Data Science with R in crystallography

R-factor and gender bias

Rita Giordano,
Royal Society of Chemistry
giordanor@rsc.org

2019-08-14

1. Royal Society of Chemistry
2. R in crystallography
3. Analysis using tidyverse package: What do the R-factors in CIF files mean?
4. Analysis using tidyverse packages: Is there a gender bias in crystallography? (RSC data)

ROYAL SOCIETY OF CHEMISTRY

Royal Society of Chemistry

Royal Society of Chemistry (RSC)

To advance science in modern technology, chemical scientists use their expertise to improve our health, environment and daily lives. Collaboration is essential. We connect scientists with each other and society as a whole, so they can do their best work and make discoveries and innovation happen.

We publish new research. We develop, recognise and celebrate professional capabilities. We bring people together to spark new ideas and new partnerships. We support teachers to inspire future generations of scientists. And we speak up to influence the people making decisions that affect us all.

We are a catalyst for the chemistry that enriches our world.

ROYAL SOCIETY OF CHEMISTRY

Data Science at RSC

We help other teams to make evidence-based decisions.

We research ways in which the chemical community can easily find our articles and compound data.

ROYAL SOCIETY OF CHEMISTRY

R in crystallography

R Programming language

- R is a functional programming language for statistics. It is widely used in data science and statistics.
- It is a flexible toolkit and can runs on a wide array of platform.
- R give you unlimited possibility to analyse your data.
- The R packages are a collection of R functions, data and compiled codes.

R packages for crystallography

Package for crystallography
- CRONE
- cRy
- bio3d

Useful package for data science in crystallography
- tidyverse

CRONE

R package for 1D crystallography for undergraduate and graduate students to learn the theory of crystallography[1], developpend by James Foadi form the university of Bath.

[1] Emily Smith et al. 2017 Eur. J. Phys. 38 065501

ROYAL SOCIETY OF CHEMISTRY

cRy packages 2009

c is for **crystallography**

R is for 

y is for **yes**

cRy is a package developed to make statistical analysis for macromolecular crystallography. It was created in 2009 by James Foadi.

Aims

1. To create an interface between crystallographic objects and file formats, and the R objects.
2. To allow crystallographic operations to be performed on the R platform.
3. To carry out all major crystallographic calculations.

cRy packages 2019

New author team: James Foadi, David Waterman, Rita Giordano

cRy is in continuous development. Nowadays we want to create a complete package that crystallographers can use to analyse their data. The new features are:

1. cRy can now read the most used data format, not only .mtz, but also .HKL and .ahkl.
2. It includes data visualization tool, based on the package ggplot2, to plot all output from the crystallographic software (currently only SHELX).

ROYAL SOCIETY OF CHEMISTRY

Text analysis of log file

Read log file from protein crystallography software to extract numbers and table from text to an R data frame using regular expression.

```
library(cry)
shelxc_log <- readSHELXlog('shelxc.log')
```

ROYAL SOCIETY OF CHEMISTRY

Text analysis of log file

```
 ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
 +  SHELXC - Create input files for SHELXD and SHELXE - Version 2016/1  +
 +  Copyright (c) George M. Sheldrick 2003-16                           +
 +  fae_kappa                       Started at 10:47:15 on 13 Mar 2017  +
 ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

  623217 Reflections read from SAD file ep_XDS_ASCII.HKL

  138919 Unique reflections, highest resolution  1.390 Angstroms
   110.1 Friedel pairs used on average for local scaling

Resl.   Inf.  7.45  4.58  3.45  2.82  2.41  2.12  1.91  1.74  1.60  1.49  1.39
N(data)    1038  3235  5827  8580 11268 14077 15911 19188 22043 22436 15316
Chi-sq     0.90  0.55  0.47  0.51  0.55  0.54  0.54  0.55  0.53  0.53  0.57
<I/sig>    60.4  52.2  48.7  35.9  23.4  16.9  10.5   5.9   2.9   1.7   0.8
%Complete  85.7  87.5  93.3  95.4  95.5  96.3  94.5  95.5  93.8  89.4  50.7
Multipl.    4.5   4.5   4.4   4.6   4.4   4.6   4.4   4.6   4.5   4.5   4.4
R(pim)%    2.82  2.91  2.49  3.20  4.43  5.31  7.60 12.62 23.83 40.25 75.83
Ranom%    12.51 11.49  8.94 11.57 14.77 16.69 21.11 31.81 55.43 91.23 174.7
<d"/sig>   5.91  4.50  3.13  2.76  2.17  1.73  1.33  1.09  0.89  0.79  0.67
CC(1/2)    97.7  96.1  90.5  85.9  78.7  66.6  50.9  36.7  19.5   5.4  -3.8
```

Filter

	Res	N_data	Chi_sq	I_sig	Complete	d_sig	CC1_2
1	7.45	1038	0.90	60.4	85.7	5.91	97.7
2	4.58	3235	0.55	52.2	87.5	4.50	96.1
3	3.45	5827	0.47	48.7	93.3	3.13	90.5
4	2.82	8580	0.51	35.9	95.4	2.76	85.9
5	2.41	11268	0.55	23.4	95.5	2.17	78.7
6	2.12	14077	0.54	16.9	96.3	1.73	66.6
7	1.91	15911	0.54	10.5	94.5	1.33	50.9
8	1.74	19188	0.55	5.9	95.5	1.09	36.7
9	1.60	22043	0.53	2.9	93.8	0.89	19.5
10	1.49	22436	0.53	1.7	89.4	0.79	5.4
11	1.39	15316	0.57	0.8	50.7	0.67	-3.8

ROYAL SOCIETY OF CHEMISTRY

Data visualization: ggplot2

ggplot2 is grammar for graphics. The philosophy behind is:

"Instead of spending time making your graph look pretty, you can focu in greating a graph that best reveals the messages in your data".

It works adding layer to the initial plot Example of ggplot code:

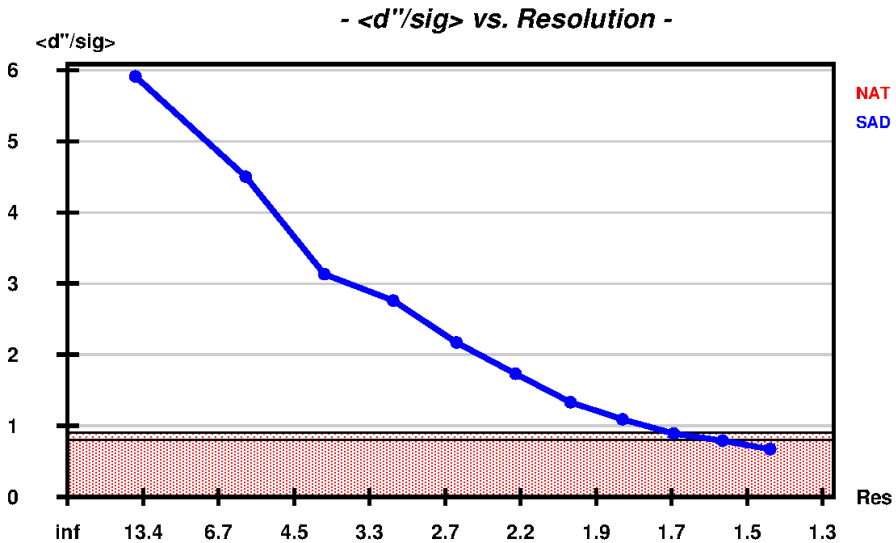
```
library(ggplot2)
ggplot(aes(x, y, color)) + geom_point() +
  geom_line() + theme_bw()
```

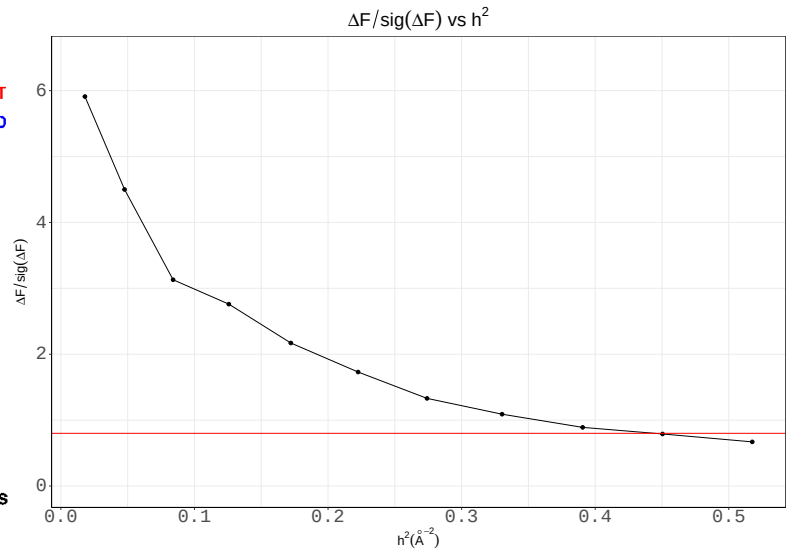
ROYAL SOCIETY OF CHEMISTRY

Data visualisation: cRy

```
plot(shelxc_log, shelxc$d_sig)
```

Visualization of results using ggplot2.





readCORRECT and alien reflectiosn

Search for alien reflection in the CORRECT.LP. The dataframe containing the aliens will be written on a file named REMOVE.HKL.

```
library(cry)
xds <- system.file("extdata/XDS/",package="cry")
correct_lp <- paste0(xds,'CORRECT.LP',sep = "")
correct <- readCORRECT(correct_lp, alien = TRUE)
```

REMOVE.HKL

REMOVE.HKL

14	15	39	2.33	14.01	604.9	18.38	"alien"
19	48	5	1.88	11.96	137.3	7.507	"alien"
11	61	9	1.69	11.58	52.55	5.706	"alien"
25	4	28	2.19	11.27	408	9.425	"alien"
5	58	22	1.74	11.25	70.87	6.47	"alien"
26	24	32	1.87	10.58	121.5	6.138	"alien"
2	57	8	1.88	10.56	121.2	8.811	"alien"
13	19	53	1.86	10.13	116.3	4.077	"alien"

ROYAL SOCIETY OF CHEMISTRY

Future perspectives

- Include more functionality for crystallography analysis.
- Read data from small molecule format.
- Improve the data visualisation also for other software log file.

ROYAL SOCIETY OF CHEMISTRY

Tidyverse

It is a set of packages, which works in harmony.

Packages included in tidyverse:

```
library(tidyverse)
tidyverse_packages(include_self = TRUE)
```

```
##  [1] "broom"       "cli"         "crayon"      "dplyr"
##  [6] "forcats"     "ggplot2"     "haven"       "hms"
## [11] "jsonlite"    "lubridate"   "magrittr"    "modelr"
## [16] "readr"       "readxl\n(>=" "reprex"      "rlang"
## [21] "rvest"       "stringr"     "tibble"      "tidyr"
## [26] "tidyverse"
```

ROYAL SOCIETY OF CHEMISTRY

Analysis using tidyverse package: What do the R-factors in CIF files mean?

The CIF Pandora's box

ROYAL SOCIETY OF CHEMISTRY

R-factors in CIF file

- _refine_ls_R_factor_gt
- _refine_ls_wR_factor_gt
- _refine_ls_R_factor_ref
- _refine_ls_wR_factor_ref (all reflections)
- ...

R-factor definition from IUCr CIF dictionary

_refine_ls_hydrogen_treatment (*char*)
Treatment of hydrogen atoms in the least-squares refinement.
The data value must be one of the following:

refall	refined all H-atom parameters
refxyz	refined H-atom coordinates only

_refine_ls_R_factor_all (*numb*)
Residual factor for all reflections satisfying the resolution limits established by **_refine_ls_d_res_high** and **_refine_ls_d_res_low**. This is the conventional R factor. See also **_refine_ls_wR_factor_** definitions.

$$R = \frac{\sum |F_{obs} - F_{calc}|}{\sum |F_{obs}|},$$

where F_{obs} = the observed structure-factor amplitudes, F_{calc} = the calculated structure-factor amplitudes and the sum is taken over the specified reflections.
The permitted range is $0.0 \rightarrow \infty$. **[refine]**

_refine_ls_R_factor_gt (*numb*)
Residual factor for the reflections (with number given by **_reflns_number_gt**) judged significantly intense (*i.e.* satisfying the threshold specified by **_reflns_threshold_expression**) and included in the refinement. The reflections also satisfy the resolution limits established by **_refine_ls_d_res_high** and **_refine_ls_d_res_low**. This is the conventional R factor. See also **_refine_ls_wR_factor_** definitions.

$$R = \frac{\sum |F_{obs} - F_{calc}|}{\sum |F_{obs}|},$$

where F_{obs} = the observed structure-factor amplitudes, F_{calc} = the calculated structure-factor amplitudes and the sum is taken over

_refine_ls_R_I_factor (*numb*)
Residual factor $R(I)$ for significantly intense reflections (satisfying **_reflns_threshold_expression**) and included in the refinement. This is most often calculated in Rietveld refinements against powder data, where it is referred to as R_B or R_{Bragg}.

$$R(I) = \frac{\sum |I_{obs} - I_{calc}|}{\sum |I_{obs}|},$$

where I_{obs} = the net observed intensities, I_{calc} = the net calculated intensities and the sum is taken over the specified reflections.
The permitted range is $0.0 \rightarrow \infty$. **[refine]**

_refine_ls_restrained_S_all (*numb*)
The least-squares goodness-of-fit parameter S' for all reflections after the final cycle of least-squares refinement. This parameter explicitly includes the restraints applied in the least-squares process. See also **_refine_ls_goodness_of_fit_** definitions.

$$S' = \left(\frac{\sum |w|Y_{obs} - Y_{calc}|^2| + \sum_r |w_r|P_{calc} - P_{targ}|^2|}{N_{ref} + N_{restr} - N_{param}} \right)^{1/2},$$

ROYAL SOCIETY OF CHEMISTRY

Methods

- More than 250 000 CIF files submitted to RSC journals were used.
- To search the different R factor I used regular expression.
- The data were analysed using tidyverse (More in the tutorial).

wRfactor_gt.csv

```
E:\Shares\LContent-SuppData\an\b1\b105136a\,crystallography_data.txt,18 KB,Text Document,06/08/2001 13:27:56,23/01/2013 10:40:30,23/01/2013 10:40:30,1,235,_refine_ls_wR_factor_gt
E:\Shares\LContent-SuppData\an\b2\b212649d\,200427.txt,25 KB,Text Document,26/02/2003 14:00:23,23/01/2013 10:40:33,23/01/2013 10:40:33,1,160,_refine_ls_wR_factor_gt 0.1138
E:\Shares\LContent-SuppData\an\b7\b706258c\,b706258c.txt,24 KB,Text Document,13/08/2007 11:42:29,23/01/2013 10:40:39,23/01/2013 10:40:39,1,150,_refine_ls_wR_factor_gt 0.1493
E:\Shares\LContent-SuppData\an\b7\b712231d\,b712231d.txt,28 KB,Text Document,30/08/2007 10:11:02,23/01/2013 10:40:38,23/01/2013 10:40:38,2,166,_refine_ls_wR_factor_gt 0.1001
E:\Shares\LContent-SuppData\an\b7\b712231d\,b712231d.txt,28 KB,Text Document,30/08/2007 10:11:02,23/01/2013 10:40:38,23/01/2013 10:40:38,2,556,_refine_ls_wR_factor_gt 0.0904
E:\Shares\LContent-SuppData\an\b8\b807581f\,b807581f.txt,27 KB,Text Document,28/10/2008 13:42:08,23/01/2013 10:40:45,23/01/2013 10:40:45,1,167,_refine_ls_wR_factor_gt 0.1648
E:\Shares\LContent-SuppData\an\b8\b823360h\,b823360h.txt,53 KB,Text Document,13/07/2009 14:25:42,23/01/2013 10:40:45,23/01/2013 10:40:45,2,162,_refine_ls_wR_factor_gt 0.1713
E:\Shares\LContent-SuppData\an\b8\b823360h\,b823360h.txt,53 KB,Text Document,13/07/2009 14:25:42,23/01/2013 10:40:45,23/01/2013 10:40:45,2,868,_refine_ls_wR_factor_gt 0.1769
E:\Shares\LContent-SuppData\an\c0\c0an00404a\,c0an00404a.txt,22 KB,Text Document,08/09/2010 14:14:58,23/01/2013 10:41:14,23/01/2013 10:41:14,1,162,_refine_ls_wR_factor_gt 0.0925
E:\Shares\LContent-SuppData\an\c0\c0an00804d\,c0an00804d.txt,24 KB,Text Document,05/01/2011 14:52:25,23/01/2013 10:41:17,23/01/2013 10:41:17,1,155,_refine_ls_wR_factor_gt 0.1705
E:\Shares\LContent-SuppData\an\c1\c1an15155j\,c1an15155j.txt,26 KB,Text Document,20/06/2011 15:06:39,23/01/2013 10:41:40,23/01/2013 10:41:40,1,155,_refine_ls_wR_factor_gt 0.1616
E:\Shares\LContent-SuppData\an\c1\c1an15609h\,c1an15609h.txt,16 KB,Text Document,23/09/2011 10:53:47,23/01/2013 10:41:28,23/01/2013 10:41:28,1,144,_refine_ls_wR_factor_gt 0.1202
E:\Shares\LContent-SuppData\an\c1\c1an15987a\,c1an15987a.txt,124 KB,Text Document,24/11/2011 14:40:48,23/01/2013 10:41:39,23/01/2013 10:41:39,1,254,_refine_ls_wR_factor_gt 0.1297
E:\Shares\LContent-SuppData\an\c2\c2an16197d\,c2an16197d.txt,21 KB,Text Document,20/03/2012 12:03:04,23/01/2013 10:42:22,23/01/2013 10:42:22,1,163,_refine_ls_wR_factor_gt 0.0804
E:\Shares\LContent-SuppData\an\c2\c2an35258c\,c2an35258c.txt,25 KB,Text Document,28/05/2012 12:47:46,23/01/2013 10:42:29,23/01/2013 10:42:29,1,188,_refine_ls_wR_factor_gt 0.1129
E:\Shares\LContent-SuppData\an\c2\c2an35524h\,c2an35524h.txt,13 KB,Text Document,18/06/2012 10:56:13,23/01/2013 10:42:26,23/01/2013 10:42:26,1,150,_refine_ls_wR_factor_gt 0.1476
E:\Shares\LContent-SuppData\an\c2\c2an35481k\,c2an35481k.txt,60 KB,Text Document,16/07/2012 17:02:16,23/01/2013 10:42:16,23/01/2013 10:42:16,2,159,_refine_ls_wR_factor_gt 0.0790
E:\Shares\LContent-SuppData\an\c2\c2an35481k\,c2an35481k.txt,60 KB,Text Document,16/07/2012 17:02:16,23/01/2013 10:42:16,23/01/2013 10:42:16,2,1021,_refine_ls_wR_factor_gt 0.1429
E:\Shares\LContent-SuppData\an\c2\c2an35560d\,c2an35560d_2.txt,12 KB,Text Document,08/06/2012 12:23:00,23/01/2013 10:42:17,23/01/2013 10:42:17,1,136,_refine_ls_wR_factor_gt 0.115
E:\Shares\LContent-SuppData\an\c2\c2an35560d\,c2an35560d.txt,12 KB,Text Document,08/06/2012 12:23:00,23/01/2013 10:42:17,23/01/2013 10:42:17,1,180,_refine_ls_wR_factor_gt 0.1154
E:\Shares\LContent-SuppData\an\c2\c2an35752f\,c2an35752f.txt,12 KB,Text Document,03/08/2012 09:36:20,23/01/2013 10:42:11,23/01/2013 10:42:11,1,148,_refine_ls_wR_factor_gt 0.1477
E:\Shares\LContent-SuppData\an\c2\c2an35860c\,c2an35860c.cif,15 KB,CIF File,22/10/2012 14:16:47,23/01/2013 10:42:23,23/01/2013 10:42:23,1,141,_refine_ls_wR_factor_gt 0.1310
E:\Shares\LContent-SuppData\an\c2\c2an35860c\,c2an35860c.txt,16 KB,Text Document,29/10/2012 09:03:58,23/01/2013 10:42:23,23/01/2013 10:42:23,1,170,_refine_ls_wR_factor_gt 0.1310
E:\Shares\LContent-SuppData\an\c2\c2an35901d\,c2an35901d.txt,35 KB,Text Document,30/08/2012 10:58:33,23/01/2013 10:42:24,23/01/2013 10:42:24,1,160,_refine_ls_wR_factor_gt 0.1870
E:\Shares\LContent-SuppData\an\c2\c2an35999e\,c2an35999e.txt,15 KB,Text Document,13/09/2012 14:33:44,23/01/2013 10:42:16,23/01/2013 10:42:16,1,152,_refine_ls_wR_factor_gt 0.1787
E:\Shares\LContent-SuppData\an\c2\c2an36076d\,c2an36076d.txt,30 KB,Text Document,19/10/2012 11:32:44,23/01/2013 10:42:30,23/01/2013 10:42:30,2,152,_refine_ls_wR_factor_gt 0.2076
E:\Shares\LContent-SuppData\an\c2\c2an36076d\,c2an36076d.txt,30 KB,Text Document,19/10/2012 11:32:44,23/01/2013 10:42:30,23/01/2013 10:42:30,2,582,_refine_ls_wR_factor_gt 0.1918
E:\Shares\LContent-SuppData\an\c2\c2an36588j\,c2an36588j.cif,30 KB,CIF File,03/01/2013 14:10:57,23/01/2013 10:42:30,23/01/2013 10:42:30,1,142,_refine_ls_wR_factor_gt 0.2123
E:\Shares\LContent-SuppData\an\c3\c3an00087g\,c3an00087g.txt,26 KB,Text Document,05/04/2013 14:57:16,05/04/2013 14:57:16,05/04/2013 14:57:16,1,163,_refine_ls_wR_factor_gt 0.2285
E:\Shares\LContent-SuppData\an\c3\c3an00279a\,c3an00279a.txt,33 KB,Text Document,22/03/2013 14:58:53,22/03/2013 14:58:53,22/03/2013 14:58:53,1,157,_refine_ls_wR_factor_gt 0.1182
E:\Shares\LContent-SuppData\an\c3\c3an01750h\,c3an01750h.txt,17 KB,Text Document,07/11/2013 15:01:46,07/11/2013 15:01:46,07/11/2013 15:01:46,1,170,_refine_ls_wR_factor_gt 0.0581
E:\Shares\LContent-SuppData\an\c3\c3an36388k\,c3an36388k.txt,42 KB,Text Document,23/01/2013 13:01:19,24/01/2013 08:31:18,24/01/2013 08:31:18,2,147,_refine_ls_wR_factor_gt 0.1393
E:\Shares\LContent-SuppData\an\c3\c3an36388k\,c3an36388k.txt,42 KB,Text Document,23/01/2013 13:01:19,24/01/2013 08:31:18,24/01/2013 08:31:18,2,777,_refine_ls_wR_factor_gt 0.1568
E:\Shares\LContent-SuppData\an\c3\c3an36887d\,c3an36887d.txt,27 KB,Text Document,28/03/2013 13:34:50,28/03/2013 13:34:50,28/03/2013 13:34:50,1,157,_refine_ls_wR_factor_gt 0.1742
```

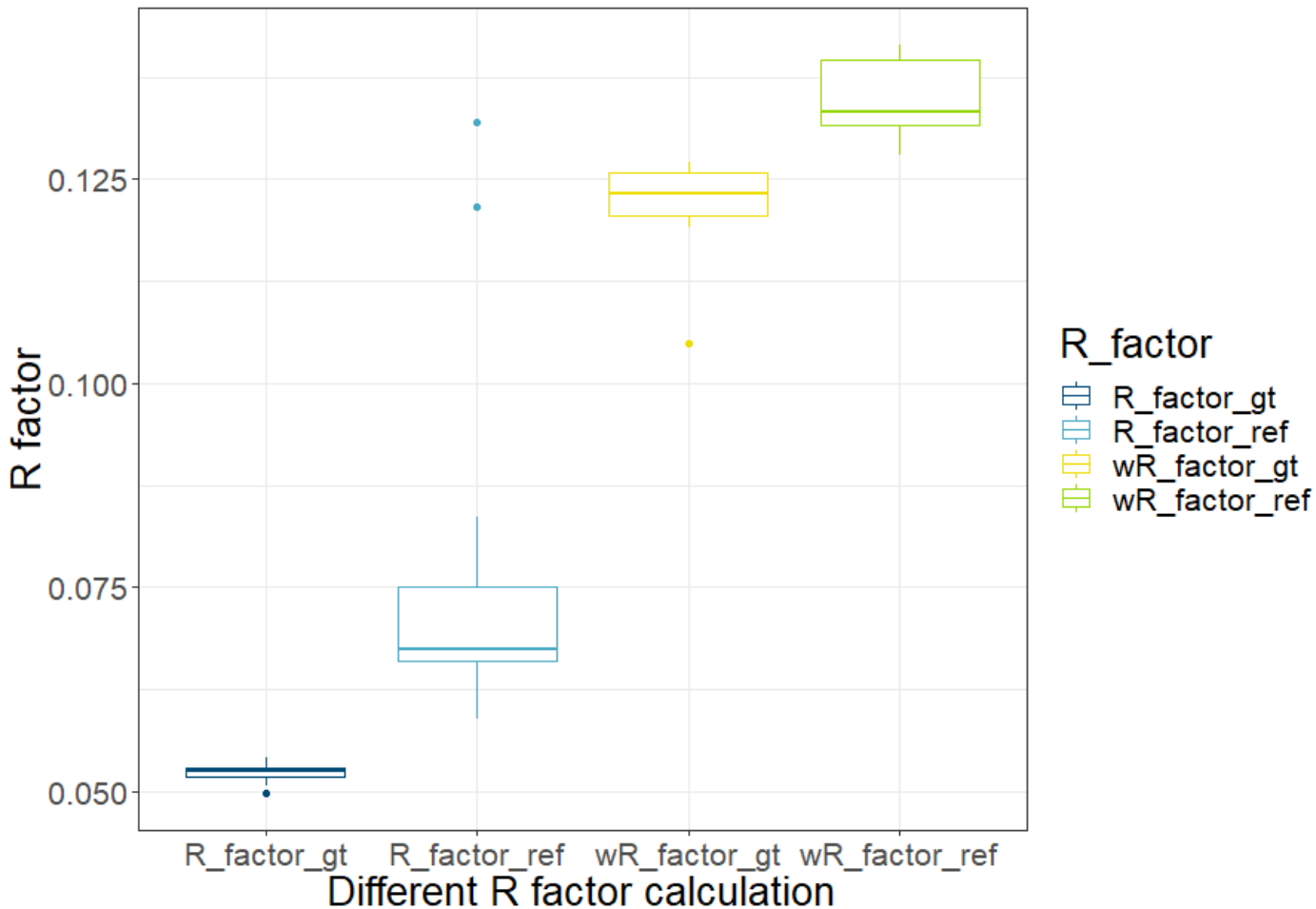
ROYAL SOCIETY OF CHEMISTRY

Tidy-analysis

Long regex expression to search for different R-factor values.

```
pattern1 = 'E:([\\])Shares([\\])LContent-SuppData([\\])\\w+
patter2 =  '_refine_ls_R_factor_gt|_refine_ls_r_factor_gt'
r_factor_search_gt <- cif_data_all_gt %>%
                      mutate(msid = gsub(pattern1,
                                         "",
                                         Folder),
                             R_factor = gsub(pattern2,
                                             "",
                                             R_Factor))
```

Analysis of R-factor and weighted R-factor



ROYAL SOCIETY OF CHEMISTRY

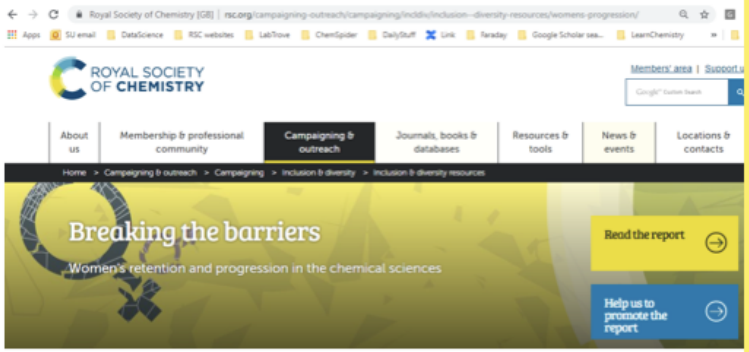
Open questions

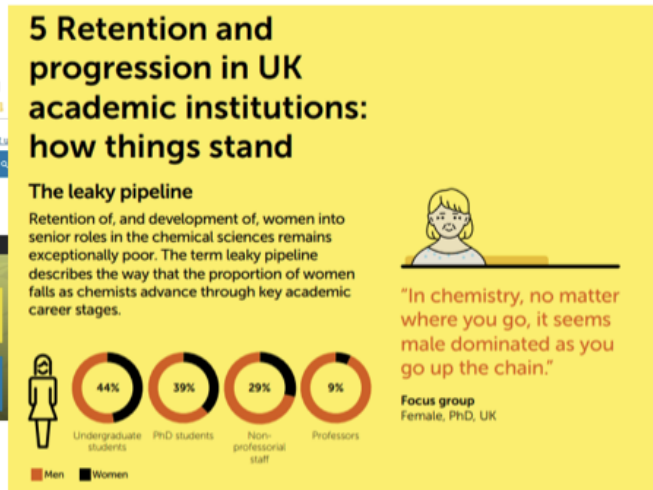
- What are the different R-factors for?
- Can we discriminate between them?

Analysis using tidyverse packages: Is there a gender bias in crystallography? (RSC data)

RSC gender bias reports (see RSC website)

- "Diversity landscape of the chemical sciences"
- "Breaking the barriers"





- Further analyses and reports in preparation...

ROYAL SOCIETY OF CHEMISTRY

Methods

- The data were analysed using tidyverse (More in the tutorial tomorrow).
- Gender of authors assigned based on first names[2].
- Single variable significance testing - Binomial significance testing of female proportions of subsets compared with baseline[3].

[2]https://github.com/OpenGenderTracking/globalnamedata - MIT PhD research by Matias, N

[3]https://en.wikipedia.org/wiki/Binomial_test

ROYAL SOCIETY OF CHEMISTRY

Picking crystallography articles

- Articles with CIF files in the supplementary information.
- Articles classified in the "crystallography" subcategory.

ROYAL SOCIETY OF CHEMISTRY

Tidyverse for selecting articles

```
articles <- mongo_collection_find(db, collection, url)%>%
            select("authors",
                   "esi_files",
                    'cats',
                    'subcats') %>%
             filter(esi_files %in% grep('Crystal',
                                        esi_files,
                                       value = TRUE) |
                    subcats %in% grep('Crystallography',
                                      subcats,
                                      value = TRUE))
```

ROYAL SOCIETY OF CHEMISTRY

Gender assignment methods

Gender inferred from first name by comparison of list with name/gender source list:

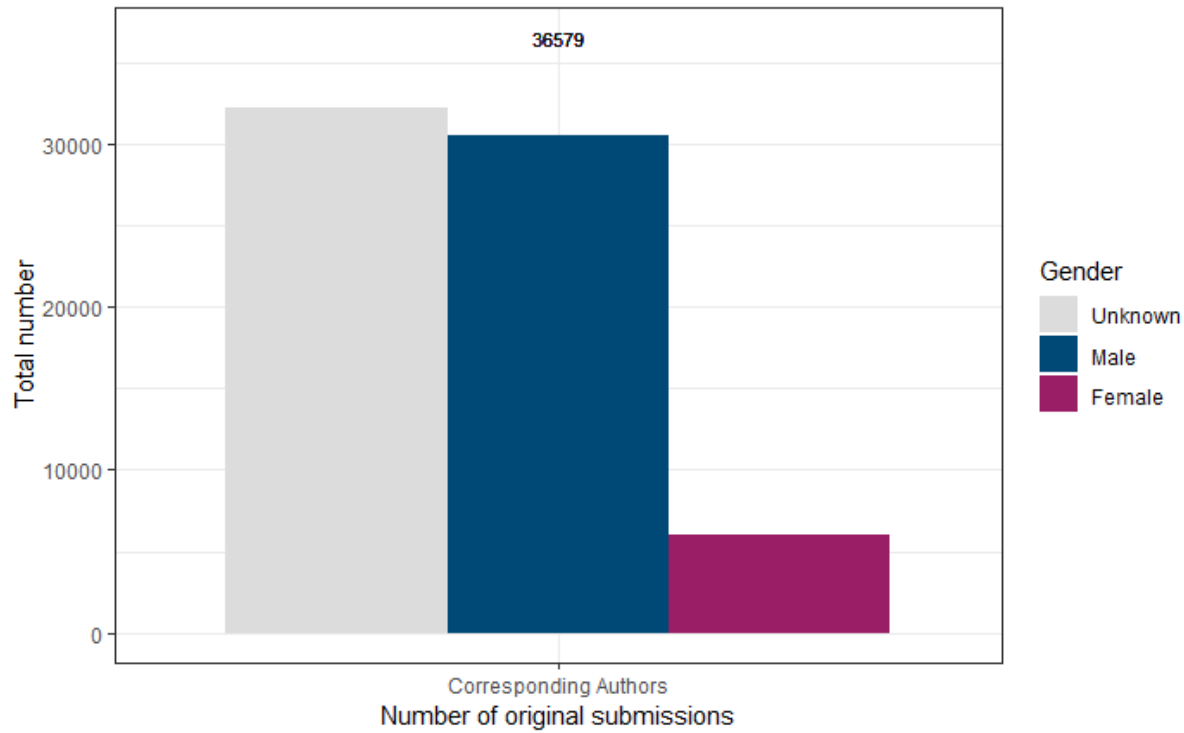
- name/gender source list and methodology originally devised by Matias[4] based on data from US Social Security Administration and the UK Office for National Statistics (ONS).
- Source list enhanced with data sources and refined.
- R program and source list used available on https://bitbucket.org/rscapplications/genderdiversity

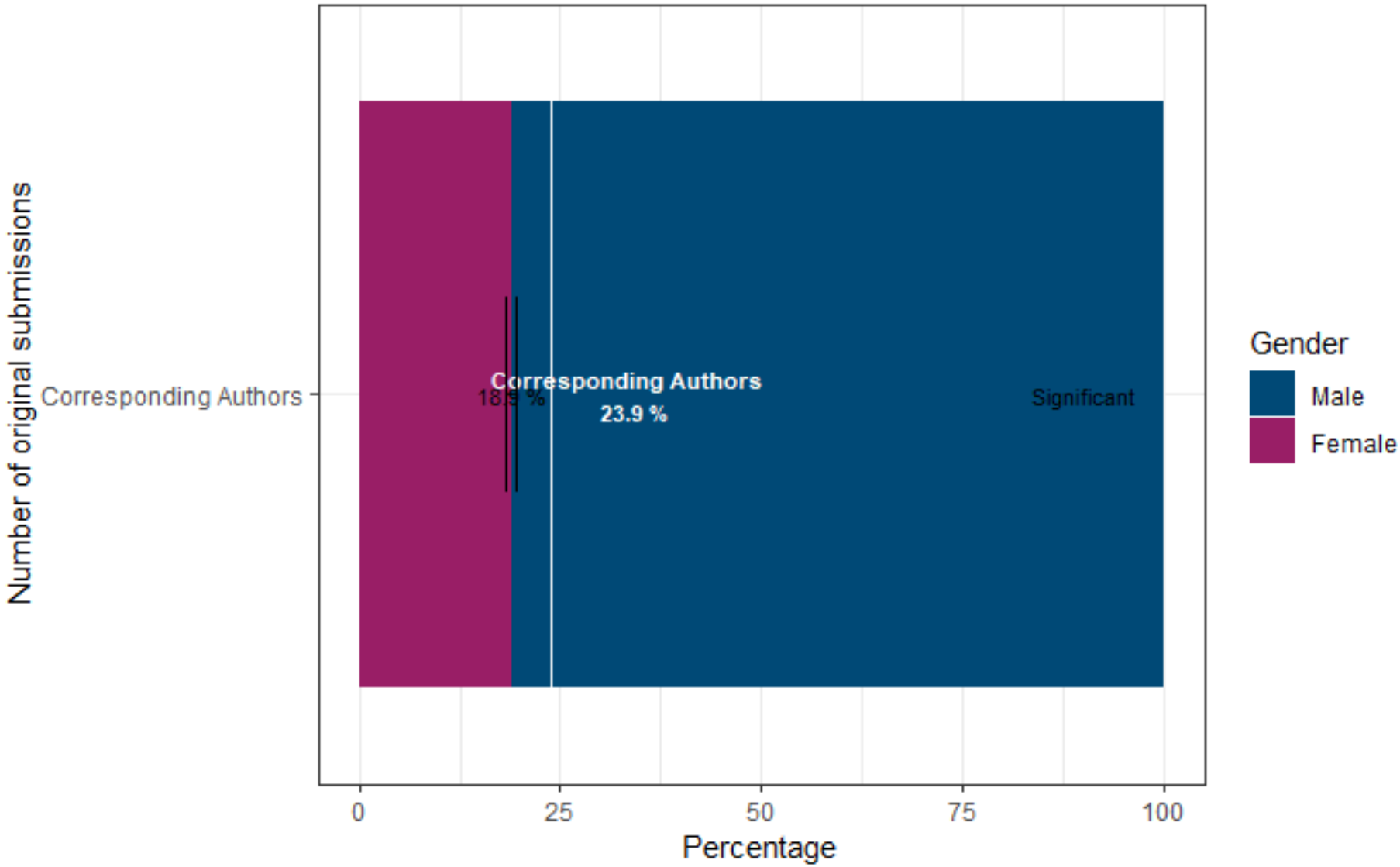
[4] https://github.com/OpenGenderTracking/globalnamedata - MIT PhD research by Matias, N

ROYAL SOCIETY OF CHEMISTRY

Gender analysis

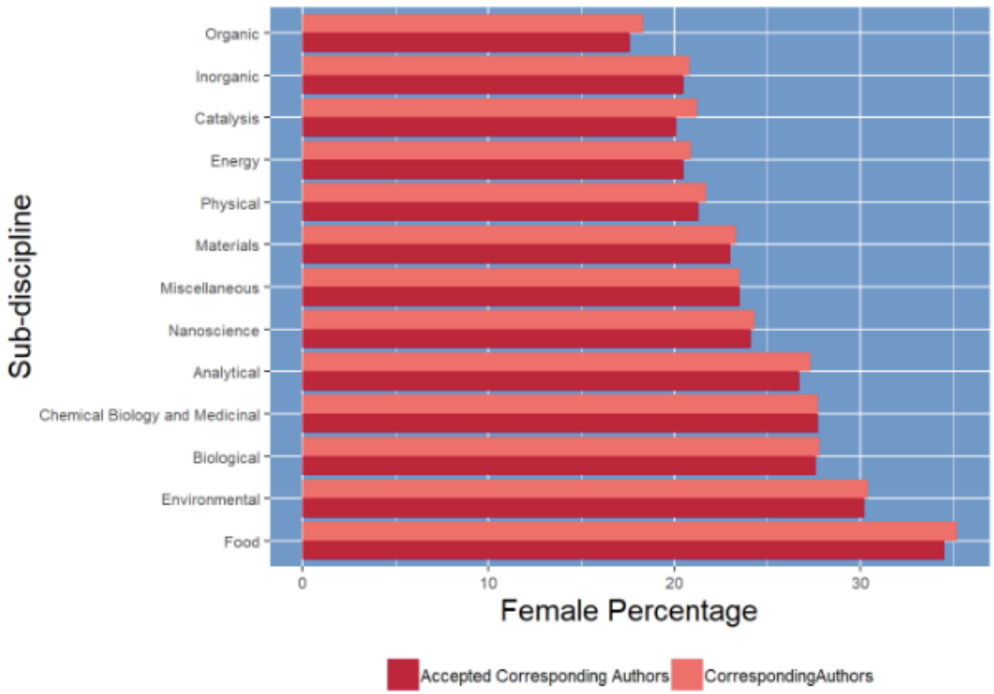
- Is there any gender bias in Crystallography?
- Do men publish more as corresponding author than women?




Number of original submissions
Corresponding Authors
Corresponding Authors
18.9 %
23.9 %
Significant
0
25
50
75
100
Percentage
Gender
Male
Female

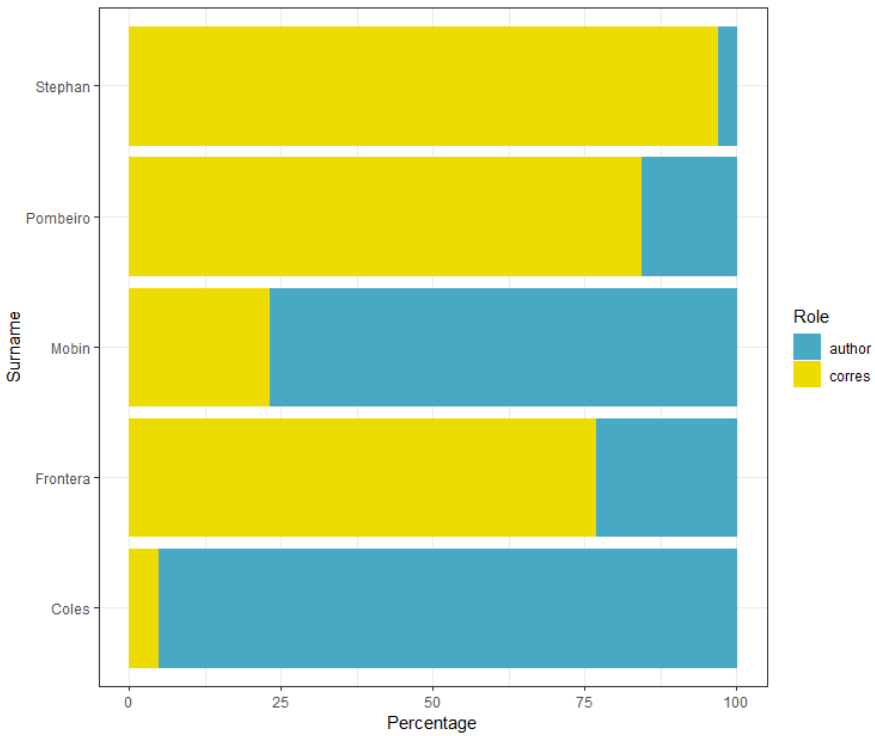
Female percentage of accepted submissions

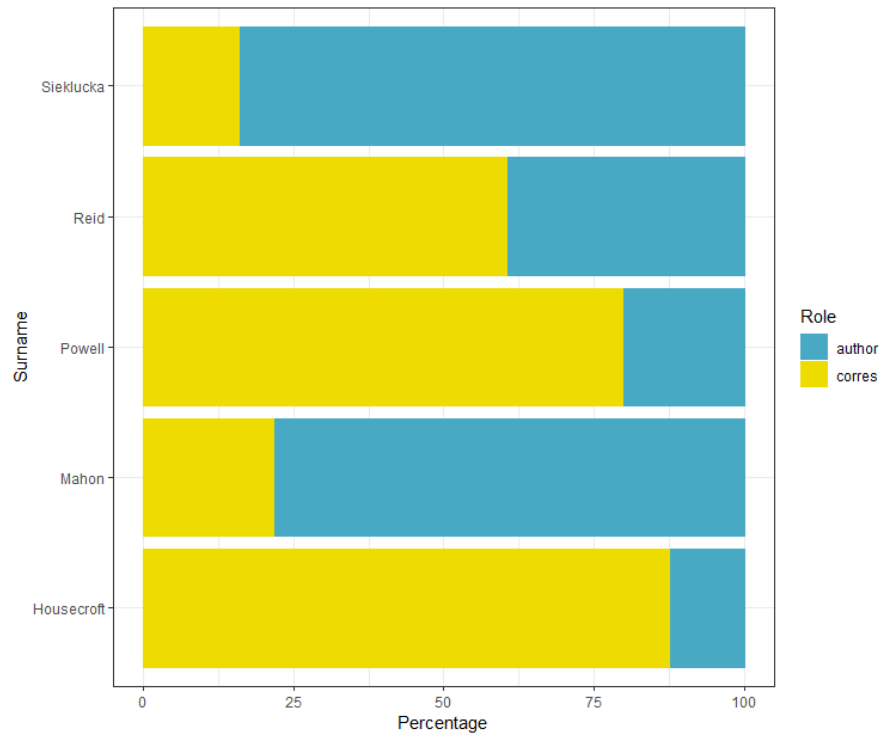
Chemistry sub-discipline. Data from the paper "Is there a gender bias in Chemical Sciences scholarly communication?" Aileen Day et al. in preparation



ROYAL SOCIETY OF CHEMISTRY

Male/Female corresponding author





ROYAL SOCIETY OF CHEMISTRY

Conclusion

- R is a powerful programming language with many package for tackling better solution for statistics and data science.
- We found wide variation in R-factor to be further investigated and discussed
- There is evidence of gender bias in crystallography that is similar to related field in chemistry.

Acknowledgement

- Colin Batchelor - Royal Society of Chemistry
- John Boyle - Royal Society of Chemistry
- Aileen Day - Royal Society of Chemistry
- James Foadi - University of Bath

ROYAL SOCIETY OF CHEMISTRY

Thank you for your attention

Basic proportions and significance testing

- Subset data set into different populations.
- Discount "Unknown" gender.
- Calculate basic proportions (Female/Male).
- Compare with baseline female proportion.
- Significance calculated by exact Binomial test[5] with confidence level of 95%.
- What background female baseline should we use for Chemistry? (23.9%)

[5] https://en.wikipedia.org/wiki/Binomial_test, R documentation binom.test function

ROYAL SOCIETY OF CHEMISTRY