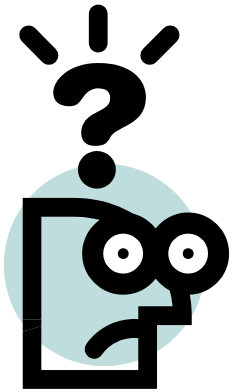
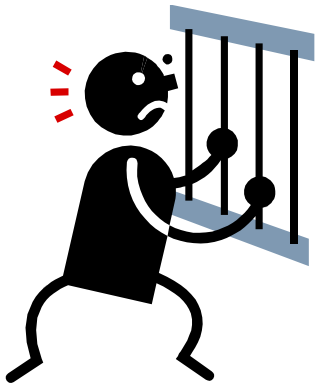
New Structures for Old: A Cautionary Tale of Fraud in Small Molecule Crystallography



Jim Simpson

Department of Chemistry
University of Otago



ISSN 1600-5368

Volume 66

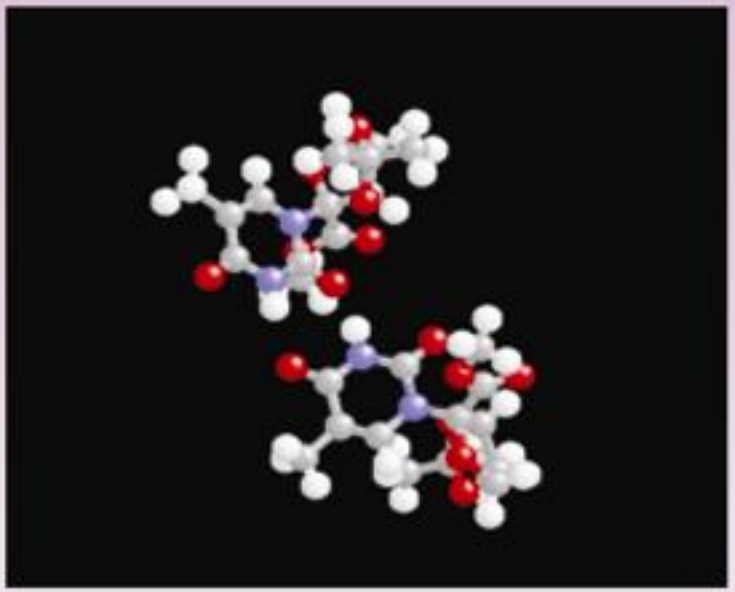
Part 2

February 2010

Acta Crystallographica Section E

Structure Reports Online

Editors: W. T. A. Harrison, J. Simpson and M. Weil



Inorganic compounds

Metal-organic compounds

Organic compounds

journals.iucr.org

International Union of Crystallography • Chester

Background

- Acta Cryst E
  - First published 2001
  - Since November 2008 the first Open Access journal from the IUCr
  - In 2010 published 4113 papers each reporting an individual small molecule structure.

Background

- Acta Cryst E
  - Simple format – an abstract, scheme, related literature section and an optional comment, plus references and information on the structure determination.
  - Designed to encourage publication of all structures – particularly the “orphans” that would not be readily included in a more substantial paper
  - This makes the journal very attractive to authors with a poor command of English or for whom English is not their first language

Top 10 authorship by country 2010

- China 38%
- Malaysia 12%
- India 8%
- Pakistan & USA 5%
- Germany 4%
- Korea 3%
- Turkey, Iran, Morocco 2%

Validation procedures – pre 2009

- CheckCIF – based on PLATON
  - Checks that all required information is present.
  - Information is internally self consistent.
  - Data and structure quality tests
  - Until 2009 this was the only validation procedure conducted on structures submitted for publication in IUCr journals
  - Considered by most authors to be the most rigorous of all the procedures adopted by journals reporting crystal structures.

And yet!!!!!!!

- In January 2010 an Acta E editorial announced: “Regrettably, this editorial is to alert readers and authors of *Acta Crystallographica Section E* and the wider scientific community to the fact that we have recently uncovered evidence for an extensive series of scientific frauds involving papers published in the jounal, principally during 2007.

  ….the extent of these problems is significant with at least 70 structures demonstrated to be falsified and meanwhile acknowledged by the authors as such. Our work is ongoing and it is likely that this figure will rise further.”

- Retracted total to date - **140 and rising**

How was the problem discovered?

- Ton Spek continually upgrades PLATON and the CheckCIF procedures.
- He uses CIF files picked at random from Acta E or C papers to test program updates
- In the process of upgrading Hirshfeld test checks he came across two dubious structures, clearly involving metal swapping, and alerted the Editors to the problems.
- Both structures had the same corresponding author.

Investigations begin

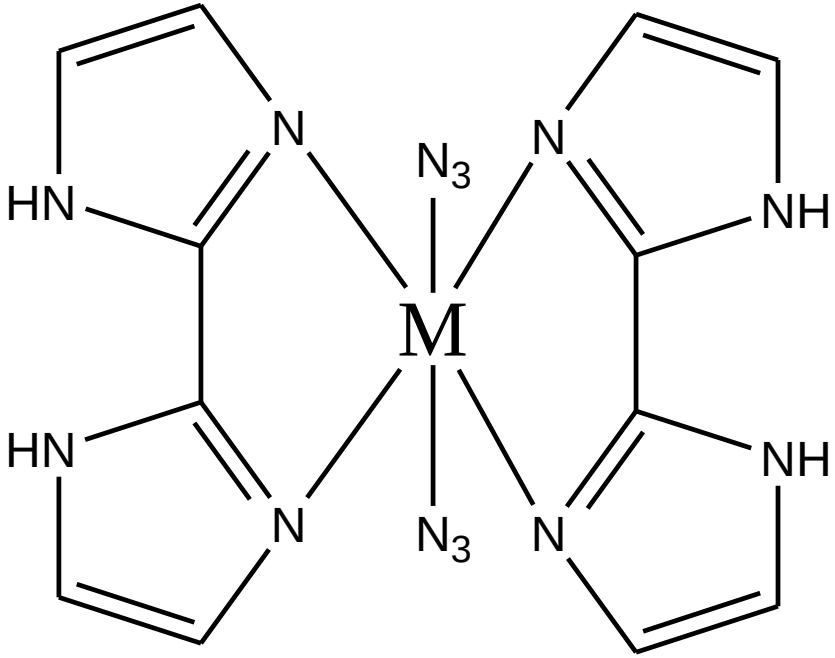
- A large number of other articles in the Journal by the same corresponding author were found when we ran checks
- Many of these showed similar problems.
- Checks were then run on other papers submitted to Acta E or C from the same University.
- Another set of structures with similar serious problems immediately showed up from a second corresponding author.

Three major strategies

- Metal swapping in coordination complexes –
- Element swapping in organic compounds
- Metal swapping accompanied by element swapping in the ligands of coordination complexes, particularly of the lanthanide elements.

Serial metal swapping

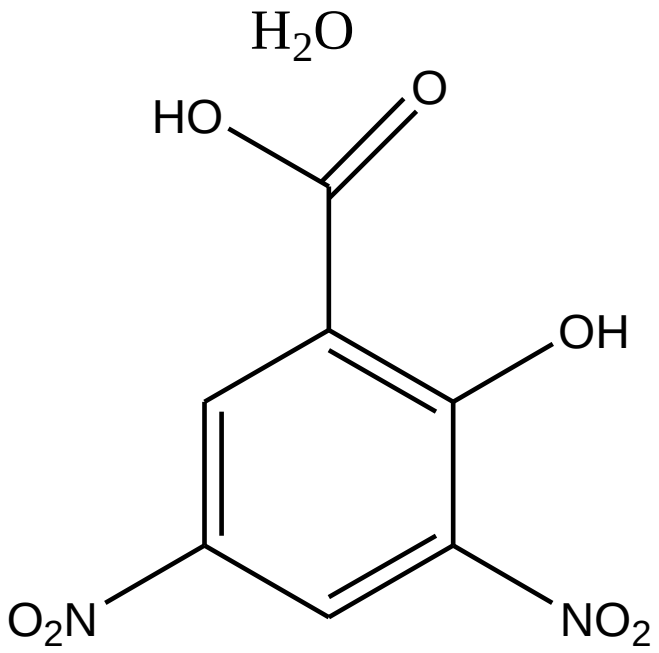
- All 5 of these 2,2'-biimidazole complexes were in fact derived from a single data set – that of the Co complex
- Came from 5 different sets of authors in 5 different institutions!



M – Mn, Fe, **Co**, Ni, Cu

Case 2 – element swapping in organic compounds

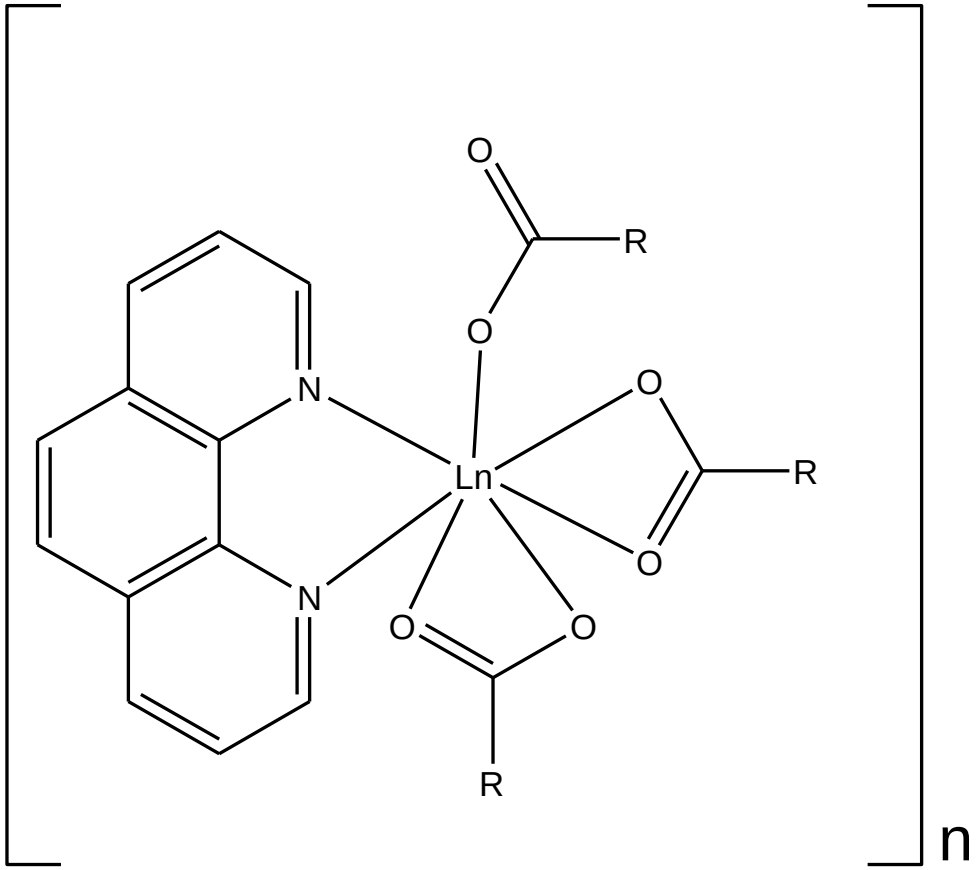
- In 1995 an Australian group reported the structure of this compound
- During 2007 no fewer than 10 look-alikes appeared



ZAJGUM

Case 3 – metal and element swapping

- These frauds involve an extensive series of Ln coordination polymers
- Ln atoms vary
- 9,10-phenantholine (phen) ligand common to all
- Acetato ligands also varied significantly
- Each reported structure derived from the same data set



Case 3 – metal and element swapping

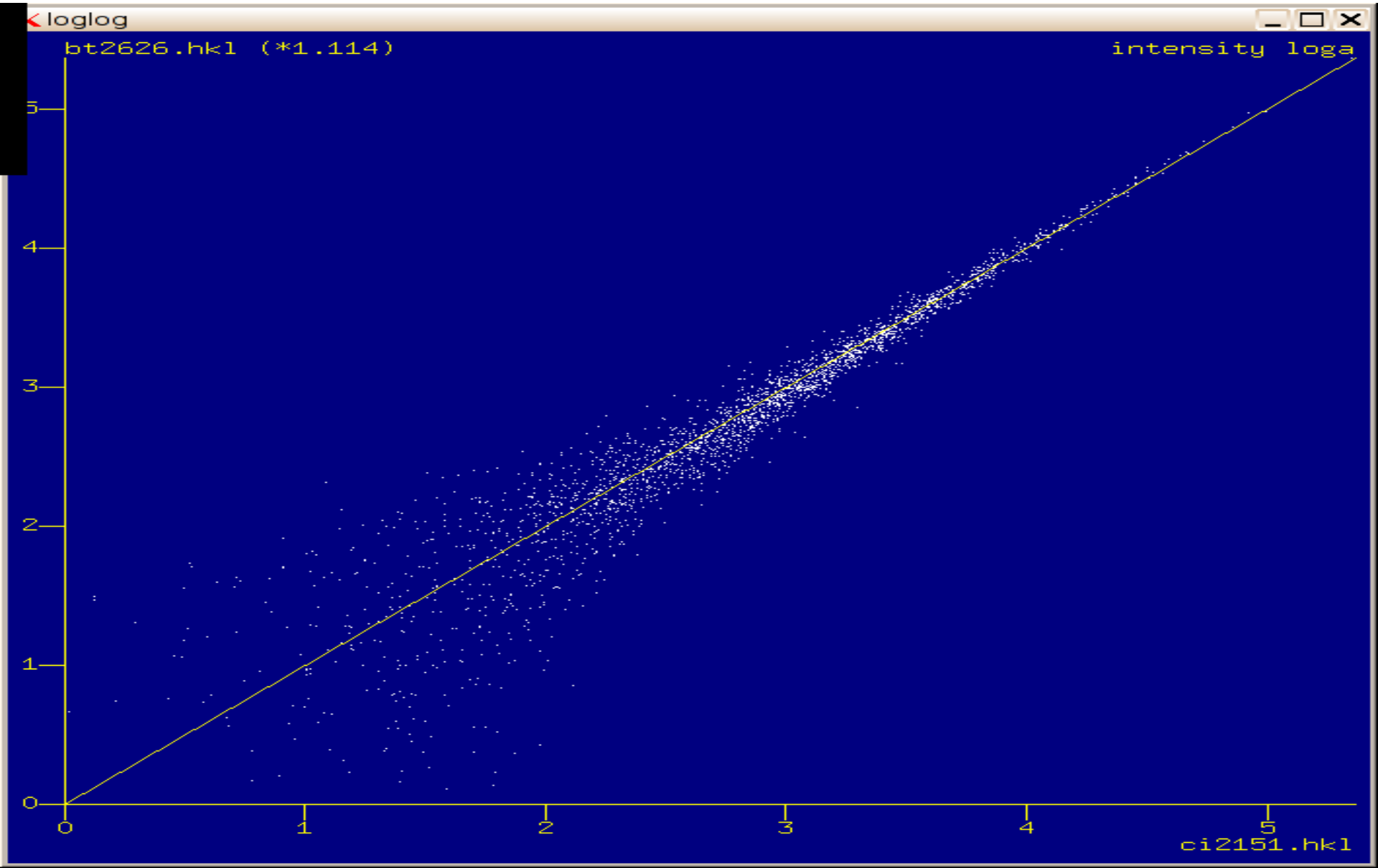
La	phenoxyacetate	$[La(C_8H_7O_3)_3(phen)]_n$
Ce	phenoxyacetate	$[Ce(C_8H_7O_3)_3(phen)]_n$
Pr	phenoxyacetate	$[Pr(C_8H_7O_3)_3(phen)]_n$
Nd	phenoxyacetate	$[Nd(C_8H_7O_3)_3(phen)]_n$
La	3-phenylpropanoate	$[La(C_9H_9O_2)_3(C12H8N2)]_n$
Nd	3-phenylpropanoate:	$[Nd(C_8H_7O_3)_3(C12H8N2)]_n$
La	2-(phenylamino)acetate	$[La(C_8H_8O_2N)_3(phen)]_n$
Nd	2-(phenylamino)acetate	$[Nd(C_8H_8O_2N)_3(phen)]_n$
Sm	2-(phenylamino)acetate	$[Sm(C_8H_8O_2N)_3(phen)]_n$
Eu	2-(phenylamino)acetate	$[Eu(C_8H_8O_2N)_3(phen)]_n$
Ce	(2-(phenylamino)acetyl)amido	$[Ce(C_8H_8ON_2)_3(phen)]_n$
Pr	(2-(phenylamino)acetyl)amido	$[Pr(C_8H_8ON_2)_3(phen)]_n$
Sm	(2-(phenylamino)acetyl)amido	$[Pr(C_8H_8ON_2)_3(phen)]_n$
La	2-(pyridin-2-yloxy)acetate	$[La(C_7H_6O_3N)_3(phen)]_n$
Pr	2-(pyridin-2-yloxy)acetate	$[Pr(C_7H_6O_3N)_3(phen)]_n$
Nd	2-(pyridin-2-yloxy)acetate	$[Nd(C_7H_6O_3N)_3(phen)]_n$

- Each carboxylate ligand has 11 C, N and/or O atoms
- 16 ‘different’ compounds generated by a mix and match process
- Data sets for each determination were shown absolutely to be essentially identical

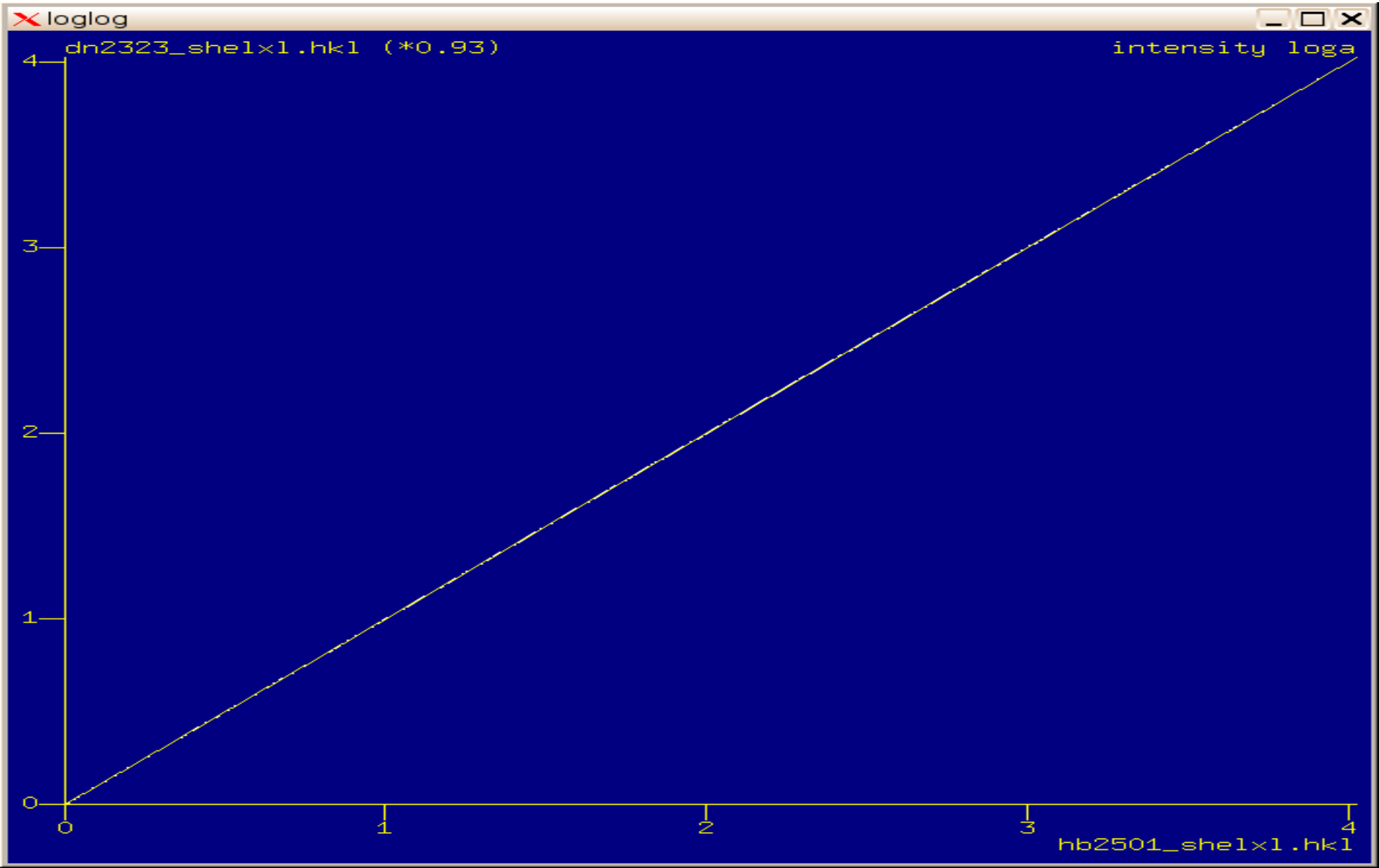
Checking for identical data-sets

- All submissions to Acta journals must deposit the X-ray data file in CIF format, known as an FCF file so that, if necessary, an hkl file can be generated from it. Only one other Journal currently requires this.
- Ton Spek commissioned a program from one of his colleagues to allow direct comparison of two hkl files.

If the files are different



But if they are the same



The retraction process

- Corresponding authors are contacted and given a detailed error report written by the investigating crystallographer.
- Asked for comments on the findings.
- If they admit the fraud, all other authors are contacted and asked to agree to the retraction.
- Article retracted either with agreement of the authors or by the Journal
- Structures reported in retracted articles are removed with the following update of the Cambridge Crystallographic Database

The aftermath

- The Editorial certainly caused a furore!!!
- Reported in most of the major Chinese newspapers including the influential “People’s Daily” and “China Youth Daily”
- Made BBC, BBC World and National Public Radio
- Articles and editorials commenting on the retractions appeared in Nature, Science, Chemistry World, even the Lancet!
- Messages of support, anger and frustration came from crystallographers worldwide.

中国青年报
中青在线 WWW.CYOL.NET

国内统一刊号
CN11-0051
邮发代号 1-9

《晶体学报》再度回应井冈山大学两名讲师学术造假事件

无意为研究者出名提供“捷径”

本报记者 来扬 雷宇

中青在线 - 中国青年报 2010-01-11 [打印] [关闭]

Simpson告诉记者，作者在向《晶体学报》E分卷投稿时，需要提交两个特定格式的ASCII文件，包括相关的图形和数据。在《晶体学报》E分卷上发表的文章由两部分组成，一部分是包括摘要在内的文字信息，另一部分是测量得到的用来描述晶体结构和测定质量的数据。在作者提交后，相关图形和数据都会接受系统的一系列自动检查。首先，系统会检查提交的文章是否具备发表所要求的所有信息；其次，系统会检查相关数据是否自洽；最后，系统会检查根据文章所提供的数据能否“重构”通过X射线衍射测定的晶体结构。检查结束后，如果发现有问题，系统会生成一份问题清单即时反馈给作者。因此，对作者来说，检查程序没有“秘密”可言。

除此之外，期刊编辑部还会对已发表的文章进行复查。复查的工作会在文章发表后的一年内进行，主要是对文章的数据与此前已有的数据进行对比。复查工作能够有效检验文章内容是否造假。正是在最近进行的复查工作中，编辑部发现了中国井冈山大学两名教师过去两年内发表的70篇文章存在造假。Simpson表示，这些文章现在都已被退回，并且，期刊编辑部已将不断检测复查的流程、不断升级复查软件列为日常工作之一。因此，读者对《晶体学报》E分卷和晶体学会下辖的其他期刊中刊登的晶体结构的质量会越来越有信心。

And the fraudsters?

- Sacked from their University positions
- Thrown out of The Party!
- Made to repay the ~$US800 per article that they were paid by their University for each article published in an international journal.
- As far as we know they weren’t shot!!!!

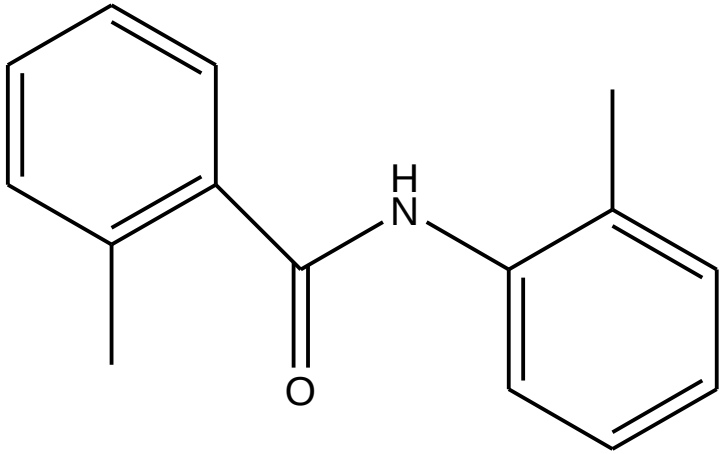
Has validation improved subsequently?

- We certainly believe so!
- The validation process for each submitted structure now converts CIF + FCF into INS and HKL files and repeats the SHELXL refinement
- Any hand altering of R factors etc thus immediately detected
- Many other criteria tightened and tests for specific substitutions such as NO_2 to CO_2^- have been introduced
- Co-editors alert to Hirshfeld problems

So how easy is it to get away with such behaviour now?

- I put this question to the test recently by converting an organic structure I published two years ago into four closely related frauds.
- Took about 90 minutes to get 4 reasonable refinements and related CIF files.

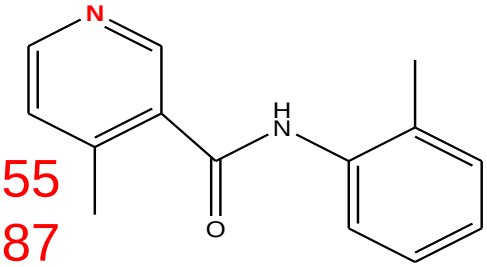
It was seemingly all too easy



2-methyl-*N*-*o*-tolylbenzamide

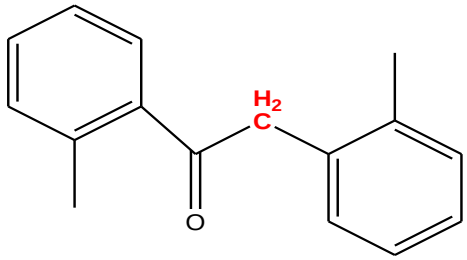
- A genuine structure I published in 2009
- Could equally well have downloaded the structure factors and CIF from someone else's B, C or E submission to generate .INS and .HKL files
- Swapped the odd C for N and *vice versa*
- Cell constants on the 'clones' were also varied somewhat in an attempt to escape detection
- R factors were reported only as the refined values

Ringing the changes



R1 = 0.0655
wR2 = 0.2087

4-methyl-*N*-*o*-tolylnicotinamide

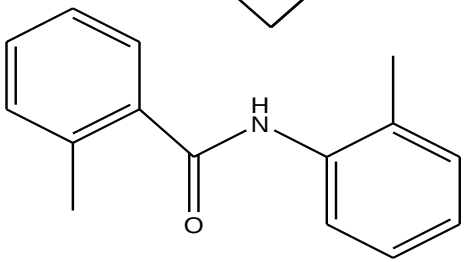


R1 = 0.0744
wR2 = 0.2397

1,2-di*o*-tolylethanone

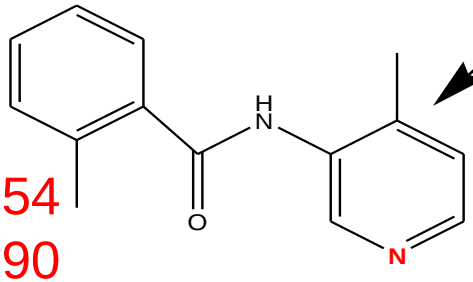
COpy

NCH_2



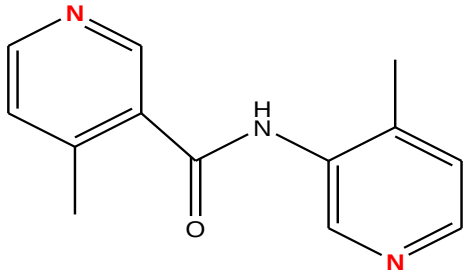
R1 = 0.0549
wR2 = 0.1678

2-methyl-*N*-*o*-tolylbenzamide



R1 = 0.0654
wR2 = 0.2090

2-methyl-*N*-(4-methylpyridin-3-yl)benzamide



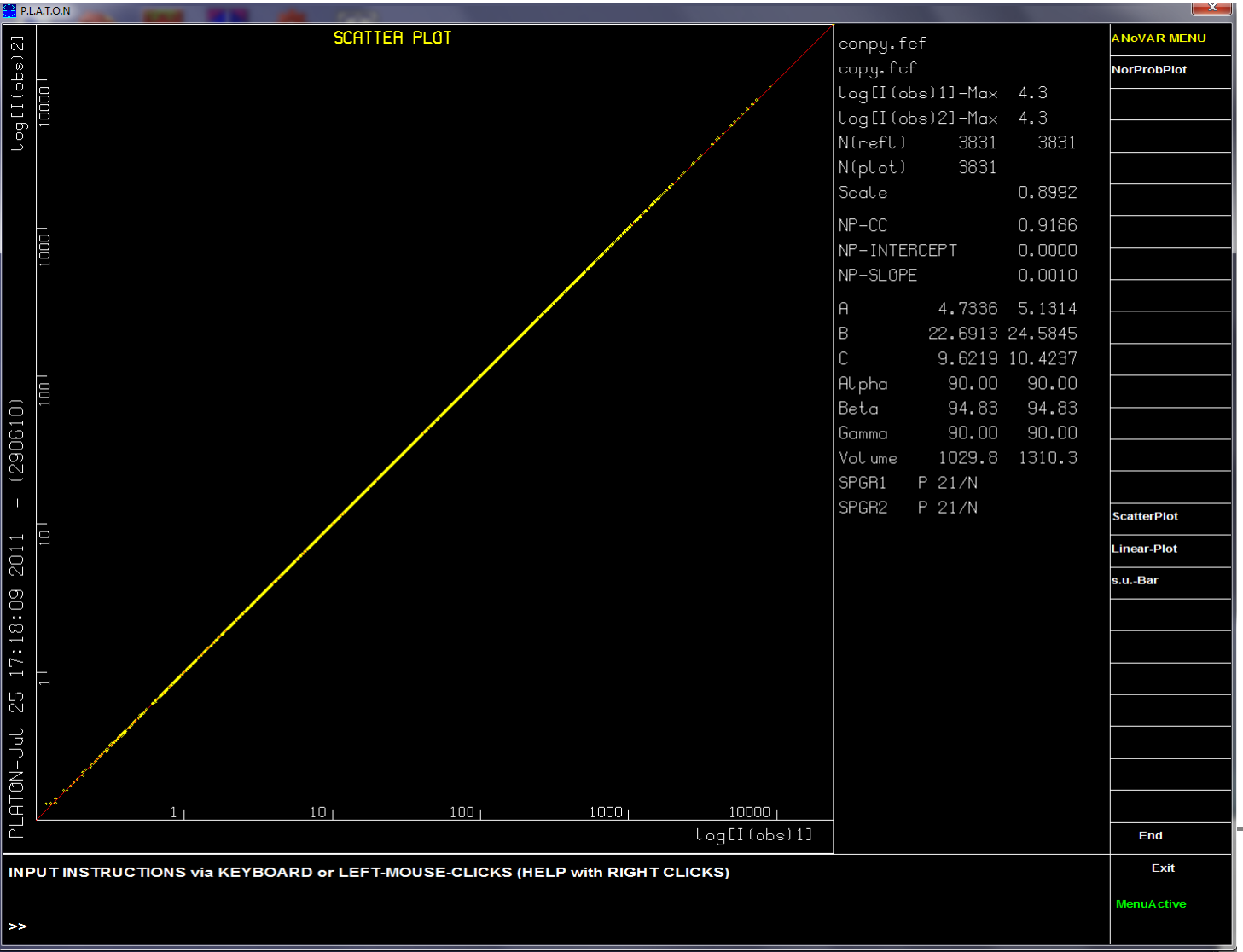
R1 = 0.0718
wR2 = 0.2349

4-methyl-*N*-(4-methylpyridin-3-yl)nicotinamide

Npy

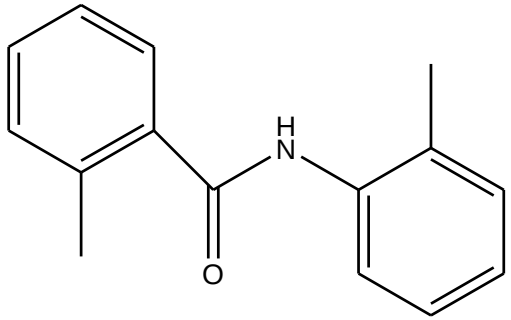
CONpy

Certainly the .FCF files for each of the clones were identical



- But such comp-arisons are unlikely to be done normally

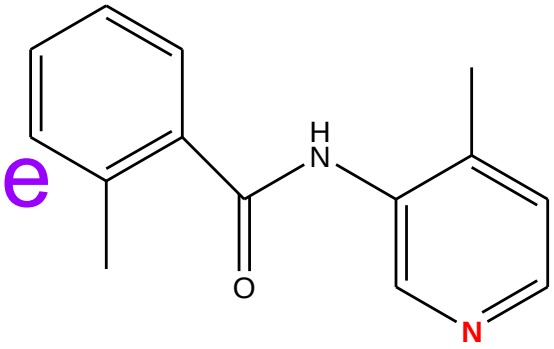
How easy is it now that CheckCIF tests have tightened appreciably?



- The original CIF gave only trivial C alerts
- But attempts to falsely improve the residuals now produce clear warnings!

- PLAT921_ALERT_1_B R1 in the CIF and FCF Differ by ..... -0.0200
- PLAT922_ALERT_1_B wR2 in the CIF and FCF Differ by ..... -0.0200
- PLAT926_ALERT_1_B Reported and Calculated R1 Differ by ..... -0.0200
- PLAT927_ALERT_1_B Reported and Calculated wR2 Differ by ..... -0.0200

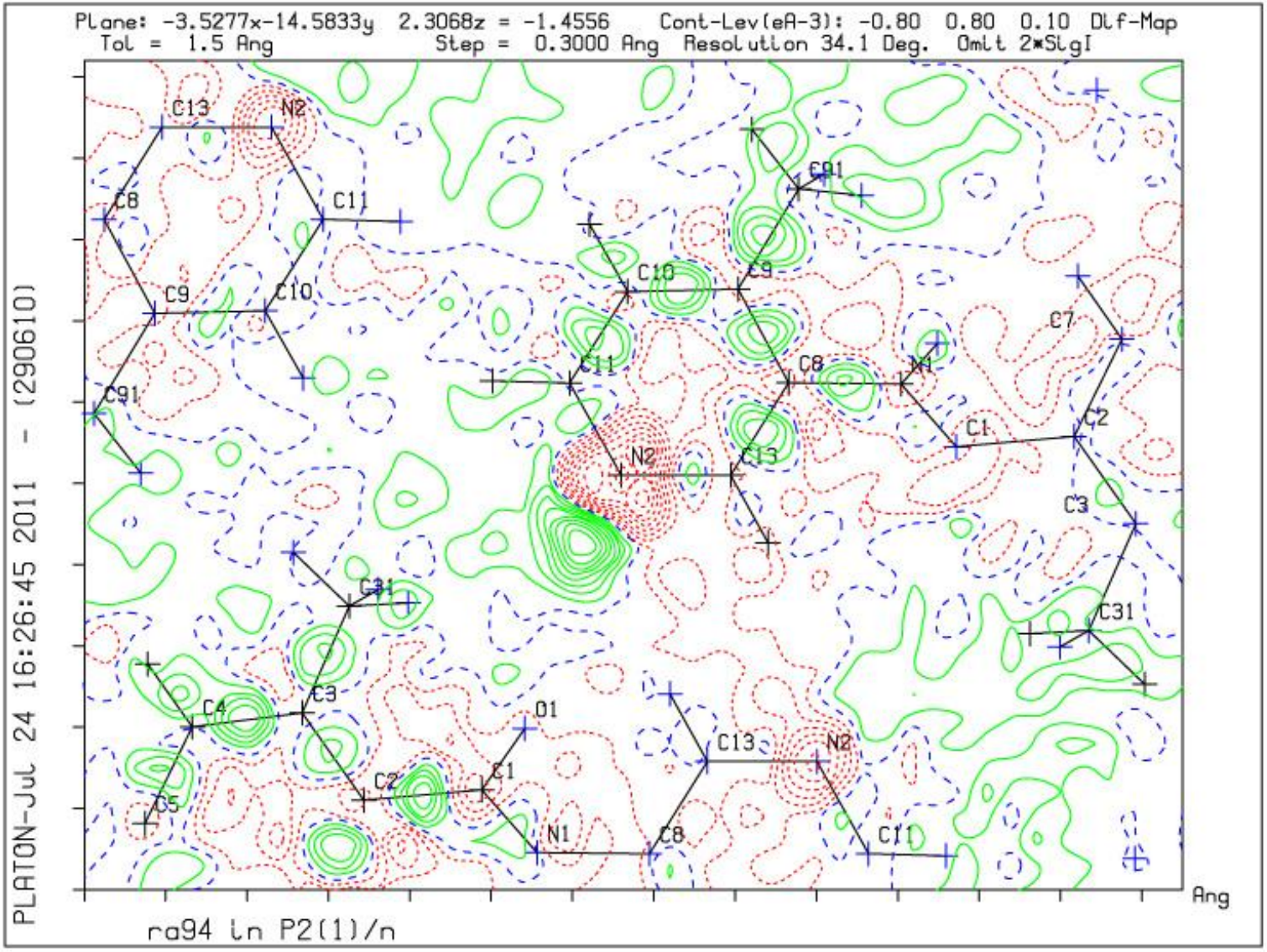
CheckCIF asks some probing questions even for the “best” clone



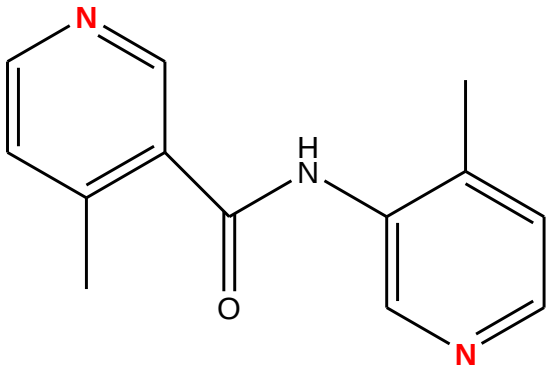
- **Alert level B**
- DIFMX01_ALERT_2_B The maximum difference density is > 0.1*ZMAX*1.00 _refine_diff_density_max given = 1.006 PLAT097_ALERT_2_B Large Reported Max. (Positive) Residual Density 1.01 eA-3
- PLAT230_ALERT_2_B Hirshfeld Test Diff for N2 -- C13 .. 7.4 su

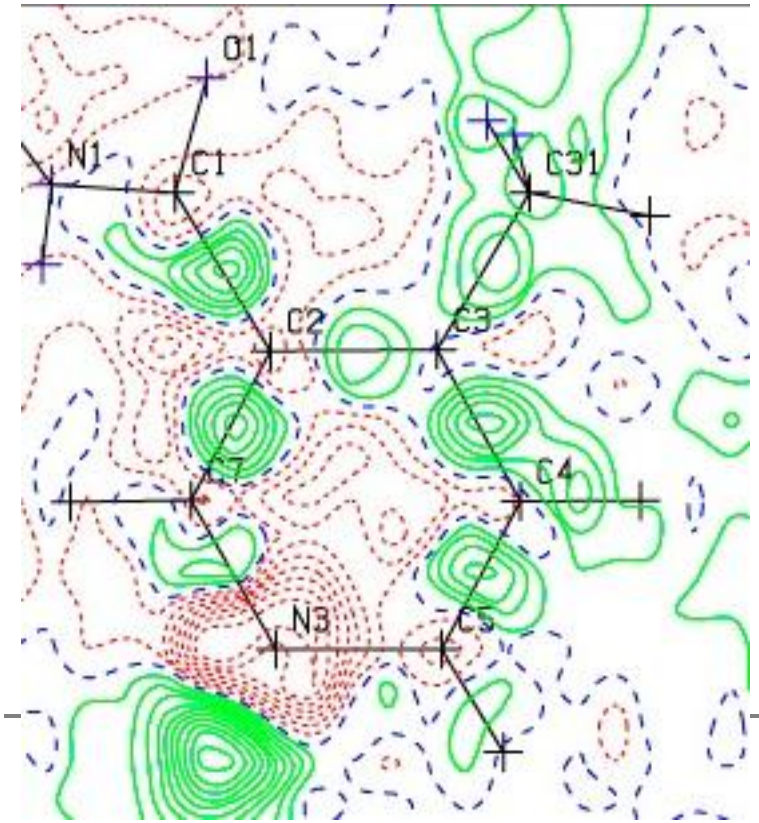
- **Alert level C**.
- DIFMX02_ALERT_1_C The maximum difference density is > .1*ZMAX*0.75 The relevant atom site should be identified.
- PLAT230_ALERT_2_C Hirshfeld Test Diff for N2 -- C11 .. 5.9 su
- These should alert the co-editor even if there were no associated attempts to fiddle the residuals

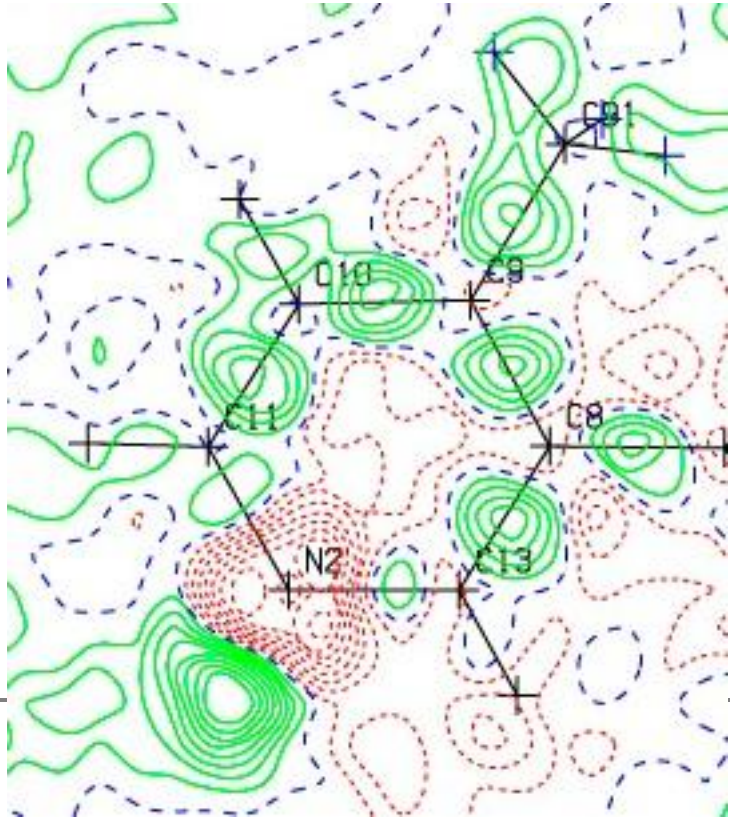
A difference Fourier map clearly shows why!



Even more so when 2 N atoms are added!







Changing the unit cell dimensions also has CIF consequences

- **Alert level G**
- REFLT03_ALERT_1_G ALERT: Expected hkl max differ from CIF values
- From the CIF: _diffrn_reflns_theta_max 34.13
- From the CIF: _reflns_number_total 3831
- From the CIF: _diffrn_reflns_limit_ max hkl 7. 22. 14.
- From the CIF: _diffrn_reflns_limit_ min hkl -7. -33. -14.
- TEST1: Expected hkl limits for theta max
- Calculated maximum hkl 7. 35. 15.
- Calculated minimum hkl -7. -35. -15.
- These alerts disappear from each of the clones if the unaltered unit cell dimensions are used.

Duplication or similarity checks

- A “Check for similar reduced cells” was introduced into the E and C submission system recently.
- These will further assist to alert us to potentially problematic structures.

Thanks to

- My fellow Section Editors, Bill Harrison and Matthias Weil
- Ton Spek
- George Ferguson
- Peter Strickland & Team Chester