

The increasing diversity of small molecule data: can one size fit all?

Simon Coles

<u>www.ncs.ac.uk</u> info@ncs.ac.uk



Small molecule crystallography 'Quality Framework'

- The basis
 - >100 years of instrumentation development
 - >50 years of building highly valuable and curated results databases
 - >40 years of trusted common refinement processes
 - >30 years of agreed and maintained standards
 - >20 years of validation tools
- Well understood statistics for the suitability of a dataset and the fit of a model
- Automation e.g. CheckCIF / PLATON and Mogul

In the service of the



Small molecule crystallography 'Quality Framework'

- Current validation and quality assessment processes for publication of crystal structures is largely based on:
 - (service) crystallography of the 1990's
 - the final derived result i.e. submitted CIF





More recent challenges

- More challenging samples smaller, not ideally/usually crystalline
- Need to quickly answer questions to further chemistry research and development
- New methods for structure determination
 - Powder, NMR Crystallography, Electron Diffraction, XFEL, Crystal Sponge
- Dynamic and in-situ crystallography
 - High pressure, porous materials, photo-excitation, electric stimulus
- Established quality framework pushed beyond its limits...





Electron diffraction - about to go viral

• A dedicated 'electron diffractometer'











'Routinely' generating Chemical Crystallography results







ov_exp		a uto \Chris\Glasgool	ex2_exp_146	2_auto\ov_e	P21 xp_1462_aut	2121 to.res
C12H25O1	13		1%		TWINS	, , ,
a = 7.59(10)	a = 90°	Z = 4	Ē	R.	14 03	%
b = 12.25(13)	β = 90°	Z' = 1		111	14.00	
c = 18.0(3)	V = 90°	V = 1671(3	37) 11.0	wR ₂	36.18	%
d min (0.0251) 2⊙=1.8°	0.80 ^{I/σ(I)}	3.2	Rint 30	.01% ^{Ful}	l 1.7°	79.1
^{Shift} -1.01	16 Max Peak	0.1 Min Peak	-0.2 GOOF	1.034	Hooft 1	(23)







Ca 300 ED structures in the CSD (1.25Mth an ED structure)



97091 reflections measured ($0.202^{\circ} \le 2\Theta \le 1.438^{\circ}$), 9201 unique ($R_{int} = 0.3529$, $R_{sigma} = 0.1495$)

The final R_1 was 0.2607 ($I > 2\sigma(I)$) and wR_2 was 0.6034 (all data).

Pearce, N., Reynolds, K.E.A., Kayal, S. *et al.* Selective photoinduced charge separation in perylenediimide-pillar[5]arene rotaxanes. *Nat Commun* **13**, 415 (2022). https://doi.org/10.1038/s41467-022-28022-3





exp_2014_auto

CoH11NO3

A typical approach to structure determination

First try



nth try (tries) exp_2014 exp_2016 exp_2017 exp_2019 exp_2020 exp_2024 exp_2024



exp_2025	ED							P21	2,2,1
C ₃₆ H ₄₄ N ₄ O ₁₂	20			%	/	<u> </u>	TWINS		*OK
a = 5.64(10)	a =	90°	Z = 4			R	40	40	%
b = 6.83(18)	β =	90°	Z' = 1			1.1	10	.45	
c = 20.34(10)	γ =	90°	V = 783	3(26)	8.03	WR_2		26.90	%
d min (0.0251) 2⊖=1.8°	0.80	l/σ(l)	5.7	Rint	16.	66% ^F	ull 1.7°	1	73.3
Shift 0	.081	Max Peak	0.1	Min Peak		-0.1	iooF	1.	117

 Future reprocessing of raw ED data – multiple (dynamic) scattering, radiation damage to improve current models



Crystal Sponge

- A new and accessible approach to chemical / structural characterisation
- Determining *molecular* structures by encapsulating in porous materials
- Miniscule amounts of material (far less than required to grow a single X-ray size crystal)
- Determine structure of oils, gases, etc to 1000ths of an angstrom precision

Well-resolved result



phenolphthalein



2021ncs7137c_jbo1a C				
C _{53.1} H _{39.4} C	I ₆ N ₁₂ O ₃ Zn ₃		🎽 😑 🔟 🐝	Ron 💢
a = 34.2990(4) b = 14.4926(1) c = 30.8022(4)	$\alpha = 90^{\circ}$ $\beta = 102.795(1)^{\circ}$ $\alpha = 90^{\circ}$	Z = 8 Z' = 1 V = 14931.0(3)		.33 %
d min (Cu∖a) 0 2Θ=136.5° 0 Shift 0 (.83 ^{Ι/σ(Ι)}	55.9 Rint m=5.38	2.46% CAP 136.	^{3°} 99.8 1 060



11

Less-well resolved result

#

1



2021ncs	7156c_jbo	91a		C2/c
C ₃₈ H _{27.4} Cl ₆ N	₁₂ 01Zn3 🌍	/ 🦉	🎽 😑 🕻	🗍 🐭 Pon 💢
a = 33.1231(3)	α = 90°	Z = 8		6 68 %
b = 14.4869(2)	β = 99.827(1)°	Z' = 1		0.00 📾
c = 30.4948(4)	V = 90°	V = 14418.3(3)	21.9 WR ₂	22.61 %
d min (Cu\a) 2⊖=136.5° 0).83 ^{Ι/σ(Ι)}	36.1 Rint m=5.32	3.53%	CAP 136.3° 99.8
Shift 0.0	001 Max Peak	1.2 Min Peak	-0.5	Goof 1.060







Further quality insight from raw data?





'Good' structure







What reliable information can be taken from this?

Analyte	Sponge Type	Sample	Exchange Site 1 Dihedral Angle / °	Exchange Site 2 Dihedral Angle / °
		А	-61.1 (17)	-67.2 (31)
		В	-59.0 (18)	-67.6 (18)
	Znl ₂ cHEX	С	58.9 (19)	-64.5 (18)
		D	-61.0 (12)	-63.5 (14)
		E	62.4 (14)	62.7 (16)
		F	61.7 (13)	65.0 (15)
		А	60.4 (16)	-66.0 (19)
		В	61.3 (15)	63.0 (18)



Data should be fit for purpose...

- Original purpose of structure determination
 - What have I made?
 - What is the reaction by-product?
 - How has my structure changed?
 - How does this material manifest these properties?
 - What are the driving forces behind structure formation, how do I control them?
- How others will reuse the results
 - Do chemicals like this exist (connectivity)
 - Starting point for follow on calculations (conformation)
 - Highly accurate structural features (precise structure)



- Crystal Sponge structures give a comparable accuracy to gold standard crystal structures
- Kept, e.g. in CSD, alongside gold standard crystal structures
- Are these structures used for follow-on research in the same way as / with gold standard crystal structures? What discernment going on?
- Need to extend the current quality framework
- A structure grading system
 - For validation/publication, particularly to enable reuse non-experts & data science
- Quantitative analysis of restraints / constraints applied
- Include properties of the primary / raw data?



Weighted sum averaged – 'good' example



SADI and DANG restraints are colour-coded to indicate the relevant relationships.

• Well defined pore (2 guests, 11 solvents, low level of unmodelled electron density

2022N E:\Simon IU0	2022NCS7050a_RC1 E:\Simon IUCr Data Quality\2022NCS7050a_RC1\2022NCS7050a_RC1.res						P_{2}^{2}	2/n	
C _{50.7} H ₄₆	716N12	0 _{1.4} Zn ₃		/	% {	<u> </u>	TWINS	PTON	, , ,
a = 31.5077(5 b = 14.9991(7 c = 34.4289(6	5) α = 1) β = 5) V =	90° 102.086(1)° 90°	Z = 8 Z' = 2 V = 159	10.0(4)	15.0	R₁ wR₂	7. 2	<mark>08</mark> 2.08	% %
d min (Cu∖a) 2⊖=136.5° Shift	0.83 0.004	l/σ(l) Max Peak	27.8 1.0	Rint m=3.73 Min Peak	4.	08% ⁰ -0.9	CAP 136.3° GooF	9 1.(9.6 059

• How to grade a structure which has two guests of different quality?





How to include a factor from the raw data?



0kl

h0l

hk0



Weighted sum averaged – poorer example



 Znl₂ poorly defined with high residual density

2023NCS E:\Simon IUCr Da	57015a_R ata Quality\2023NC	C1 S7015a_RC1\2023	NCS7015a_RC	C2/C
C ₇₈ H ₈₀ N ₁₂ C	D ₆ Zn₃l ₆ ⊖	/	🎽 😑 (🗍 🐖 Pon 💢
a = 77.257(6)	α = 90°	Z = 24		13 03 04
b = 15.0021(8)	β = 100.342(5)°	Z' = 3		13.05 %
c = 41.5124(18)	γ = 90°	V = 47332(5)	18.7 WR2	43.63 %
d min (Cu∖a) 2⊖=136.5° 0	.83 ^{Ι/σ(Ι)}	13.9 Rint m=3.72	8.19%	Full 135.4° 99% to 136.5° 99.6
Shift -0.0	006 Max Peak	1.4 Min Peak	-1.3	Goof 1.337

- Multiple exchanges
- Provides more insight into different conformations
- However, all require quite considerable restraints





Quantitatively poorer raw data...











hk0





Minimising number of restraints / constraints



Prior

2022NC E:\Simon IUCr D	S7050a_R	C1 57050a_RC1\2022N	ICS7050a_RC	P2/n
C _{50.7} H _{46.7} I	₆ N ₁₂ O _{1.4} Zn ₃	/ 🦉	🎽 😑 🕻	🗍 🐨 Pon 💢
a = 31.5077(5)	α = 90°	Z = 8		7 09 %
b = 14.9991(1)	β = 102.086(1)°	Z' = 2		1.00
c = 34.4289(6)	Y = 90°	V = 15910.0(4)	15.0 WR ₂	22.08 %
d min (Cu∖a) 2⊖=136.5° (0.83 ^{I/σ(I)}	27.8 Rint m=3.73	4.08%	CAP 136.3° 99.6
Shift 0.	.004 Max Peak	1.0 Min Peak	-0.9	Goof 1.059

Minimal

2022NCS C:\Users\Aaron\	67050a_R	C1		P2/n
C _{50.7} H _{46.7} I ₆	N ₁₂ O _{1.4} Zn ₃	2 🗾	🎽 😑 📋	TWINS PRON 💢
a = 31.5077(5)	α = 90°	Z = 8	Ē P.	7 09 %
b = 14.9991(1)	β = 102.086(1)°	Z' = 2		1.00
c = 34.4289(6)	V = 90°	V = 15910.0(4)	15.0 WR ₂	22.17 %
d min (Cu∖a) 2⊖=136.5° 0	.83 ^{Ι/σ(Ι)}	27.8 Rint m=3.73	4.08% ^{Ful}	^{l 135.4°} 99.7
Shift -0.0	001 Max Peak	1.0 Min Peak	-0.9 ^{Go}	^{oF} 1.060

• Statistically significant changes in the model not reflected in conventional metrics.



Minimising number of restraints / constraints



2023NCS E:\Simon IUCr Dat	7015a_R	C1 57015a_RC1\2023	NCS7015a_RC1.res	<i>C</i> 2/ <i>c</i>
C ₇₈ H ₈₀ N ₁₂ C	₀Zn₃l₀ <i>C</i>	<u> </u>	🎽 🚞 💓	Pron 💢
a = 77.257(6) b = 15.0021(8)	α = 90° β = 100.342(5)°	Z = 24 Z' = 3	R1 13	.03 %
c = 41.5124(18) $d \min (Cu a)$ $2\Theta = 136.5^{\circ}$	v = 90° 83 ^{Ι/σ(I)}	V = 47332(5) 13.9 Rint m=3.72	18.7 WR ₂ 8.19% Full 135.4 99% to 13	43.63 %
Shift -0.0	06 Max Peak	1.4 Min Peak	-1.3 GooF	1.337

- Sites B and C do not have dramatic visual shifts but geometric features undergo statistically significant change
- Lower occupancy tends to indicate greater reliance on restraints/constraints - the impact of removing them results in greater statistical shifts (~5/6o)



Conclusions and Further Work

- Increasing number of situations where conventional crystallographic metrics don't truly portray the quality of a model
- An extension to the established quality framework is necessary and viable particularly to enable *appropriate* reuse of results
- Refinement of grading system necessary particularly review of contributing factors
- How to combine contributing factors?
- How generally applicable is this approach?
- How to quantitatively include factors from raw data?
- Thanks to Rob Carroll and Aaron Horner for application of initial grading scheme