DOI: 10.1002/prot.1234

RESEARCH ARTICLE

PROTEINS WILEY

The Shameon enthalpy function P.Leu(P) illustrates the dangers of overzealous protein structure validation

J Lange | G Vriend

CMBI, Nijmegen, The Netherlands

Correspondence
To whom insults should be addressed:
Email: vriendgert@gmail.com

Abstract

Validation and re-refinement of PDB files have long been considered an abhorrent way to remove the spread of data in the PDB and consequently the volume of information that can be extracted from this beautiful database. In this paper we describe how we applied Shameon's information enthalpy law on leucine rotamers to find examples in the PDB to illustrate the harm that can be done by rigorous protein structure validation. We are happy that many crystallographers still avoid this harm by resisting having useful information validated away from their structure models before they deposited their coordinates in the PDB.

1 | INTRODUCTION

In 1993 Vriend,[3] Laskowski,[4] Sippl,[5] and their colleagues, introduced in rapid succession three methods to compare individual protein structure parameters with averaged protein structure parameters. Hardly could these authors foresee that a few years later Hooft *et al*[6] would abuse these ideas to design the first ever comprehensive protein structure validation software package[6,7] The concepts used by this protein structure vigilante are simple: a) take a parameter that is computationally accessible like packing, hydrogen bonds, bond lengths, rotamers, etc., b) determine the average value for that parameter over a few PDB files, c) call this average a standard, and d) flag all proteins that do not adhere to this standard as errors.

We believe the latter caused significant harm. Would it have stayed with just flagging interesting structural features, it would have been possible to leave all the validated, boring structures out of any analysis. But unfortunately, we celebrate this year already the 10 year anniversary of the PDB-REDO project[8] that enforces the validation dictates. It modifies structure models to reduce their information content P.Leu(P) with P being the natural spread in parameters such as bond angles, interatomic distances, rotamers, etc.

According to modern-day FB-based science[9] and opinions of important politicians, validation reports are fake news[10] just like reports on climate change,[11] or bad things written about the famous Dr Wakefield.[12] PDB-REDO thus is much like the messenger that gave notice of Lucullus' coming.

In this paper we concentrate on an example: Rotamers. Rotamers are preferred side chain conformations, albeit it actually never became clear by whom they are preferred. NMR spectroscopists believed already in the late 70's that rotamers were a useful tool when determining and disseminating protein structures.[13] Crystallographers tend to be a bit more careful, and it took till the early 90's before they started to use this concept too.[14] Studies on rotamers suggest that each residue type has its own rotamer distribution,[15] and the structure validation[6] and re-refinement fields use these rotamer distributions just as much as science fiction writers do.[16,17]

2 | RESULTS

We casually looked at a few PDB files and concluded that the rotamer distribution for the residue type leucine essentially contains ten possible χ1-χ2 angle combinations. These are shown in Figure 1.

Figure 2 illustrates that rotamer 10 can be observed in PDB files solved by X-ray diffraction at high resolution. These files were randomly selected from the PDB.

We extracted from the PDB 9653 files that have exactly the same protein exactly twice in the crystallographic asymmetric unit with as extra constraint that these two proteins needed to be related by NCS (non-crystallographic symmetry). From these we took the subset in which all amino acids in both databases were present, and intact. We chose these files because in practice coordinates of proteins related by NCS get averaged far less often than those related

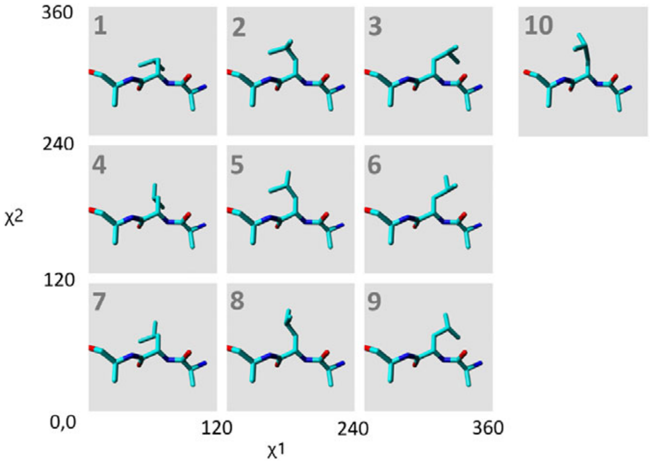

FIGURE 1 ***The ten rotamers of leucine.*** *Rotamers are shown for leucine as the middle residue in an ala-leu-ala tri-peptide in a β-strand like conformation. For reasons that will become clear further down in this article, rotamer 10 is shown a bit separate from the others. Its (χ1,χ2) torsion angles are about (210, 210). The relative locations of the nine other rotamers are the same as in the (χ1,χ2) frequency plots* (Figures 3). *Gray numbers correspond to the 120*120 boxes that were drawn arbitrarily in those (χ1,χ2) plots*

by crystallographic symmetry. We now apply the Shameon information enthalpy law[1] H = P.Leu(P) to all leucine rotamer distributions. P is set to p*(1-p) in which p is the fraction of leucine pairs (one leucine in the one protein; the other leucine at the equivalent position in its NCS related protein) that differ by more than 60° in either χ1 or χ2. The Leu function is simply the normalized logarithm, which is commonly used when applying either Shannon's entropy function[18] or Shameon's enthalpy function.[1]

It has to be kept in mind that the PDB subset consists of files that have two identical molecules in the same asymmetric unit, and thus in the same crystal. Both molecules, and thus both leucines, are therefore in the same biophysical environment. The fact that we often observe both rotamer 8 and rotamer 10 in the same crystal demonstrates that Figure 2 does not merely lists some exceptions, but that rotamer 10 really exists and is not just the result of Murthy's law[19,20] that "everything that was done wrong might look wrong". It is depressing to see that rotamer 10 is almost never observed in PDB-REDO files. Figure 3 shows the χ1-χ2 distribution for all leucines observed in the dataset in both the original PDB files and in the PDB-REDO files.

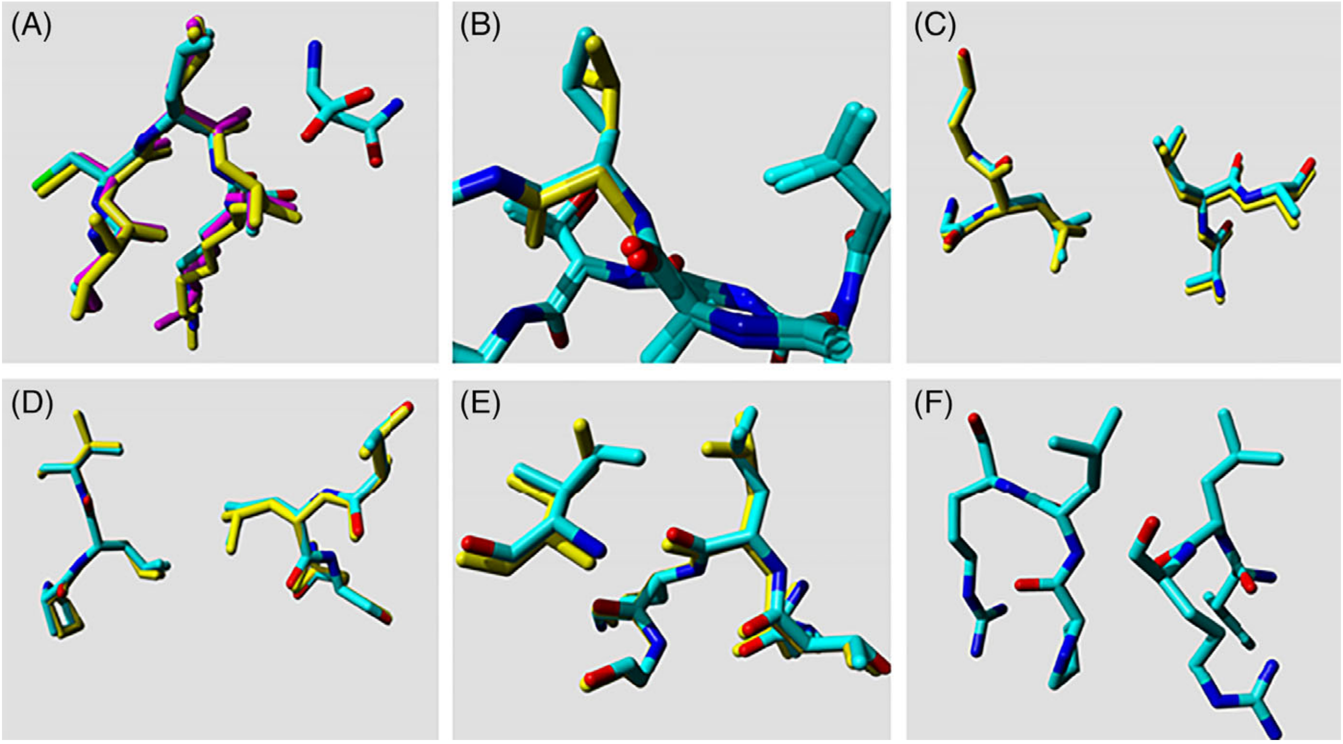

FIGURE 2 **Six examples of leucines in rotamer 10.** *All six structures are solved at 1.5-1.9 A resolution, so no doubt can exist about the quality of the coordinates. In A-E multiple NCS related copies were found, and the other chains are coloured yellow or purple and superposed on the A-chain. A) Leu 53 in 1gto; B) Leu 91 in 1gmq; C) Leu 223 and 339 in 1n1e; D) Leu 58 and 171 in 1fj2; E) Leu 191 in 1te6; F) Leu 1114 and 1122 in 1dll. In panel A the presence of the C-terminal Asn (absent in the two NCS related chains) correlates with the presence of rotamer 10. Several panels seem to suggest that the presence of rotamer 10 seems correlated with atomic contacts with a leucine in rotamer 6 or 8; panel B shows that this must be accidental. In panel C we see two pairs of leucines; in the one NCS copy rotamer 10 (to the right) contacts a leucine with rotamer 8, in the other a leucine with rotamer 10. In panel E the two NCS related leucines each touch an isoleucine with a different rotamer.*[15] *Panel E illustrates the cascading effect of validation dictates*

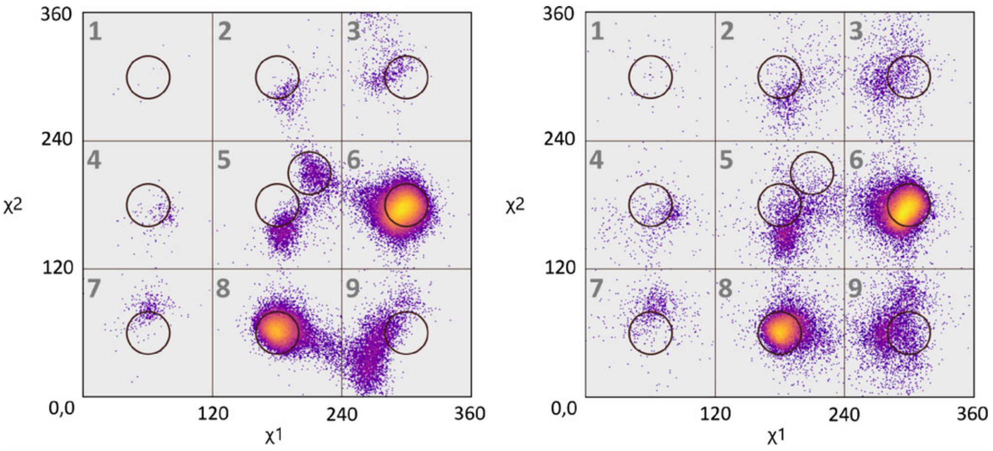

FIGURE 3 ***χ*1-*χ*2 distributions in PDB files (left) and in PDB-REDO files (right).** *The circles are drawn arbitrarily around the centres of the nine 120∗120 sectors in χ1-χ2 space. They have a 20° radius. The circle in the top-right of the central sector 5 corresponds to rotamer 10 (see also* Figure 1*). Only every 10-th point is actually displayed in order not to overload the user with information*

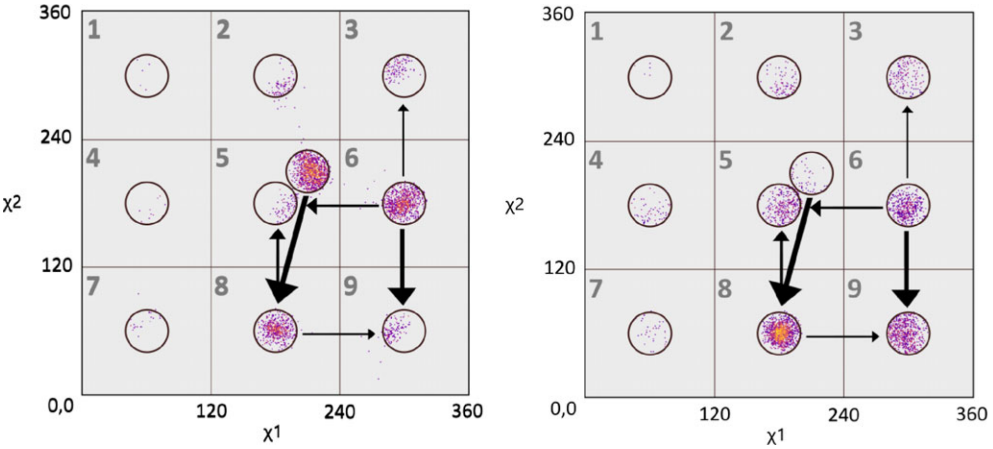

FIGURE 4 ***Rotamer transitions performed by PDB-REDO.*** *These plots are identical to* Figure 3*, but now only residues are shown that fall inside any of the arbitrary circles in a PDB-REDO file. Arrows are added, the thickness of which relates to the number of rotamers that PDB-REDO converted*

3 | DISCUSSION

Figure 3 not only illustrates that rotamer 10 is essentially absent in the PDB-REDO database; only a couple dozen residues evaded REDO's attention; but it also illustrates several other striking aspects. In the PDB plot (left) many rotamer islands are connected by continuous paths of observable rotamers, indicating that dynamic rotamer transfers are possible in Molecular Dynamics (MD) simulations. If we were to believe PDB-REDO then most of these conversions cannot take place without rather extreme forms of quantum tunnelling.

"To be or not to be a leucine, that is the question". This famous soliloquy was spoken a long time ago already for rotamer 10 in the PDB-REDO database. Is rotamer 10, or isn't it? We asked where did these rotamer 10 leucines go, and found that REDO not only converts nearly all leucines from rotamer 10 to rotamer 8, but it does a series of other conversions too. These are illustrated in Figure 4. It is striking that PDB-REDO moves many residues from rotamer 10 to the next-door rotamer 5 only via rotamer 8, which we believe to be a very inefficient way of doing things.

Like baboon droppings,[21] science is full of surprises, and we should not use prior assumptions in a posterior way. This article shows how dangerous that can be because protein structure validation and re-refinement would have caused us to entirely lose sight on leucine's rotamer 10. The second law of thermodynamics really puts the nail in PDB-REDO's coffin. This law says that when structures change from one form to another form, or rotamers move freely, entropy (disorder) in a closed system increases, and our P.Leu(P) analysis of PDB-REDO is in disagreement with this fundamental law of the universe.

ACKNOWLEDGEMENTS

We thank the many anonymous crystallographers for depositing their un-validated data in the PDB, and apologize to those whose PDB file we did not include in this study because the large number of excellent submissions unfortunately required that we had to make hard choices. The authors thank Rob Hooft and Robbie Joosten for critically reading the manuscript.

REFERENCES

1. https://ieeexplore.ieee.org/document/5231753
2. The wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *NAR*. 2019;47(D1):D520-528-D528. .Protein Data Bank: the single global archive for 3D macromolecular structure data
3. Vriend G, Sander C. Quality control of protein models: directional atomic contact analysis. *J. Appl. Cryst*. 1993;26:47-60.
4. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst*. 1993;26:283-291.
5. Sippl MJ. Recognition of Errors in Three-Dimensional Structures of Proteins. *Proteins*. 1993;17:355-362.

6. Hooft RWW, Vriend G, Sander C, Abola EE. Errors in protein structures. *Nature*. 1996;381:272.
7. Who checks the checkers? Four validation tools applied to eight atomic resolution structures. EU 3-D Validation Network. *J Mol Biol*. 1998;276:417-436.
8. Joosten RP, Salzemann J, Bloch V, et al. PDB-REDO: automated re-refinement of X-ray structure models in the PDB. *J Appl. Cryst*. 2009; 42:376-384.
9. https://www.wired.com/story/facebook-snapchat-and-the-dawn-of-the-post-truth-era/
10. Jones TA, Kleywegt GJ, Brunger AT. Storing diffraction data. *Nature*. 1996;383:18-19.
11. https://www.bbc.com/news/world-us-canada-46351940
12. Wakefield AJ et al. Ileal–lymphoid–nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet*. 1998;351:637-641.
13. Nicholls LJ, Jones CR, Gibbons WA. Proton magnetic resonance study of conformational dynamics, coordinated internal motions, and chemical shifts of tocinamide. *Biochemistry*. 1977;16:2248-2254.
14. Jones TA, Zou JY, Cowan SW, Kjelgaard M. Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in these Models. *Acta Cryst A*. 1991;A47:110-119.
15. Berntsen KRM, Vriend G. Anomalies in the refinement of isoleucine. *Acta Cryst D*. 2014;D70:1037-1049.
16. Rodriguez R, Chinea G, Lopez N, Pons T, Vriend G. Homology modelling, model and software evaluation: three related resources. *Bioinformatics*. 1998;14:523-528.
17. Venselaar H, Joosten RP, Vroling B, et al. Homology modelling and spectroscopy, a never-ending love story. *Eur. Biophys J*. 2010;39: 551-563.
18. http://www.eoht.info/page/Neumann-Shannon+anecdote
19. https://en.wikipedia.org/wiki/Murphy's_law
20. Dauter Z, Baker EN. Black sheep among the flock of protein structures. *Acta Cryst D*. 2010) D66;1.
21. https://www.cbc.ca/radio/asithappens/as-it-happens-the-friday-edition-1.4955963/journal-removes-poop-drawing-with-donald-trump-s-face-but-offers-no-explanation-1.4955972
22. https://simple.m.wikipedia.org/wiki/Second_law_of_thermodynamics

How to cite this article: Lange J, Vriend G. The Shameon enthalpy function P.Leu(P) illustrates the dangers of overzealous protein structure validation. *Proteins*. 2019;1–4. https://doi.org/10.1002/prot.1234