Citius, Altius, Fortius

J Lange[1,4], G Vriend[2,3,4]
1) CMBI, Radboudumc, Nijmegen, The Netherlands
2) BIPS, Baco, Mindoro, Philippines
3) To whom insults should be addressed
4) These authors contributed equally little to this work

Abstract

2020 is a leap year. That means that we have one day extra and, if the Olympic games had survived the corona crisis, we would all be watching television and ask the eternal question whether Olympic records will for ever be broken and broken again, or that there are limits to human biology[1]. In this article we ask the same question, but rather than discussing aspects of Citius, Altius, and Fortius of athletes we will discuss them for macromolecules. It is remarkable how many parallels can be found between Olympic records in these two seemingly different worlds.

People involved in structure validation and re-refinement try to make us believe that most aspects of macromolecular structures can be caught by a number that has some constant value with little variation around it. We will show here that the PDB[2] databank proves this idea to be wrong. In the protein structure world, it holds for many that "participating is more important than winning", but some, fortunately, still go for the record books.

Introduction

Protein structure quality evaluation has been plaguing science ever since Bränd´en and Jones told the crystallographic community that they should deposit their macromolecular coordinates without errors[3]. This concept *without errors* was widely misinterpreted as *without 6σ deviations*[4]. This validation plague started with three articles in the early 90's[5,6,7], and has taken such grotesque proportions that it became "The war of tools"[8].

All validation tools use the dictate of Engh and Huber[9] for bond lengths and bond angles. Tronrud *et al*[10] showed that using results from the refinement of one protein in the refinement of another protein makes those proteins look like each other, and Touw and Vriend[11] even claimed that they understood the molecular basis for the circularity of the reasoning behind some of the Engh and Huber parameters.

The PDB[2] nowadays provides a validation server (https://validate-rcsb-1.wwpdb.org/) that checks that protein structures have been refined using the Engh and Huber dictates. This server makes rather much the same mistakes as Hooft *et al*[12] who made a list of a million errors in the PDB that we will show to merely be Olympic records of the protein world.

Methods

Cheating is a favourite pass-time for many, especially when feeling that we can get away with it (e.g. https://en.wikipedia.org/wiki/Tax_returns_of_Donald_Trump). But cheating happens everywhere else too;

like in the Olympics (https://www.britannica.com/list/8-olympic-cheating-scandals) and, amazingly, even in crystallography[13,14,15,16]. The Olympic games have been marred by a large number of doping abuse cases, and the number of athletes caught increase from games to games (https://en.wikipedia.org/wiki/Doping_at_the_Olympic_Games). The systematic country-wide doping abuse of East-Germany, though, remained undetected too long to be backed up by physical evidence. Something similar is going on in crystallography. In pre-history, structures were built by hand (see e.g. Figure 1) and cheating was difficult because one could always check the conclusions by travelling to the lab that built the model, and remeasure everything.

Figure 1. Protein models as they were built in the good old days; before computers came around to spoil the fun. *These metal models had one big problem, all residues of a certain type always had the same bond lengths and bond angles. (Figure courtesy A Finkelstein).*

At some moment, though, computers became available, and from then on crystallographers could cheat much more eloquently by using refinement software with restraints and constraints, and parameter sets like those of Engh and Huber. Fortunately, not all crystallographers do this, so reality is not completely hidden and the Olympic records for bond lengths can still be determined.

Proprietary software[17] was used to scan the PDB for a whole series of different aspects, such as bond lengths and angles, fraction of a certain residue type, the number of water molecules, ions, co-factors, and many more. Several PDB Olympic records will be discussed here, others will be discussed in the next leap year.

year	100m	discus	pole	mara.	jump	1500m	jave.
1896	12.0	29.15	3.30	2:58		4.33	
1900	11.0				7.19	4.06	
1904		39.28	3.50			4.05	
1908	10.8	40.89	3.71	2:55	7.48	4.03	54.83
1912		45.21	3.95	2:36		3.56	60.64
1920				2:32			
1924	10.6	46.16				3.53	62.96
1928					7.73	3.53	
1932	10.3	49.49	4.32	2:31		3.51	72.71
1936				2:29	8.06	3.47	
1948							
1952		55.03	4.55	2:32		3.45	73.78
1956				2:25		3.41	
1960	10.2	59.18	4.70	2:15	8.12	3.35	84.64
1964	10.0	61.00	5.10	2:12			
1968	9.9	64.78	5.40		8.90	3.34	90.10
1972							
1976		67.50		2:09			94.58
1980			5.78				
1984				2:09		3.32	
1988		68.82	5.90				
1992							
1996	9.8	69.40	5.92				
2000						3.32	
2004		69.89	5.95				
2008	9.7		5.96	2:06			
2012	9.6		5.97				
2016			6.03				

Fig 2. Olympic records over the years. *Times are in seconds, minutes, or hours; distances are in meters. Data was extracted with some difficulty from the olympic.org website that the Olympic committee has meticulously maintained since 1896. Not all events took place at all Olympic games. Non-pandemic man-made catastrophes prevented holding the games three times.*

Figure 2 lists the Olympic records for seven arbitrarily selected events. Some records are broken often, some only occasionally. Some records, like discus or javelin throw, improved dramatically over the years, others, like the 100-meter running, only improved by a few seconds. Many entries in this figure can be discussed or criticized; like the East-German records in the '70s and '80s, or a whole series of other controversies (https://en.wikipedia.org/wiki/List_of_Olympic_Games_scandals_and_controversies). Figure 2, however, is merely meant to illustrate the trend that records keep getting broken, some occasionally, some almost every year. At some Olympic games many records are broken, at other games almost none. Amazingly, the Figures 2 and 4 show that the same happens in the protein world; records keep getting broken; some of them often, some of them occasionally. Some proteins break just one record while other proteins break many records.

Results

Proteins can be caught in numbers in too many different ways to exhaustively enumerate[18]. Given the limit on the number of pages a Proteins-reader is willing to read[19], we restricted the results to five representative examples.

Protein backbone bond lengths. Often, strain is exerted on a protein's backbone, and the more strain, the longer (or shorter) bonds can become. Figure 3 shows which four bonds we followed over the years. Figure 4 shows how these backbone bond length records kept improving over the years, both in terms of being the longest and being the shortest.

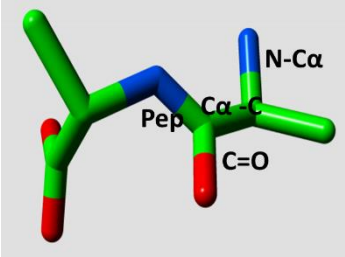


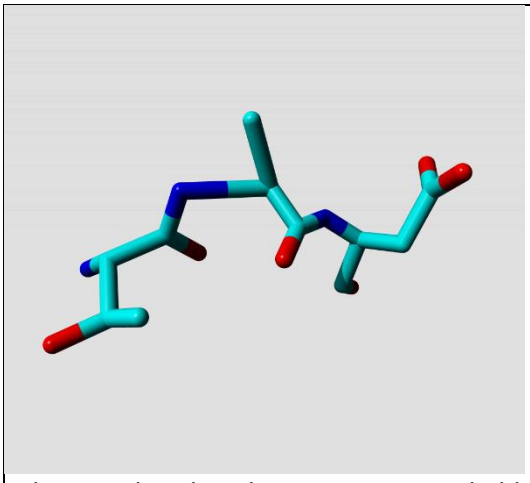
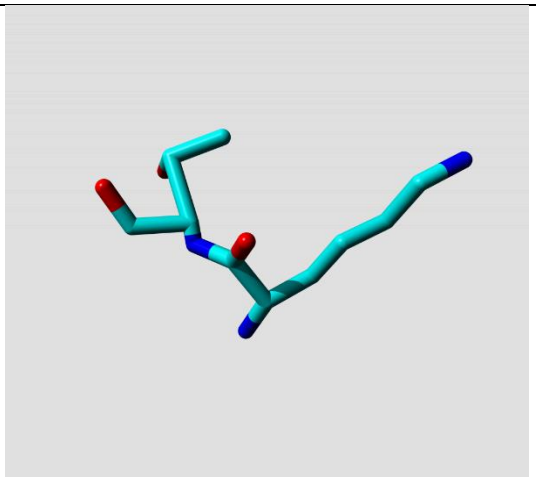
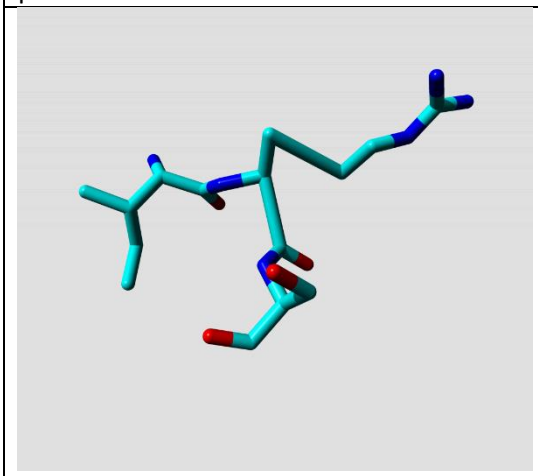
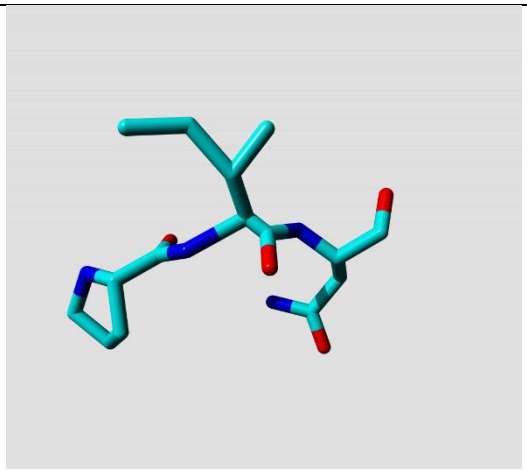
Figure 3. The four backbone bond lengths for which the records from year to year are listed in Figure 4. *Pep stands for the peptide bond.*

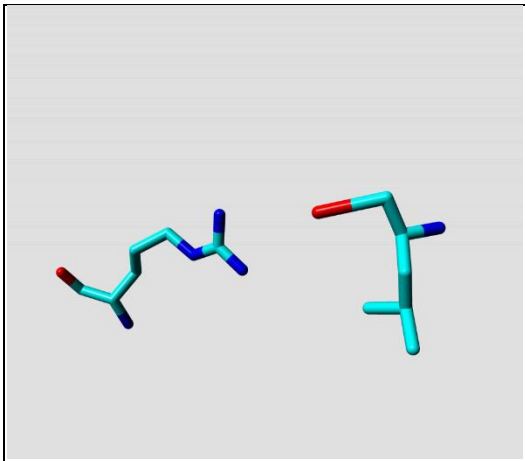
Year	N-Ca	Ca-C	C=O	peptide
1973	1.57 1mbn	1.64 1mbn	1.37 1mbn	1.44 1mbn
1975	1.65 3lyz			
1976	1.71 2sbt	1.75 2sbt	1.37 2sbt	1.89 2sbt
1981			1.61 1pfc	
1982	1.85 2taa		1.65 1hbs	
1983				2.06 3bp2
1984		1.87 4rxn		
1990		2.00 6gpb	2.16 6gpb	2.39 2hmq
1993				2.48 1bdm
1994	1.90 1aic			
1995	1.99 1nch			
1996	2.16 1dkx	2.10 1qbb		
1998	2.37 1bn6			
2001			2.26 1ihj	
2002		2.39 1lmi		
2006			2.33 2j3u	
2009		2.46 2wes		
2011	2.45 2yda			

Year	N-Ca	Ca-C	C=O	peptide
1973	1.37 1mbn	1.42 1mbn	1.16 1mbn	1.22 1mbn
1976	1.15 2sbt	1.35 2sbt	1.01 2sbt	0.75 2sbt
1979	0.92 1tgb			
1982		1.31 1hbs	1.00 1hbs	
1990		1.27 6gpb		0.54 3ypi
1991		1.19 2rcr		
1995		0.77 1nch		
1996			0.98 1bka	
1997			0.89 1ld9	0.51 1a12
1998			0.42 1tr1	
1999	0.91 1qun			
2001				0.30 1k39
2002				0.19 1m7d
2011	0.58 2y8c			

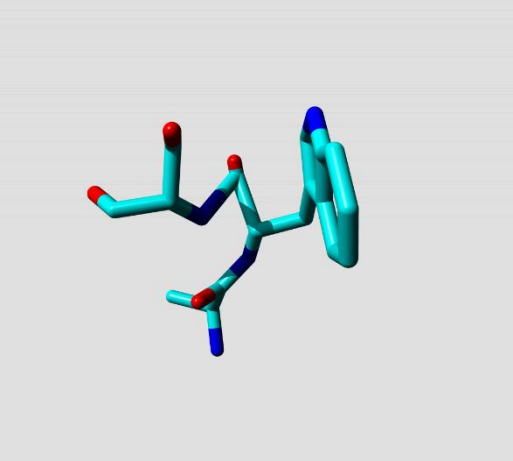
Figure 4. Records for the longest (left) and shortest (right) backbone bond-lengths over the years*. In the real Olympics every winner in 1896 was immediately a record holder. For proteins, obviously, a similar effect is seen. The WHAT IF software cannot detect bonds that are longer than 2.5 Å. So, we must apologize to the brave experimentalists who already broke this record without getting credits for their efforts. We don't know if it is coincidental that no records were broken after PDB_REDO[20] was broadly advertised and placed in a wider contex[21]. A number of PDB files were rejected because of cheating with experimental real space refinement, or similar techniques.*

The bond length records behave over the years similar to real Olympic records. 2SBT was, in 1976, a bit a Mark Spitz (http://www.famousdaily.com/history/mark-spitz-wins-7-olympic-gold-medals.html) in terms of winning everything with a record. We feel that 2SBT cheated a bit as it was refined using the software package NULL that is known to work better than all other programs. Nevertheless, all of 2SBT's records were broken later, mostly by PDB entries that were refined using the mainstream refinement packages X-Plor or CNS. Figure 5 shows the final record holders.

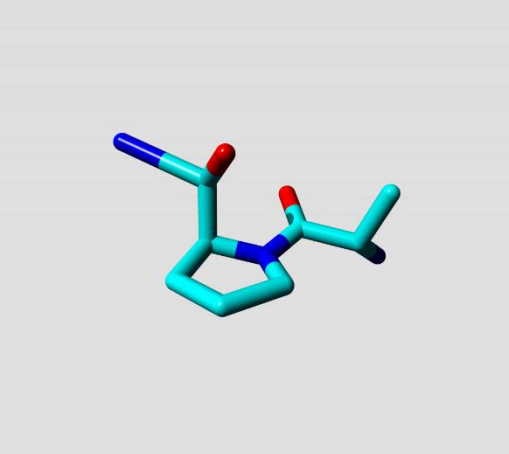
<table>
<tr>
<td> 
The N-Cα bond in Ala-166 in 1BN6 probably is so long because the occupancy refinement used so many steroids that locally many atoms are present for much more than 100%.</td>
<td> 
The N-Cα bond in Thr-21 in 2Y8C might have escaped the local attractive forces of Lys-20 because this lysine has occupancy 0.0.</td>
</tr>
<tr>
<td> 
The Cα-C bond in Arg-3D in 2WES is probably very long to keep in harmony with the Cβ-Cγ bond in its side chain.</td>
<td> 
The Ca-C bond in Ile-7B in 1NCH is probably very short to compensate for the sharp angle in the side chain.</td>
</tr>
</table>



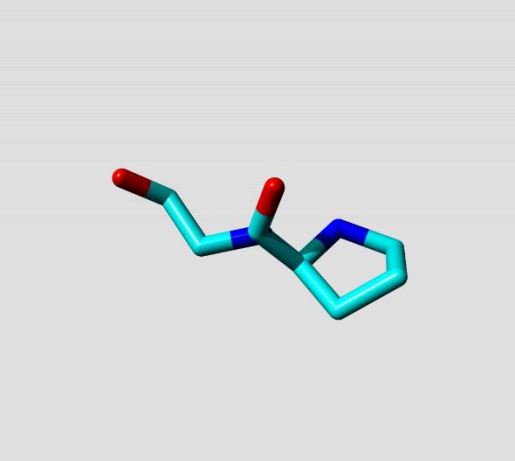
The C=O bond in Leu-74D in 2J3U got extended by the strong hydrogen bond(s) with Arg-86 in the same chain.



The C=O bond in Trp-398A in 1TR1 got shortened to compensate for the short contact distances and sharp angles in the peptide bond.



The Pro90-Arg91 peptide bond in chain A of 1BDM might be so long because the arginine consists of a single nitrogen atom only?



The Pro14-Gly14 peptide bond in the B chain of 1M7D really got short because we see a mixture of two styles of peptide bonds, the classical one and a reverse atom-order peptide bond.

Figure 5. The 8 record holders for backbone band lengths.

C-terminus to OXT distance. It is commonly known that strange things can happen at the end of a wave (e.g. https://www.youtube.com/watch?v=k6fr6GUSmAA) and similar effects are observed when waves run through the backbone of a protein. At the C-terminal end of a chain we can observe weird effects, especially for the OXT atom that is really at the very end of any chain (OXT is the non-IUPAC PDB name for the second O on the C-terminal residue). Because of the same effect as seen in the above youtube movie the OXT gets a big swing when a backbone wave arrives at the C-terminus. Figure 6 shows the C-terminal leucine in the I chain of 1YPG. We found half a dozen longer C-OXT bonds but those were all the result of cheating by swapping OXT atoms between chains or using non-canonical residues. A chemist could easily make you believe that the two oxygens at the end of a chain are equivalent, but we know better. And figure 6 shows that indeed the distances of the two C-terminal oxygens to their carbon are not identical. And, not surprisingly, the two refinement programs commonly known to work best (NULL and SHELLXL) on average have the largest difference between these two C-O distances. These are also the only two programs that detected the systematic difference between de C-O and C-OXT distances.

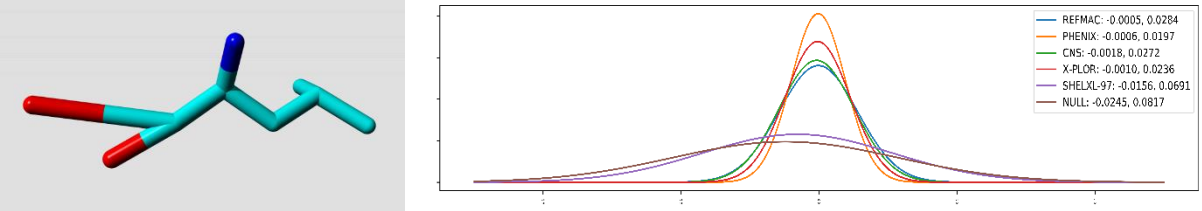


Figure 6. C-O versus C-OXT distances. *Left: The C-OXT bond in the C-terminal leucine in the I chain of 1YPG is more than 3.5 Å long. Right: Gaussian fit of the histograms of differences in the C-O and C-OXT distances in PDB files. The insert gives the average and standard deviations of these six distributions. We apologize to the authors of many other refinement software packages. Several of them could easily have made it into the record books, but low counting statistics precluded a good fit of the histograms.*

Cα-Cβ bond lengths in β-branched residues. Touw and Vriend claim to understand why the angle τ in amino acids (τ: N-Cα-C in backbone) differs in β-branched residues in many aspects from this angle in the 17 other canonical amino acid types. Figure 7, shamelessly copied from their article[10], is the linchpin around which their reasoning revolves. But, like so many[22], they were blinded by using PDB_REDO files, rather than the real data, and thereby forgot the much simpler option that the Cα-Cβ bond can get a bit longer to move the Cγ atoms away from the local backbone.

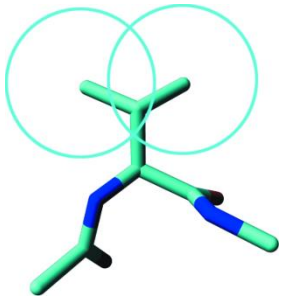


Figure 7. The reason that the τ angle gets sharper in β -branched residues. *"Both Cγ atoms in Val push against their own backbone. The two circles that are centred on the Cγ atoms have a radius of about 1.8 Å, reflecting a commonly used Van der Waals radius for these CH_3 groups".*

We decided to check this simpler solution and analysed the Cα-Cβ bond lengths in Isoleucine, Threonine, and Valine residues in all PDB files that hold 50 amino acids or more. In smaller proteins all residues are at the surface anyway and the solvent will make that the results water down too much. Figure 8 shows the three clearest examples of this much smarter solution for the atomic clashes.

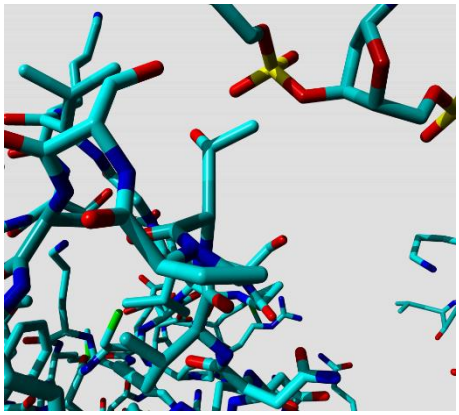
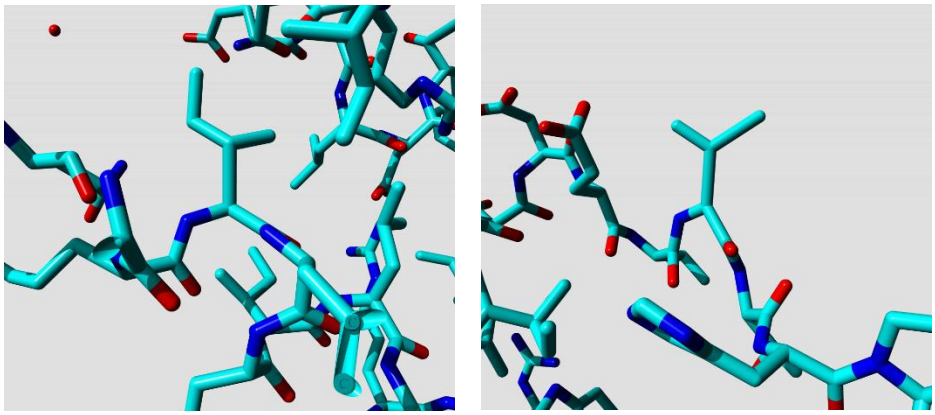


Figure 8. For each of the three β-branched residue types the clearest example of the Cα-Cβ bond extension is shown*. Top: Threonine 117E in 1R9T. It is possible that this bond elongation is supported by the attempt of the threonine's Cγ to covalently bind the phosphate group of thymine 12 in chain N. Bottom left: Isoleucine 27 in chain D in 4BPT. Bottom right: valine 11 in the A chain of 5K98. In this residue the energy needed to make the Cα-Cβ bond so much longer probably comes from the rather uneven Cβ-Cγ bond lengths.*



The most cuddable residue. We looked for the most cuddable residue. Cuddability is an important factor in Olympic events like horse- and ice-dancing, and, of course, in election years in the USA, and thus probably also in science. Not to our surprize did we find the word 'cuddable' much more often through Facebook then through Google Scholar, which taught us again to avoid fake news and make sure we supported all conclusions with solid FB-based evidence. It was remarkably difficult to come up with a good definition for cuddability of amino acids, but at the end histidine 27 in 1HCE (see Figure 9) won by a unanimous vote of the authors.

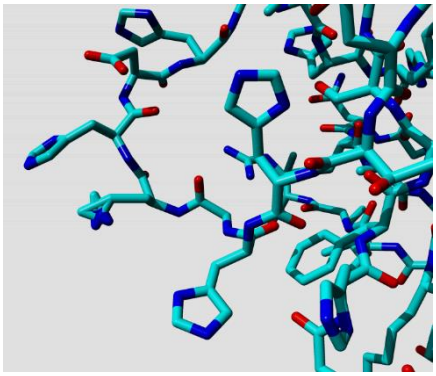


Figure 9. The most cuddable residue is histidine 27 in 1HCE*. This histidine is found in the stretch His-25, Gly-26, His-27, His-28 in the lower left of this figure.*

The shortest hydrogen bond. We have always been puzzled by the term 'hydrogen bond'. The name suggests it to be a bond between two hydrogens, but most scientists see that differently[17]. We decided to look for two kinds of hydrogen bonds, those between the hydrogens on the positive nitrogens of two lysines and of two arginines. Figure 10 shows that hydrogen bonds indeed are bonds between hydrogens.

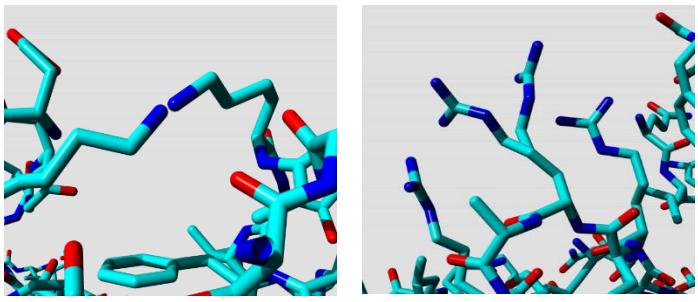


Figure 10. Lys-Lys and Arg-Arg hydrogen bonds. *Left: The N-N-distance in the lysines 53A and 104A in 5B04 is 0.69Å, which proves that the two hydrogens must be covalently bound. Right: arginine 299 in the H chain of 1A7R has two alternate conformations. One forms a 2.2Å hydrogen bond with arginine 302H, the other alternate has a 2.9 Å hydrogen bond with arginine 297H. This arginine example is not the record holder (it isn't even a covalent bond) because many shorter arginine-arginine hydrogen bonds can be found in the PDB. But we felt that this arginine really proves our point: even when alternate conformations are available then both will try to form a nice hydrogen bond.*

Discussion

One of the main errors made by modern-day crystallographers is to rely on standard values for bond lengths and angles that are extracted from the CCD[23]. As Figure 11 shows, this can lead to big errors, especially when Average, Mean, Mediocre, and Median are being confused (see https://statistically-funny.blogspot.com/2015/). So, we strongly discourage the use of the Engh and Huber dictate in coordinate refinement.

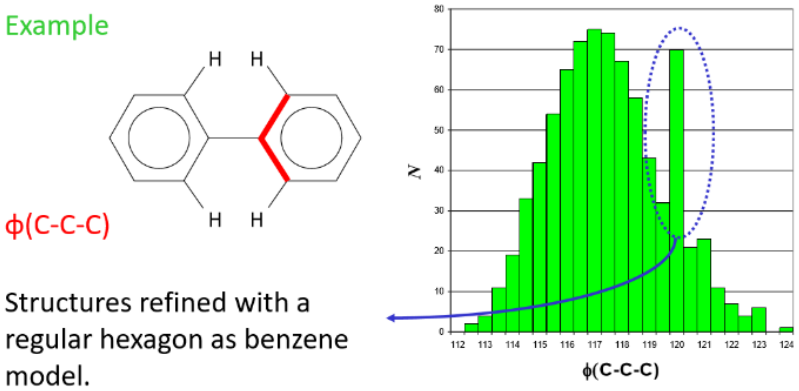


Figure 11. A search in the CCD for the angle φ (three carbons in an aromatic plane of which the centre one is bound to 'something'). *The funny peak at 120° is the result of molecules overzealously refined with this value in the library. This is one more warning against the use of libraries in refinement.*

In the end, we felt the need to determine an overall winner, a greatest Olympian of all times. After all, the article is about citius, altius, fortius. We decided that 2PDE is the greatest Olympian of the PDB. This decision was based on its many remarkable features (see Figure 12).

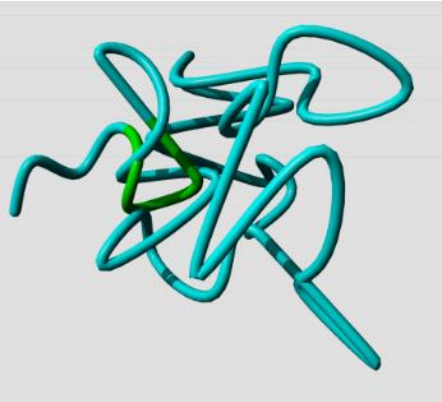


Figure 12. The winner of the PDB-Olympics is 2PDE*. This remarkable structure indeed has one residue that in the Ramachandran plot falls within the allowed regions. It proves Kim Henrick wrong when he said that "quantum chemistry breaks down in the hands of a crystallographer", because 2PDE was 'solved' by NMR. It has equally many left as right handed chiral centers. It has more severe bumps than residues. It broke the stupidly fixed, limited output format of WHAT_CHECK[12] at 6 locations. And finally, it boldly bend the backbone in ways backbones have never been bend before*.

Acknowledgments

The authors thank Robbie Joosten and Rob Hooft for critically reading this manuscript, and Huub Kooijman for help with the CCD search.

References

1. Will athletes ever stop breaking world records? The Week. https://theweek.com/articles/473488/athletes-ever-stop-breaking-world-records.

2. Berman HM. The Protein Data Bank: a historical perspective. *Acta Crystallogr A*. 2008;64(1):88-95. doi:10.1107/S0108767307035623

3. Bränd´en C-I, Alwyn Jones T. Between objectivity and subjectivity. *Nature*. 1990;343(6260):687-689. doi:10.1038/343687a0

4. Read RJ, Adams PD, Arendall WB, et al. A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure*. 2011;19(10):1395-1412. doi:10.1016/j.str.2011.08.006

5. Vriend G, Sander C. Quality control of protein models: directional atomic contact analysis. *J Appl Crystallogr*. 1993;26(1):47-60. doi:10.1107/S0021889892008240

6. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr*. 1993;26(2):283-291. doi:10.1107/S0021889892009944

7. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. *Proteins Struct Funct Bioinforma*. 1993;17(4):355-362. doi:10.1002/prot.340170404

8. Saccenti E, Rosato A. The war of tools: how can NMR spectroscopists detect errors in their structures? *J Biomol NMR*. 2008;40(4):251-261. doi:10.1007/s10858-008-9228-4

9. Engh RA, Huber R. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr A*. 1991;47(4):392-400. doi:10.1107/S0108767391001071

10. Tronrud DE, Berkholz DS, Karplus PA. Using a conformation-dependent stereochemical library improves crystallographic refinement of proteins. *Acta Crystallogr D Biol Crystallogr*. 2010;66(7):834-842. doi:10.1107/S0907444910019207

11. Touw WG, Vriend G. On the complexity of Engh and Huber refinement restraints: the angle τ as example. *Acta Crystallogr D Biol Crystallogr*. 2010;66(12):1341-1350. doi:10.1107/S0907444910040928

12. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. *Nature*. 1996;381(6580):272. doi:10.1038/381272a0

13. H.M. Krishna Murthy. Wikipedia. https://en.wikipedia.org/wiki/H.M._Krishna_Murthy.

14. Weiss MS, Einspahr H, Baker T, Dauter Z. Another case of fraud in structural biology. *Acta Crystallograph Sect F Struct Biol Cryst Commun*. 2012;68(4):365-365. doi:10.1107/S1744309112011852

15. New Structures for Old: A Cautionary Tale of Fraud in Small Molecule Crystallography. http://journals.iucr.org/services/coeditors/meetings/jcomm/slides/plenary_new_structures_for_old_sj.pdf.

16. Matthews BW. Five retracted structure reports: Inverted or incorrect? *Protein Sci*. 2007;16(6):1013-1016. doi:10.1110/ps.072888607

17. Vriend G. WHAT IF: a molecular modeling and drug design program. *J Mol Graph*. 1990;8(1):52-56, 29.

18. Petsko GA. Large cast, but no plot. *Nature*. 1992;359(6396):596-597. doi:10.1038/359596a0

19. Haslam N. Bite-Size Science: Relative Impact of Short Article Formats. *Perspect Psychol Sci*. 2010;5(3):263-264. doi:10.1177/1745691610369466

20. Joosten RP, Salzemann J, Bloch V, et al. PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J Appl Crystallogr*. 2009;42(3):376-384. doi:10.1107/S0021889809008784

21. Touw WG, Baakman C, Black J, et al. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res*. 2015;43(Database issue):D364-D368. doi:10.1093/nar/gku1028

22. Dauter Z, Weiss MS, Einspahr H, Baker EN. Expectation bias and information content. *Acta Crystallograph Sect F Struct Biol Cryst Commun*. 2013;69(2):83-83. doi:10.1107/S1744309113001486

23. Taylor R, Wood PA. A Million Crystal Structures: The Whole Is Greater than the Sum of Its Parts. *Chem Rev*. 2019;119(16):9427-9477. doi:10.1021/acs.chemrev.9b00155