
Article:
Willett, P. orcid.org/0000-0003-4591-7173, Cole, J.C. and Bruno, I.J. (2020) The impact of the Cambridge structural database and the small molecule crystal structures it contains: a bibliographic and literature study. CrystEngComm, 22 (43). pp. 7233-7241. ISSN 1466-8033

https://doi.org/10.1039/d0ce00045k

ARTICLE

The impact of the Cambridge Structural Database and the small molecule crystal structures it contains: A bibliographic and literature study

Peter Willet*[a] Jason C. Cole [b] and Ian J. Bruno [b]

A bibliographic and literature-based analysis of the impact of the Cambridge Structural Database (CSD) and the papers associated with crystal structures in the CSD has been undertaken. The analysis shows the broad impact of the CSD in the chemical sciences and also highlights how areas where the CSD has impact have changed over time. In addition, we note the changing nature of crystallography as a science, observing how crystal structures are now impactful. A brief illustration of some more unusual contributions to the CSD and publications using the CSD is also presented.

[a] *Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK*
[b] *CCDC, !2 Union Road, Cambridge, CB2 1EZ, UK*

Introduction

This special issue of *CrystEngComm* focuses on application examples of the Crystal Structure Database (CSD) in celebration of the significant milestone recently achieved by the Cambridge Crystallographic Data Centre (CCDC), namely the curation of its 1 millionth crystal structure. In this paper, we provide a contribution that is complementary in character, presenting a bibliometric and word frequency based, rather than a crystallographic, study of the CCDC and of the CSD. A study of published literature is inevitably retrospective in nature. It can, however, suggest at least some avenues for the development of the field. Most obviously, it may help to identify novel techniques or emerging centres of excellence that may be expected to increase in importance in the future, while unexpected citations (such as those exemplified below) can suggest new applications for crystallographic data.

Bibliometric studies involve analysing data such as the authors of, and the citations to, articles in the published academic literature. Analyses of this sort are used for an increasing range of applications, such as the quantification of the impact of the research conducted by an individual or an organization, the growth characteristics of the literature, and the identification of the key researchers and organizations in a discipline. While there have been many such studies conducted in the area of chemistry in general, there have been very few to date that focus on crystallography [1-6]. Two of these bibliometric studies have analysed publications relating to the CSD. Redman *et al*. reported a citation analysis of ten highly cited articles written by members of the CCDC [3], and Wong *et al*. subsequently reported a citation analysis of a total of 46 highly cited articles (these including articles written by both CCDC and non-CCDC authors) that describe the CSD or scientific research that makes use of it [5]. These papers considered the variation of citations with time, the journals in which citations occur, and the types of organization and the geographic regions that used the CSD *inter alia*, and demonstrated clearly the scientific importance of the crystal structure information that the CCDC has made available to the international research community.

This brief communication considers further the impact of the CCDC's work as reflected by publications and citations in the academic literature. The focus of the study falls into two areas. Firstly we consider the impact of what we shall refer to as the standard references, i.e., those that the CCDC suggests that authors should cite when making use of the CSD and of the core software components (rather than papers on its associated application software or on the CCDC as a research organization). Citations to these standard references are analysed with respect to the principal citing journals, countries, and subject areas, this last illustrating the CSD's impact beyond the normal crystallographic community. We also describe how these citations vary over time, and discuss a word-based, rather than citation-based, analysis of the titles of the citing articles. Secondly, we then consider the content within the CSD, which contains in excess of 1 million structures published in a variety of sources. By analysing the titles of these articles associated with these structures we can understand how small-molecule organic and organo-metallic crystallography as a field has impacted the scientific community over the years of the CSD's growth.

Experimental

Bibliographic results reported here were obtained using the bibliographic data available in the Web of Science Core Collection database produced by Clarivate Analytics [7]. The data were collected during the summer of 2019 and are based on publications and citations up to and including the end of 2018. The standard references that form the principal focus of the study are listed in Table 1, with each reference accompanied by the number of citations that it had attracted and its Digital Object Identifier (or DOI). Information on individual paper titles was extracted from CCDC's internal data deposition system which includes a record of associated paper titles where available. Titles are added for paper publications into CCDC's internal system by a combination of automated extraction from publisher feeds, manual editing and automated cross-validation and lookup using CrossRef [8]. 89.7% of the 1,009,141 crystal structures in the CSD have an associated publication title. These amount to 466,394 unique publications (as a significant proportion of papers contain greater than one structure). Information on journal submission frequency was derived directly from the CSD (December 2019 release) using a Python script built using the CCDC Python API. Natural language processing was carried out using the Python toolkit NLTK [9]. Tokenization was performed using NLTK's built-in word corpus tokenizer. For lemmatization, the built in WordNetLemmatizer was used. Standard English stop words (as available in NLTK) were ignored in the analysis. Bigram frequency analysis (a bigram being a pair of adjacent words in a text corpus) was performed using the built-in methods in NLTK. The Python script that was used to undertake this analysis is included in the supplementary material.

Results and discussion

Citation analysis of Cambridge Structural Database reference works

The twelve publications listed in Table 1 were published over a period of 37 years, during which time there have been substantial changes in the growth and the composition of the scientific literature. Thus, in 1979, when the first of the papers in Table 1 was published, a total of 728,927 items was added to the *Web of Science Core Collection*, of which 2,554 were allocated the Crystallography subject category, whereas the corresponding figures in 2016 (when the last of the papers in Table 1 was published) were 3,052,205 and 6,994. A still more striking difference is seen if we consider publications from the People's Republic of China. In 1979, it made just 472 such contributions to the Core Collection, with none of these allocated to the Crystallography category; in 2016, conversely, 447,615 Chinese publications were added to the database and 1,760 of these made up the largest single national contribution to the Crystallography category. Accordingly, rather than treating the twelve articles as a single, homogeneous whole they have been analysed here in three groups: the first three articles from Table 1 that were published in the period 1979-1997, then the four published in 2002 and 2004 (where the three 2002 articles all appeared in a special issue of *Acta Crystallographica Section B* that was devoted to crystallographic databases), and finally the remaining five articles published from 2006 to 2016.

F. H. Allen, S. Bellard, M. D. Brice, B. A. Catrwright, A. Doubleday, H. Higgs, T. Hummelink, B.-G. Hummelink-Peters, O. Kennard, W. D. S. Motherwell, J. R. Rodgers and D. G. Watson, Cambridge Crystallographic Data Center - Computer-based search, retrieval, analysis and display of information, *Acta Crystallographica Section B - Structural Science*, 1979, **B**35(10), 2331-2339. 1558 citations. DOI: 10.1107/S0567740879009249

F. H. Allen, J. E. Davies,J. J. Galloy, O. Johnson, O. Kennard, C. F. Macrae, E. M. Mitchell, G. F. Mitchell, J. M. Smith and D. G. Watson, The development of Version-3 and Version-4 of the Cambridge Structural Database system, *Journal of Chemical Information and Computer Sciences*, 1991, **31**(2), 187-204. 1545 citations. DOI: 10.1021/ci00002a004

I. J. Bruno, J. C. Cole, J. P. M. Lommerse, R. S. Rowland, R. Taylor and M. L. Verdonk, IsoStar: A library of information about nonbonded interactions, *Journal of Computer-Aided Molecular Design*, 1997, 11(6), 525-537. 227 citations. DOI: 10.1023/A:1007934413448

F. H. Allen, W. D.S. Motherwell, Applications of the Cambridge Structural Database in organic chemistry and crystal chemistry, *Acta Crystallographica Section B - Structural Science*, 2002, B**58**(3), 407-422. 495 citations. DOI: 10.1107/S0108768102004895

F. H. Allen, The Cambridge Structural Database: a quarter of a million crystal structures and rising, *Acta Crystallographica Section B - Structural Science*, 2002, B**58**(3), 380-388. 10348 citations. DOI: 10.1107/S0108768102003890

I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson and R. Taylor, New software for searching the Cambridge Structural Database and visualizing crystal structures, *Acta Crystallographica Section B - Structural Science*, 2002, B**58**(3), 389-397. 2510 citations. DOI: 10.1107/S0108768102003324

I.J. Bruno, J. C. Cole, M. Kessler, J. Luo, W. D. S. Motherwell, L. H. Purkis, B. R. Smith, R. Taylor, R. I. Cooper, S. E. Harris and A. G. Orpen, Retrieval of crystallographically-derived molecular geometry information, *Journal of Chemical Information and Computer Sciences*, 2004, **44**(6), 2133-2144. 494 citations. DOI: 10.1021/ci049780b

C. F. Macrae, P. R. Edgington, P. McCabe, E. Pidcock, G. P. Shields, R. Taylor, M. Towler and J. van De Streek, Jacco, Mercury: visualization and analysis of crystal structures, *Journal of Applied Crystallography*, 2006, **39**(3), 453-457. 3930 citations DOI: 10.1107/S002188980600731X

C. F. Macrae, I. J. Bruno, J. A. Chisholm, P. R. Edgington, P. McCabe, E. Pidcock, L. Rodriguez-Monge, R. Taylor, J. van de Streek and P. A. Wood, Mercury CSD 2.0 - new features for the visualization and investigation of crystal structures, *Journal of Applied Crystallography*, 2008, **41**(2), 466-470. 4331 citations. DOI: 10.1107/S0021889807067908.

I. R. Thomas, I.J. Bruno, J. C. Cole, C. F. Macrae, E. Pidcock and P. A. Wood, WebCSD: the online portal to the Cambridge Structural Database, *Journal of Applied Crystallography*, 2010, **43**(2), 362-366. 79 citations. DOI: 10.1107/S0021889810000452

C. R. Groom and F.H. Allen, The Cambridge Structural Database in retrospect and prospect, *Angewandte Chemie - International Edition*, 2014, **53**(3), 662-671. 723 citations. DOI: 10.1002/anie.201306438

C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, The Cambridge Structural Database, *Acta Crystallographica Section B - Structural Science, Crystal Engineering and Materials*, 2016, B**72**(2), 171-179. 1871 citations. DOI: 10.1107/S2052520616003954

Table 1. CCDC publications forming the basis for the analyses in this paper, together with the total numbers of pre-2019 citations and the Digital Object Identifier.

The three 1979-1997 standard references had attracted a total of 3,330 citations up to the end of 2018. The great majority of these citations (89.8% of them) had come from journal articles, with a still greater majority (97.5%) in English. The ten journals that contributed the greatest numbers of citations are listed in the first two columns of Table 2: at the other extreme, no less than 236 of the 474 citing publications yielded just a single citation. Such highly skewed distributions are characteristic of much bibliographic data, and in like vein the first two columns of Table 3 list the ten nations that contributed the greatest numbers of citations (with 20 of the total of 78 nations providing just a single citation).

1979-1997 references		2002-2004 references		2006-2014 references	
Journal	Citations	Journal	Citations	Journal	Citations
Acta Crystallographica Section C Crystal Structure Communications	207	Acta Crystallographica Section E Structure Reports Online	909	Acta Crystallographica Section E Crystallographic Communications	1498
Journal of the American Chemical Society	138	Acta Crystallographica Section E Crystallographic Communications	705	Acta Crystallographica Section E Structure Reports Online	817
Acta Crystallographica Section B Structural Science	119	Acta Crystallographica Section C Crystal Structure Communications	665	CrystEngComm	502
Journal of Molecular Structure	90	CrystEngComm	622	Crystal Growth & Design	485
Inorganic Chemistry	88	Crystal Growth Design	558	Journal of Molecular Structure	481
Journal of Medicinal Chemistry	82	Acta Crystallographica Section C Structural Chemistry	466	Acta Crystallographica Section C Structural Chemistry	465
Organometallics	72	Polyhedron	448	Polyhedron	375
Journal of Organometallic Chemistry	69	Dalton Transactions	435	Dalton Transactions	337
Journal of the Chemical Society Dalton Transactions	66	Inorganic Chemistry	421	Acta Crystallographica Section C Crystal Structure Communications	281
Inorganica Chimica Acta	60	Journal of Molecular Structure	393	Inorganic Chemistry	241

Table 2. The journals that cite CCDC standard references most frequently

1979-1997 references		2002-2004 references		2006-2014 references	
Nation	Citations	Nation	Citations	Nation	Citations
USA	647	USA	2073	USA	1322
UK	625	UK	1889	India	1234
Italy	299	Poland	1112	UK	1046
Germany	288	Germany	1077	Germany	802
Spain	281	Spain	881	Poland	741
Poland	191	India	853	Russia	478
Switzerland	126	Russia	787	Spain	447
France	120	People's Republic of China	684	Italy	443
India	116	France	580	People's Republic of China	441
Australia	95	Italy	566	Brazil	430

Table 3. The author nationalities that cite CCDC standard references most frequently

Analogous data for citations to the four 2002-2004 and five 2006-2016 standard references are shown in the second and third parts respectively of Tables 2 and 3, these corresponding to totals of 13,847 and 10,934 citations to the CCDC publications. It will come as no surprise that the various sections of *Acta Crystallographica* dominate all three journal rankings, and that *CrystEngComm* and *Crystal Growth & Design* (which commenced publication in 1999 and 2001 respectively) have rapidly established themselves as primary sources of citations. However, there is a lesser degree of uniformity in the nationality data shown in Table 3, with authors from India, the People's Republic of China and Brazil starting to make significant contributions to the CSD-related literature.

All of the journals listed in Table 2 belong to the core crystallographic literature but citations to the CSD standard references come from a wide range of disciplines. Considering just the most recent 2006-2016 articles, Table 4 lists the ten Web of Science subject categories providing the largest numbers of citations and these all represent mainstream areas of chemistry. Citations are also received from no less than 67 other categories, and while many of these disciplines are - as would be expected - from the chemical and life sciences, some of the citations appear on first sight to be from disciplines that are not obviously related to the activities of the CCDC. However, inspection of these serves to demonstrate the knowledge flows [10] that are taking place. For example, Simu *et al*. draw on a CSD structure in the development of a bio-mimetic material for use in dentistry or orthopaedics (a paper in the subject category Anatomy and Morphology) [11], Marchese and Marchese use images generated using the Mercury system in the evaluation of a new method for computer graphics rendering based on medieval and renaissance theories of colour (Computer Science Theory & Methods) [12], Zhang *et al*. include the CSD in a bibliometric study of the open data movement (Information Science & Library Science) [13], and Kale *et al*. report the use of Mercury in a study of the dyeing and stiffness characteristics of cellulose-coated cotton fabrics (Materials Science, Paper & Wood) [14].

The ten categories listed in Table 4 were also ranked high when the categories of the citing articles for the 1979-1997 standard references were considered, with all ten of the Table 4 categories occurring in the top fifteen ranked positions. However, immediately below these top-ranked categories there are clearly subject areas that are making more, or making less, use of the work of the CCDC (if we accept the standard assumption that citation of an article corresponds in some way to its use by the author(s) of the citing article). Examples of categories that are now citing the standard references more frequently include Nanoscience (ranked 33rd for citing the 1979-97 references but twelfth for the 2006-2016 references), Material Science Textiles (43rd to 21st), Optics (65th to 26th) and Plant Science (67th to 30th). Conversely, Cell Biology has gone from 19th in the 1979-1997 rankings to 34th in the 2006-2016 rankings, Information Science Library Science from 22nd to 58th, and Quantum Science Technology from 27th to 61st.

Subject category	Citations
Crystallography	5256
Chemistry Multidisciplinary	3275
Chemistry Inorganic Nuclear	1931
Chemistry Physical	1148
Materials Science Multidisciplinary	864
Chemistry Organic	568
Biochemistry Molecular Biology	307
Physics Atomic Molecular Chemical	203
Chemistry Medicinal	185
Spectroscopy	181

Table 4. The Web of Science subject categories that cite the 2006-2016 CCDC standard references most frequently

Of the twelve articles in Table 1, four of them (by Allen *et al*. in 1979, by Allen *et al*. in 1991, by Allen in 2002, and by Groom and Allen in 2014) not only provide standard references at the time of their publication but also illustrate the historical development of the CSD over a period of 35 years. In the great majority of scientific articles, citation counts rise from the date of publication as the community becomes aware of an article but then start to drop away as the information contained within the article is increasingly overtaken by more recent research. Figure 1 plots the citation counts for the four references above, and it will be seen that the 2002 and 2014 articles show the expected behaviour of a rapid rise, followed by a falling away (behaviour that is often referred to as obsolescence). The normal post-maximum decrease in counts for the 2002 article are probably exacerbated by the fact that the CCDC highlighted the 2014 article as the new standard reference that should be cited once it had been published. There is, however, an unusual marked dip in the approximate period 2007-2010, resulting in a bimodal distribution for this article. This may be a result of what Serenko and Dumay have called the Google Scholar Effect [15]. They found such bimodal citation-count plots in a study of 100 important articles in knowledge management, and noted that the second peak in these plots occurred in the period 2008-2012 (i.e., at around the same time as the second maximum in the plot for Allen's 2002 article). This was the time when Google Scholar was starting to gain prominence in the academic community, and Serenko and Dumay provide evidence to suggest that Google Scholar's ranking of search outputs in order of decreasing numbers of citations means that authors may decide to read, and then to cite, a top-ranked article in preference to a more recent one. This would result in a boost to the citation counts for articles that were in the downward part of the normal distribution, and could thus result in the second maximum observed for their knowledge management articles and for the 2002 Allen paper here.

The behaviour of the two earlier articles is rather different in that the plots extend over long time periods. Thus, the annual citation count for 1979 article grew for no less than 14 years before reaching its maximum value in 1993, after which point it then slowly fell away (with the advent of the new, 1991 standard reference seemingly having only a minor effect on the decline). The extended growth phase here may be due to the growth in the literature: in 1979 there were 2,554 new articles in the Crystallography subject category of Web of Science whereas there were 5,297 such articles in 1993, i.e., the normal falling away may have been partially masked by the fact that there were 107% more articles available in the pool of those that might need to cite the standard reference. For the 1991 article, conversely, the maximum count was reached in only five years, during which time the crystallography literature had increased by just 30%.

The post-maximum phase of the distributions for the 1979 and 1991 articles are both very extended. Indeed, the 1979 article still received 37 citations in the five-year period 2014-18, despite the fact that the CSD was by then totally different from that described in the cited article. Moreover, seven of these citations came from core journals that are included in the right-hand part of Table 2 (one from *Crystal Growth & Design*, *Inorganic Chemistry*, and *Journal of Molecular Structure*, and two from *Dalton Transactions* and *Polyhedron*) where the authors might have been expected to be aware of more recent CCDC material. The 1991 article shows much the same behaviour, with again seven of the 33 citations during the period 2014-2018 coming from core journals. As to why such elderly material is still being cited, one possible cause is a study showing that there are now proportionally less citations to the most recent material than used to be the case, owing to the much greater ease with which older material can now be identified using digital search engines [16].

Where do crystallographers publish their structures?

Using the CCDC CSD Python API it is possible to identify the journals that most frequently contain crystal structures deposited into the CSD (the script for this is included in the Supplementary Information). The five most common sources of crystal structures

overall in the CSD are *Inorganic Chemistry, Dalton Transactions, Organometallics, The Journal of The American Chemical Society* and *Acta Crystallographica Section E.* We can, however, analyse how journal submission frequency has changed with time to understand the changing nature of crystallographic impact.

Figure 2a shows a line graph of the growth of submissions from the 11 most frequent sources of structures in the CSD up until 2018. In Figure 2b the same information is presented for journals that have a relatively recent history and seem to be rapidly becoming significant sources of crystallographic information. The graphs show interesting trends. For most journals there is still a steady increase in submissions containing structural data, but some drift down in the more traditional data sources. One noticeable fall is the decrease in submissions to *Acta Crystallographica Section E*. The decline began when publications from this journal stopped being included in the Science Citation Index, which in turn meant that the journal had no recorded impact factor after 2011 [17]. As has been noted, a lack of a measured journal impact factor can be problematic for academics [18].

We note the increase in *CSD Communications*[19] - i.e. authors directly depositing their structures with the CCDC - and an increase in publication in journals dedicated to the understanding and control of crystal structures (namely this journal, *CrystEngComm*, and the ACS journal *Crystal Growth and Design*); both these journals are now in the top 10 of data sources for CSD information. We also note the rapid growth of two general chemistry journals from the RSC, *RSC Advances* and *Chemical Science* which are now both in the top 20.

The CSD also has some cases of journals which contain few crystal structures, but often these publications themselves are interesting. For example, there is only one CSD-compliant crystal structure in the journal *Astrobiology* [20] that helps to support a theory for the origins of some key chemicals in the origin of life. We also note other rarities such as a crystal structure that was proven relevant in the preservation of books [21], a crystal structure in an electric engineering journal [22], where the authors were interested in developing new biofuel cell mediators, and curiously a crystal structure in a journal dedicated to non-crystalline solids [23]. Crystallography truly has broad impact.

Why do crystallographers publish their structures?

The internal repository of structures at CCDC contains the DOI to all publications containing structures in the CSD, alongside the paper title of the structural source. The unique set of paper titles represents an interesting corpus of words (the title corpus) that can be analysed to understand the trends in publication. It is hence possible to easily perform word frequency analysis and bigram frequency analysis to detect common words and phrases in the title corpus.

Using this data, one question we can immediately ask is “what motivates publication of a crystal structure?”. Frequency analysis of words in the publication titles provides an indication of motivations. The 20 most common words in publication titles associated with crystal structures are shown in Table 5. Most highly ranked words are unsurprising; words such as ‘structure’, ‘crystal’, ‘molecular’ and ‘characterization’ suggest crystallography’s primary use: namely to establish the nature of chemical material. Words such as ‘synthesis’ and ‘reaction’ further show how crystallography is a key characterization technique in understanding chemical processes and outcomes. The words ‘property’ and ‘complex’ also feature highly in the list.

Word	Count	Frequency (%)
structure	115210	3.14
synthesis	109619	2.99
complex	97271	2.65
crystal	64415	1.75
ligand	40882	1.11
reaction	29285	0.80
property	27072	0.74
molecular	23878	0.65
characterization	23768	0.65
study	22573	0.61
acid	22009	0.60
structural	19963	0.54
derivative	18896	0.51
coordination	18578	0.51
x-ray	18117	0.49
novel	17604	0.48
copper	16926	0.46
compound	16057	0.44
bond	15958	0.43
metal	13575	0.37

Table 5. Word frequencies for the 20 most frequent words in the titles of publications containing CSD-compliant crystal structures

The analysis can be taken further using bigram frequency analysis. In Figure 3 a word cloud of the most common bigrams is shown. Visual inspection gives the reader a general overview of areas of impact with many bigrams related to synthesis. We see the types of property commonly associated with titles ("magnetic property" stands out most strongly), and we can also see the types of metal complex that are common in titles.

How have themes changed over time?

Word frequencies can be broken down over time periods to give a good indication of what is fashionable in each time frame. The title corpus was split into several smaller corpuses that were constrained to fixed time ranges. Each smaller corpus was next analysed for the 50 most frequent words, and these were then combined into a larger set (i.e. all of the words that occur in the top 50 in at least one of the smaller corpuses) of 93 words and the frequency of occurrence of these words in a title was extracted for each time-bound smaller corpus (the full spreadsheet result of this analysis is available in the Supplementary Information).

We note some artefacts of this analysis: for example the French word "cristalline" (sic.) does occur in the list of 93 frequent words, but "crystalline" does not feature. Reference to the primary data shows that 636 titles contain "cristalline". Most of these publications with French titles are in the earliest time-bound corpus. This corpus is quite small compared to the later corpuses, and so the frequency of this word is high there. By comparison, "crystalline" occurs only 215 times in the early corpus, even though "crystalline" occurs 6203 times overall in the title corpus. Consequently, "crystalline" does not get included in the set of 93 words. We note, however, that this does reveal a common trend in publication; since the 1980s English has become the *de facto* standard language for publication. Another artefact is, for example, "triphenylphosphine" and "pph3" being treated as separate words even though they have the same chemical meaning (we note that examples of this type are a challenge for natural language processing which would require very sophisticated chemically aware methods for detection of word equivalence beyond classical stemming and lemmatization).

In Figure 4 we can see a selection of words where we note significant movement in relative frequency. For example, "synthesis" seems to have become more prevalent while "structure" has fallen away, and words relating to catalysis and frameworks seem to have risen. We also note the rise in words ("efficient", "property", "activity", "application") that relate strongly to applied research. This perhaps hints at the changing priorities of academic research groups (and indeed funding!) over time. Words strongly relating to the crystal structure itself have fallen over time, reflecting the change in focus of publication of crystal structures, since crystal structures are now published to aid research, rather than as the aim of research.

Conclusions

This short article has presented two themes in its analysis. In the first theme, the impact of the CSD as a collection of structures has been highlighted. The research shows that the CSD is still a heavily used resource but that the application domains where the CSD is used most heavily has changed since the inception of the database with changes to the rank order of classes of journal where citation is common. Generally, it is noted that getting the most recent 'reference' article cited appears to be more challenging as evidenced by the long tails to citation of old material when new material is available. The second theme deals with the structures inside the CSD using analysis of the titles of the related publications. Several interesting observations are observed as outlined in the earlier discussion. We note that molecular crystallography is now more commonly a technique that is used for research rather than as a research end in its own right. We also note that the aims and directions of research where crystallography is used has moved from fundamental domains to applied domains. Finally, we can understand the breadth of impact that small molecule organic and organo-metallic crystallography has had through an analysis of the journal article titles associated with the 1 million+ crystal structures that are now in the CSD. .

Conflicts of interest

Dr. Jason C. Cole & Dr Ian J. Bruno both work for the Cambridge Crystallographic Data Centre, the creators of the Cambridge Structural Database.

Acknowledgements

We acknowledge all the authors of publications and crystallographers who have, by their sterling efforts, created the information that makes up the powerful resource that is the Cambridge Structural Database. We also acknowledge those scholarly publishers who help to facilitate the publication of this information.

Notes and references

1 D. T. Hawkins, *Acta Cryst*., 1980, A**36**, 475-482
2 S. C. Abrahams and R. A. Matula, *Acta Cryst*., 1988, A**44**, 401-410.
3 J. Redman, P. Willett, F. H. Allen and R. Taylor, *J. Appl. Cryst*., 2001, **34**, 375-380
4 H. Behrens and P. Luksch, *Acta Cryst*., 2006, B**62**, 993-1001
5 R. Wong, F. H. Allen and P. Willett, *J. Appl. Cryst*., 2010, **43**, 811-824.
6 J. Newman, D. R. Burton, S. Caria, S. Desbois, C. L. Gee, V. J. Fazio, M. Kvansakul, B. Marshall, G. Mills, V. Richter, S. A. Seabrook, M. B. Wu and T. S. Peat, *Acta Cryst*., 2013, F**69**, 712-718.
7 *Home - Clarivate Analytics*; https://clarivate.com/ (last accessed 13th January 2020)
8 *You are Crossref – Crossref*; https://www.crossref.org/ (last accessed 13th January 2020)
9 S. Bird, E. Loper, E. Klein (2009), *Natural Language Processing with Python.* O'Reilly Media Inc.
10 E. Yan, Y. Ding, B. Cronin and L. Leydesdorff, J. Informetrics, 2013, 7, 249-264.
11 M. R. Simu, E. Pall, T. Radu, M. Miclaus, B. Culic, A.-S. Mesaros, A. Muntean, G. A. Filip, *Tissue Cell*, 2018, **52**, 101-107.
12 F. T. Marchese and Suzanne M Marchese, *Proceedings of the 14th International Conference on Information Visualisation*, 2010, 487-493.
13 Y. Zhang,W. Hua and S. Yuan, *Learned Publishing*, 2018, **31**, 95-106.
14 B. M Kale, J., Wiener, J. Militky, S. Rwawiire, R. Mishra and A. Jabbar, *Cellulose*, 2016, **23**, 981-992.
15 A. Serenko and J. Dumay, 2015, *J. Knowledge Manag*., **19**, 1335-1355
16 A. Verstak, A. Acharya, H. Suzuki, S. Henderson, M. Iakhiaev, C. C. Y. Lin and N.Shetty, 2014, *arXiv*, 1411.0275
17 *(IUCr) impact factors for IUCr journals*; https://journals.iucr.org/services/impactfactors.html (last accessed 19 December 2019)
18 J. Bajorath*, Repositioning the Chemical Information Science Gateway;* https://f1000research.com/articles/8-976 (last accessed 19th December 2019)
19 CSD Communications, ISSN: 2631-9888, https://www.ccdc.cam.ac.uk/Community/csd-communications/ (last accessed 9th January 2020)
20 S. Fox and H. Strasdeit, *Astrobiology*, 2013, 13, 6
21 J. Stenger, E. E. Kwan, K. Eremin, S. Speakman, D. Kirby, H. Stewart, S. G. Huang, A. R. Kennedy, R. Newman, N. Khandekar, *e-PRESERVATIONScience*, 2010, 7 147-157, ISSN: 1581-9280

22 Y. Takeuchi and T. Akitsu*, Journal of Electrical Engineering*, 2016, 4, 189-195
23 O. Petrova,I. Taydakov, M. Anurova, A. Akkuzina, R. Avetisov, A. Khomyakov, E. Mozhevitina, I. Avetissov, *Journal of Non-Crystalline Solids*, 2015, 429,213-218

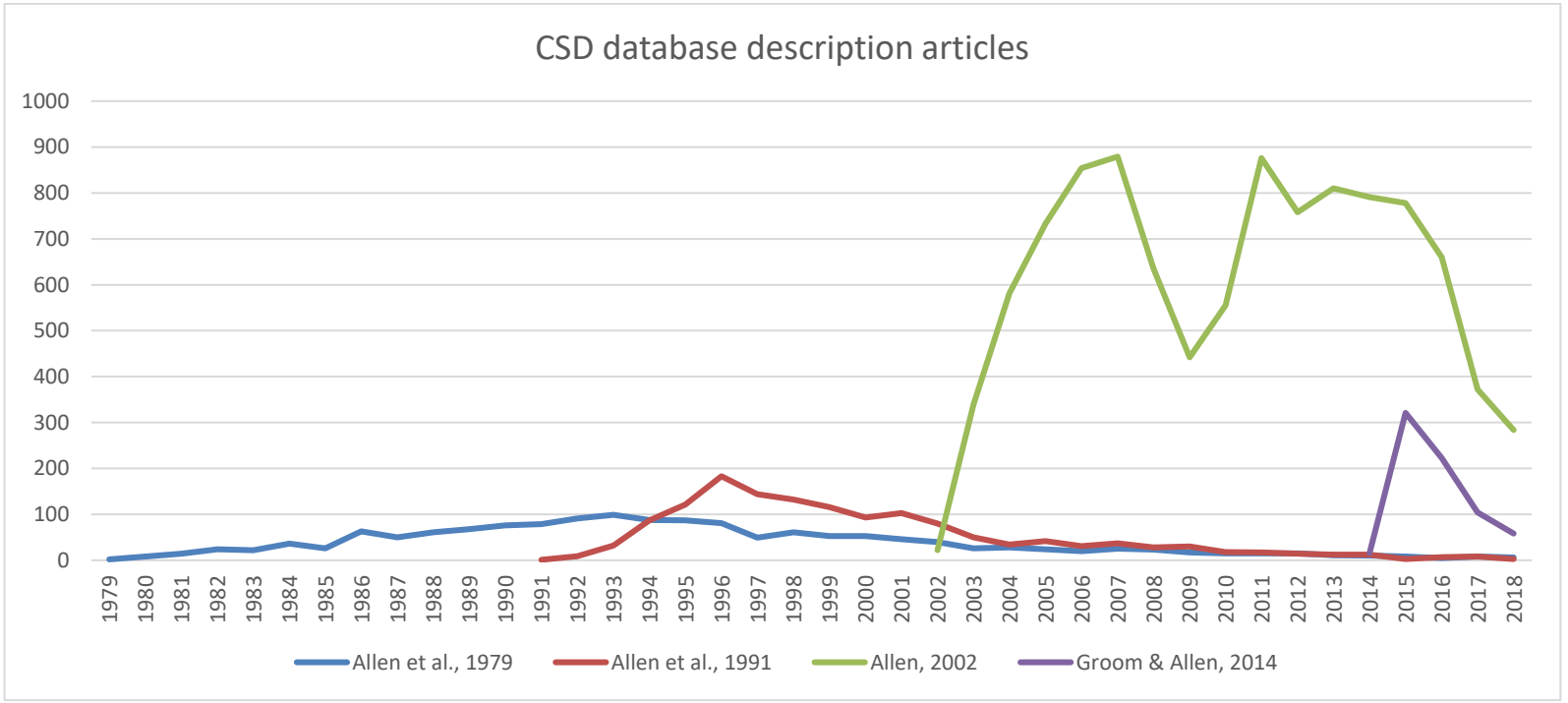
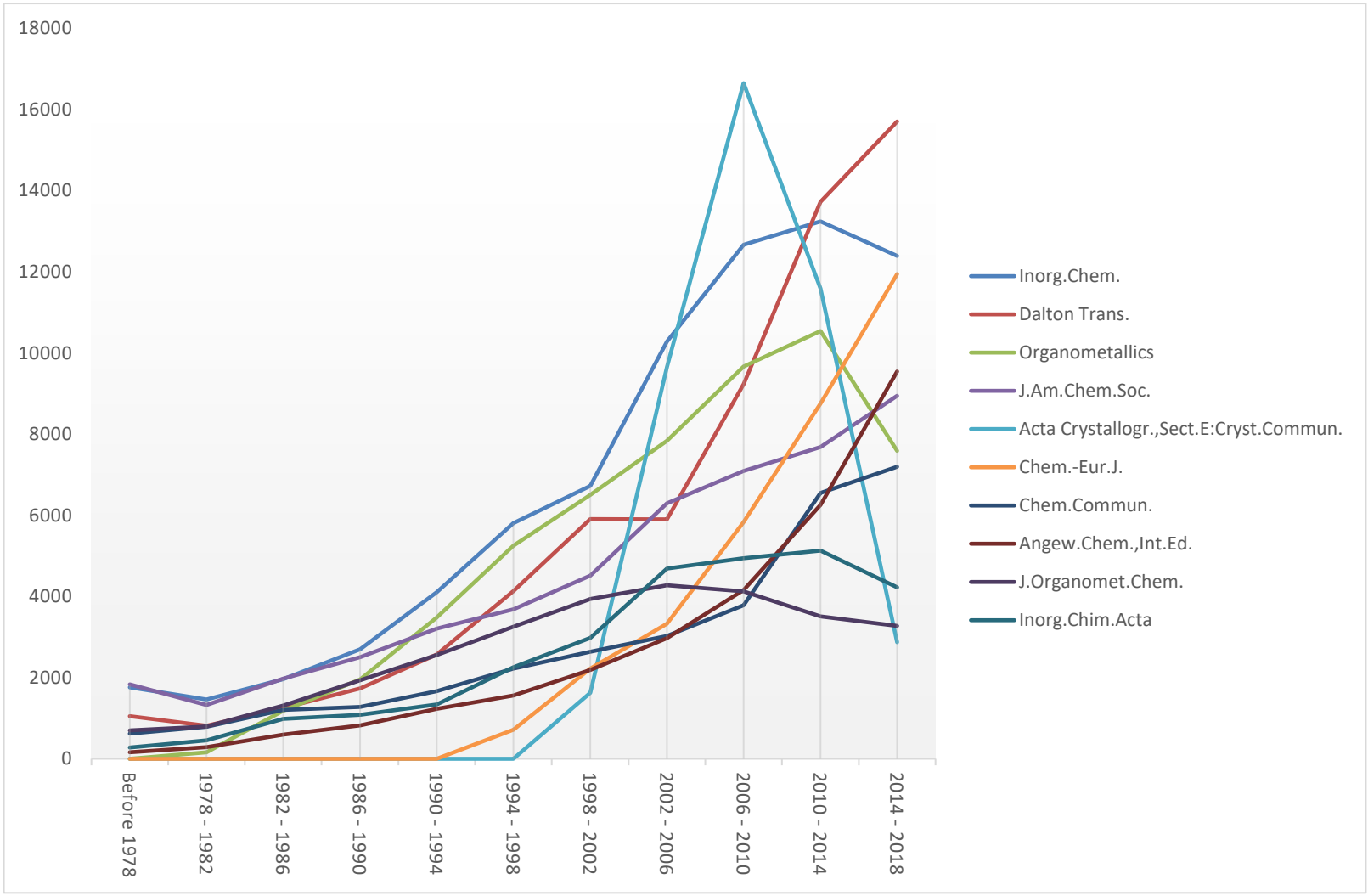
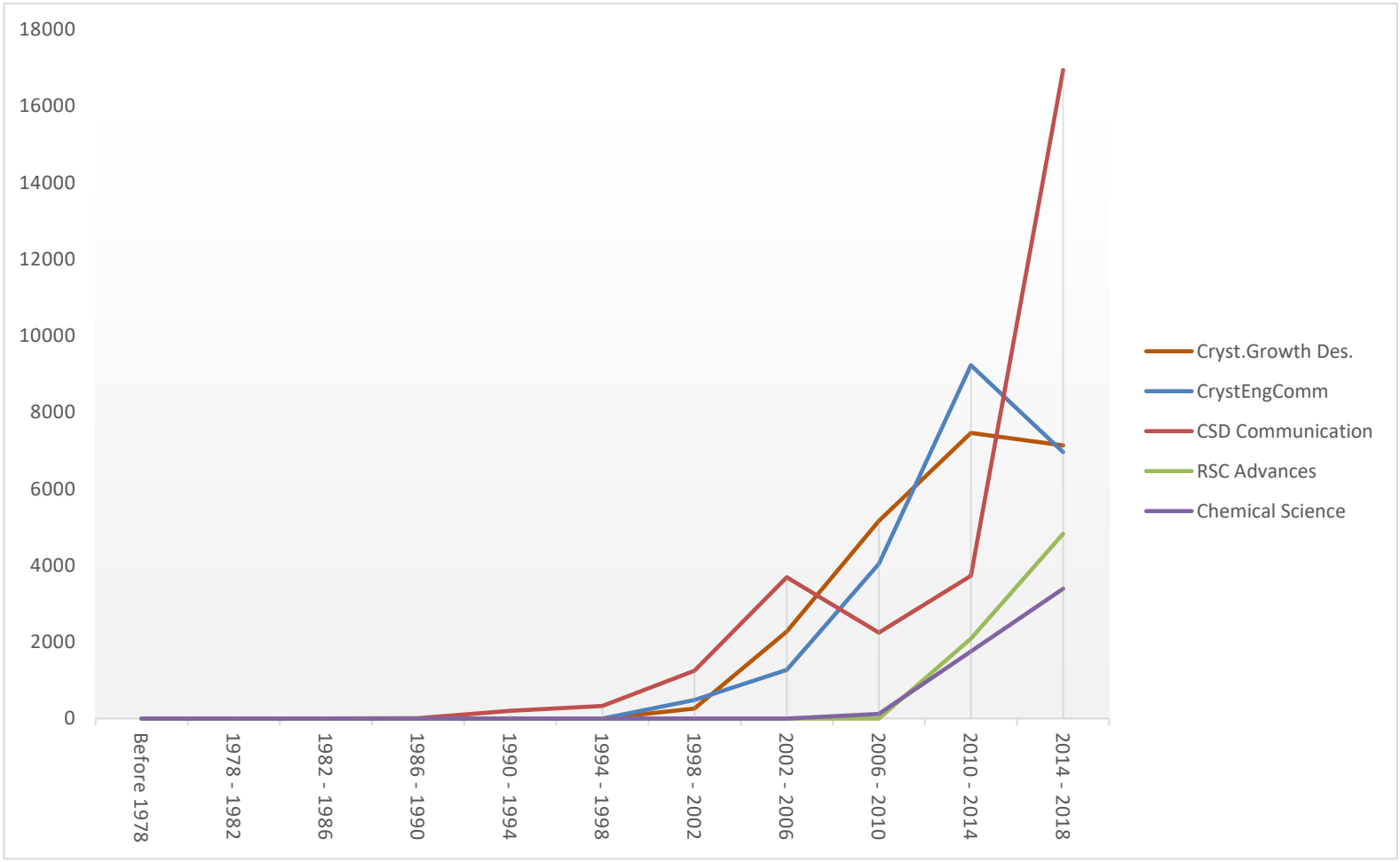


Figure 1. Obsolescence of CSD description articles


18000
16000
14000
12000
10000
8000
6000
4000
2000
0
Before 1978
1978 - 1982
1982 - 1986
1986 - 1990
1990 - 1994
1994 - 1998
1998 - 2002
2002 - 2006
2006 - 2010
2010 - 2014
2014 - 2018
Inorg.Chem.
Dalton Trans.
Organometallics
J.Am.Chem.Soc.
Acta Crystallogr.,Sect.E:Cryst.Commun.
Chem.-Eur.J.
Chem.Commun.
Angew.Chem.,Int.Ed.
J.Organomet.Chem.
Inorg.Chim.Acta

(a)



(b)

Figure 2. Growth (a) and decline (b) of data sources for CSD-compliant crystal structures

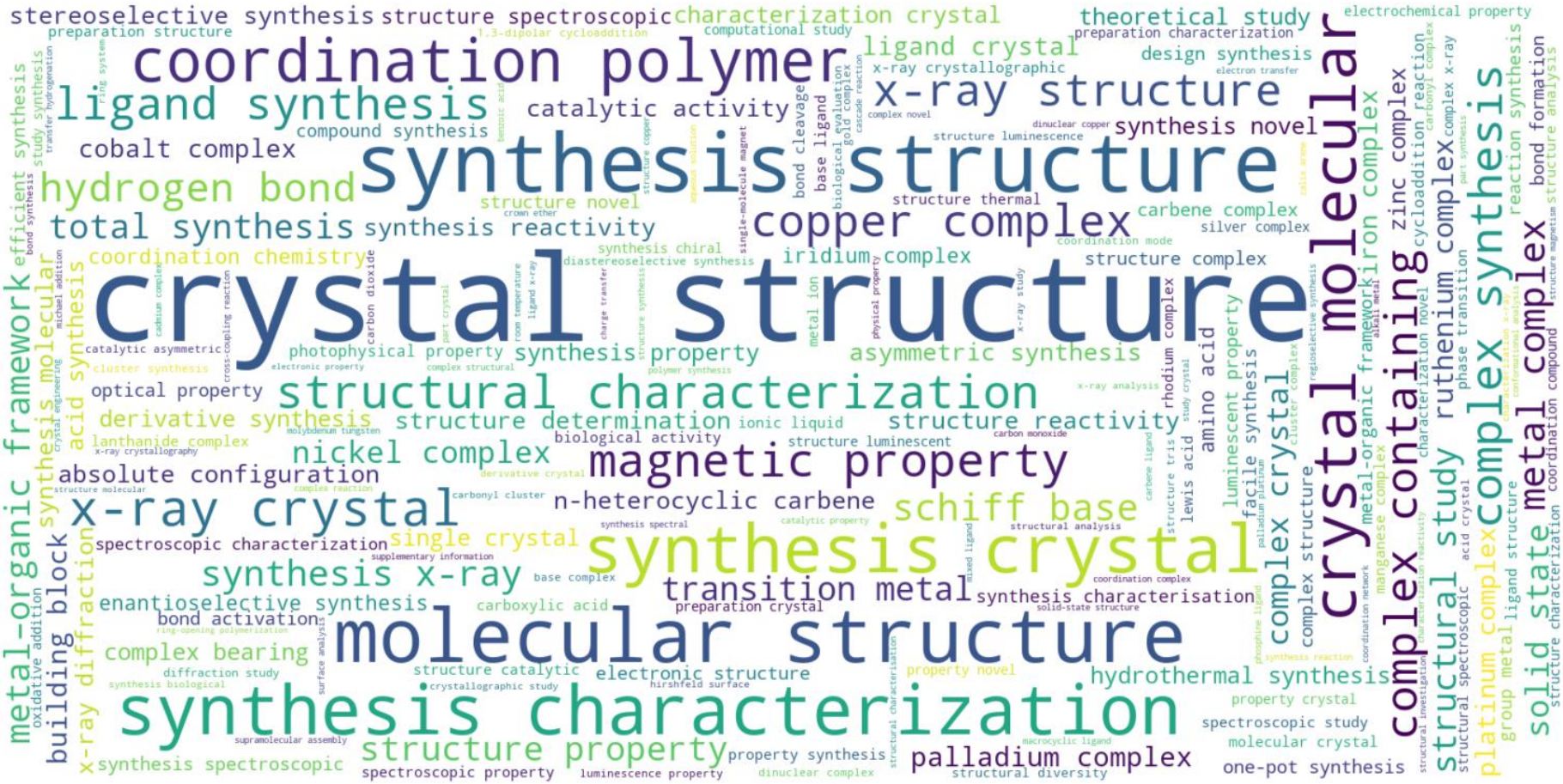
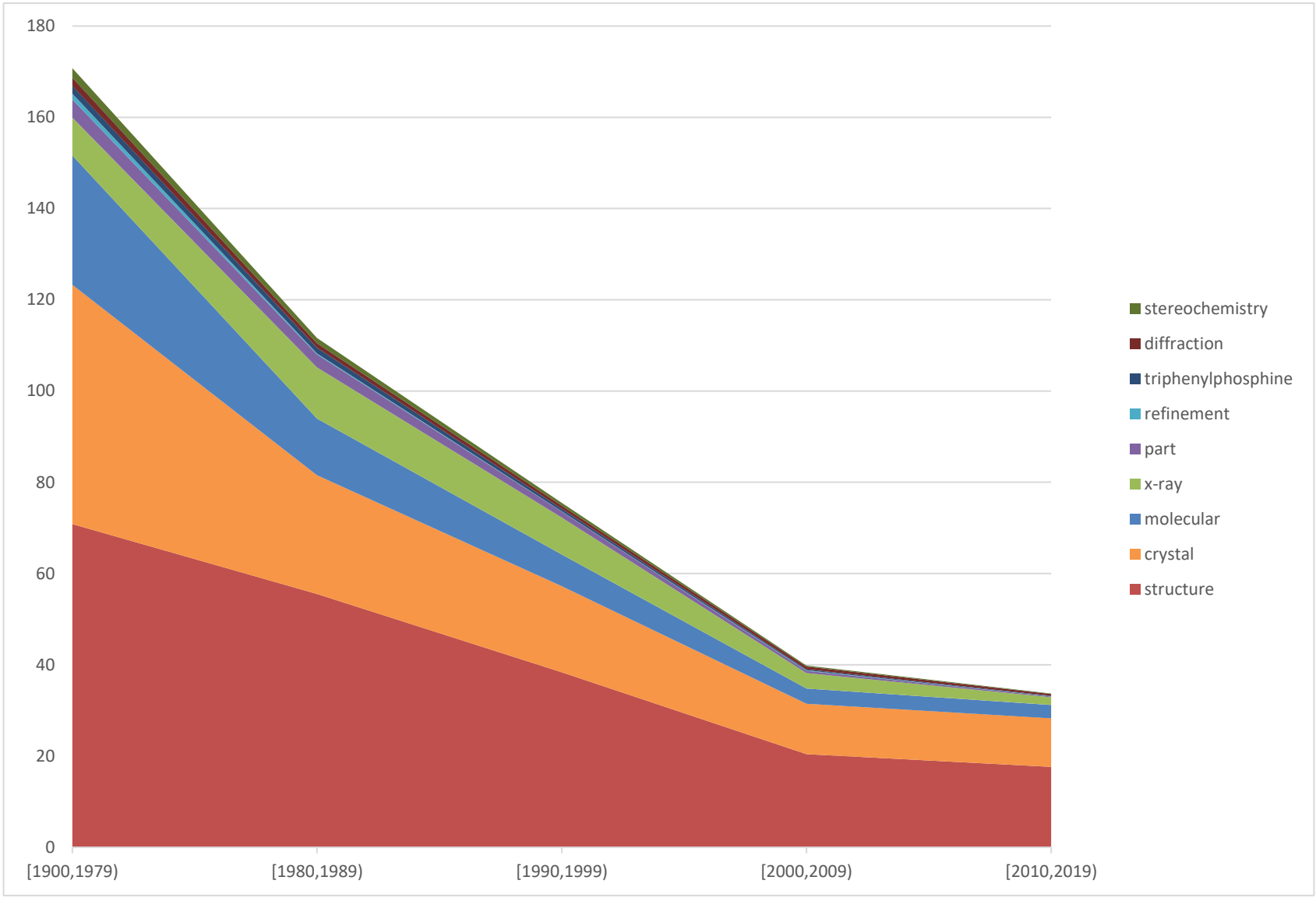
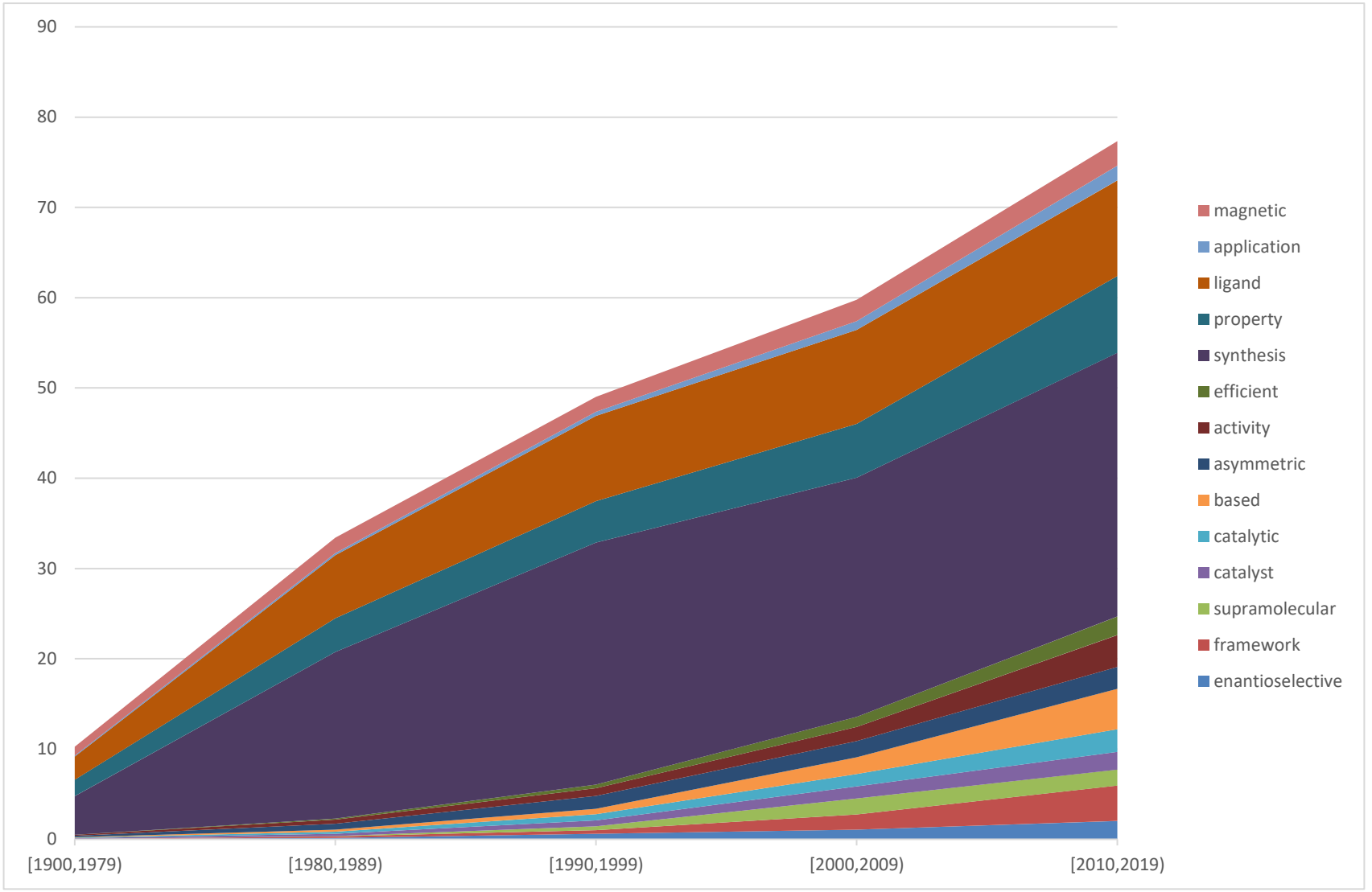


Figure 3. A word cloud of common bigrams in the titles of publications containing CSD-compliant crystal structures


180
160
140
120
100
80
60
40
20
0
[1900,1979)
[1980,1989)
[1990,1999)
[2000,2009)
[2010,2019)
stereochemistry
diffraction
triphenylphosphine
refinement
part
x-ray
molecular
crystal
structure

(a)



(b)

Figure 4. The fall (a) and rise (b) of specific words in titles of publications containing CSD-compliant structures