contoured at

330   0.9 σ (cyan) and 2.5 σ (dark blue). **(B)** Although there is diversity for this region in chain B of the asymmetric unit

331   from this structure and in chains A and B from the 0.95 Å cryogenic structure (PDB ID 3s6e, blue), none of the

332   other instances explain the electron density at RT in chain A. There is also no clear evidence for missing

333   alternative conformations in these other instances (not shown). **(C)** The RT qFit model (magenta) adds a flipped

334   peptide as an alternative conformation, which positions the Met519 sidechain differently. Collectively, these

335   changes better explain the local electron density.

336

337   In addition to selection of conformers based on fit to density for the backbone O atom for all amino

338   acids, qFit 2.0 also adds sampling based on this atom for glycine, enabling density-driven backbone

339   sampling for the most flexible amino acid. This facilitates modeling peptide flips in which one of the

340   constituent residues is a glycine, as seen in the examples above (**Figure 5**) -- but also opens the door

341   to modeling less discrete glycine flexibility. For the 489 glycines across the 15 datasets in the test set

342   (**Table 2**), qFit 1.0 cannot model more than a single conformation, but qFit 2.0 models alternative

343   conformations for 365/489 (75%) of glycines. The Cα displacements average 0.28 Å and range from

344   <0.01 Å up to 1.70 Å. Only 4 (4%) of these glycines were modeled with alternative conformations in the

345   original PDB structures. These results show that the direct sampling and selection based on electron

346   density for glycine backbone atoms in qFit 2.0 successfully identify conformational heterogeneity that

347   was formerly unrecognized. For example, a small, glycine-rich loop in PDB ID 3ie5 is modeled with a

348   single conformation in the deposited structure and qFit 1.0 model (**Figure 7A**). By contrast, qFit 2.0

349   recognizes the anisotropy of the electron density for each of the three glycine O atoms in the loop, so

350   models them with alternative conformations that collectively shift the entire mini-loop region (**Figure**

351   **7B**).

352
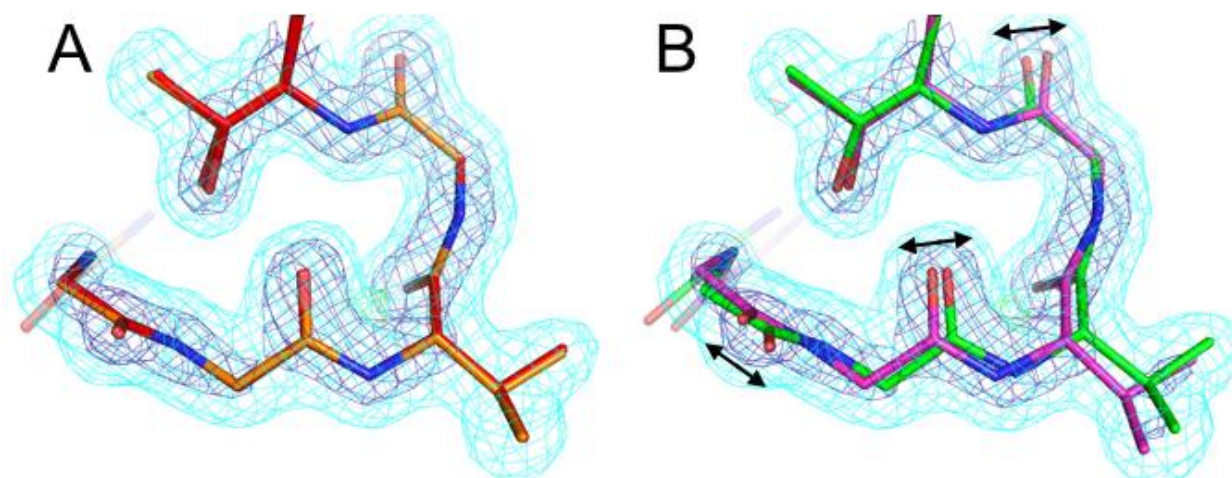
353

354 **Figure 7: qFit 2.0 identifies alternative glycine conformations.** This small loop in the 1.69 Å structure of Hyp-

355 1 protein from St. John's wort (PDB ID 3ie5) includes several glycines: 49, 50, and 52. **(A)** The deposited

356 structure (orange) depicts these glycines with single conformations. The qFit 1.0 model (red) does the same,

357 because it cannot sample alternative glycine conformations. **(B)** The qFit 2.0 model identifies alternative

358 conformations (green/purple) for the entire loop, including all three glycines, based on subtly anisotropic

359 backbone O atoms (arrows). 2mFo-DFc electron density contoured at 1.0 σ (cyan) and 3.0 σ (blue); mFo-DFc

360 electron density contoured at +3.0 σ (green) and -3.0 σ (red).

361

362 Selecting conformers based on fit to density for the backbone O atom helps find alternative

363 conformations not only for glycines, but also more generally for other amino acids. In many cases, this

364 additional data-driven aspect to conformer selection drives the identification of subtle, non-discrete

365 backbone motions that are coupled to larger, discrete sidechain changes. Indeed, for the 15 proteins in

366 **Table 2**, qFit 2.0 shifts the Cα more than does qFit 1.0 for 52% of residues, but the reverse is true for

367 only 20% of residues (the remaining residues are not moved by either version) (**Figure 8A**).

368 Furthermore, for 63% of the residues for which qFit 2.0 finds a new sidechain rotamer that qFit 1.0 does

369 not, qFit 2.0 also moves the Cα more (**Figure 8B**). These results imply that the backbone sampling by

370 qFit 2.0 not only increases backbone heterogeneity in and of itself, but also drives discovery of

371 sidechain conformational heterogeneity. As one specific example, Thr157 in cyclophilin A is modeled

372 with alternative backbone and rotamer conformations in the deposited structure (**Figure 8A**). qFit 1.0

373 fails to find the alternative rotamer because it maintains a single backbone conformation (**Figure 8B**),

374 but, driven by carbonyl O anisotropy, qFit 2.0 identifies the alternative backbone conformations,

375 allowing it to discover the second rotamer (**Figure 8C**).

376

377
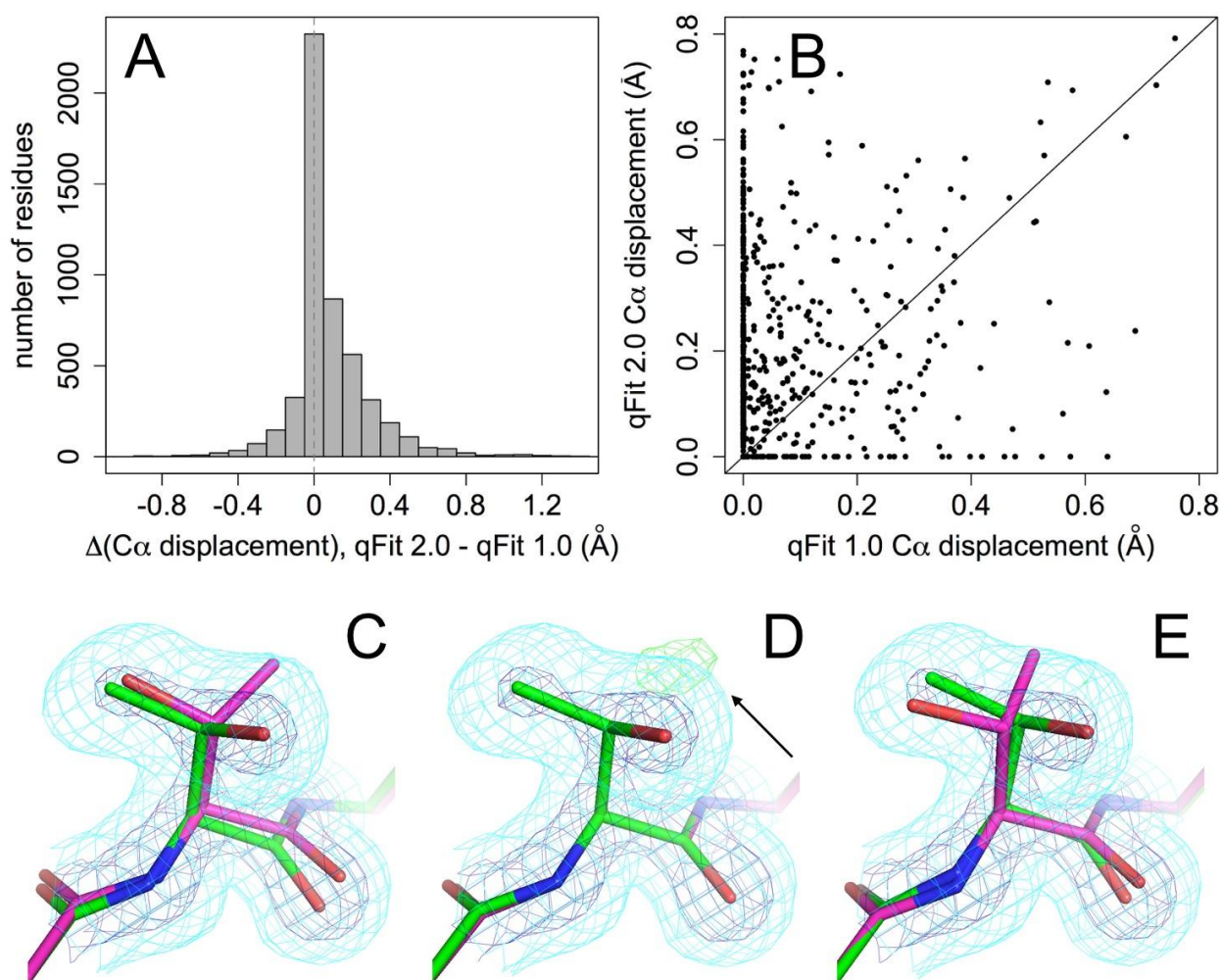
**Figure 8: Extra backbone heterogeneity in qFit 2.0 helps discover new sidechain heterogeneity.**

(A) Histogram of difference in maximum Cα displacement across all combinations of alternative conformations between qFit 2.0 and qFit 1.0 for the test set. Vertical dotted line at 0 difference. **(B)** Maximum Cα displacements for qFit 2.0 vs. 1.0 for residues with a newly discovered sidechain rotamer in the qFit 2.0 model but not in the qFit 1.0 model. Many of these residues fall above the diagonal line, meaning the Cα moves more in the qFit 2.0 model than in the qFit 1.0 model. **(C-D)** Thr157 in cyclophilin A at room temperature (PDB ID 3k0n). 2mFo-DFc electron density is contoured at 0.5 σ (cyan) and 3.0 σ (blue); mFo-DFc difference electron density is contoured at +3.1 σ (green) and -3.1 σ (red). **(C)** The deposited structure has alternative rotamers that were correctly manually modeled. **(D)** qFit 1.0 does not move the backbone and misses the alternative rotamer, as evidenced by a peak of +mFo-DFc density (arrow). **(E)** qFit 2.0 does move the backbone (note especially the backbone carbonyl displacement), and successfully identifies the alternative rotamer.

389

Newly identified peptide flips in the "flap" region of HIV protease

We also observed hidden peptide flips for the Ile50-Gly51 tight turn in the "flap" region of HIV-1 protease. HIV-1 protease is a homodimer, with residue numbers often denoted by 1-99 and 1'-99'. The flap region consisting of residues 46-56 is an antiparallel β-sheet and tight turn at the interface of the dimer (**Figure 9A**). In most of the hundreds of crystal structures of HIV-1 protease, the two tight turns

395 (Leu50-Gly51 and Leu50'-Gly51') adopt an asymmetric conformation, with one flap in a single type I

396 conformation and the other in a single type II conformation. However, NMR relaxation data suggest

397 that these flips can undergo chemical exchange on a slow (~10 µs) timescale in solution [34].

398 Mutational data also linked collective conformational exchanges of these flips to catalytic rates [35]. In

399 line with these solution studies, we noticed that for many HIV-1 protease crystal structures, the electron

400 density maps actually reveal strong evidence for alternative conformations related by dual peptide flips.

401 For example, in one high-resolution inhibitor-bound structure (PDB ID 3qih), the Leu50-Gly51 and

402 Leu50'-Gly51' flaps are modeled with single asymmetric conformations, but strong positive mFo-DFc

403 electron density coincides with potentially flipped states (**Figure 9B**). Strikingly, qFit 2.0 automatically

404 identifies dual "flap flips", suggesting the flaps actually populate two different asymmetric states (green

405 vs. purple in **Figure 9C**) in this particular inhibitor complex. More generally, this result suggests that

406 these inhibitor-gating flaps in HIV-1 protease sample multiple conformations more often than previously

407 recognized across many inhibitor complexes, which may motivate further investigation of the effects

408 that protein and inhibitor flexibility have on binding affinity, efficiency of catalytic inhibition, and arisal of

409 drug resistance in this biomedically important target.

410
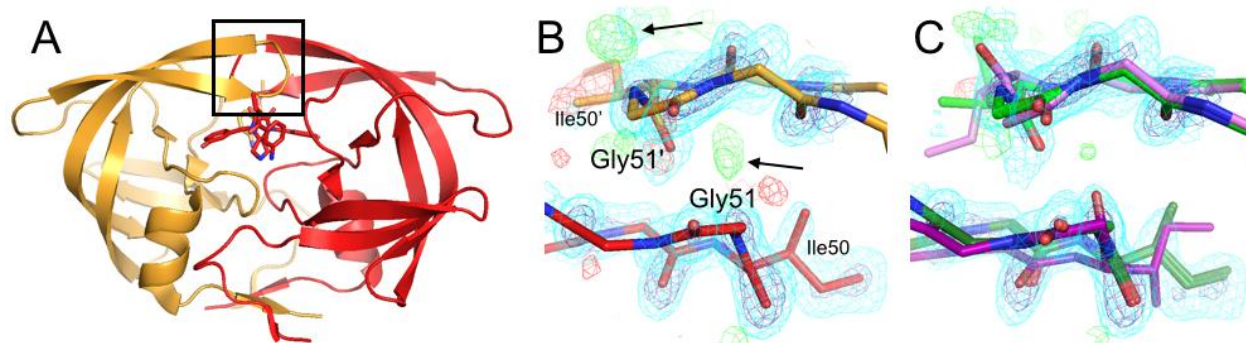
411

412 **Figure 9: Hidden unmodeled peptide flips in the inhibitor-gating "flaps" of HIV-1 protease. (A)** In the 1.39

413 Å structure of a mutant of HIV-1 protease bound to a novel inhibitor (PDB ID 3qih), the Ile50-Gly51 tight turn

414 interacts with the dimer-related copy of itself, Ile50'-Gly51' (boxed region). Chain A in orange, chain B in red. The

415 inhibitor (sticks) binds in two overlapping poses immediately adjacent to these flaps. **(B)** This dimer interface,

416 viewed as if from above in **(A)**, is asymmetric in the deposited structure: both copies of the peptide point

417 downwards in this view. However, positive difference electron density (arrows) suggest unmodeled

418 conformations. **(C)** qFit 2.0 models this region with coupled asymmetric peptide flips, such that both copies of the

419 peptide point down (~70%, green) or both point up (~30%, purple) in this view. The multiconformer model has

420 diminished difference electron density peaks, suggesting it is a better local fit to the data. Residual difference

421 peaks may reflect unmodeled partial-occupancy waters that are mutually exclusive with the new protein

422 alternative conformations. 2mFo-DFc contoured at 1.2 σ (cyan) and 3.0 σ (blue); mFo-DFc contoured at +3.0 σ

423 (green) and -3.0 σ (red).

424

425

426 **Discussion**

427 The ruggedness of protein energy landscapes leads to conformational heterogeneity even in folded

428 globular proteins. Evidence for these alternative conformations is remarkably prevalent in high-

429 resolution (<2 Å) crystallographic electron density maps [6]. However, because these alternative

430 conformations are difficult and/or time-consuming to model manually using existing graphics and

431 refinement tools, they are underrepresented in the PDB [6]. qFit is a computational approach to

432 overcoming these problems, by automatically identifying "hidden" alternative conformations and using

433 quadratic programming to select a parsimonious subset that collectively best explains the diffraction

434 data. Here we have demonstrated a new version of this algorithm, called qFit 2.0, with several

435 enhancements to handling flexible backbone -- most notably, automated detection of discrete peptide

436 flips and explicit fitting of backbone atoms for glycines.

437

438 qFit has previously captured different types of backbone motion that can occur in secondary structure.

439 For example, it correctly identifies the backrub motion [14] that helps Ser99 transition between

440 sidechain rotamers in the active-site β-sheet network of CypA [15, 16], and also identifies a previously

441 hidden α-helix winding/unwinding or "shear" motion [14, 29] (**Figure S1**). However, qFit 2.0 can now

442 model larger backbone motions in which the backbone change itself is discrete, instead of inherently

443 continuous but coupled to discrete sidechain rotamer changes. Specifically, it models peptide flips,

444 which occur outside of helices and sheets and involve discrete jumps over a larger energetic barrier.

445

446 Peptide flips have important implications for understanding protein function. For example, our results for

447 HIV-1 protease (**Figure 9**) strongly suggest that conformational heterogeneity, in particular peptide

448 flips, may play underappreciated roles in protein-inhibitor complexes. Previously, molecular dynamics

449 simulations identified a large-scale "curling" motion of these flaps that is maintained by drug-resistance

450 mutations and therefore seems important for substrate access [36]. Although this motion is more

451 dramatic than the peptide flaps at the tips of the flaps that we observe, it underlines that flap flexibility --

452 potentially across multiple length scales -- is central to protease function and viral propagation. The

453 peptide flip acts as a key conformational switch between type I/II turns, rearranging its environment

454 beyond its immediate sequence neighbors and enabling alternative sidechain conformations with

455 implications for function. However, the large number of unmodeled turns in HIV protease structures

456 illustrates the challenge of distinguishing alternative conformations in electron density maps, even at

457 high resolution. As an additional example which unfortunately lacks deposited structure factors, the

458 active-site Gly57-Asp58 peptide in *C. beijerinckii* flavodoxin adopts distinct peptide flip states in concert

459 with the oxidation state of the FMN prosthetic group [19]. The N137A mutation removes artificial lattice

460 contacts that otherwise influence the conformation of the Gly57-Asp58 peptide, which results in a

461 mixture of these peptide conformations simultaneously populated in the crystal; this suggests these

462 multiple flip states may also coexist in solution [19].

463

464 Beyond the specific improvements to peptide flips, qFit 2.0 now fits conformations for each residue

465 based on both sidechain (beyond Cβ) and backbone (carbonyl O) atoms. Although we originally

466 envisioned this change for modeling glycines, we observed that it results in dramatically more extensive

467 backbone conformational heterogeneity across the protein (**Figure 8**). R-factors are similar or better

468 (**Figure 4**), suggesting the new models with more heterogeneity are at least as good an explanation of

469 the experimental data. Notably, these new backbone shifts drive discovery of many more alternative

470 sidechain rotamers (**Figure 8**). Our results suggest that sidechain and backbone degrees of freedom in

471 proteins are tightly coupled, in agreement with previous reports that even subtle backbone motions can

472 facilitate rotamer changes [14], open up breathing room for natural mutations [37], and expand

473 accessible sequence space in computational protein design [30, 38].

474

475 Future work will investigate an armamentarium of methods for modeling larger backbone

476 conformational change in qFit, including helix shear motions [29], adjustments of entire α-helices [39,

477 40], correlated β-sheet flexing [28], automated loop building algorithms such as Xpleo [9], and pre-

478 knowledge of conformational differences between homologous structures. While these future steps will

479 move us closer to capturing the full hierarchy of protein conformational substates [41], they will also

480 dramatically increase the computational cost of automated multiconformer model building. Many

481 aspects of qFit are parallelizable; however, the total computational cost for reproducing the data in this

482 manuscript is approximately 10^5 CPU hours. As cloud-computing capabilities of 10^8 CPU hours can

483 now be leveraged for pure simulation data [42], we envision that marshalling similar computational

484 capabilities will become increasingly important for analysis of experimental X-ray data. Such data-

485 driven computational approaches to studying the dynamic relationship between protein structure and

486 function will be especially powerful when applied to series of datasets in which the protein is subjected

487 to perturbations that modulate conformational distributions, such as ligand binding or temperature

488 change [23].

489

490 **Materials and Methods**

491

492 Learning peptide flip geometries

493

494 To define possible relative geometries between flipped peptide conformations, we searched for

495 trustworthy peptide flips modeled as alternative conformations in the Top8000 database. This database

496 contains ~8000 (7957) quality-filtered protein chains from high-resolution crystal structures, each with

497 resolution < 2 Å, MolProbity score [33] < 2, nearly ideal covalent geometry, and <70% sequence

498 identity to any other chain in the database [43 ISBN: 978-981-4449-15-1]. We searched the Top8000

499   for peptides with carbonyl C-O bonds pointed away from each other (O-O distance > C-C distance + 1
500   Å) and rotated by at least 90°, and for which both flanking Cα atoms reconverged to < 1.5 Å. Although
501   peptide rotations of < 90° also occur, they occur more often in irregular loop regions, have less well-
502   converged backbone for flanking residues, and are generally more diverse and difficult to simply
503   categorize. By contrast, in this study we investigate the class of localized peptide rotations with well-
504   converged backbone for both flanking residues. These are either very small rotations, or large flips with
505   a rotation nearer to 180° -- the latter being the focus here. To identify test cases for qFit 2.0, we
506   curated the resulting dataset by removing examples with more than two alternative peptide
507   conformations; a *cis* rather than *trans* conformation for either state; or obvious errors based on steric
508   clashes, strained covalent geometry, or torsional outliers from MolProbity [33]. This resulted in 104
509   examples, from which we kept a randomly selected 79 for a geometry training set (**Table S1**). We
510   combined a subset of the remaining 25 peptide flips with a few other known examples for a test set of
511   18 examples (**Table 1**). The resolution range is 0.92-1.95 Å for the training set and 0.80-1.85 Å for the
512   test set.

514   Next we characterized the geometry of peptide flips by clustering the coordinates of the flipped
515   alternative conformation (labeled "B") in the training set after superimposing onto a reference peptide.
516   We used the k-means algorithm with RMSD between the five heavy atoms of the peptide backbone
517   (Cα1, C1, O1, N2, and Cα2) for different values of k. We selected k = 4 because we observed cluster
518   centroids with approximately 180°, +120°, and -120° rotations and for k > 4 no other significantly
519   different rotations were identified. Notably, all four cluster centroids featured translations of the flanking
520   Cα atoms of >0.2 Å, and as much as >0.9 Å for one cluster ("tweaked down", red in **Figure 2**). The
521   transformation matrices relating the flipped peptide cluster centroids to the reference peptide were used
522   in qFit 2.0 to sample plausible alternative conformations, with subsequent refinement adjusting the
523   atomic positions away from the centroid geometry.

525   Tight turns and glycine enrichment

527   We defined tight turns as having a mainchain-mainchain hydrogen bond between *i*-1 carbonyl C=O and
528   *i*+2 amide N-H that was detectable by the program Probe [44]. This definition is somewhat
529   conservative; several more examples also were visually similar to tight turns. Enrichment of glycines at
530   the two positions involved in a peptide flip was assessed for different peptide flip clusters within the
531   training set relative to a large set of 337 randomly selected structures containing 6,092 total glycines
532   out of 78,094 total amino acid residues. The statistical significance of this enrichment was assessed
533   using a one-tailed Fisher's exact test based on the hypergeometric distribution [45].

535   qFit

537   *qFit part 1: Preparing each residue for qFit*

539   qFit exhaustively examines a vast number of interpretations of local electron density, and
540   deterministically selects a small ensemble that optimally explains the density. The method starts from
541   an initial single-conformer model. The occupancies of all atoms in a residue, *k*, beyond the Cβ atom
542   are set to zero with phenix.pdbtools, and the model is refined with phenix.refine. Refinement uses
543   anisotropic B-factors for all residues if the resolution is better than 1.45 Å, or just for residue *k*
544   otherwise. Finally, all atoms in residue *k* beyond the Cβ atom are removed. These steps result in two
545   inputs to qFit: (1) an omit map and (2) starting coordinates with an anisotropic tensor for the Cβ atom.

546

547 *qFit part 2: Peptide flips and backbone sampling*

548

549 Next, the peptide from residue k to $k+1$ is aligned to the centroids identified from clustering the
550 Top8000 dataset (see above). We calculate local coordinate frames for the peptide and cluster centers
551 by orthogonalizing the $C\alpha_i$-$C\alpha_{i+1}$ and $C\alpha_i$-O_i vectors and taking their cross-product. Each centroid
552 conformation is then transformed onto the starting peptide using a homogeneous coordinate
553 transformation, resulting in a candidate flipped alternative conformation.

554

555 Peptide flips do not occur in canonical secondary structure due to steric constraints, so qFit 2.0 does
556 not attempt them in helices and sheets, as detected by the CCP4 MMDB library [46]. This both avoids
557 false positives and affords a computational speedup by reducing combinatorics in the selection steps
558 (see below).

559

560 Next, for each residue k, a fragment of length 7 centered on residue k is extracted. For each candidate
561 conformation (one unflipped plus four flipped), the $C\beta$ atom is moved along the major and minor axes of
562 the ellipsoid (six total directions) by a distance determined by the ellipsoid eigenvectors and a scale
563 value provided by the user. Here we used 0.1, 0.2, and 0.3 for this scale value, and 0.05 for the
564 optional value for random additions to scale. For glycines, which lack a $C\beta$ atom, the backbone O atom
565 is used to define the anisotropic ellipsoid. To preserve the exact geometry of the fragment, we use
566 inverse kinematics to deform the fragment. The gradient of the distance function is projected onto the
567 nullspace spanned by the dihedral degrees of freedom of the fragment [9, 10]. These motions further
568 position backbone atoms to accommodate rotameric sidechain conformations.

569

570 *qFit part 3: Sidechain sampling*

571

572 For small sidechains (Ala, Asn, Asp, Cys, Gly, Iso, Leu, Pro, Ser, Thr, Val), a 40° neighborhood of each
573 rotameric χ dihedral angle, starting at -20°, is sampled in 10° increments on each of the 35 backbone
574 conformations. To avoid a combinatorial explosion, large sidechains (Arg, Glu, Gln, His, Lys, Met, Phe,
575 Trp, Tyr) are sampled hierarchically. First, the backbone and first dihedral angle are sampled similarly
576 to small sidechains. A larger neighborhood of 50° is sampled in 4.5° increments to avoid missing
577 conformations that are initially suboptimal but can accommodate better fits for subsequent χ angles.
578 This set is then subjected to the selection procedure, which returns a handful of conformations that fit
579 the density up to the Cγ atom. These selected conformations form the basis for sampling the next χ
580 angle using the same parameters. This procedure is repeated until the entire sidechain is built.

581

582 *qFit part 4: Conformer selection*

583

584 For each of the N conformations sampled for each residue, we calculate an electron density map $\boldsymbol{\rho}_i^c$.
585 We scale the observed electron density map $\boldsymbol{\rho}^o$ to $\boldsymbol{\rho}^c$. We then subject the weighted sum of $\boldsymbol{\rho}_i^c$ to a
586 quadratic program (QP) to determine a vector of occupancies w^T that minimizes the least squares
587 residuals with respect to the observed electron density:

$$\min_{w} \ \left\| \rho^o - \sum_{i} w_i \, \rho_i^c \right\|_2$$
$$\text{s.t.} \quad w_i \geq t_{d_{min}}, \ i = 1, \ldots, n$$
$$0 \leq \sum_{i} w_i \leq 1.$$

588
589   The residuals are calculated over regularly spaced voxels that are within a resolution-dependent radius
590   r of any of the sidechain (Cβ and beyond) or carbonyl O atoms. The radius r (in Å) is determined by $r =$
591   0.7 + (d - 0.6)/3.0 if d < 3.0Å, and 0.5 d if d ≥ 3.0Å, where d is the resolution in Å.
592
593   The vast number of conformations results in a system that is generally underdetermined. We therefore
594   enforce sparsity of the solution by introducing a resolution-dependent *threshold* constraint $0 < t_{dmin} \leq 1$
595   for the occupancies; i.e., $w_i > t_{dmin}$ for all i. The threshold constraint prevents overfitting by suppressing
596   arbitrarily small occupancies that model noise. Together with the constraint that the total occupancy
597   cannot exceed unity, the threshold also enforces a *cardinality* constraint; i.e., the number of non-zero
598   occupancies is bounded by the integer part of $1/t_{dmin}$. In effect, the threshold constraint enforces
599   selection of an optimal subset in the regression. Note that the two constraints imply $w_i \in \{0\} \cup [t_{dmin}, 1]$.
600   Introducing binary variables $z_i \in \{0,1\}$, we can rewrite the optimization problem as a mixed integer
601   quadratic program (MIQP):

$$\min_{w} \ \left\| \rho^o - \sum_{i} w_i \, \rho_i^c \right\|_2$$
$$\text{s.t.} \quad t_{d_{min}} z_i \leq w_i \leq z_i, \ i = 1, \ldots, n$$
$$z_i \in \{0, 1\}^n$$
$$0 \leq \sum_{i} w_i \leq 1.$$

602
603   This optimization problem belongs to the class of convex quadratic problems, for which solvers can
604   find a globally optimal solution. An MIQP is NP-hard. We therefore pre-fit conformers with QP, and
605   subject all conformations with non-vanishing occupancies to MIQP. While in theory this no longer
606   guarantees an optimal solution, practice tests on small sets of conformers did not show an effect of pre-
607   fitting.
608
609   *qFit part 5: Putting the model back together*
610
611   Assembling a consistent, multiconformer model from individually fitted residues requires two steps.
612   First, backbone heterogeneity can extend over multiple, consecutive residues, each with slightly
613   different occupancies and/or numbers of conformers. To synchronize the number of conformers and
614   their occupancies over a fragment of length K residues consisting of consecutive backbone
615   multiconformers, we again rely on conformational selection by MIQP. We enumerate all possible
616   connections between all conformations C_i of residues $i = 1, \ldots, K$ to obtain $C_f = \Pi^{K}_{i=1} C_i$
617   conformations to model this fragment. We subject all C_f conformations to the MIQP, which selects a
618   parsimonious ensemble of at most $1/t_{dmin}$ conformations based on optimal fit to the observed electron
619   density, each with identical occupancy for all atoms in the fragment. Note that C_f can be quite large,

620 even for modestly long fragments. To avoid a combinatorial explosion, for long fragments we
621 implemented a divide-and-conquer procedure that fits segments of each fragment with MIQP. The
622 fitted segments are then combinatorially recombined and again subjected to MIQP to obtain the final
623 set of conformations for the fragment. The peptide-bond geometry of the output model at this stage of
624 qFit can be distorted. A later refinement stage with phenix.refine corrects the geometry.
625
626 Second, conformations encoding collective motions are often mutually exclusive.
627 In crystal structures, each internally consistent set of residues is labeled with an alternative
628 conformation or "altloc" identifier -- a capitalized letter ("A", "B", etc.) for multiple conformations or a
629 blank space (" ") for a single conformation. However, the initial model from the preceding steps in the
630 qFit pipeline has random labels. To identify internally consistent labels, we use a simple downhill
631 Monte Carlo optimization protocol. The program Label minimizes a simple Lennard-Jones score by
632 randomly swapping labels between conformations and accepting the change if the score improves.
633 This is repeated 10,000 times per trial over 10 trials, and the model with the best score is used for
634 subsequent steps.
635
636 To finalize the model, we first refine the relabeled model with phenix.refine. Next, we remove
637 conformations that are now indistinguishable from other conformations within predicted coordinate
638 error, and reset occupancies to sum to unity for each atom. Finally, we refine again, using anisotropic
639 B-factors if the resolution is better than 1.45 Å.
640
641 Hydrogen treatment
642
643 Hydrogens were placed at nuclear positions for Label in qFit 1.0 and at electron-cloud positions for
644 Label in qFit 2.0. Correspondingly, for Label in qFit 2.0, hydrogen van der Waals radii were taken from
645 the new values in Reduce [47], which are intended to match those used in PHENIX. Hydrogens were
646 absent for all other steps in qFit, including the final refinement step; however, the user is encouraged to
647 add hydrogens to the final qFit model for their protein of interest and proceed to other analyses. Future
648 work will update programs for downstream analysis of qFit models such as CONTACT [16] to also use
649 electron-cloud instead of nuclear hydrogen positions.
650
651 Generating synthetic datasets
652
653 To generate synthetic datasets for testing qFit, we used the protein chains containing the four peptide
654 flip cluster centroids (3mcw B 101-102, 2ior A 159-160, 2g1u A 51-52, 3g6k F 172-173). We first used
655 phenix.pdbtools to convert any anisotropic B-factors to isotropic, added 10 $Å^2$ to each B-factor per Å of
656 resolution worse than the original structure's resolution to roughly simulate the general rise of B-factors
657 with resolution, and placed the chain in a P1 box that comfortably encompassed it. Next we used
658 phenix.fmodel to calculate structure factors (with the "k_sol=0.4" and "b_sol=45" bulk solvent
659 parameters, and also generating 5% R-free flags) and added 10% noise in complex space with the
660 sftools utility in CCP4 [46]. This process was repeated for every simulated resolution from 0.9 to 2.0 Å
661 with a 0.1 Å step size.
662
663 Evaluating true and false positives
664
665 qFit uses an input parameter (MC_AMPL) to scale the magnitude of movements of the Cβ (or O for
666 glycines) along the directions dictated by its thermal ellipsoid. As in previous work [10, 16, 26], we

667 explored multiple values for this parameter: 0.1, 0.2, and 0.3. For evaluating results such as true vs.
668 false positive peptide flips and rotamers here, we considered all three resulting qFit models for each
669 dataset. This is sensible because an end user of qFit 2.0 will likely reproduce this same protocol (with
670 a few MC_AMPL values) and thus have a choice of models to use for developing insights into
671 conformational heterogeneity and its connection to function. For other analyses, we used the minimum-
672 R_{free} qFit model model unless otherwise noted.
673
674 Re-refinement with water picking
675
676 To compare R-factors between the deposited models and qFit 2.0, we finalized both models with
677 phenix.refine for 10 macro-cycles using the same parameters, including the "ordered_solvent=true"
678 flag. The resulting R-factors for qFit 2.0 models are similar or slightly better (**Figure 4**).
679
680 Programs and databases
681
682 PHENIX version 1.9-1692 (the most recent official release) [48] was used for all steps of both qFit 1.0
683 and 2.0. Coordinates and structures factors were obtained from the Protein Data Bank [49]. qFit uses
684 the following libraries: IBM's ILOG CPLEX solver for QP and MIQP, which is available free of charge for
685 academic use, and LoopTK for inverse kinematics calculations [50]. qFit is implemented in parallel; it is
686 capable of sampling and evaluating conformations for each residue as an independent job on a Linux
687 cluster. We have implemented job management for qFit on both Oracle/Sun Grid Engine and LSF
688 Platform.
689
690 **Supplementary Figures**
691
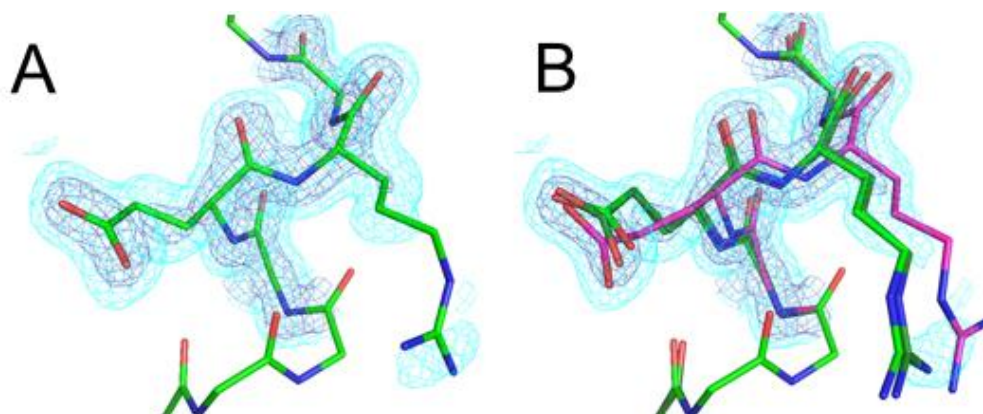
692
693 **Figure S1: qFit detects a shear backbone motion in a room-temperature crystal structure of cyclophilin A.**
694 **(A)** Residues 142-145 in CypA are modeled with a single conformation in the single-conformer structure (PDB ID
695 3k0n). The model is a reasonable fit to the 2mFo-DFc electron density contoured at 1.0 σ (cyan) and 2.5 σ (dark
696 blue), which is slightly anisotropic for the central carbonyl oxygen. **(B)** The multiconformer qFit model, on the
697 other hand, includes three alternative conformations with backbones related by a shear-like motion to explain the
698 electron density. Each shear end-state (greens vs. purple) is allocated about 50% occupancy. The
699 multiconformer model adds a second rotamer (purple) in addition to the original rotamer (greens) for Glu143 (left-
700 hand-side of panel) and sweeps the Arg144 sidechain sideways (right-hand-side of panel).
701
702
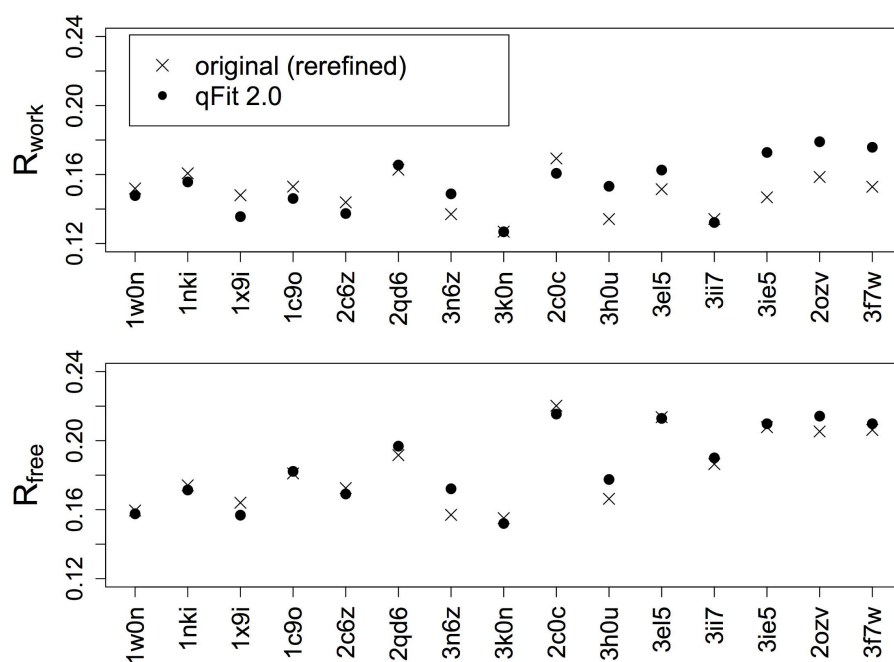
Figure S2: Multiconformer modeling with qFit does not result in better crystallographic R-factors before solvent picking. R_{work} and R_{free} are plotted vs. PDB ID sorted from high to low resolution. X's indicate original structures rerefined without automated addition and removal of water molecules, and filled circles indicate qFit 2.0 models.

Table S1: List of peptide flip examples from Top8000 used as training set.

References

1.   Smock RG, Gierasch LM. Sending signals dynamically. Science. 2009;324(5924):198-203. doi: 10.1126/science.1169377. PubMed PMID: 19359576; PubMed Central PMCID: PMC2921701.

2.   van den Bedem H, Fraser JS. Integrative, dynamic structural biology at atomic resolution-it's about time. Nat Methods. 2015;12(4):307-18. doi: 10.1038/nmeth.3324. PubMed PMID: 25825836.

3.   Woldeyes RA, Sivak DA, Fraser JS. E pluribus unum, no more: from one crystal, many conformations. Curr Opin Struct Biol. 2014;28:56-62. doi: 10.1016/j.sbi.2014.07.005. PubMed PMID: 25113271; PubMed Central PMCID: PMC4253534.

4.   Smith JL, Hendrickson WA, Honzatko RB, Sheriff S. Structural heterogeneity in protein crystals. Biochemistry. 1986;25(18):5018-27. PubMed PMID: 3768328.

5.   Kuzmanic A, Pannu NS, Zagrovic B. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. Nat Commun. 2014;5:3220. doi: 10.1038/ncomms4220. PubMed PMID: 24504120; PubMed Central PMCID: PMC3926004.

6.   Lang PT, Ng HL, Fraser JS, Corn JE, Echols N, Sales M, et al. Automated electron-density sampling reveals widespread conformational polymorphism in proteins. Protein Sci. 2010;19(7):1420-31. doi: 10.1002/pro.423. PubMed PMID: 20499387; PubMed Central PMCID: PMC2974833.

7.   Lang PT, Holton JM, Fraser JS, Alber T. Protein structural ensembles are revealed by redefining X-ray electron density noise. Proc Natl Acad Sci U S A. 2014;111(1):237-42. doi: 10.1073/pnas.1302823110. PubMed PMID: 24363322; PubMed Central PMCID: PMC3890839.

8.   Burnley BT, Afonine PV, Adams PD, Gros P. Modelling dynamics in protein crystal structures by ensemble refinement. Elife. 2012;1:e00311. doi: 10.7554/eLife.00311. PubMed PMID: 23251785; PubMed Central PMCID: PMC3524795.

9.   van den Bedem H, Lotan I, Latombe JC, Deacon AM. Real-space protein-model completion: an inverse-kinematics approach. Acta Crystallogr D Biol Crystallogr. 2005;61(Pt 1):2-13. doi: 10.1107/S0907444904025697. PubMed PMID: 15608370.

10.   van den Bedem H, Dhanik A, Latombe JC, Deacon AM. Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers. Acta Crystallogr D Biol Crystallogr. 2009;65(Pt 10):1107-17. doi: 10.1107/S0907444909030613. PubMed PMID: 19770508; PubMed Central PMCID: PMC2756169.

11.   Lovell SC, Davis IW, Arendall WB, 3rd, de Bakker PI, Word JM, Prisant MG, et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins. 2003;50(3):437-50. doi: 10.1002/prot.10286. PubMed PMID: 12557186.

12.   Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins. 2000;40(3):389-408. PubMed PMID: 10861930.

13.   Sharp KA, O'Brien E, Kasinath V, Wand AJ. On the relationship between NMR-derived amide order parameters and protein backbone entropy changes. Proteins. 2015;83(5):922-30. doi: 10.1002/prot.24789. PubMed PMID: 25739366; PubMed Central PMCID: PMC4400257.

14.   Davis IW, Arendall WB, 3rd, Richardson DC, Richardson JS. The backrub motion: how protein backbone shrugs when a sidechain dances. Structure. 2006;14(2):265-74. doi: 10.1016/j.str.2005.10.007. PubMed PMID: 16472746.

15.   Fraser JS, Clarkson MW, Degnan SC, Erion R, Kern D, Alber T. Hidden alternative structures of proline isomerase essential for catalysis. Nature. 2009;462(7273):669-73. doi: 10.1038/nature08615. PubMed PMID: 19956261; PubMed Central PMCID: PMC2805857.

16.   van den Bedem H, Bhabha G, Yang K, Wright PE, Fraser JS. Automated identification of functional dynamic contact networks from X-ray crystallography. Nat Methods. 2013;10(9):896-902. doi: 10.1038/nmeth.2592. PubMed PMID: 23913260; PubMed Central PMCID: PMC3760795.

17.   Jones TA, Zou JY, Cowan SW, Kjeldgaard M. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A. 1991;47 (Pt 2):110-9. PubMed PMID: 2025413.

18.   Kleywegt GJ, Jones TA. Efficient rebuilding of protein structures. Acta Crystallogr D Biol Crystallogr. 1996;52(Pt 4):829-32. doi: 10.1107/S0907444996001783. PubMed PMID: 15299648.

764   19.   Ludwig ML, Pattridge KA, Metzger AL, Dixon MM, Eren M, Feng Y, et al. Control of oxidation-
765   reduction potentials in flavodoxin from Clostridium beijerinckii: the role of conformation changes.
766   Biochemistry. 1997;36(6):1259-80. doi: 10.1021/bi962180o. PubMed PMID: 9063874.
767   20.   Milner-White JE, Watson JD, Qi G, Hayward S. Amyloid formation may involve alpha- to beta
768   sheet interconversion via peptide plane flipping. Structure. 2006;14(9):1369-76. doi:
769   10.1016/j.str.2006.06.016. PubMed PMID: 16962968.
770   21.   Richardson JS, Richardson DC. Amino acid preferences for specific locations at the ends of
771   alpha helices. Science. 1988;240(4859):1648-52. PubMed PMID: 3381086.
772   22.   Yan BX, Sun YQ. Glycine residues provide flexibility for enzyme active sites. J Biol Chem.
773   1997;272(6):3190-4. PubMed PMID: 9013553.
774   23.   Keedy DA, Kenner LR, Warkentin M, Woldeyes R, Thompson MC, Brewster AS, et al. Mapping
775   the Conformational Landscape of a Dynamic Enzyme by XFEL and Multitemperature
776   Crystallography2015 2015-01-01 00:00:00.
777   24.   Arnlund D, Johansson LC, Wickstrand C, Barty A, Williams GJ, Malmerberg E, et al. Visualizing
778   a protein quake with time-resolved X-ray scattering at a free-electron laser. Nat Methods.
779   2014;11(9):923-6. doi: 10.1038/nmeth.3067. PubMed PMID: 25108686; PubMed Central PMCID:
780   PMC4149589.
781   25.   Fraser JS, van den Bedem H, Samelson AJ, Lang PT, Holton JM, Echols N, et al. Accessing
782   protein conformational ensembles using room-temperature X-ray crystallography. Proc Natl Acad Sci U
783   S A. 2011;108(39):16247-52. doi: 10.1073/pnas.1111325108. PubMed PMID: 21918110; PubMed
784   Central PMCID: PMC3182744.
785   26.   Keedy DA, van den Bedem H, Sivak DA, Petsko GA, Ringe D, Wilson MA, et al. Crystal
786   cryocooling distorts conformational heterogeneity in a model Michaelis complex of DHFR. Structure.
787   2014;22(6):899-910. doi: 10.1016/j.str.2014.04.016. PubMed PMID: 24882744; PubMed Central
788   PMCID: PMC4082491.
789   27.   Fonseca R, Pachov DV, Bernauer J, van den Bedem H. Characterizing RNA ensembles from
790   NMR data with kinematic models. Nucleic Acids Res. 2014;42(15):9562-72. doi: 10.1093/nar/gku707.
791   PubMed PMID: 25114056; PubMed Central PMCID: PMC4150802.
792   28.   Fenwick RB, Orellana L, Esteban-Martin S, Orozco M, Salvatella X. Correlated motions are a
793   fundamental property of beta-sheets. Nat Commun. 2014;5:4070. doi: 10.1038/ncomms5070. PubMed
794   PMID: 24915882.
795   29.   Hallen MA, Keedy DA, Donald BR. Dead-end elimination with perturbations (DEEPer): a
796   provable protein design algorithm with continuous sidechain and backbone flexibility. Proteins.
797   2013;81(1):18-39. doi: 10.1002/prot.24150. PubMed PMID: 22821798; PubMed Central PMCID:
798   PMC3491125.
799   30.   Smith CA, Kortemme T. Backrub-like backbone simulation recapitulates natural protein
800   conformational variability and improves mutant side-chain prediction. J Mol Biol. 2008;380(4):742-56.
801   doi: 10.1016/j.jmb.2008.05.023. PubMed PMID: 18547585; PubMed Central PMCID: PMC2603262.
802   31.   Hayward S. Peptide-plane flipping in proteins. Protein Sci. 2001;10(11):2219-27. doi:
803   10.1110/ps.23101. PubMed PMID: 11604529; PubMed Central PMCID: PMC2374056.
804   32.   Gunasekaran K, Gomathi L, Ramakrishnan C, Chandrasekhar J, Balaram P. Conformational
805   interconversions in peptide beta-turns: analysis of turns in proteins and computational estimates of
806   barriers. J Mol Biol. 1998;284(5):1505-16. doi: 10.1006/jmbi.1998.2154. PubMed PMID: 9878367.
807   33.   Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity:
808   all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr.
809   2010;66(Pt 1):12-21. doi: 10.1107/S0907444909042073. PubMed PMID: 20057044; PubMed Central
810   PMCID: PMC2803126.
811   34.   Nicholson LK, Yamazaki T, Torchia DA, Grzesiek S, Bax A, Stahl SJ, et al. Flexibility and
812   function in HIV-1 protease. Nat Struct Biol. 1995;2(4):274-80. PubMed PMID: 7796263.
813   35.   Torbeev VY, Raghuraman H, Hamelberg D, Tonelli M, Westler WM, Perozo E, et al. Protein
814   conformational dynamics in the mechanism of HIV-1 protease catalysis. Proc Natl Acad Sci U S A.
815   2011;108(52):20982-7. doi: 10.1073/pnas.1111202108. PubMed PMID: 22158985; PubMed Central
816   PMCID: PMC3248522.

817   36.   Scott WR, Schiffer CA. Curling of flap tips in HIV-1 protease as a mechanism for substrate entry
818   and tolerance of drug resistance. Structure. 2000;8(12):1259-65. PubMed PMID: 11188690.
819   37.   Keedy DA, Georgiev I, Triplett EB, Donald BR, Richardson DC, Richardson JS. The role of local
820   backrub motions in evolved and designed mutations. PLoS Comput Biol. 2012;8(8):e1002629. doi:
821   10.1371/journal.pcbi.1002629. PubMed PMID: 22876172; PubMed Central PMCID: PMC3410847.
822   38.   Georgiev I, Keedy D, Richardson JS, Richardson DC, Donald BR. Algorithm for backrub
823   motions in protein design. Bioinformatics. 2008;24(13):i196-204. doi: 10.1093/bioinformatics/btn169.
824   PubMed PMID: 18586714; PubMed Central PMCID: PMC2718647.
825   39.   Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone
826   freedom. Science. 1998;282(5393):1462-7. PubMed PMID: 9822371.
827   40.   Deis LN, Pemble CWt, Qi Y, Hagarman A, Richardson DC, Richardson JS, et al. Multiscale
828   conformational heterogeneity in staphylococcal protein a: possible determinant of functional plasticity.
829   Structure. 2014;22(10):1467-77. doi: 10.1016/j.str.2014.08.014. PubMed PMID: 25295398; PubMed
830   Central PMCID: PMC4191857.
831   41.   Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins.
832   Science. 1991;254(5038):1598-603. PubMed PMID: 1749933.
833   42.   Kohlhoff KJ, Shukla D, Lawrenz M, Bowman GR, Konerding DE, Belov D, et al. Cloud-based
834   simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. Nat Chem.
835   2014;6(1):15-21. doi: 10.1038/nchem.1821. PubMed PMID: 24345941; PubMed Central PMCID:
836   PMC3923464.
837   43.   Richardson JS, Keedy DA, Richardson DC. "THE PLOT" THICKENS: MORE DATA, MORE
838   DIMENSIONS, MORE USES. Biomolecular Forms and Functions: A Celebration of 50 Years of the
839   Ramachandran Map. 2013:46.
840   44.   Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, et al. Visualizing and
841   quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol.
842   1999;285(4):1711-33. doi: 10.1006/jmbi.1998.2400. PubMed PMID: 9917407.
843   45.   Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a
844   class of genes: which test? Bioinformatics. 2007;23(4):401-7. doi: 10.1093/bioinformatics/btl633.
845   PubMed PMID: 17182697.
846   46.   Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the
847   CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr. 2011;67(Pt 4):235-42. doi:
848   10.1107/S0907444910045749. PubMed PMID: 21460441; PubMed Central PMCID: PMC3069738.
849   47.   Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using
850   hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol. 1999;285(4):1735-47.
851   doi: 10.1006/jmbi.1998.2401. PubMed PMID: 9917408.
852   48.   Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a
853   comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol
854   Crystallogr. 2010;66(Pt 2):213-21. doi: 10.1107/S0907444909052925. PubMed PMID: 20124702;
855   PubMed Central PMCID: PMC2815670.
856   49.   Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The Protein Data
857   Bank. Acta Crystallogr D Biol Crystallogr. 2002;58(Pt 6 No 1):899-907. PubMed PMID: 12037327.
858   50.   Yao P, Dhanik A, Marz N, Propper R, Kou C, Liu G, et al. Efficient algorithms to explore
859   conformation spaces of flexible protein loops. IEEE/ACM Trans Comput Biol Bioinform. 2008;5(4):534-
860   45. doi: 10.1109/TCBB.2008.96. PubMed PMID: 18989041; PubMed Central PMCID: PMC2794838.
861