VOLUME THIRTEEN

COMPUTATIONAL CRYSTALLOGRAPHY NEWSLETTER

JANUARY MMXXII

AlphaFold 2, cis-nonPro

Table of Contents

Editor

Nigel W. Moriarty, NWMoriarty@LBL.Gov

Phenix News

Announcements

New Phenix Release Imminent

The latest version of Phenix – 1.20.1 – was released in January 2022. The list of changes and updates include:

1.20.1 Changes

- Add backwards compatibility for new solvent masking algorithm
- Bug fix SHELX HKLF format output
- Bug fix for phenix.dock_and_rebuild where no model is obtained
- Bug fix for map and model from phenix.douse not aligning
- Add --without-dials option to installation script

1.20 Changes

- New tools and methods
  - phenix AlphaFold2 notebook: Run AlphaFold on Google Colab from Phenix GUI
  - phenix.process_predicted_model: Identify useful domains in AlphaFold model
  - phenix.dock_predicted_model: Dock domains of AlphaFold model into cryo-EM
  - phenix.rebuild_predicted_model: Rebuild AlphaFold model in cryo-EM map using docked domains
  - phenix.dock_and_rebuild : Process, dock and rebuild AlphaFold model with cryo-EM map
  - phenix.model_completion: Connect fragments and fill in gaps based on a map
  - phenix.rebuild_model: Rebuild a model using a map and keeping connectivity
  - phenix.replace_with_fragments_from_pdb: Rebuild a model using fragments from PDB
  - phenix.search_and_morph: SSM search PDB; morph to match target
  - phenix.fragment_search: Search for a fragment in PDB matching target
  - phenix.reverse_fragment: Reverse chain direction of a fragment
  - phenix.superpose_and_morph: SSM or least-squares superpose one model on another; optionally trim and morph to match
  - phenix.voyager.casp_rel_ellg: Calculate relative eLLG score for predicted model quality
- phenix.match_maps:
  - Bug fix (superposed map was not matching target map)
- phenix.real_space_refine:
  - Symmatry multiprocessing aware individual ADP and occupancy refinement

- Multiple changes to improce runtime (for certain refinement strategies)
- Make NQH flips symmetry aware

- phenix.superpose_pdbs:
  - Add feature to transform additional models with matrix found with moving model
- phenix.dock_in_map:
  - Allow splitting model into domains based on chain ID from phenix.process_predicted_model
- Restraints
  - GeoStd updated with 12k plus entity restraints files
  - cis-PRO default updated to EH99
- phenix.fetch_pdb, iotbx.cif_as_mtz:
  - Bug fix: Multiple datasets with different unit cells in a cif file now preserved as multiple crystals in mtz file.

Please note that the latest publication should be used to cite the use of Phenix:

Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB, Croll TI, Hintze B, Hung LW, Jain S, McCoy AJ, Moriarty NW, Oeffner RD, Poon BK, Prisant MG, Read RJ, Richardson JS, Richardson DC, Sammito MD, Sobolev OV, Stockwell DH, Terwilliger TC, Urzhumtsev AG, Videau LL, Williams CJ, Adams PD: Acta Cryst. (2019). D75, 861-877.

Downloads, documentation and changes are available at phenix-online.org

Expert advice

Fitting Tip #22 – Places where you should never fit a *cis*-nonPro peptide

Jane Richardson, Christopher Williams, and Vincent Chen, Duke University

Cis-nonPro Background

Pre-proline peptides are *cis* about 5% of the time, but only about 0.03%, or 1 in 3000, of peptides preceding an amino acid other than Pro are *cis*. There is a larger energy gap between *trans* and *cis* for nonPro than for Pro, and a *cis*-nonPro is harder for surrounding structure to hold in place. Those rare, genuine *cis*-nonPros are almost always important either to biological function or to 3D structure (Williams 2018) whilst occuring only in well-ordered parts of the protein (with the rare exception of a *cis* vicinal disulfide (Richardson 2017) where the SS bond can constrain a *cis* conformation even on a relatively mobile loop).

Not on poorly resolved loops

At overall resolution of 2.5Å or worse, a *cis*-nonPro should not be assigned unless it is present in a closely related structure at high resolution, or paired Bayesian refinements show enough better fit to data for *cis* than for *trans* to balance the 8 log units of disfavored *cis*-nonPro prior probability. The same argument applies to loops with poor, or even absent, density, which effectively have low local resolution (see Figure 1). It is tempting to fit unjustified *cis*-nonPro peptides and is done much more often than random, because they are more compact than *trans* peptides and actually match somewhat better (but incorrectly) to weak, patchy density. From about 2006 to 2015 there was an epidemic of *cis*-nonPro overuse by as much as 2 orders of magnitude (Croll 2015; Williams 2015). It was rapidly cured, once discovered, by hard-to-miss visual markup inside the trapezoidal *cis* backbone shape (see Figure 1) and outlier flags in validation reports implemented in MolProbity (Williams 2018), Phenix (Liebschner 2019), Coot (Emsley 2010), Isolde (Croll 2018) and other systems. However, AlphaFold (Jumper 2021) is now

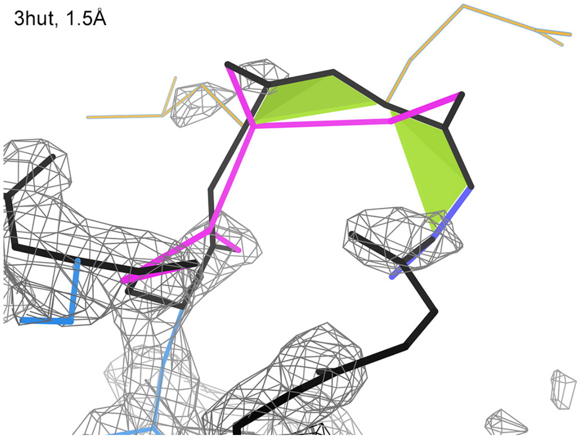


Figure 1: Two successive *cis*-nonPro peptides (green trapezoids) modeled at high resolution but in a loop with almost no density (contours at 1.2σ). The loop also contains 2 CaBLAM outliers (magenta), a bond angle outlier (blue), and 2 rotamer outliers (gold).

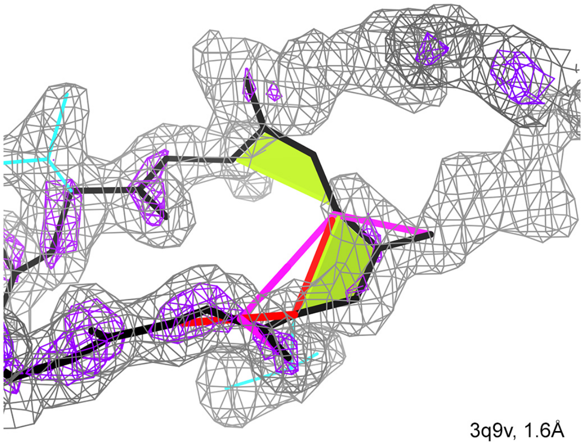


Figure 2: Two incorrect *cis*-nonPro peptides that jump across between the strands of a clear β hairpin that continues into the neighboring molecule. It is actually a chain-swap dimer rather than the deposited compact monomer.

producing large numbers of *cis*-nonPro and twisted peptides in its low confidence regions (see accompanying article on AlphaFold predicted models), which fortunately are seldom taken seriously.

Not two-in-a-row

Two (or more) successive *cis*-nonPro, as seen in Figure 1, are always both incorrect. To test this and other cases, we compared unfiltered vs residue-level quality filtered occurrences in our high resolution Top2018 reference dataset (Williams 2022). There were 14 cases of successive *cis*-nonPro peptides, all of which were eliminated by B, density, and/or clash filters. 6esr has been superseded by 6qe3, with a *trans* fit. The most serious and interesting example is 3q9v, where the two *cis*-nonPros jump across between a pair of β strands that clearly continue as a chain-swap with the neighboring molecule, producing a compact structure for the wrong biological unit.

As discussed in the accompanying article on AlphaFold predicted models (Williams et al. 2022), successive incorrect *cis*-nonPro (or twisted) peptides are extremely common in low-confidence, low pLDDT regions.

Not at chain ends in the model

In the unfiltered Top2018 there are 200 *cis*-nonPro either at chain termini or at the ends of unmodeled loops. These incorrect fittings are enabled by the fewer constraints and typically lower map density at such positions. 197 of the 200 failed the residue-level filters. Manual examination of the three that passed showed that 2 were unjustified because of barely passing but ambiguous density, and the third in a quite different, unexpected way involving too-high rather than too-low map density. Figure three shows the chain N-terminus of the copper-transport protein 4f2f, where an adventitious, partially occupied Cu site at a crystal contact (well above the purple 3σ contour) was incorrectly modeled as the N-terminal N atom. This case also exemplifies a wrong fitting choice at a branch point in the covalent connectivity, similar to the switch of sidechain vs mainchain described in Fitting Tip #6 (Richardson 2013). Here the switch is between the carbonyl O vs the Cα (as labeled in Figure 3), which changes the peptide from

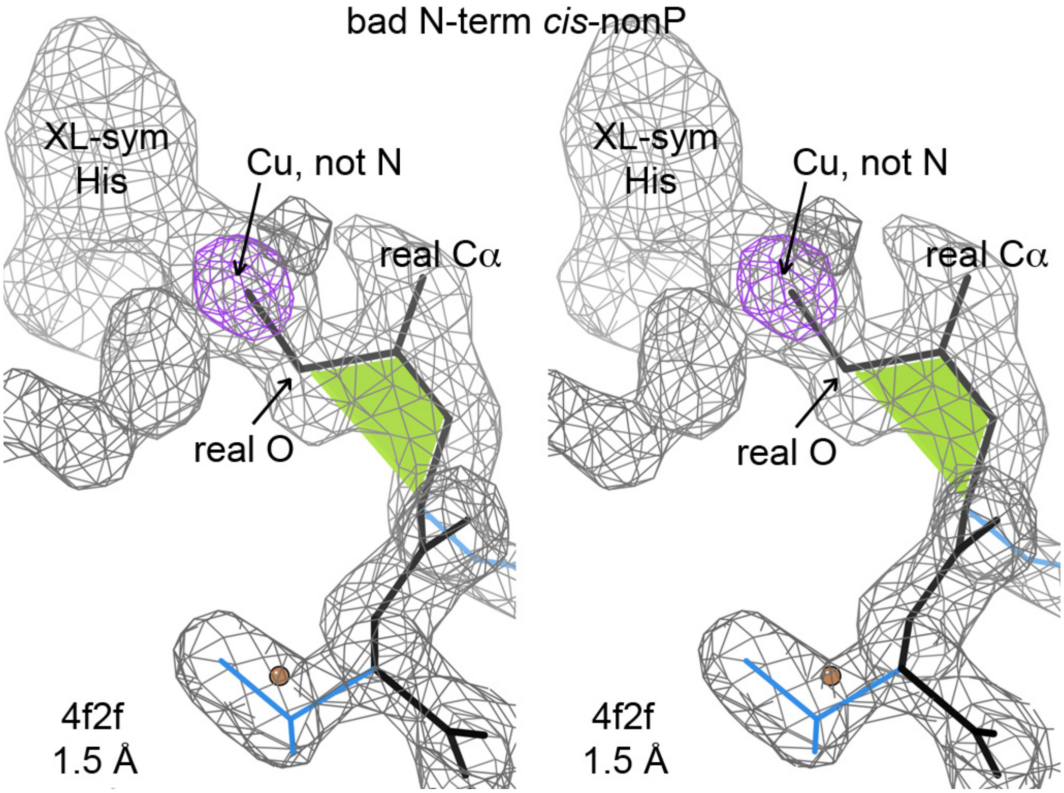


Figure 3: Stereo of an incorrect N-terminal *cis*-nonPro peptide that was confused by an unrecognized adventitious Cu site at a crystal contact.

trans to *cis*. Of 9 nearly-passing chain-terminal *cis*-nonPro, three show this O-Cα switch.

Often genuine in transmembrane proteins

We also surveyed *cis*-nonPro occurrence in the rather atypical environment of the transmembrane regions of integral membrane proteins. In the unfiltered list, only 10 membrane proteins had a *cis*-nonPro, and a *cis*-nonPro in 6 of those proteins passed the residue-level filters. This is an extremely small sample, but does imply some useful conclusions. All 19 *cis*-nonPros were examined in context of the local map density and MolProbity outliers. 16 of them are well outside any transmembrane region, of which three passed the filters and were judged clearly correct, 2 were marginal, and 11 failed the filters and were definitely unjustified by low, patchy density. Several of them (e.g., Asp-AsnN21three in 3q7m) showed a revealing model feature of sidechains truncated down to the Cβ, which means the backbone is also not strongly held in place.

Three of the 19 *cis*-nonPros are within a transmembrane region: one between chain pairs in a trimer, one in the plane of the membrane surface, and one in the middle of a transmembrane β strand. All three passed the filters, have strong, clear density for their *cis* conformation, and make favorable contacts. In contrast to the overwhelmingly incorrect cases described above, *cis*-nonPro peptides in transmembrane regions seem usually to lie in well-ordered structure and can be convincingly identified and modeled at resolutions of 2Å or better. Figure 4 shows the genuine transmembrane β-strand *cis*-nonPro in the 3q7m BamB structure at 1.65Å (Noinaj 2011), which turns and bulges the strand locally to form two backbone H-bonds

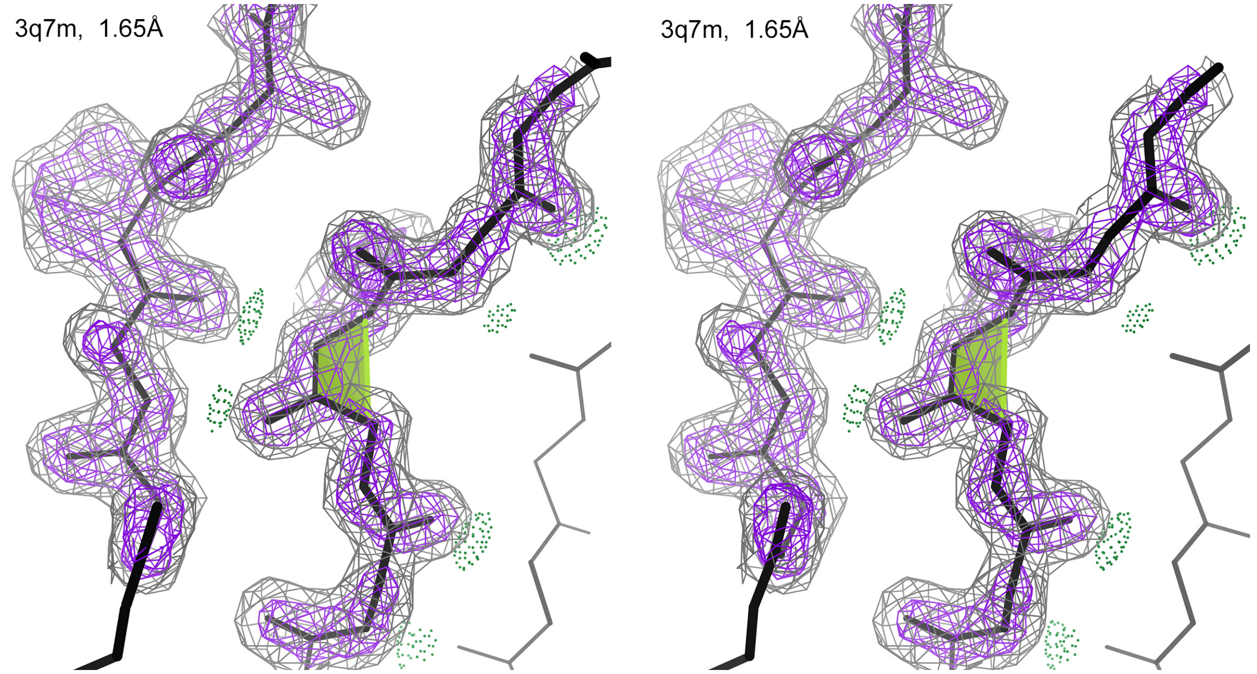


Figure 4: Stereo of a genuine *cis*-nonPro (green trapezoid) on a transmembrane β -strand in 3q7m, in context of its clear map density. Backbone H-bonds (pillows of green dots) are shown for the *cis*-nonPro containing strand. Density contours are shown at 1.2σ (gray) and 3σ (purple).

with the edge strand of the neighboring β-propeller sheet. It forms a type of β-bulge we've never seen before, using a *cis* peptide between a wide pair of antiparallel β H-bonds.

The bottom line

In general, *cis*-nonPro peptides occur only in well-ordered regions of a protein and should show good density themselves to be believable. However, this Tip gives some rules you can use to identify incorrect examples just from a model, either yours or someone else's: a *cis*-nonPro is almost certainly wrong if 1) It is one of two or more *cis*-nonPro in a row, 2) It is on an external loop with many geometry outliers, very high relative B-factors, or truncated sidechains, 3) It is at a chain terminus or at an end of an unmodeled loop.

References:

Croll TI (2015) The rate of *cis-trans* conformation errors is increasing in low-resolution crystal structures, *Acta Crystallogr* **D71**: 706-709

Croll TI (2018) ISOLDE: a physically realistic environment for model building into low-resolution electron density maps, *Acta Crystallogr* **D74**: 519-530

Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot, *Acta Crystallogr* **D66**: 486-501

Jumper J, Evans R, Pritzel A, Green T, Figurnov M et al. (2021) Highly accurate protein structure prediction with AlphaFold, *Nature* **596**: 583-589

Liebschner D, Afonine PV, Baker ML, Bunkoczi G, Chen VB, Croll TI. Hintze BJ, Hung L-W, Jain S, McCoy AJ, Moriarty NW, Oeffner RD, Poon BK, Prisant MG, Read RJ, Richardson JS, Richardson DC, Sammito MD, Sobolev OV, Stockwell DH, Terwilliger TC, Urzhumtsev AG, Videau LL, Williams CJ, Adams PD (2019) Macromolecular structure determination using X-rays, neutrons, and electrons: Recent developments in Phenix, *Acta Crystallogr* **D75**: 861-877

Noinaj N, Fairman JW, Buchanan SK (2011) The crystal structure of BamB suggests interactions with BamA and its role within the BAM complex, *J Molec Biol* **407**: 248-260

Richardson J, Arendall B (2013) "Fitting Tips #6: Potential misfitting by a switch of sidechain vs mainchain", *Comput Cryst Newsletter* **4**: 30-31

Richardson JS, Videau LL, Williams CJ, Richardson DC (2017) Broad analysis of vicinal disulfides: Occurrences, conformations with *cis* or with *trans* peptides, and functional roles including sugar binding, *J Molec Biol* **429**: 1321-1335

Williams CJ, Richardson JS (2015) "Fitting Tips #9: Avoid excess *cis* peptides at low resolution or high B", *Comp Cryst Newsletter* **6**: 2-6

Williams CJ, Videau LL, Hintze BJ, Richardson JS, Richardson DC (2018) *Cis*-nonPro peptides: Genuine occurrences and their functional roles, *bioRxiv*, doi: 10.1101/324517

Williams CJ, Richardson DC, Richardson JS (2022) The importance of residue-level filtering and the Top2018 best-parts dataset of high-quality protein residues, *Protein Sci* **31**: 290-300

Events

American Crystallography Association Annual Meeting 2022. Crystallographic and cryo-EM Structure Solution with Phenix. A workshop will be conducted in-person on Friday, July 29, 2022 @ 8:30 AM PT. Phenix personnel will also be participating in various meeting activities.

Gordon Conference: Diffraction Methods in Structural Biology, Bates College, Lewiston, ME, 24-29 July, 2022. Phenix personnel will also be participating in various meeting activities.

Extreme backbone outlier patterns when AlphaFold gives up

Christopher J. Williams, David C. Richardson and Jane S. Richardson

Department of Biochemistry, Duke University, Durham NC USA

Correspondence email: dcrjsr@kinemage.biochem.duke.edu

Introduction

We, and the rest of the Phenix developers, are primarily concentrating on how the high-confidence AlphaFold (Jumper 2021) predictions can best be used to make experimental structure solution easier and more accurate. But our group has also become interested in the wildly differing features we see within low-confidence (low pLDDT) regions and when one can still gain useful information from those predicted models. Since that requires large-number statistics, we are using the EBI AlphaFold DataBase site (alphafold.ebi.ac.uk), which makes available predictions for entire genomes of common model organisms. Here we are using their data for the human, *E. coli*, and *Methanocaldococcus jannaschii* genomes, in order to sample all three domains of life. The features we are describing can presumably occur in low-pLDDT regions of AlphaFold models run from any site or specifications, but relative occurrence could vary, because we have sometimes experienced different output model results (not just different speeds) with and without templates or relaxation, with different multiple-sequence-alignment methods, with AlphaFold version, and even with the user's level of Colab subscription. So, an orthogonal point to keep in mind is that if you want your use of an AlphaFold model to be reproducible, you need to record and report all details of how the prediction was run.

Bimodal behavior of low-pLDDT parts of predicted models

The AlphaFold pLDDT measure provides a good estimate of prediction accuracy, both overall and across regions within a structure. However, as described here, there seem to be two quite distinct types of behavior that occur in the low pLDDT region from about 20 to 65, but that are not separable by pLDDT value.

The potentially useful regions show a plausible, protein-like model. In spite of very low confidence and usually a high incidence of sidechain clashes, they still can often be close to the right answer. We propose to call them "near-folded". One such example is model 3 in the MMSeqs2 ColabFold (Mirdita 2021) prediction for the functionally and conformationally unusual large catalytic domain of PDB 2vov (Helland 2008), run with templates (of which there are only 2) and without relaxation. It is

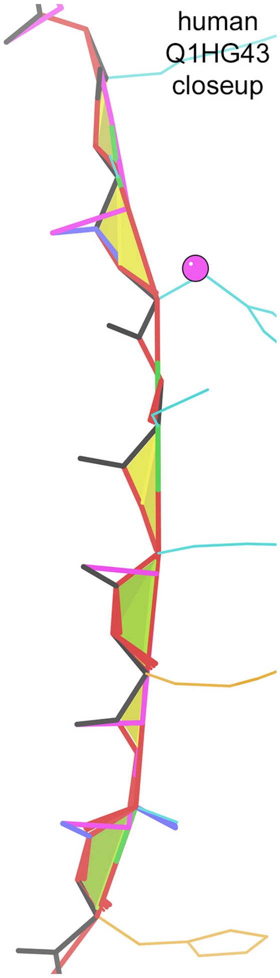


Figure 1: Closeup of a short stretch (327-335) from the EBI AlphaFold Database predicted model of human protein Q1HG43 (Uniprot ID), in a 55-residue non-protein-like barbed-wire region that has pLDDT mainly between 30 and 40. The five types of backbone conformational outliers described in the text are flagged, as well as backbone covalent-angle outliers (red or blue fans).

shown in Figure 2a below. The clashscore is 82 and pLDDT is very low (30 to 40), but it has a majority of the β-sheet structure correct and H-bonded.

In contrast, model regions showing what we will call "barbed-wire" behavior are probably excellent predictions of actual disorder. The model in those regions does not form physically, chemically or conformationally possible protein structure, but looks as if default residue positions were just lined up next to each other in sequence order. It seems that in these regions AlphaFold has given up on making an actual prediction. Figure 1 shows a closeup of a short segment from such a region. These barbed-wire loops or termini are placed away from any high-confidence core and from each other, and so have neither favorable H-bond and Van der Waals interactions nor even very many clashes. A distinctive hallmark of many barbed-wire regions is an unprecedentedly high level of backbone outliers that are sensitive to local peptide geometry such as *cis*-nonPro (lime trapezoids) and twisted (yellow) peptides, CaBLAM (magenta) and Cα-geometry (red) outliers (Prisant 2020), and Ramachandran (green) outliers. A 55-residue barbed-wire region at pLDDT of 30-40 in the predicted model for protein Q1HG43 has 72 of those backbone outliers, or 1.3 per residue, and a 17-residue barbed-wire in the P22455 model has 2.2 of those outliers per residue. In comparison, the plausibly-folded domain of 2vov has only 0.11 such backbone geometry outliers per residue. Notice that nearly all carbonyl groups point left and nearly all sidechains point right in this physically impossible, flat, extended chain of Figure 1.

Surprisingly, given the abundance of backbone geometry outliers in the barbed-wire regions of some AlphaFold predictions, other AlphaFold predictions from the same datasets contain barbed-wire regions with

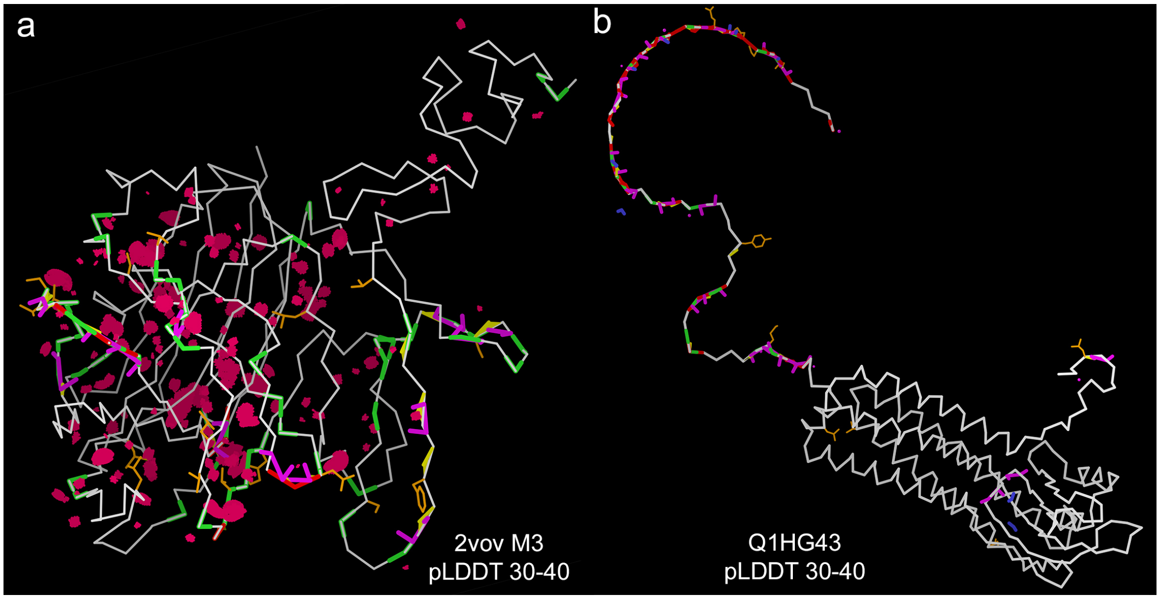


Figure 2: Overviews of MolProbity multi-kin validation (Williams 2020) for differently behaving AlphaFold predicted model regions with pLDDT in the 30-40 range. a) A near-folded ColabFold model that approximates the catalytic domain of the unusual protein in PDB 2vov. b) The AlphaFold Database model of human protein Q1HG43, with a high-confidence helical domain and a non-protein-like, barbed-wire C-terminal tail.

essentially no backbone outliers. These barbed-wire regions still contain many fewer Hydrogen bonds and other contacts than do plausibly-folded regions. Backbone outliers are a powerful and obvious identifier of much barbed-wire, but their absence is merely necessary, not sufficient, to identify plausible, near-folded regions.

Figure 2 compares overall predicted models with similarly low pLDDT score regions of near-folded versus barbed-wire behavior. Figure 2a shows the 2vov model 3 described above, with plausibly-folded features for the main large domain: relatively poor packing, with a high level of sidechain-sidechain clashes, but a compact, protein-like conformation. The large β barrel is the right topology and well H-bonded, and if suitably trimmed would almost certainly give a successful molecular-replacement hit. (Note that models 1 and 2 have high confidence, each used one of the only two templates in the wwPDB, and they are extremely similar to each other and to the experimental structure. Models 4 and 5 are quite similar to model 3.)

Figure 2b shows the overall predicted model of human protein Q1HG43, with excellent, high-confidence prediction for the main helical domain and barbed-wire behavior for the C-terminal region. That C terminus is quite sure to actually be disordered, but the AlphaFold model for it provides no conformational information at all and would better be shown as a dotted line. It is physically impossible and could not be a valid member of a conformational ensemble. Within a long loop or tail that mostly shows probably-disordered behavior, there are often places with nearly outlier-free backbone but still no contacts with anything else, such as a completely isolated helix or the 5-residue vertical stretch in Fig 2b around the gold rotamer-outlier tyrosine.

Statistics from the EBI Database

An overall conclusion is that pLDDT values fall into three ranges with very distinct, although not sharp, boundaries between them at 55-65 and at 20-25 pLDDT scores. Figure 3a plots pLDDT for the 346,319 AlphaFold-predicted *cis*-nonPro peptides in the human genome, defined as omega angle -30° to +30°. The distribution of *cis*-nonPro over-use emphasizes the very different prediction behavior in each of the 3 distinct ranges of pLDDT parameter. Above a local pLDDT score of 65, the *cis*-nonPro occurrence rate is 0.0078%, suitably conservative relative to the 0.03% rate of genuine *cis*-nonPro peptides found for high-quality residues in high-quality crystal structures (Williams 2018b). Between pLDDT 20 and 65 the plot fairly suddenly becomes 1000-fold denser with 7.6% of the non-Pro peptides modeled as *cis*. For *E. coli* the excess *cis*-nonPro at pLDDT 20-65 is still very high, but less by an order of magnitude than human at 0.72%. There are very few predicted residues of any type with pLDDT below 20, so we cannot say anything about that range.

Genuinely "twisted" peptides (>30° non-planar) are extremely rare (Berkholz 2012; Williams 2018a), but they are even more over-used than *cis*-nonPro at low pLDDT. Figure 3b plots the strikingly asymmetric omega distribution for the million twisted peptides at low pLDDT in the human genome predictions. Essentially all of them are in the barbed-wire regions where peptides are arbitrarily lined up next to each other, presumably in non-random and anisotropic starting positions.

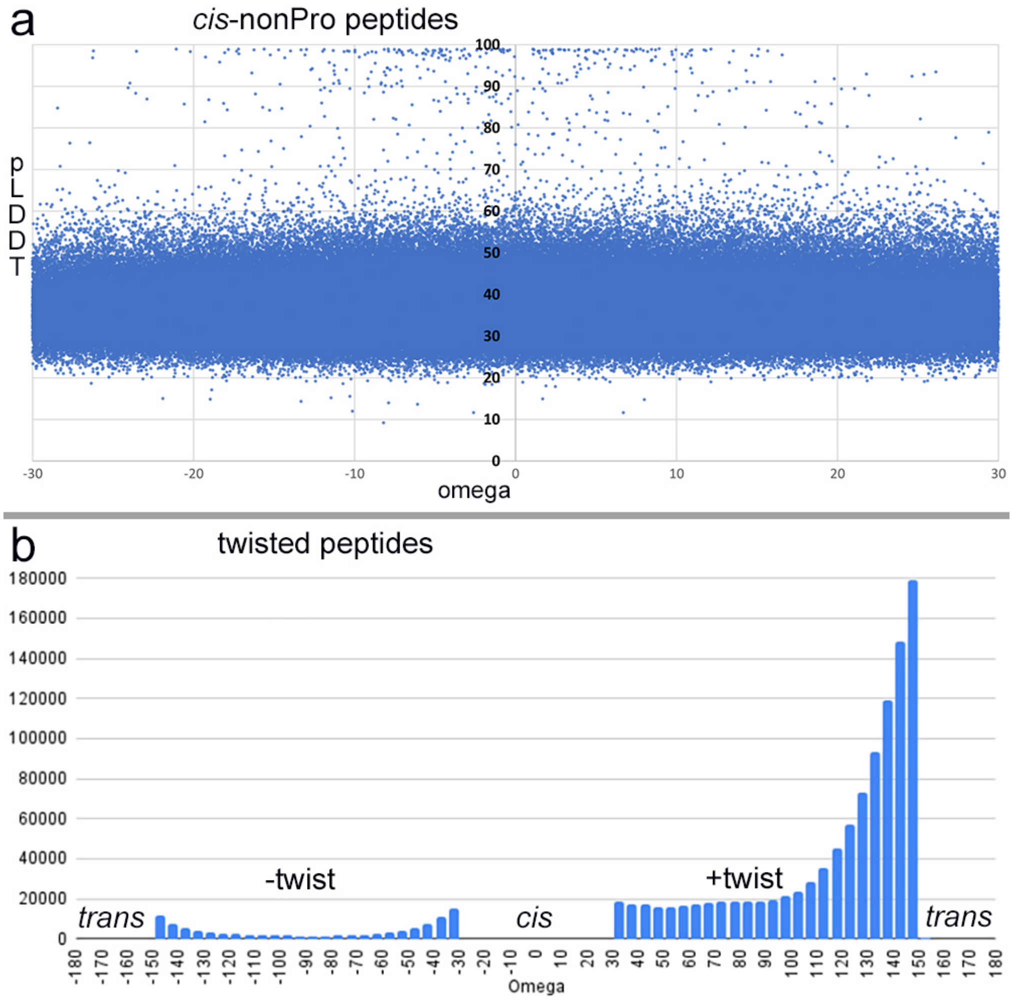


Figure 3: Peptide statistics in AlphaFold Database models for the human genome. a) All *cis*-nonPro peptides as a function of pLDDT score, plotted vs omega angle (+/-30°). Above pLDDT ~65 there are somewhat less *cis*-nonPro than expected, many of which are indeed correct. Between pLDDT 20 and 65 they suddenly become 1000 times more frequent, saturating the plot. b) Twisted peptides (>30° from planar) are implausibly common even up to 90°, and are highly asymmetrical, strongly preferring positive over negative values.

The human genome predictions show the highest content of low-pLDDT regions (33%); *E. coli* and *M. jannaschii* each show about 5% low pLDDT. Within the low-confidence regions the sum of *cis*-nonPro and twisted residues is 30.4% for human, 7.2% for *E. coli*, and 4.7% for *M. jannaschii*. Those are huge numbers, even for the bacterial and archaeal genomes and could certainly seriously distort data from incautious automated statistics. Barbed-wire and near-folded prediction behavior is found up as high as pLDDT 50-70, sometimes within one structure such as the gp39 capsid protein of the Syn5 virus (Matt Baker, personal communication). As noted above, we are developing discriminators for possibly useful near-folded vs non-protein-like barbed-wire behavior, mostly below pLDDT 65, to be available both in Phenix and on the MolProbity website.

An all-barbed-wire example

The AlphaFold Database prediction for human Q86YZ3 has 66.4% Rama outliers and a Rama-Z score of -8, with almost all phi values positive, and psi tightly at +110° +/- 30° (Figure 4c). The protein is a splice variant entirely made up of 13-residue near-repeats with almost half Ser or Gly and extremely few hydrophobics, so disorder would be expected. If backbone bond angles are included it has an average of 5 backbone outliers per residue across the entire model, and provides an extreme example that well illustrates why we call such regions barbed-wire, as shown by the comparison in Figure 4a and b. Figure 4c shows the Ramachandran plot for the Q86YZ3 model. Psi is the only backbone dihedral determined for an isolated residue, so this plot suggests that the randomly positioned starting residues in AlphaFold have a conformation with psi at 110°, which are then only somewhat modified when they are translated to join up in sequence order. Almost none of the resulting peptides have plausible conformations. That hypothesis would also explain why twisted peptides are ubiquitous and highly asymmetric in barbed-wire regions.

The bottom line

In the low-pLDDT regions of AlphaFold models, neither believing everything, nor throwing away everything is a good strategy. The “barbed-wire” regions at low-confidence mean that AlphaFold has given up on producing a protein-like model, presumably because it saw essentially no evolutionary covariance in that region. The barbed-wire part of the sequence is almost certainly disordered in the actual molecule, at least as a monomer, and the specific conformation modeled is arbitrary and usually impossible.

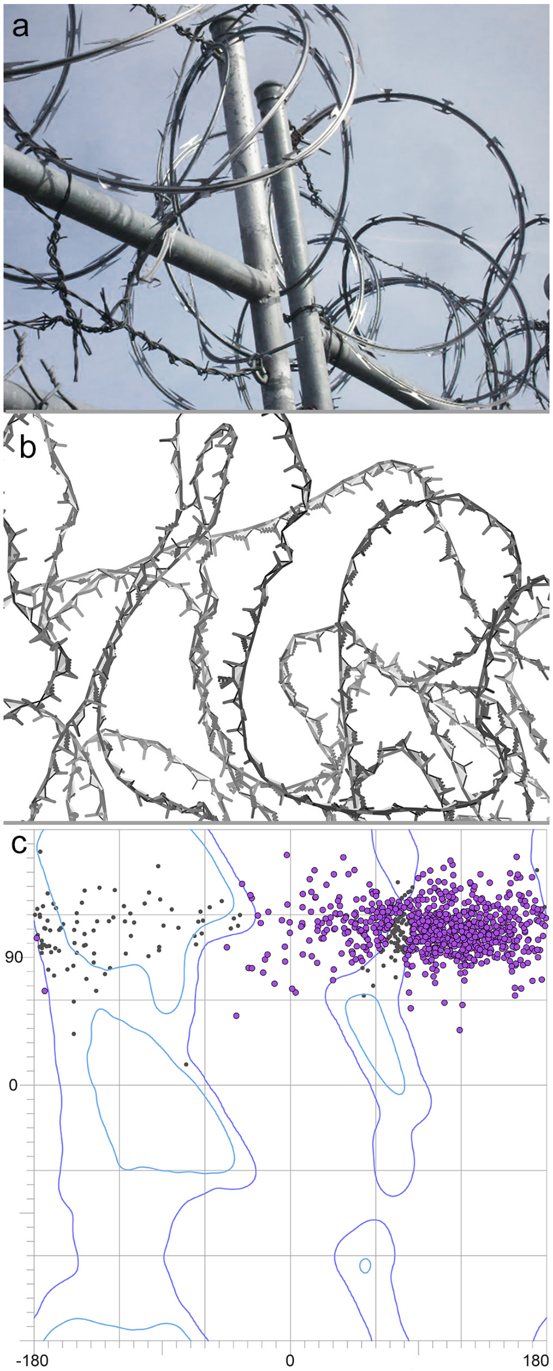


Figure 4: Analyses of the all barbed-wire prediction for human Q86YZ3. a) A fence topped with mixed coils of razor wire and barbed wire. b) A large section of the AlphaFold Database backbone model for Q86YZ3. c) The Ramachandran plot for the Q86YZ3 model, with outliers in purple.

Other low-confidence regions, however, really can be usable 3D structure predictions, and such cases have been noted and successfully used to solve individual experimental structures. We find that typically they are compact but a bit under-packed and with clashing sidechains, but often sufficient for molecular replacement or for fitting into cryo-EM density. Either informed examination by eye or the patterns of backbone conformational, geometric and steric outliers described here can usually distinguish these barbed-wire vs near-folded regions quite clearly, although there are marginal cases and further complications. Apparently, lower organisms have fewer predicted barbed-wire regions, so it might be that more of their low-pLDDT regions contain useful structural information.

References:

Berkholz DS, Driggers CM, Shapovalov MV, Dunbrack RL, Karplus PA (2012) Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. Proc Natl Acad Sci USA 109:449–453.

Helland R, Fjellbirkeland A, Karlsen OA, Lillehaug JR, Jensen HB (2008) An oxidized tryptophan facilitates copper binding in Methylococcus capsulatus-secreted protein Mope, *J Biol Chem* **283**: 13897

Jumper J, Evans R, Pritzel A, Green T, Figurnov M et al. (2021) Highly accurate protein structure prediction with AlphaFold, *Nature* **596**: 583-589

Mirdita M, Ovchinnikov S, Steinegger M (2021) ColabFold – Making protein folding accessible to all, bioRxiv, doi: 10.1101/2021.08.15.456425

Prisant MG, Williams CJ, Chen VB, Richardson JS, Richardson DC (2020) New Tools in MolProbity validation: CaBLAM for cryoEM backbone, Undowser to rethink “waters”, and NGL Viewer to recapture online 3D graphics, *Protein Sci* **29**: 315-329

Williams CJ, Richardson JS (2015) "Fitting Tips #9: Avoid excess *cis* peptides at low resolution or high B", *Comp Cryst Newsletter* **6**: 2-6

Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL et al. (2018a) MolProbity: More and better reference data for improved all-atom structure validation, Protein Sci 27: 293-315

Williams CJ, Videau LL, Hintze BJ, Richardson JS, Richardson DC (2018b) *Cis*-nonPro peptides: Genuine occurrences and their functional roles, *bioRxiv*, doi: 10.1101/324517

Difference Density Tracer: A Tool for Rapid Visualization of Difference Densities in Proteins to identify Dynamics

Medhanjali Dasgupta[a], Asmit Bhowmick[a] and Jan F. Kern[a]

[a] *Molecular Biophysics & Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.*

Correspondence email: mdasgupta@lbl.gov, abhowmick@lbl.gov

Introduction

An important way of identifying protein dynamics using X-ray diffraction data is by looking at the relevant difference features in the isomorphous difference density (Fo-Fo) maps, at a specific map contour level (3σ being the usual threshold for difference maps). While scanning an entire protein crystal structure manually for significant difference features (any features visible in the map at greater than or equal to +/- 3σ) is technically possible, it is cumbersome, time consuming and risks missing out on features in regions of the protein that we don't expect and therefore don't inspect. Moreover, this is not feasible for very large protein structures. Herein we present a completely automated new tool, called *difference density tracer*, modified from Wickstrand *et al*'s original work[1], that computes all positive and negative isomorphous difference (Fo-Fo) intensities around an imaginary sphere of user given radius, focused around a set of user specified amino acid residues in the structure of interest. The analysis method is also extendable to mF_o-DF_c maps. The output is a 2-D plot of the measured positive and negative difference intensities (Y axis) around each atom of a specified set of amino acids (X-axis), going in order from Cα to Cß to Cγ and finally the sidechain atoms, if any. This provides a simplified representation of the appearance and progression of significant difference features in a protein in multiple conditions/timepoints, providing an alternate way to analyze difference density maps.

Methods

Our method to calculate the positive and negative isomorphous difference density amplitudes (henceforth referred to as Fo-Fo amplitudes) about an atom follows very closely the method outlined in Wickstrand et. al. However we have condensed all the different steps in that publication which made use of different softwares into one simple python script making full use of the CCTBX library. Following are the primary inputs to the script -

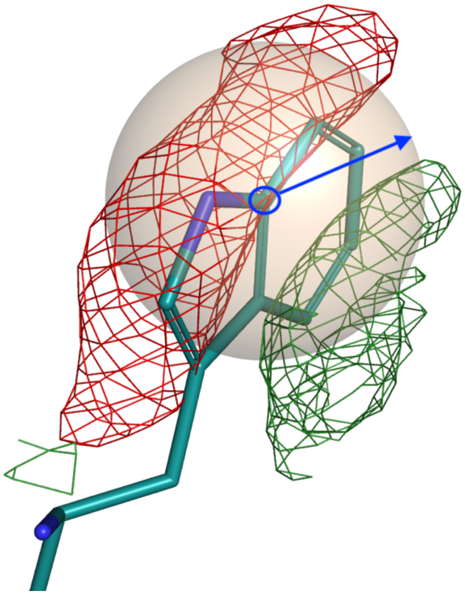
a) An MTZ file which contains the Fo-Fo difference map coefficients. This can be generated in Phenix.

b) The PDB file of the reference state (typically the resting state in time-resolved studies).

c) The residue numbers and chain names of interest.

d) The radius of the sphere about which to calculate the positive and negative difference intensities (Default is 2.0Å)

The script, named *difference_density_tracer.py*, first calculates the map using default settings specified in the CCTBX library and sums up separately the positive and negative difference amplitudes for each atom of the specified residues within a sphere of user specified radius (default is 2.0Å) as illustrated in Fig 1A. We can also set a minimum threshold sigma level if needed to filter out noise in a map and that can also be provided as input (default is 1.0). The final output is a plot of the difference density amplitudes



(green=positive, red=negative) on the Y-axis and the atom name/residue number on the X-axis as shown in Fig 1B.

Code Availability

The python script can be found at https://github.com/asmit3/eden and needs to be run using *phenix.python* as distributed with the command line version of the Phenix software (requires version 1.19.2 and above).

Running the script

An example usage of running the script is as follows (used to generate figure 1B) –

Figure 1: **(A)** Overview of the method used to calculate summed amplitude values. All positive and negative map values within a fictitious sphere (in yellow) of specified radius (2Ȧ) about an atom (highlighted with blue circle) are separately summed to give the respective values. In the figure the difference map is contoured at 3σ (green=positive, red=negative) but the summation can be done above any threshold sigma level (default is 1.0).

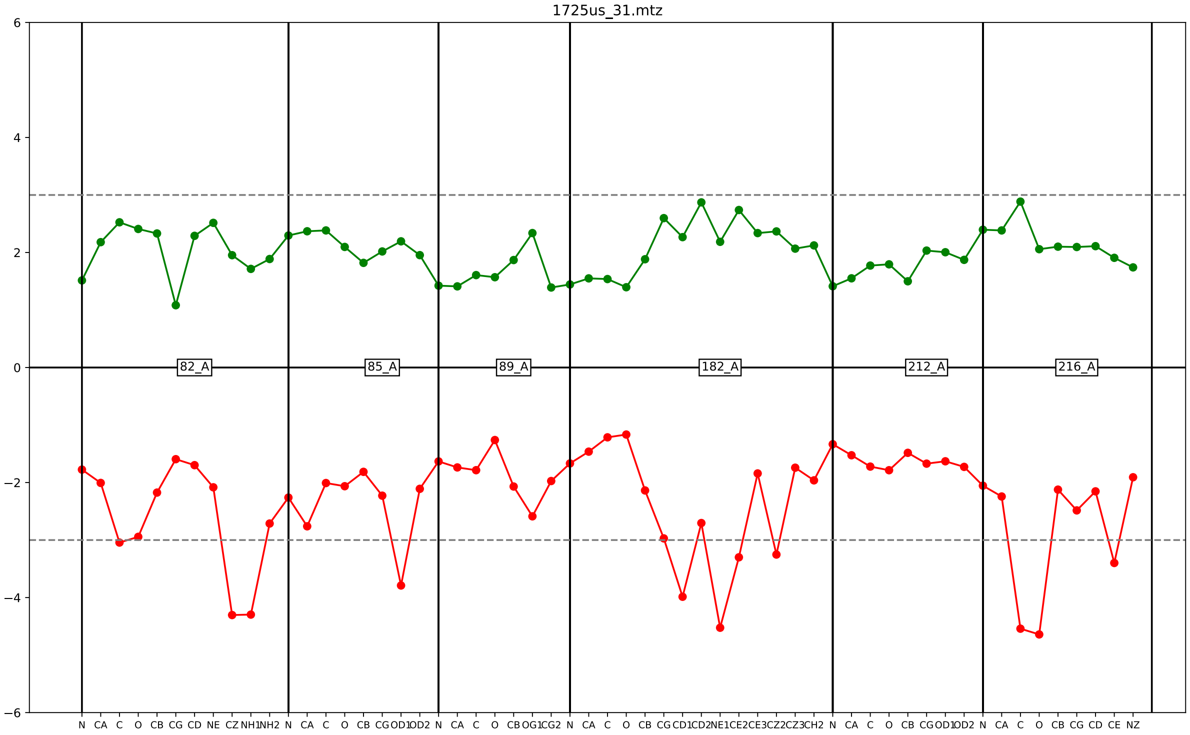


Figure 1: **(B)** Output display of the tool with positive (green) and negative (red) map values plotted on the Y-axis. On the X-axis are the atoms at which the map values were calculated. Residue numbers and chain names are shown in the middle of the plot for each residue. The data used for the plot was taken from Wickstrand et. al with the 1.725ms timepoint features for selected residues being shown. Model and map generated in PyMOL.

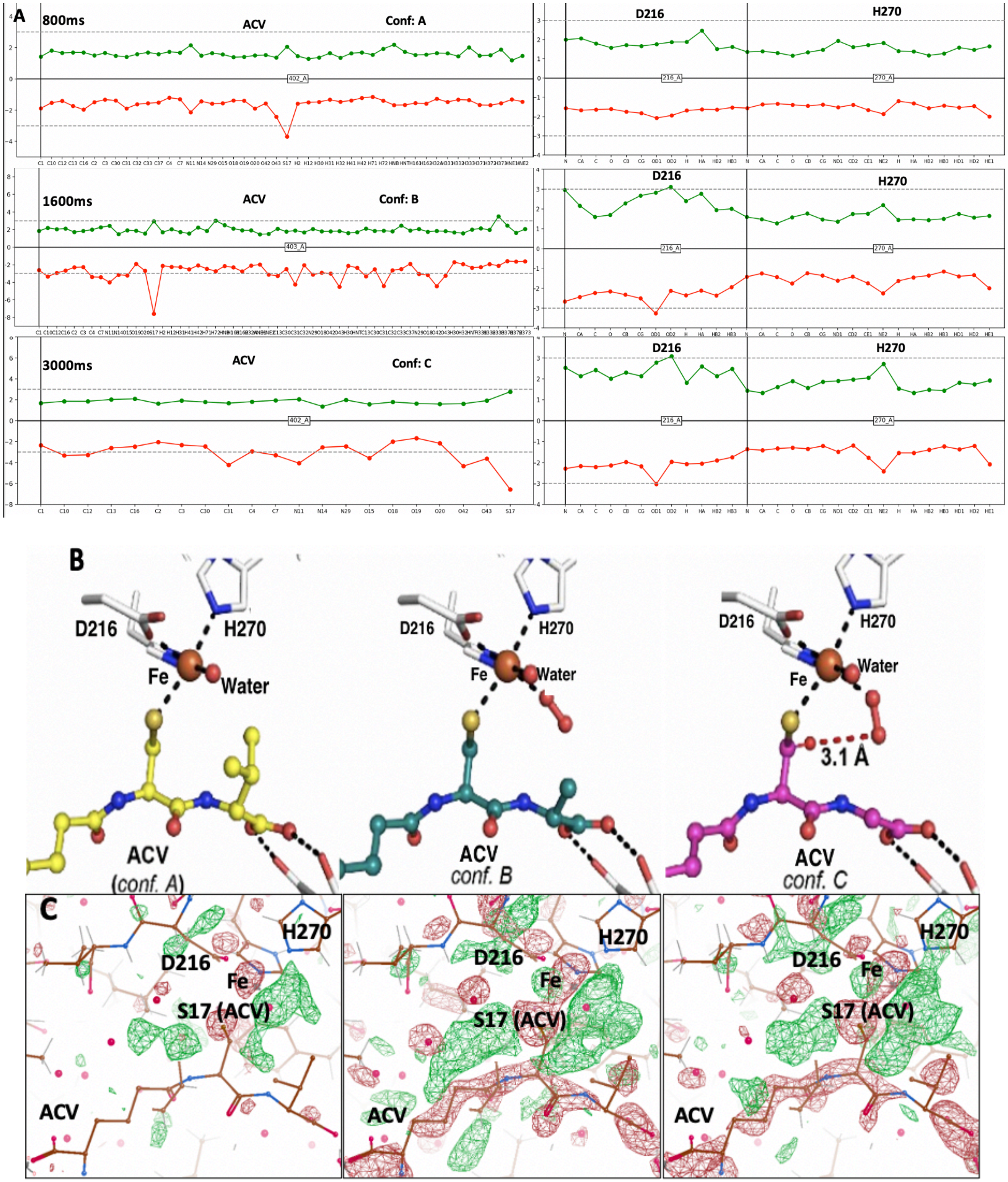


Figure 2: **(A)** The new density tracer tool shows elevated difference features develop at 3σ (dashed line), around ACV and its environment (H270, D216, Fe) at timepoints 1600ms and 3000ms, following oxygen incubation of IPNS-ACV cocrystals **(B)** shows the ACV modeled in different conformations at the mentioned time points following oxygen incubation. Figure modified from ref. [2]. **(C)** Isomorphous difference maps were generated as timepoint_minus_ground state, model used is 6ZAE. Isomorphous difference features around the ACV molecule and its surroundings (H270, D216) at the time points studied, are modeled at 3σ track with the difference density tracer tool in (A). Isomorphous difference maps were generated in PHENIX. Maps and Models displayed with Coot and Chimera.

```
phenix.python difference_density_tracer.py
model_filename=5b6v-H_protein_ret_h2o.pdb
mtz_filename=1725us_31.mtz
residue_numbers=82,85,89,182,212,216
```

Further inputs can be given in terms of chain names etc. For all the available command line inputs run the script as *phenix.python difference_density.py_tracer.py -h*

One of the biggest advantages of this tool is a totally automated, rapid and accurate identification of any indication of dynamics in the entire protein structure, ranging from subtle sidechain flexibilities, to sampling of completely new conformations. It is often very easy to miss these subtle difference features in large protein structures across different timepoints when screening for them manually in a visualization software such as Coot. Moreover, manual screening does not provide a clear picture of any correlated motions occurring outside of our regions of interest where we do not necessarily know to probe intuitively. This new tool allows the user to screen through the entire protein structure and visualize every positive and negative difference feature round every and any selection of amino acids/ligands one is interested in, in a structure of interest. Moreover, the calculation is fast, mainly reliant on the number of residues as well as the number of atoms in the residue, averaging at 3 seconds per Tryptophan residue ($C_{11}N_2O$) on a Macbook Pro with 2.8 GHz Quad-Core Intel Core i7 processor where the number of atoms is 14.

Ideal use cases are i) large systems where one cannot manually screen around each residue feasibly. ii) new protein systems where we don't know where or when to expect dynamics. iii) Tracking changes for dynamics across multiple timepoints/conditions at specific sites.

Applications/Use cases

Tracking correlated motions in Iso Penicillin N Synthase (IPNS) during catalysis: In collaboration with Rabe *et al*[2], we have previously collected serial femtosecond crystallography (SFX) datasets on co-crystals of IPNS with its substrate, amino-adipoyl-L-cysteinyl-D-valine (ACV), undergoing catalysis, at various timepoints (400ms, 500ms, 800ms, 1600ms and 3000ms) following O_2 exposure. Interestingly, we have noticed conformation changes in substrate ACV, coordinated to IPNS residues Histidine270, Aspartate216 and Histidine 214 and the metal Fe atom, as we prolong O_2 exposure (most noticeable 2Fo-Fc and Fo-Fc features develop at 800ms, 1600ms and 3000ms following O_2 treatment[2]).

Using our difference density tracer tool, we were able to observe significant Fo-Fo features around the ACV molecule at 3σ Fo-Fo map contour level (spherical probe radius used = 2Å) at 1600ms and 3000ms following oxygen incubation (Figure 2A). This is clearly reflected in the Fo-Fo maps when manually screening for these features at this particular region (Figure 2C), indicating sampling of alternate conformations by the ACV molecule at these mentioned timepoints following reaction triggering with O_2 treatment. This is consistent with the published results that indicate flexible sampling of conformations by ACV at the indicated timepoints.

Moreover, these SFX datasets also show an interesting increase in conformational disorder in IPNS, specifically around the α3 helix (residues 47-64) and the β11sheet (residues 280-283), with O_2 binding[2]. We see

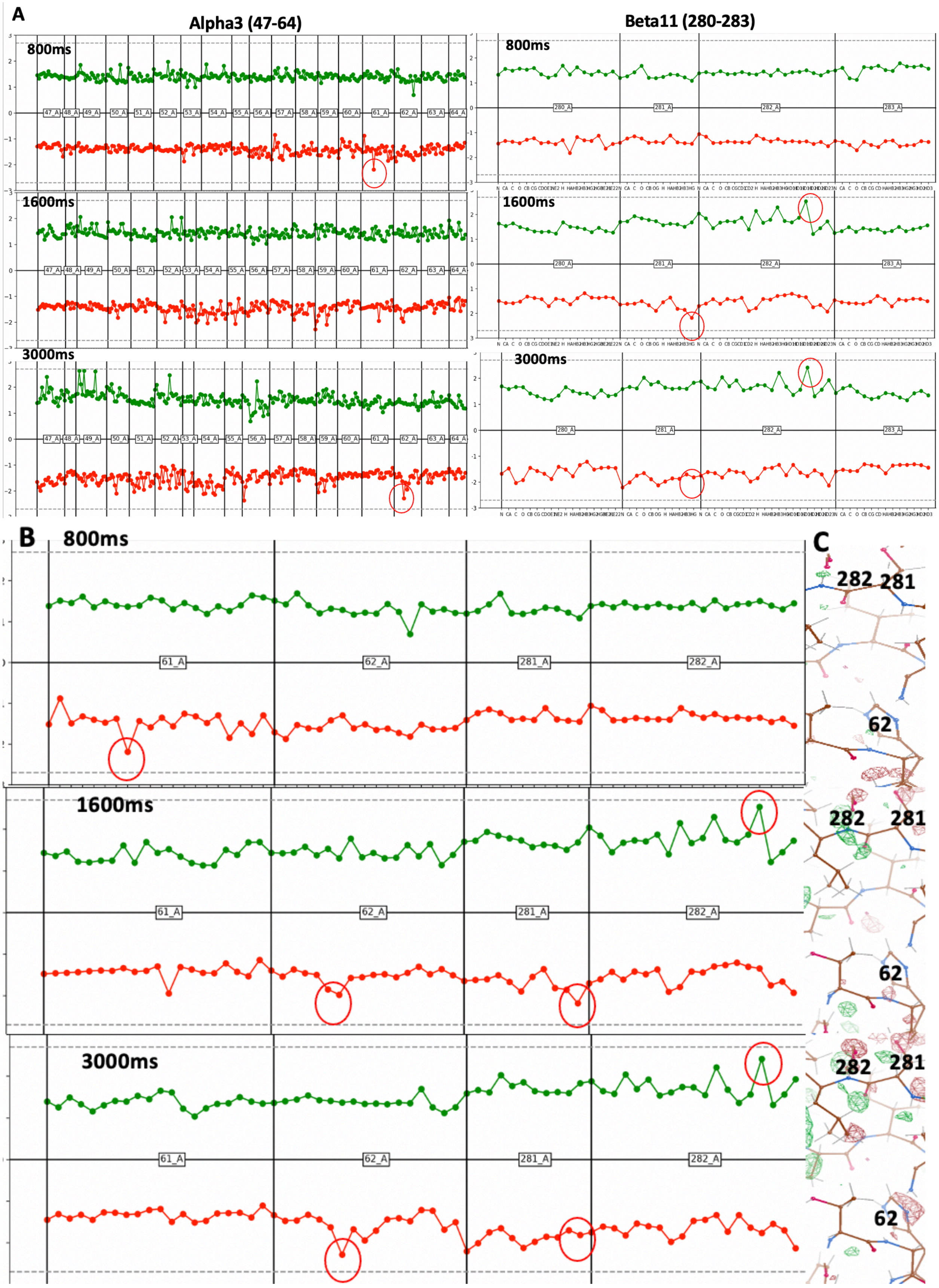


Figure 3: **(A) (Left)** Difference density tracer tools shows Isomorphous difference features (Fo-Fo maps made with timepoint minus 6ZAE ground state and σ threshold shown at 2.7 with dotted lines) around

the α3 region of the IPNS enzyme, comprising residues 47-64. Elevated features around residue 61 in 800ms following O_2 exposure at 2.6 σ (red circle) is diminished in the 1600ms dataset and complete gone by the 3000 dataset. Reside 62 shows the reverse trend, that is, negative difference features show up around 62 in the 3000ms dataset at 2.7 σ (red circle), which was previously not present in the 800ms and 1600ms datasets. **(Right)** Co-ordinated difference features around the β11 region in IPNS, specifically around residues 281 and 282 in 1600ms and 3000ms (red circles) following O_2 exposure. **(B)** Difference density traces of residues 61 and 62 (from α3 helix) and 281 and 282 (from β11 sheet) show dynamic Fo-Fo positive features at 800, 1600 and 3000ms following O_2 treatment, indicating flexible conformational sampling by these residues as IPNS catalysis progresses. **(C)** Fo-Fo difference features around residues 62, 281 and 282 show developing flexibility around these residues at 1600ms following catalysis triggering and these are well developed at 3000ms. Isomorphous difference maps are contoured at 3 σ and were generated in PHENIX. Models/maps displayed with Coot. The atoms on the X-axis are not shown for ease of viewing the plots in this paper.

clear mFo-DFc features at 2.6σ map contour levels, at these regions of interest in IPNS, in 800ms (PDB: 6ZAH), 1600ms (PDB: 6ZAI) and 3000ms (PDB: 6ZAJ) following reaction triggering by O_2 treatment. These features are not present in the anaerobic ground state dataset (PDB: 6ZAE). We used our new difference density tracer tool on the IPNS 800ms, 1600ms and 3000ms Isomorphous Difference Density maps (with 6ZAE as the ground state mtz), to replicate these results. Unsurprisingly, we see evidence of dynamics starting to build in the later O_2 exposure timepoints (1600ms and 3000ms) compared to the early 800ms timepoint and the anaerobic ground state model in the β11/α3 region of the IPNS enzyme, specifically residues 63 (α3 helix residue) Ser281 and Leu282 (β11 sheet residues), as it undergoes catalysis (Figure 3). Fo-Fo maps contoured at 3σ show significant difference features around the IPNS Histidine 62 residue, at 1600ms and 3000ms following O_2 exposure (Figure 3C). These elevated difference features are also picked up by our tracer tool (Figure 2A-B) where we see somewhat coordinated features in the nearby β11 sheet residues Ser281 and Leu282. The CD_2 atom of L282 is ~ 5.0 Å away from the His62 CE_1 atom in the anaerobic ground state model (PDB: 6ZAE); in the 3000ms dataset, we see significant negative features around the His62 sidechain, indicating increased flexibility, perhaps to accommodate a shifted conformation of the Ser281-Leu282 residues, which shows coordinated difference features developing at 3σ contour levels, at the 1600 and 3000ms timepoints (Figure 3B-C). This demonstrates the utility of the new tool in picking up obscure correlated features within the entire protein molecule, at user specified map contour levels and spherical probe radii.

Conclusions

We presented a new tool called *difference density tracer* that allows quantification of positive and negative isomorphous difference density features between 2 states of a protein and allows for rapid identification of important structural changes without having to scan through an entire map. Since Fo-Fo maps can be generated much faster than difference maps that involve refinement of the model, it can be used to rapidly detect dynamics in an enzyme system undergoing catalysis. This tool further expedites that process by allowing us to scan broad regions of a protein as well as multiple timepoints together by using a simplified representation of the difference map. In time-resolved crystallography where prompt decisions need to be made based on whether the timepoint/conditions being probed are showing changes in difference maps, this tool can help experimentalists keep track of the changes as more data is collected.

The tool is intended currently for use in time-resolved structural studies but we are looking to extend its use in other research topics involving map analysis. For example, this tool can also be used to analyze mF_o-DF_c maps in the same manner. Scientists looking to use this tool are encouraged to reach out to us if it needs to be adapted to their specific use case and we hope to incorporate new features based on the input.

References

1. Wickstrand, C. *et al.* A tool for visualizing protein motions in time-resolved crystallography. *Struct. Dyn.* **7**, 024701 (2020).

2. Rabe, P. *et al.* X-ray free-electron laser studies reveal correlated motion during isopenicillin *N* synthase catalysis. *Sci. Advances.* **7,** 34 (2020).