Impact Analysis of “Berman HM et al., (2000), The Protein Data Bank”

Rutgers University

May 2017

Contact details

Jodi Basner, PhD (jodi.basner@clarivate.com)

Scientific Analyst

Clarivate Analytics

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 EXECUTIVE SUMMARY

This report was commissioned by **Rutgers University** to obtain insight into the impact of The Protein Data Bank (PDB) on the research community. In 2000, Berman et al. published a paper in the journal *Nucleic Acids Research* that discussed the data, goals and expected future developments of the world-wide archive of structural data of biological macromolecules, The Protein Data Bank. Applying this seminal paper as a proxy, Rutgers University is seeking to characterize the citation-based impact of the more than 14,000 papers whose authors have indicated by citing the Berman et al. paper that The PDB is essential to their research. This report provides insight into the relative performance of the citing papers with perspectives on their scientific fields, geographic origin and funding support.

1.1 Key Findings

The PDB motivated high-quality research throughout the world.

- The Berman paper is among the Top 100 most highly-cited papers in all of Clarivate Analytics Web of Science™.
- The citing papers originated from 106 countries. United States of America (USA), United Kingdom (UK), Germany and India each contributed over 1,000 citing papers.
    - Nearly all of the Top 25 citing countries had a citation-based impact greater than the world-average.
- Among the Top 100 papers published since 2000, the Berman paper is ranked 5th out of the Top 11 with 13% highly-cited citing papers.

The PDB has maintained impact on a wide variety of scientific fields over the 17 years since publication.

- The greatest contribution of the citing papers was to the scientific fields of Biology & Biochemistry, Chemistry, Molecular Biology & Genetics and Computer Science with over a 1,000 papers in each research area.
- The greatest impact was from the 1,351 Computer Science papers with a citation-based impact twice the world-average.
- The PDB demonstrated high-quality impact with extensive reach across nearly all scientific research areas.
    - Citing papers had a citation-based impact exceeding the world-average in 16 scientific fields including Biology & Biochemistry, Computer Science, Plant & Animal Sciences, Physics, Environment/Ecology, Mathematics and Geosciences.

Papers citing The PDB were primarily funded by world-wide government agencies.

- Supporting a third of the citing research, the USA government agencies National Institutes of Health (NIH), National Science Foundation (NSF) and Department of Energy (DOE) were the most frequently acknowledged among the Top 25 funding organizations.
    - These USA government agencies funded high-quality research in a wide-breadth of research areas including Biology & Biochemistry, Chemistry, Molecular Biology & Genetics, Computer Science, Immunology, Clinical Medicine, Plant & Animal Science, Physics and Mathematics.

2 INTRODUCTION

2.1 Background

In 2000, Berman et al. published a paper in the journal *Nucleic Acids Research* that discussed the data, goals and expected future developments of the world-wide archive of structural data of biological macromolecules, The Protein Data Bank. Rutgers University is seeking to characterize the citation-based impact of the more than 14,000 papers whose authors have indicated by citing the Berman et al. paper that The Protein Data Bank (PDB) is essential to their research. This report provides insight into the relative performance of the citing papers with perspectives on their scientific fields, geographic origin and funding support.

2.2 About the client[1]

The PDB archive is the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. These are the molecules of life that are found in all organisms including bacteria, yeast, plants, flies, other animals, and humans. Understanding the shape of a molecule deduce a structure's role in human health and disease, and in drug development. The structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome.

The PDB archive is updated weekly and available at no cost to users.

The PDB was established in 1971 at Brookhaven National Laboratory under the leadership of Walter Hamilton and originally contained 7 structures. After Hamilton's untimely death, Tom Koetzle began to lead the PDB in 1973, and then Joel Sussman in 1994. In 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) became responsible for the management of the PDB. In 2003, the wwPDB was formed to maintain a single PDB archive of macromolecular structural data that is freely and publicly available to the global community. It consists of organizations that act as deposition, data processing and distribution centers for PDB data.

In addition, the RCSB PDB supports a website where visitors can perform simple and complex queries on the data, analyze, and visualize the results. Details about the history, function, progress, and future goals of the RCSB PDB can be found in the Annual Reports and Newsletters.

2.3 About Clarivate Analytics – Formerly the IP & Science Business of Thomson Reuters

Clarivate Analytics, formerly the IP & Science business of Thomson Reuters, provides reporting and consultancy services within research analytics using customized analyses to bring together several indicators of research performance to enable customers to rapidly make sense of and interpret a wide-range of data points, facilitating research strategy decision-making. We have extensive experience with databases on research inputs, activity and outputs and have developed innovative analytical approaches for benchmarking, interpreting and visualization of international, national and institutional research impact.

For over half a century we have pioneered the world of citation indexing and analysis, helping to connect scientific and scholarly thought around the world. Today, academic and research institutions, governments, not-for-profits, funding agencies, and all others with a stake in research, need reliable, objective methods for managing and measuring performance.

Our consultants have up to 20 years of experience in research performance analysis and interpretation. In addition, the Clarivate regional sales team provides effective project management and on-site support to maximize values of our projects and meet the expectations of Rutgers University.

[1] From 'About the PDB Archive and the RCSB PDB': http://www.rcsb.org/pdb/static.do?p=general_information/about_pdb/index.html

2.4 Scope of this report

Clarivate Analytics conducted a set of bibliometric analyses for all citing publications to Dr Berman's 2000 seminal paper. The citing papers to the Berman paper are proxy for understanding the citation-based impact of the PDB on the research community. The analysis consists of:

1. Overview landscape analysis
2. Research fields analysis
3. Funding organizations analysis

Together, these analyses provide a rich picture of the impact that was made by the PDB as indicated by Dr Berman's foundational work.

3 METHODOLOGY AND DATA SOURCES

In this section definitions and descriptions of the data sources, data, and metrics used throughout the report are provided.

3.1 Definitions and Metrics

Papers/publications: Clarivate Analytics abstracts publications including research journal articles, editorials, meeting abstracts and book reviews. The terms "paper" and "publication" are often used interchangeably to refer to printed and electronic outputs of many types. In the analyses presented here, the term "paper" is used exclusively to refer to substantive journal articles, reviews and some proceedings papers and excludes editorials, meeting abstracts or other types of publication. **Papers** are the subset of publications for which citation data are most informative and which are used in calculations of citation impact.

Research field: Standard bibliometric methodologies use Web of Science journal categories or **Clarivate Analytics InCites™: Essential Science Indicators℠** fields as a proxy for research fields.[2] Essential Science Indicators aggregate data at a higher level than the journal categories – there are only 22 Essential Science Indicators research fields compared to 252 journal categories. Journals are assigned to one or more categories, and every article within that journal is subsequently assigned to that category. Papers from prestigious, "multidisciplinary" and general medical journals such as *Nature*, *Science*, *The Lancet*, *The BMJ*, *The New England Journal of Medicine and the Proceedings of the National Academy of Sciences* (PNAS) are assigned to specific categories based on the journal categories of the references cited in the article. The selection procedures for the journals included in the citation databases are documented at the Clarivate Analytics master journal list website.

Highly-cited papers: Highly cited work is recognized as having a greater impact and Clarivate Analytics has shown that high citation rates are correlated with other qualitative evaluations of research performance, such as peer review. In the analysis presented here, publications that are in the top 10% in terms of citation frequency, taking into account year of publication and field, are considered to be highly cited. This threshold was selected after the review of a number of previous analyses showed this to be a useful value for general management purposes.

Category-normalized citation impact: Citation rates vary between research fields and with time. Consequently, analyses must take both field and year into account. In addition, the type of publication will influence the citation count. The standard normalization factor is the world average citations per publication for the year and journal category in which the publication was published, as well as document type to which the publication was assigned. This normalization is also referred to as 'rebasing' the citation count. The world-average is set to a value of one. Therefore, a paper with a Category-normalized citation impact of one is equal to the world-average.

3.2 Descriptions of Data Source

For the work described in this document, bibliometric data will be sourced from databases underlying the **Web of Science**, which gives access to conference proceedings, patents, websites, and chemical structures, compounds and reactions in addition to journals. It has a unified structure that integrates all data and search terms together and therefore provides a level of comparability not found in other databases. It is widely acknowledged to be the world's leading source of citation and bibliometric data. The **Web of Science** Core Collection is part of the Web of Science and focuses on research published in journals and conferences in science, medicine, arts, humanities and social sciences. The authoritative, multidisciplinary content covers over 12,500 of the highest impact journals worldwide, including Open Access journals and over 170,000 conference proceedings. Coverage is both current and retrospective in the sciences, social sciences, arts and humanities, in some cases back to 1900. Within the research community these data are often still referred to by the acronym 'ISI'. Clarivate Analytics has extensive experience with databases on research inputs, activity and outputs and has developed innovative analytical approaches for benchmarking and interpreting international, national and institutional research impact.

[2] Essential Science Indicators are defined by a unique grouping of journals with no journal being assigned to more than one field. These fields are focused on the science, technology, engineering and medicine subjects. Arts & humanities subjects are excluded. Customized analyses, however, can be designed to include these as an additional category.

Additionally, **InCites** will be also used in this project to generate relevant metrics and indicators across research activities and proposed analyses. **InCites** is a customized, citation-based research evaluation tool enabling analysis of productivity and benchmarking of output against peers worldwide, with underlying data drawn from Web of Science Core Collection.

The analysis will cover all publications that cited "Berman HM et al., (2000), The Protein Data Bank" from its publication to the end of December 2016. This will cover 15,647 research publications, and will include citations counts for those records to the end of December as updated in **InCites**.

3.3 Bibliometrics and citation analysis

Research evaluation is increasingly making wider use of bibliometric data and analyses. Bibliometrics is the analysis of data derived from publications and their citations. Publication of research outcomes is an integral part of the research process and is a universal activity. Consequently, bibliometric data have a currency across subjects, time and location that are found in few other sources of research-relevant data. The use of bibliometric analysis, allied to informed review by experts, increases the objectivity of, and confidence in, evaluation.

Research publications accumulate citation counts when they are referred to by more recent publications. Citations to prior work are a normal part of publication and reflect the value placed on a work by later researchers. Some papers get cited frequently and many remain uncited. Highly cited work is recognized as having a greater impact and Clarivate Analytics has shown that high citation rates are correlated with other qualitative evaluations of research performance, such as peer review. This relationship holds across most science and technology areas and, to a limited extent, in social sciences and even in some humanities subjects.

Indicators derived from publication and citation data should always be used with caution. Some fields publish at faster rates than others and citation rates also vary. Citation counts must be carefully normalized to account for such variations by field. Because citation counts naturally grow over time, it is essential to account for growth by year. Normalization is usually done by reference to the relevant global average for the field and for the year of publication.

Bibliometric indicators have been found to be more informative for core natural sciences, especially for basic science, than they are for applied and professional areas and for social sciences. In professional areas the range of publication modes used by leading researchers is likely to be diverse as they target a diverse, non-academic audience. In social sciences there is also a diversity of publication modes and citation rates are typically much lower than in natural sciences.

Bibliometrics work best with large data samples. As the data are disaggregated, so the relationship weakens. Average indicator values (e.g., of citation impact) for small numbers of publications can be skewed by single outlier values. At a finer scale, when analyzing the specific outcome for individual departments, the statistical relationship is rarely a sufficient guide by itself. For this reason, bibliometrics are best used in support of, but not as a substitute for, expert decision processes. Well-founded analyses can enable conclusions to be reached more rapidly and with greater certainty, and are therefore an aid to management and to increased confidence among stakeholders, but they cannot substitute for review by well-informed and experienced peers.

4 GLOBAL LANDSCAPE ANALYSIS

The Berman paper is among the Top 100 most Highly-cited papers in all of Web of Science. Since 2000, when the paper titled 'The Protein Data Bank' authored by Berman et al (Berman paper) was published, researchers on more than 15,000 publications have included the Berman paper among their bibliographic references. The term *publication* refers to many types of documents including conference proceedings and abstracts. Not all types of documents are characterized by the Clarivate Analytics bibliometrics therefore from this point forward the analysis presented in the report will be based on *papers*, a subset of publications that includes only Journal Articles and Reviews. The Berman paper has had extensive influence on the research community with 14,860 citing papers from 2000-2016 compared to an average of 16.4 citing papers per paper across the Web of Science.

In this section the overall citation-based impact on the world-wide research community of the more than 14,000 papers citing the Berman et al. paper is presented. Analysis results demonstrate that the PDB is an essential foundation of world-wide, high-quality research.

4.1 Performance of Citing Papers by Country

Papers citing the Berman paper originate from 106 countries with 31 countries each having contributed greater than 100 citing papers. Figure 1 presents a geographic map color coded by the count of citing papers from the Top 25 countries. With 5,528 papers, the USA had the greatest number of papers citing the Berman paper. The Top 10 countries, presented in Figure 2, had greater than 500 citing papers each. Nine of the top ten countries had a Category-normalized citation impact on par or greater than the world average, with India as the exception (0.71).

Figure 1 Geographical origin of citing papers by country (2000-2016). Top 25 countries based on the count of papers.

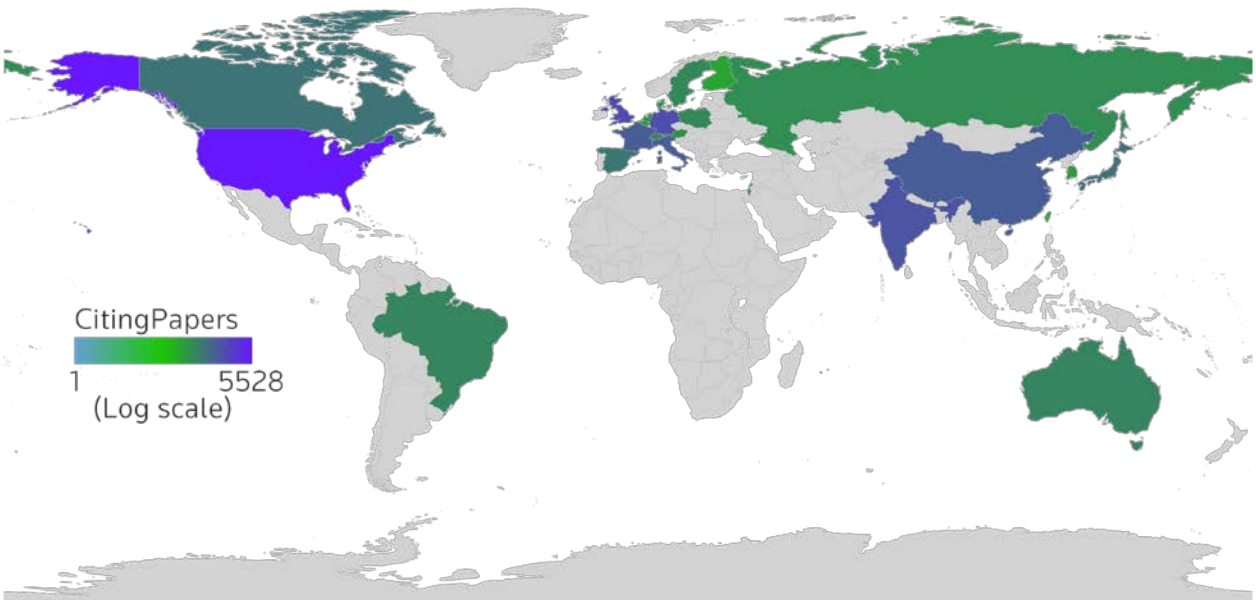


Figure 2 Performance of citing publications by country (2000-2016). Top 10 countries based on the count of papers.

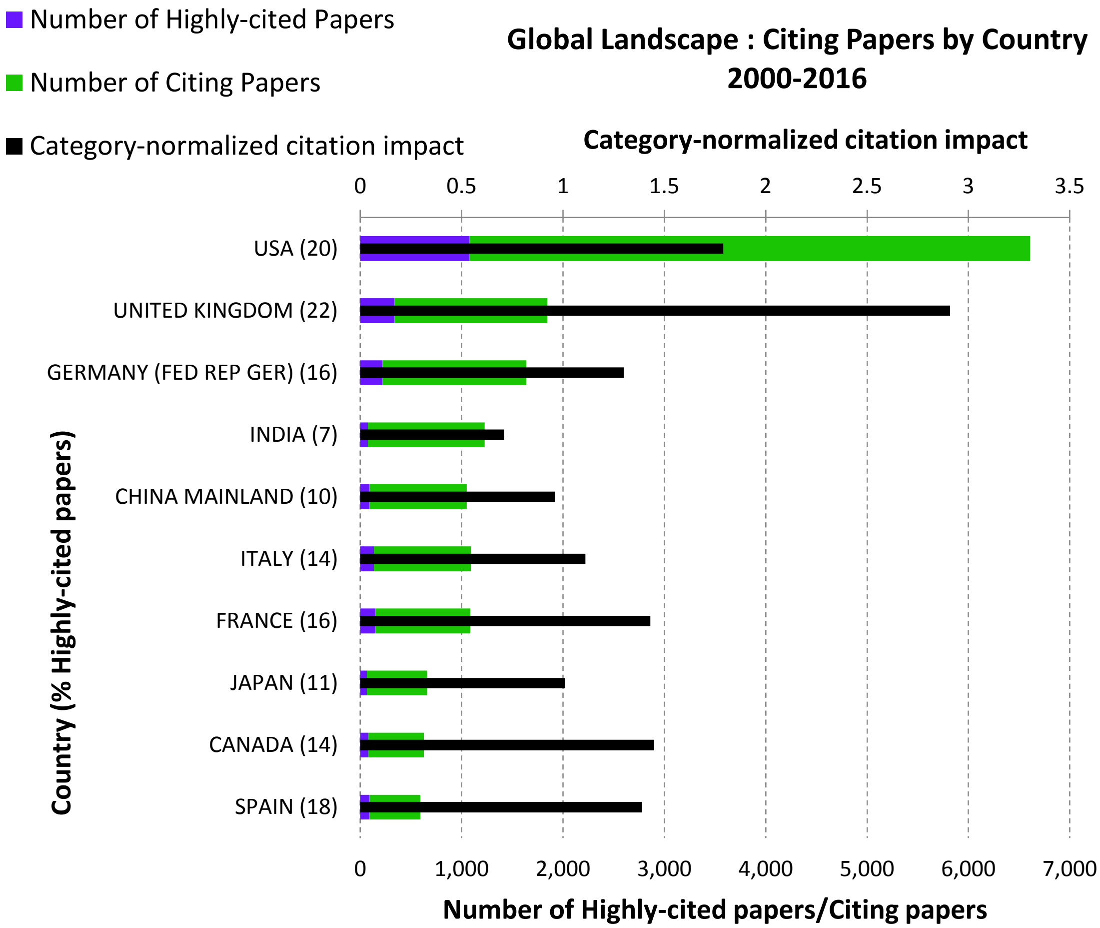


Table 1 presents the performance results for the Top 25 countries based on the count of papers. Among the Top 25, 21 countries had at least 10% of Highly-cited papers with Netherlands, UK, Denmark and the USA having at least 20% Highly-cited papers. Eighteen out of the Top 25 countries achieved a Category-normalized citation impact greater than the world-average (Category-normalized citation impact=1). The PDB has motivated high-quality research throughout the world.

Table 1 Performance of citing papers by country (2000-2016). Top 25 countries based on count of papers.

Country	Number of Citing Papers	Category-normalized citation impact	% Highly-cited Papers
USA	5,528	1.79	20
UK	1,510	2.91	22
GERMANY (FED REP GER)	1,417	1.3	16
INDIA	1,151	0.71	7
CHINA MAINLAND	959	0.96	10
ITALY	956	1.11	14
FRANCE	934	1.43	16

Country	Number of Citing Papers	Category-normalized citation impact	% Highly-cited Papers
JAPAN	594	1.01	11
CANADA	549	1.45	14
SPAIN	502	1.39	18
AUSTRALIA	362	1.1	15
BRAZIL	351	0.8	6
SWITZERLAND	347	1.76	18
POLAND	328	0.95	11
SWEDEN	326	1.33	15
RUSSIA	295	1.06	10
AUSTRIA	262	1.42	17
ISRAEL	258	1.49	19
NETHERLANDS	255	1.82	25
DENMARK	209	1.74	21
SOUTH KOREA	209	0.79	5
BELGIUM	201	1.18	12
FINLAND	164	1.01	13
TAIWAN	162	0.82	7
SINGAPORE	158	0.92	11

4.1 Performance of Citing Papers by Journal

The journal where a paper is published can influence its impact on the research community due to the journal's level of accessibility, visibility, respectability and interdisciplinarity. Overall, the citing papers were published in 1,726 different journals in 154 out of the 252 more granular Web of Science Journal Subject Categories. The Top 20 journals based on the count of papers presented in Table 2 indicate that the primary scientific fields impacted by the Berman paper include biophysics, molecular biology, biochemistry, and bioinformatics. Two multidisciplinary journals which reached a broader scientific audience, *PLOS ONE* and *PNAS*, are among the Top 20. The Journal Impact Factor is provided but we recommend, due to the known variability of citation accumulation by field and year (see Appendices), to consider the Journal Impact Factor cautiously. When possible, it is more appropriate to use the field-normalized bibliometrics.

Table 2 Performance of citing papers by journal (2000-2016). Top 20 journals based on the count of papers.

Journal Title	Number of Citing Papers	Journal Impact Factor	Category-normalized citation impact
PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS	710	2.499	1.15
NUCLEIC ACIDS RESEARCH	570	9.202	2.81
JOURNAL OF MOLECULAR BIOLOGY	476	4.517	1.43
PLOS ONE	433	3.057	0.77

Journal Title	Number of Citing Papers	Journal Impact Factor	Category-normalized citation impact
JOURNAL OF CHEMICAL INFORMATION AND MODELING	419	3.657	1.81
BIOINFORMATICS	416	5.766	1.97
JOURNAL OF BIOLOGICAL CHEMISTRY	322	4.258	0.97
BMC BIOINFORMATICS	311	2.435	1.34
JOURNAL OF MEDICINAL CHEMISTRY	306	5.589	2.97
ACTA CRYSTALLOGRAPHICA SECTION D-BIOLOGICAL CRYSTALLOGRAPHY	287	n/a	6.80
BIOCHEMISTRY	270	2.876	0.77
PROTEIN SCIENCE	246	3.039	1.04
JOURNAL OF PHYSICAL CHEMISTRY B	207	3.187	0.89
BIOPHYSICAL JOURNAL	200	3.632	1.39
JOURNAL OF COMPUTATIONAL CHEMISTRY	162	3.648	2.74
JOURNAL OF BIOMOLECULAR STRUCTURE & DYNAMICS	160	2.3	0.67
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA (PNAS)	159	9.423	2.33
JOURNAL OF MOLECULAR MODELING	156	1.438	0.42
JOURNAL OF COMPUTER-AIDED MOLECULAR DESIGN	147	3.199	1.49
JOURNAL OF MOLECULAR GRAPHICS & MODELLING	146	1.674	1.23

The Top 20 journals based on the Journal Impact Factor are presented in Table 3. Among these Top 20 journals, there were 63 papers from the *Nature* journal collection with an average Category-normalized citation impact of 4.96.

The Category-normalized citation impact of the two journal sets are also presented in Table 2 and Table 3. The citation-based impact of most of the journals in the set ranked by Journal Impact Factor is much greater than the world-average as well as the set of journals ranked by the Number of Citing Papers. This demonstrates the high-achieving performance of papers citing the PDB both where citing papers are most commonly published as well as those citing papers contained within some of the most highly-influential journals.

Table 3 Performance of citing papers by journal (2000-2016). Top 20 journals ranked by Journal Impact Factor.

Journal Title	Number of Citing Papers	Journal Impact Factor	Category-normalized citation impact
NATURE REVIEWS DRUG DISCOVERY	10	47.12	11.22
NATURE BIOTECHNOLOGY	8	43.113	6.64
NATURE REVIEWS MOLECULAR CELL BIOLOGY	3	38.602	7.68
NATURE	14	38.138	7.38
CHEMICAL REVIEWS	23	37.369	1.72
NATURE REVIEWS GENETICS	5	35.898	0.70

Journal Title	Number of Citing Papers	Journal Impact Factor	Category-normalized citation impact
NATURE NANOTECHNOLOGY	1	35.267	9.28
SCIENCE	13	34.661	4.86
CHEMICAL SOCIETY REVIEWS	5	34.09	0.44
NATURE GENETICS	3	31.616	3.47
NATURE MEDICINE	1	30.357	4.00
NATURE REVIEWS NEUROSCIENCE	1	29.298	0.34
CELL	3	28.71	8.44
NATURE CHEMISTRY	2	27.893	5.34
NATURE METHODS	9	25.328	5.28
NATURE REVIEWS MICROBIOLOGY	2	24.727	0.29
IMMUNITY	1	24.082	1.65
CANCER CELL	1	23.214	4.38
ACCOUNTS OF CHEMICAL RESEARCH	7	22.003	0.61
NATURE IMMUNOLOGY	4	19.381	2.84

4.2 Overall Performance of Citing Papers

With a Category-normalized citation impact of 1.39 and 13% highly-cited citing papers the Berman paper has performed above the world-average. To provide some additional context on the relative achievement of the Berman paper compared to other papers with a strong citation-based impact we examined the 11 other papers among the Top 100 most cited papers of Web of Science published since 2000. With a range between 6 to 31% highly-cited citing papers, the Berman paper ranked fifth among these 11 high impact papers demonstrating a strong performance among the top performers[3].

4.3 Global Landscape Review

The Berman paper has been cited by more than 14,000 original research papers from 106 countries across 1,726 journals. Eighteen of the Top 25 countries contributed high-quality research inspired by the Berman paper indicated by a citation-based impact greater than the world-average. The global landscape of the Berman paper demonstrates that the PDB has been a fundamental resource influencing research in the world-wide community.

[3] While the highly-cited metric is normalized, a comparison of the papers among the Top 100 most cited papers in Web of Science published across many different years and categories is likely also influenced by exogenous factors.

5 RESEARCH FIELDS ANALYSIS

In this section, the citation-based impact on the research community of the more than 14,000 papers citing the Berman et al paper is characterized by **Research field** using the broad Essential Science Indicator (ESI) subject categories. As presented in Figure 3, there are 22 (Multidisciplinary category is not shown) broad subject categories by which each paper indexed in the Web of Science is categorized. Analysis results demonstrate that the PDB maintained impact on the research community over the 17 years since publication on a wide variety of these scientific disciplines.

Figure 3 ESI broad subject categories.

Clinical Medicine	Materials Science	Psychiatry/Psychology
Chemistry	Engineering	Microbiology
Physics	Immunology	Agricultural Sciences
Biology & Biochemistry	Social Sciences, General	Space Science
Molecular Biology & Genetics	Environment/Ecology	Computer Science
Neuroscience & Behavior	Pharmacology & Toxicology	Economics & Business
Plant & Animal Science	Geosciences	Mathematics

5.1 Performance of Citing Papers by ESI Subject Category

An investigation of the subject categories of the citing papers provides perspectives on the reach and impact of the PDB across scientific research areas. Figure 4 presents the Category-normalized citation impact for the Top 10 ESI Subject Categories based on a count of papers. For seven of the Top 10 categories the citation impact was greater than the world-average, demonstrating that the PDB has had high-quality impact with extensive reach across scientific disciplines.

Figure 4 Performance of citing papers by ESI subject category (2000-2016). Category-normalized citation impact for the Top 10 ESI based on a count of papers.

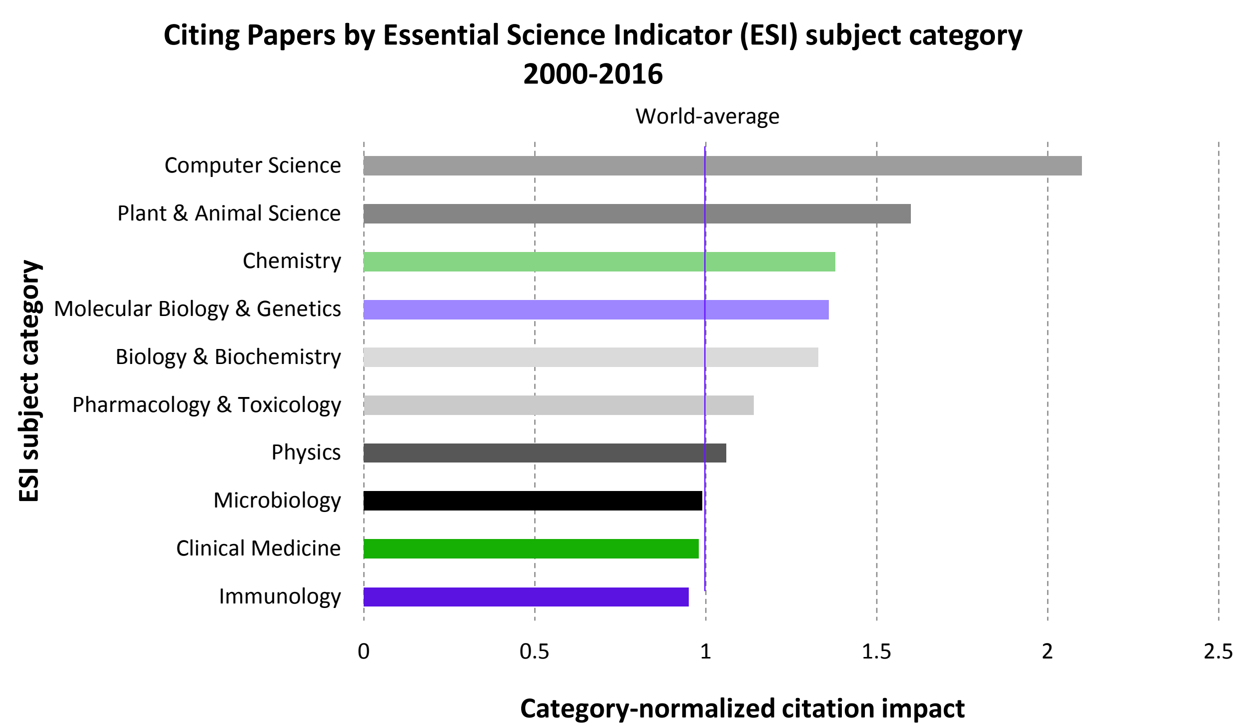
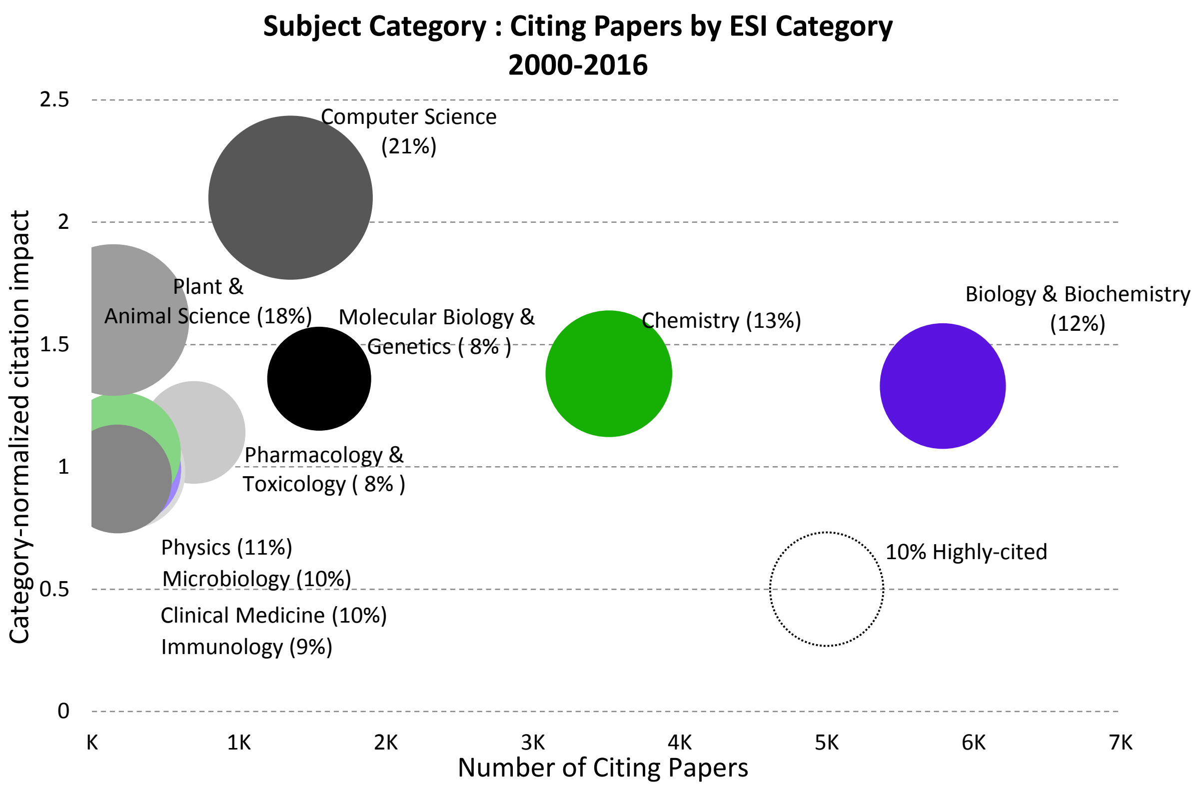


Figure 5 presents the Category-normalized citation impact and Number of Citing Papers along with the %Highly-cited papers for the Top 10 ESI based on a count of papers. The greatest counts of citing papers were from Biology & Biochemistry (5,790 papers) and Chemistry (3,518). The greatest impact was from the greater than 1,000 Computer Science papers indicated both by the Category-normalized citation impact (2.1) and %Highly-cited papers (21%). With a Category-normalized citation impact of 1.6 and 18% %Highly-cited papers, Plant & Animal Sciences ranked second in citation-based impact among the Top 10 ESI.

Figure 5 Performance of citing papers by ESI subject category (2000-2016). Top 10 ESI based on count of papers.



As presented in Table 4, the PDB reached nearly all ESI categories with high-quality, a remarkable accomplishment. Of the 22 ESI categories, the PDB is represented in all but one – Space Science. For most of the ESI categories, except Psychiatry/Psychology, the citation-based impact indicated by the Category-normalized citation impact was at least the world-average. The distribution of the number of citing papers across the broad ESI categories, as was indicated in Table 2 Performance of citing papers by journal (2000-2016). Top 20 journals based on the count of papers., supports the perspective that the major contribution of the citing papers was to the scientific fields of Biology & Biochemistry, Chemistry, Molecular Biology & Genetics and Computer Science. However, the Berman paper had extensive reach in that it received many citations from research fields not only related to Biology including, as already noted, Computer Science as well as Physics, Plant & Animal Sciences, Agricultural Sciences, and Mathematics.

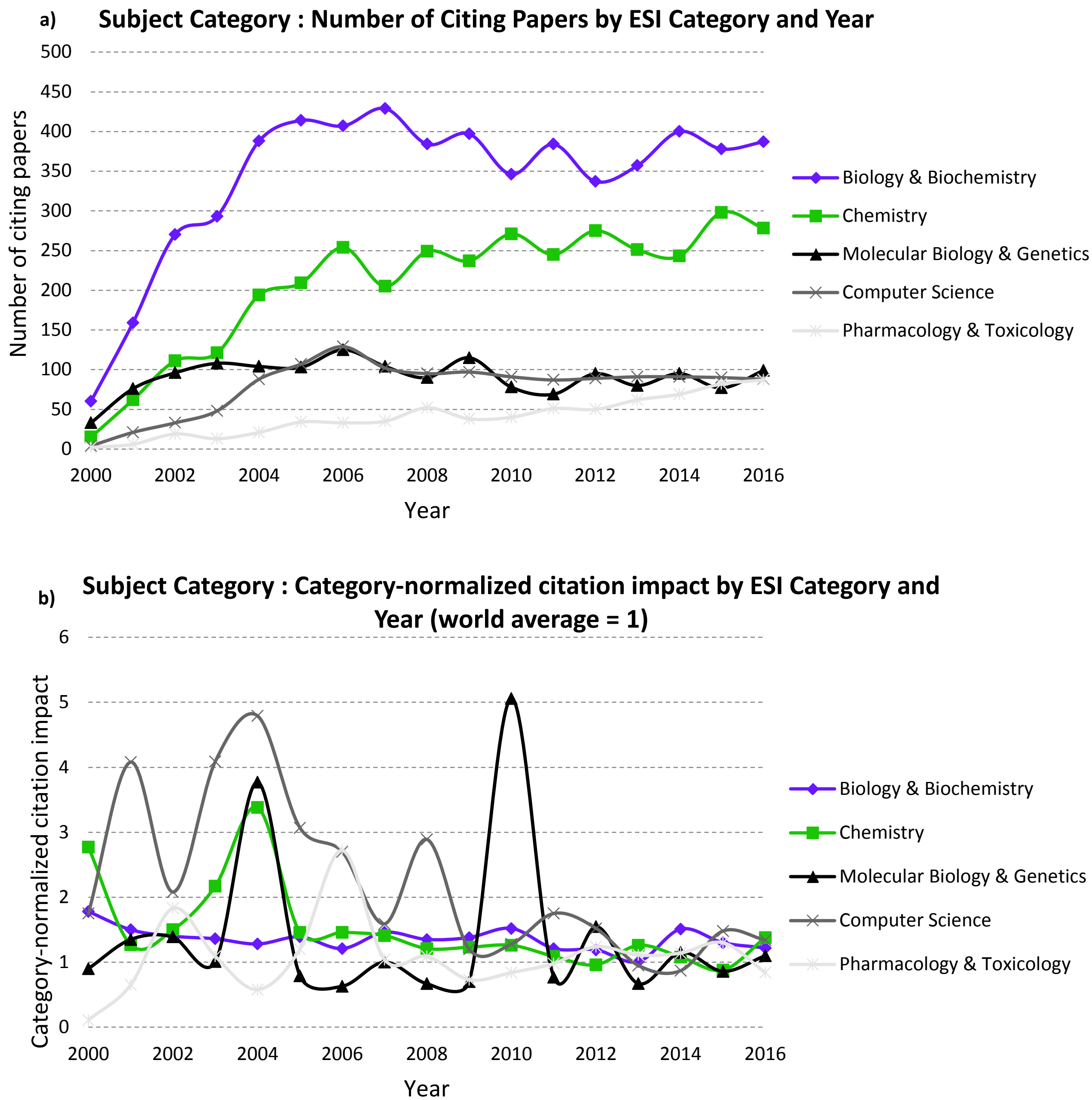
Table 4 Performance of citing papers by ESI subject category (2000-2016). All categories except Space Science are represented.

ESI Category	Number of Citing Papers	Category-normalized citation impact	% Highly-cited Papers
Biology & Biochemistry	5,790	1.33	12
Chemistry	3,518	1.38	13
Molecular Biology & Genetics	1,547	1.36	8
Computer Science	1,351	2.1	21
Pharmacology & Toxicology	694	1.14	8
Clinical Medicine	246	0.98	10
Microbiology	226	0.99	10
Physics	196	1.06	11
Immunology	174	0.95	9
Plant & Animal Science	145	1.6	18
Engineering	122	0.99	7
Materials Science	71	1.01	7
Agricultural Sciences	48	1.64	21
Neuroscience & Behavior	44	1.09	16
Mathematics	39	2.05	28
Environment/Ecology	26	1.22	12
Multidisciplinary	25	1.08	12
Social Sciences, general	19	2.65	11
Economics & Business	2	1.01	0
Psychiatry/Psychology	2	0.49	0
Geosciences	1	4.17	100

5.2 Trends in Performance of Citing Papers by ESI Subject Category

An investigation of trends in citation-based impact reveals if performance has been consistent over time. In each of the Top 5 ESI categories, presented in Figure 6, the Berman paper demonstrated a strong performance with a steady increase in the number of citing papers in the first five years of publication and then sustaining citation accumulation to 2016, rather than the more typical pattern of tapering off after the sharp initial rise. For eight of the Top 10 categories (data not shown for categories ranked 5-10) the Category-normalized citation impact was at or exceeded world-average for more than half of the 17 years since publication; Biology & Biochemistry (17 years), Chemistry, Computer Science, Plant & Animal Science (15 years), Pharmacology & Toxicology (10 years), Molecular Biology & Genetics, Physics, and Immunology (nine years).

Figure 6 a) Number of citing papers by ESI Category and year (2000-2016). b) Category-normalized citation impact trend of citing papers by ESI Category and year (2000-2016). Top 5 categories based on count of papers.



The ESI categories are broad definitions of scientific fields. The Web of Science Journal Subject Category groups are more granular. To further highlight the wide-breadth of the impact of the Protein Data Bank, Table 5-7 present the Journal Subject Categories of the Top 10 journals of the ESI categories Computer Science, Physics and Plant & Animal Sciences. These tables provide some insight into the sub-topics among the broader ESI categories of the citing papers. Within Computer Science, sub-topics included mathematical biology, mathematics, and method development. Within Physics sub-topics included multidisciplinary physics, condensed matter, and optics. Within Plant & Animal Sciences sub-topics included plant sciences, cell biology, multidisciplinary, and evolutionary biology.

Table 5 Journal Subject Categories of the Top 10 journals of ESI category Computer Science.

Computer Science	
Subject Categories of Top 10 Journals	**Number of Journals**
Mathematical & Computational Biology	7
Biochemical Research Methods	6
Computer Science, Interdisciplinary Applications	5
Biology	2
Biotechnology & Applied Microbiology	2
Biochemistry & Molecular Biology	1
Chemistry, Multidisciplinary	1
Computer Science, Information Systems	1
Crystallography	1
Engineering, Biomedical	1
Mathematics, Interdisciplinary Applications	1
Statistics & Probability	1

Table 6 Journal Subject Categories of the Top 10 journals of ESI category Physics.

Physics	
Subject Categories of Top 10 Journals	**Number of Journals**
Physics, Multidisciplinary	3
Physics, Condensed Matter	2
Optics	2
Physics, Mathematical	2
Physics, Fluids & Plasmas	1
Instruments & Instrumentation	1
Physics, Atomic, Molecular & Chemical	1
Computer Science, Interdisciplinary Applications	1
Physics, Applied	1

Table 7 Journal Subject Categories of the Top 10 journals of ESI category Plant & Animal Sciences.

Plant & Animal Sciences	
Subject Categories of Top 10 Journals	**Number of Journals**
Plant Sciences	7
Biochemistry & Molecular Biology	3
Cell Biology	3
Multidisciplinary Sciences	2
Evolutionary Biology	1
Genetics & Heredity	1

To further illustrate the topics explored by high-impact research citing the Protein Data Bank, Table 8 presents the Top 5 citing papers based on the Category-normalized citation impact for ESI categories Computer Science, Physics and Plant & Animal Sciences. This presents perspectives on sub-topics at the paper-level. From Computer Science the top citing papers were building new algorithms and software based on the Protein Data Bank. From Physics the PDB supported continued research on understanding protein folding, improved algorithms and software for structure determination, and understanding radiation damage. From Plant & Animal Sciences the top citing papers reported on basic research in genomics and proteomics of plants and animals.

Table 8 Top 5 citing papers based on the Category-normalized citation impact for ESI categories Computer Science, Physics, and Plant & Animal Sciences.

ESI Category	Title of Paper	Publication Year	Category-normalized citation impact
Computer Science	MUSCLE: a multiple sequence alignment method with reduced time and space complexity	2004	194.64
Computer Science	I-TASSER server for protein 3D structure prediction	2008	96.14
Computer Science	Intrinsically disordered protein	2001	57.77
Computer Science	PISCES: a protein sequence culling server	2003	48.59
Computer Science	Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites	2005	42.03
Physics	Using DelPhi Capabilities to Mimic Protein's Conformational Reorganization with Amino Acid Specific Dielectric Constants	2013	16.21
Physics	DNA sensing by silicon nanowire: Charge layer distance dependence	2008	10.06
Physics	Recent developments in the PHENIX software for automated crystallographic structure determination	2004	9.2
Physics	Small-world view of the amino acids that play a key role in protein folding	2002	8.7
Physics	A beginner's guide to radiation damage	2009	5.93
Plant & Animal Science	The structural basis for the difference in absorbance spectra for the FMO antenna protein from various green sulfur bacteria	2009	11.62
Plant & Animal Science	A RhoGDP dissociation inhibitor spatially regulates growth in root hair cells	2005	8.78
Plant & Animal Science	Genome-Wide Analysis of the bZIP Transcription Factors in Cucumber	2014	8.17

ESI Category	Title of Paper	Publication Year	Category-normalized citation impact
Plant & Animal Science	Comprehensive Genome-Wide Survey, Genomic Constitution and Expression Profiling of the NAC Transcription Factor Family in Foxtail Millet (Setaria italica L.)	2013	8.15
Plant & Animal Science	Plant membrane assays with cytokinin receptors underpin the unique role of free cytokinin bases as biologically active ligands	2015	7.51

5.3 Research Fields Review

The PDB has influenced high-quality work across the scientific research spectrum. The more than 14,000 papers citing the Berman paper generated research in nearly all scientific areas as indicated by the papers being classified into 21 of the 22 ESI subject categories. Over 1,000 citing papers were contributed to each of the fields of Biology & Biochemistry, Chemistry, Molecular Biology & Genetics and Computer Science. Top cited papers in Computer Science, Physics, and Plant & Animal Sciences reported on research, with the PDB as the foundation that developed algorithms to progress our understanding of protein folding, function and integrity. For eight out of the Top 10 scientific fields the citation-based impact was at least the world-average for greater than 50% of the 17 years since publication demonstrating consistent impact on a wide variety of research areas.

6 Funding Organization Analysis

In this section the citation-based impact on the research community of the papers citing the Berman et al paper is characterized by Funding Organization. Funding organization data are available in Web of Science starting in mid-2008 and primarily based on the funding agencies noted within the acknowledgements of a publication. The total count of distinct citing papers for the analysis in this section is 5,227.[4] The PDB has been positively influencing research supported by government funding agencies throughout the world.

6.1 Performance of citing papers by funding organization

The Top 25 funding organizations were identified based on a count of the citing papers. Table 9 presents the count of papers, field-normalized bibliometrics and the percent of international collaborations of the citing papers for the Top 25 funding organizations acknowledged on the citing papers. Thirty-four percent (1,796 distinct papers) of the research published citing the Berman paper acknowledged funding by the USA Federal Agencies National Institutes of Health (NIH), National Science Foundation (NSF) and Department of Energy (DOE). The NIH, NSF, and DOE were among the top 10 highest impact funding organizations with Category-normalized citation impacts of 2.22, 1.55, and 3.66 respectively. This suggests that research supported by these agencies and including use of the PDB achieves an impact greater than the world average.

Papers funded by 14 of the Top 25 funding organizations have accumulated as many or more citations than other world-wide papers relative to their field and publication year as indicated by the Category-normalized citation impact. Papers acknowledging funding from the Wellcome Trust had the greatest citation-based impact with a Category-normalized citation impact of 8.22 and 38% Highly-cited papers. Additionally, at 62%, papers supported by Wellcome Trust had the second greatest percentage of international collaborations. Wellcome Trust, a research charity, is the only non-governmental organization among the Top 25 funding organizations acknowledged on the citing papers.

Table 9 Number of funded papers and field-normalized bibliometrics by funding organization (2008-2016). Top 25 funding organizations based on count of papers that had acknowledged the funding organization.

Funding Organization	Number of Papers	Category-normalized citation impact	% Highly-cited Papers	% International Collaborations
NIH - USA	1,462	2.22	21	32
NSF	674	1.55	17	25
National Natural Science Foundation of China	468	0.97	9	28
DOE	238	3.66	19	35
German Research Foundation (DFG)	233	1.21	15	40
European Union (EU)	226	1.91	18	53
Council of Scientific & Industrial Research (CSIR) - India	172	0.74	7	13
National Council for Scientific and Technological Development (CNPq)	161	0.78	6	24
Government of India	158	0.62	4	15

[4] Fifty-nine percent of the citing papers from 2008-2016 acknowledged a funding source.

Funding Organization	Number of Papers	Category-normalized citation impact	% Highly-cited Papers	% International Collaborations
Department of Science & Technology	138	0.65	4	15
Department of Biotechnology (DBT)	134	0.6	4	19
Biotechnology and Biological Sciences Research Council (BBSRC)	127	3.21	28	42
Natural Sciences and Engineering Research Council of Canada	116	1.7	18	49
Scientific & Technological Research Council of Turkey (TUBITAK)	115	0.93	12	41
Wellcome Trust	100	8.22	38	62
Ministry of Education, Culture, Sports, Science and Technology, Japan	99	0.64	4	27
Ministry of Science and Innovation, Spain (MICINN)	96	1.46	22	46
National Basic Research Program of China	96	0.85	8	24
Coordination for the Improvement of Higher Education Personnel (CAPES)	95	0.97	7	29
Swiss National Science Foundation	87	1.81	23	64
University Grants Commission, India	85	0.86	9	8
The Ministry of Education, University and Research (MIUR)	84	1.17	18	52
Canadian Institutes of Health Research	83	2.13	22	39
Swedish Research Council	79	1.35	18	58
Portuguese Foundation for Science and Technology	70	1.29	21	59

6.2 Performance by funding organizations and ESI category

The Top 3 ESI categories, based on a count of citing papers, were determined for each of the Top 25 acknowledged funding organizations. Identifying the top ESI categories in terms of publication counts is a proxy of the subject areas that have been prioritized by each of the funding agencies. Six ESI categories were represented among the Top 3 for the Top 25 funding organizations. Table 10 presents the number of funding organizations (among the Top 25) that had a particular ESI among their Top 3. All of the Top 25 funding organizations supported papers in Biology & Biochemistry, and all but one supported papers in Chemistry. Of the Top 25 funding organizations, only the University Grants Commission, India, had research support in the field of Clinical Medicine among their Top 3 ESI categories. The analysis throughout Section 6.2 includes information focused on the Top 3 ESI of the Top 25 organizations[5].

Table 10 Summary of the Top 3 ESI Categories of the citing papers from the Top 25 funding organizations (2008-2016).

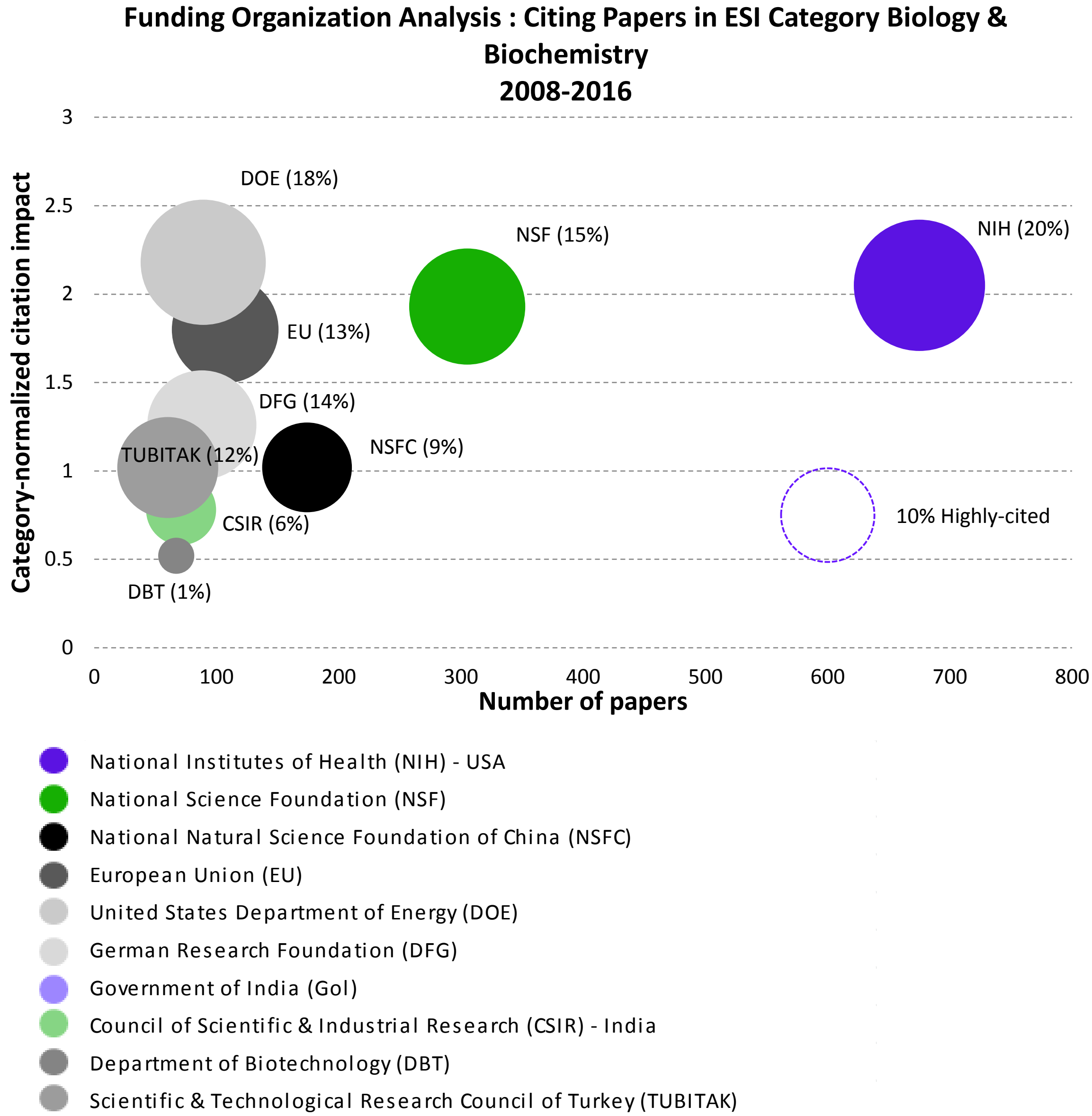
ESI Category	Number of Organizations with ESI Category among the Top 3 most funded
Biology & Biochemistry	25
Chemistry	24
Molecular Biology & Genetics	16
Computer Science	7
Pharmacology & Toxicology	2
Clinical Medicine	1

6.2.1 Biology & Biochemistry

The count of citing papers and field-normalized bibliometrics for the funding organizations that supported research in Biology & Biochemistry are presented in Figure 7 (Top 10 funding organizations) and in Table 11 (Top 25 funding organizations). High-impact research in Biology & Biochemistry was strongly supported by NIH, NSF, DOE, EU, DFG, National Natural Science Foundation of China and the Scientific & Technological Research Council of Turkey. These funding organizations ranked among the Top 10 organizations based on the count of citing papers and had a Category-normalized citation impact greater than the world average. The Top 25 funding organizations supported 77% of the citing papers in Biology & Biochemistry (1,620 out of the 2,116 distinct papers in Biology & Biochemistry with funding organization information).

[5] Therefore, if one of the Top 25 funding organizations supported research in an ESI subject category that was ranked greater than third, based on a count of citing papers, this information is not included in Section 6.2.

Figure 7 Performance of citing papers in ESI category Biology & Biochemistry by funding organization (2008-2016). Top 10 funding organizations based on the count of citing papers. The label provides the Funding Organization (% Highly-cited papers). This analysis focused on the Top 3 ESI of the Top 25 funding organizations[6].



[6] Organizations with 0% highly-cited papers will not appear on the bubble graphs.

Table 11 Performance of citing papers in ESI category Biology & Biochemistry by funding organization (2008-2016). The 25 organizations among the Top 25 funding organizations that supported research in Biology & Biochemistry. This analysis focused on the Top 3 ESI of the Top 25 funding organizations.

Funding Organization	Number of Papers	Category-normalized citation impact	% Highly-cited Papers
NIH - USA	675	2.05	20
NSF	305	1.93	15
National Natural Science Foundation of China	174	1.02	9
EU	107	1.8	13
DOE	89	2.18	18
DFG	88	1.26	14
Government of India	75	0.48	0
CSIR- India	71	0.78	6
DBT	67	0.52	1
TUBITAK	60	1.02	12
Department of Science & Technology (India)	57	0.49	2
BBSRC	55	3.42	16
Natural Sciences and Engineering Research Council of Canada	53	2.07	11
Wellcome Trust	46	5.61	35
Canadian Institutes of Health Research	40	2.77	15
Ministry of Education, Culture, Sports, Science and Technology, Japan	40	0.78	5
MICINN	39	1.37	15
National Basic Research Program of China	39	1.09	10
CNPq	36	0.52	3
University Grants Commission, India	32	1.13	9
Swiss National Science Foundation	31	2.28	23

6.2.2 Chemistry

The count of papers and field-normalized bibliometrics for the funding organizations that supported research in Chemistry are presented in Figure 8 and Table 12. The Top 24 funding organizations with high activity in Chemistry supported 75% of the citing papers in Chemistry (1,001 papers out of the 1,332 papers in Chemistry with funding organization information). High-impact research in Chemistry was strongly supported by NIH, NSF, DOE and the DFG. These funding organizations ranked among the Top 10 funding organizations based on the count of citing papers and had a Category-normalized citation impact greater than the world average. The Top 24 funding organizations with high activity in Chemistry supported 75% of the citing papers in Chemistry (1,001 papers out of the 1,332 papers in Chemistry with funding organization information).

Figure 8 Performance of citing papers in ESI category Chemistry by funding organization (2008-2016). Top 10 funding organizations based on the count of citing papers. The label provides the Funding Organization (% Highly-cited papers). This analysis focused on the Top 3 ESI of the Top 25 funding organizations.

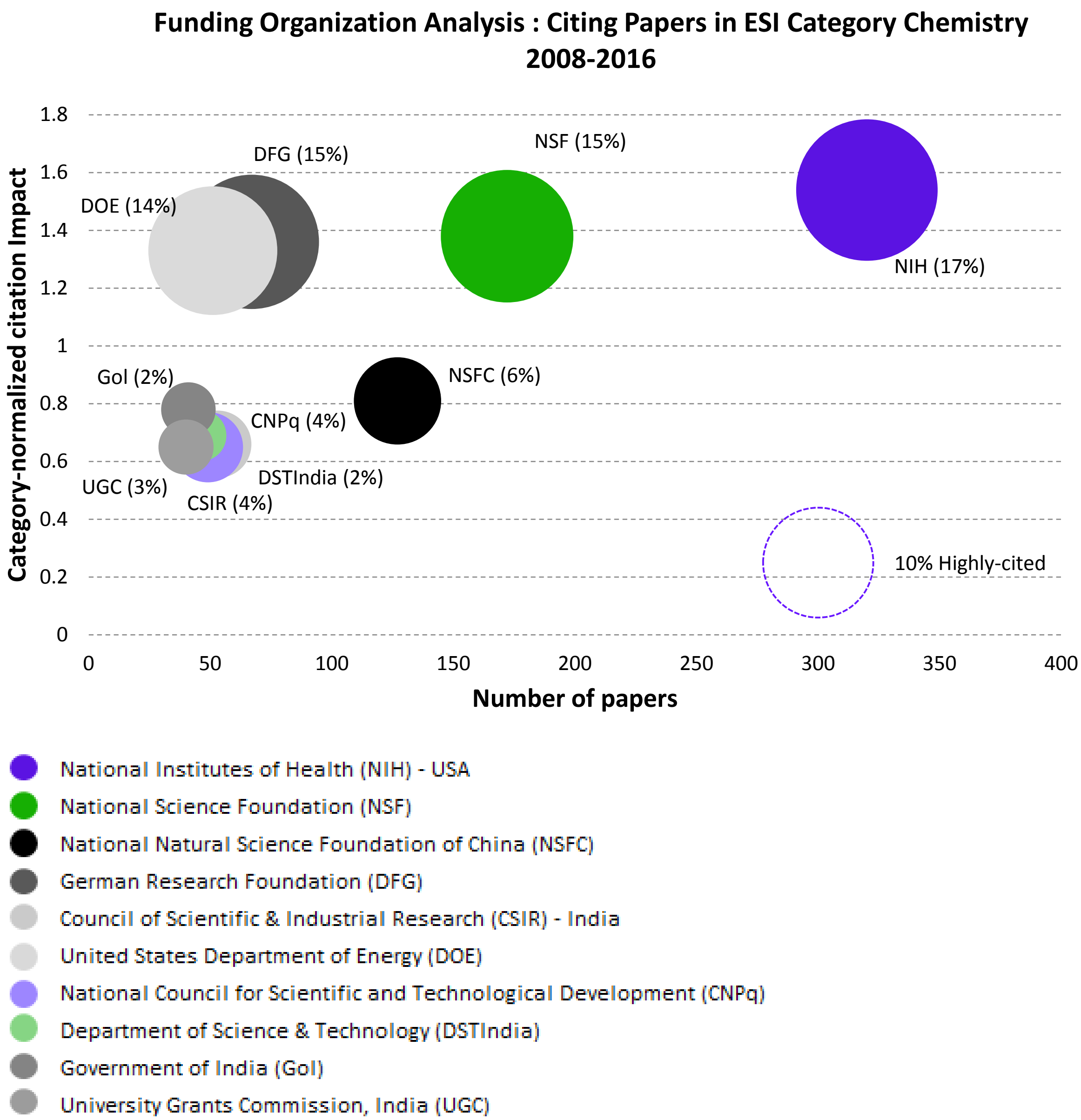


Table 12 Performance of citing papers in ESI category Chemistry by funding organization (2008-2016). The 24 organizations of the Top 25 funding organizations that supported research in Chemistry. This analysis focused on the Top 3 ESI of the Top 25 funding organizations.

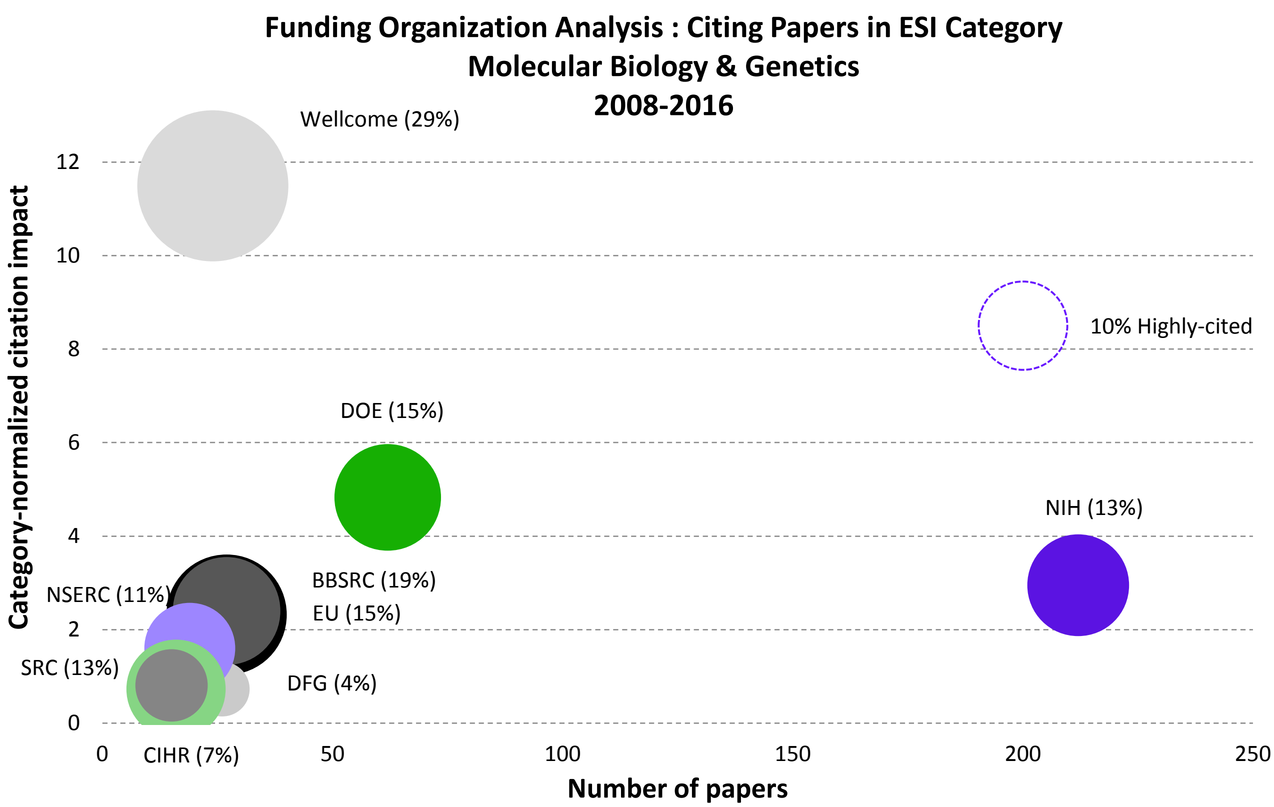
Funding Organization	Number of Papers	Category-normalized citation impact	% Highly-cited Papers
NIH - USA	320	1.54	17
NSF	172	1.38	15
National Natural Science Foundation of China	127	0.81	6
DFG	67	1.36	15
CSIR - India	53	0.66	4
DOE	51	1.33	14
CNPq	49	0.65	4
Department of Science & Technology (India)	46	0.69	2
Government of India	41	0.78	2
University Grants Commission, India	40	0.65	3
EU	39	1.88	21
MIUR	36	1.44	22
Ministry of Education, Culture, Sports, Science and Technology, Japan	35	0.66	3
DBT	30	0.72	0
CAPES	29	0.74	7
TUBITAK	29	0.68	3
Portuguese Foundation for Science and Technology	28	1.19	11
MICINN	24	1.37	17
Swiss National Science Foundation	24	1.19	4
National Basic Research Program of China	24	0.63	4
Swedish Research Council	23	2.01	22

Funding Organization	Number of Papers	Category-normalized citation impact	% Highly-cited Papers
Natural Sciences and Engineering Research Council of Canada	16	1.31	13
Canadian Institutes of Health Research	15	1.99	27
Wellcome Trust	6	4.63	50

6.2.3 Molecular Biology & Genetics

The count of papers and field-normalized bibliometrics for the funding organizations that supported research in Molecular Biology & Genetics are presented in Figure 9 and Table 13. High-impact research in Molecular Biology & Genetics was strongly supported by the Wellcome Trust, NIH, and DOE. These funding organizations ranked among the Top 10 funding organizations based on the count of citing papers and had a mean Category-normalized citation impact greater than the world average. The 16 funding organizations of the Top 25 with high activity in Molecular Biology & Genetics supported 64% of the citing papers this field (350 papers out of the 550 citing papers).

Figure 9 Performance of citing papers in ESI category Molecular Biology & Genetics by funding organization (2008-2016). Top 10 funding organizations based on the count of citing papers. The label provides the Funding Organization (% Highly-cited papers). This analysis focused on the Top 3 ESI of the Top 25 funding organizations.



- National Institutes of Health (NIH) - USA
- United States Department of Energy (DOE)
- Biotechnology and Biological Sciences Research Council (BBSRC)
- European Union (EU)
- German Research Foundation (DFG)
- Wellcome Trust
- Natural Sciences and Engineering Research Council of Canada (NSERC)
- Swedish Research Council (SRC)
- Canadian Institutes of Health Research (CIHR)
- Department of Biotechnology (DBT)

Table 13 Performance of citing papers in ESI category Molecular Biology & Genetics by funding organization (2008-2016). The 16 funding organizations among the Top 25 funding organizations that supported research in Molecular Biology & Genetics. This analysis focused on the Top 3 ESI of the Top 25 funding organizations.

Funding Organization	Number of Papers	Category-normalized citation impact	% Highly-cited Papers
NIH - USA	212	2.95	13
DOE	62	4.83	15
EU	27	2.4	15
BBSRC	27	2.32	19
DFG	26	0.73	4
Wellcome Trust	24	11.49	29
Natural Sciences and Engineering Research Council of Canada	19	1.6	11
Swedish Research Council	16	0.73	13
Canadian Institutes of Health Research	15	0.81	7
Swiss National Science Foundation	12	1.4	17
Ministry of Education, Culture, Sports, Science and Technology, Japan	12	0.37	0
DBT	12	0.31	0
MICINN	11	2.56	9
TUBITAK	11	1.01	9

Funding Organization	Number of Papers	Category-normalized citation impact	% Highly-cited Papers
National Basic Research Program of China	11	0.52	0
Portuguese Foundation for Science and Technology	10	1.25	20

6.2.4 Computer Science

The count of papers and the field-normalized bibliometrics for the funding organizations that supported research in Computer Science are presented in Figure 10 and Table 14. High-impact research in Computer Science was strongly supported by the NSF, National Natural Science Foundation of China, BBSRC, and the Government of India. All four funding organizations had a mean Category-normalized citation impact greater than the world average. About 35% of the citing papers in Computer Science were funded by four organizations among the Top 25 funding organizations (163 citing papers out of 472 papers).

Figure 10 Performance of citing papers in ESI category Computer Science by funding organization (2008-2016). The four organizations among the Top 25 funding organizations that supported research in Computer Science. The label provides the Funding Organization (% Highly-cited papers). This analysis focused on the Top 3 ESI of the Top 25 funding organizations.

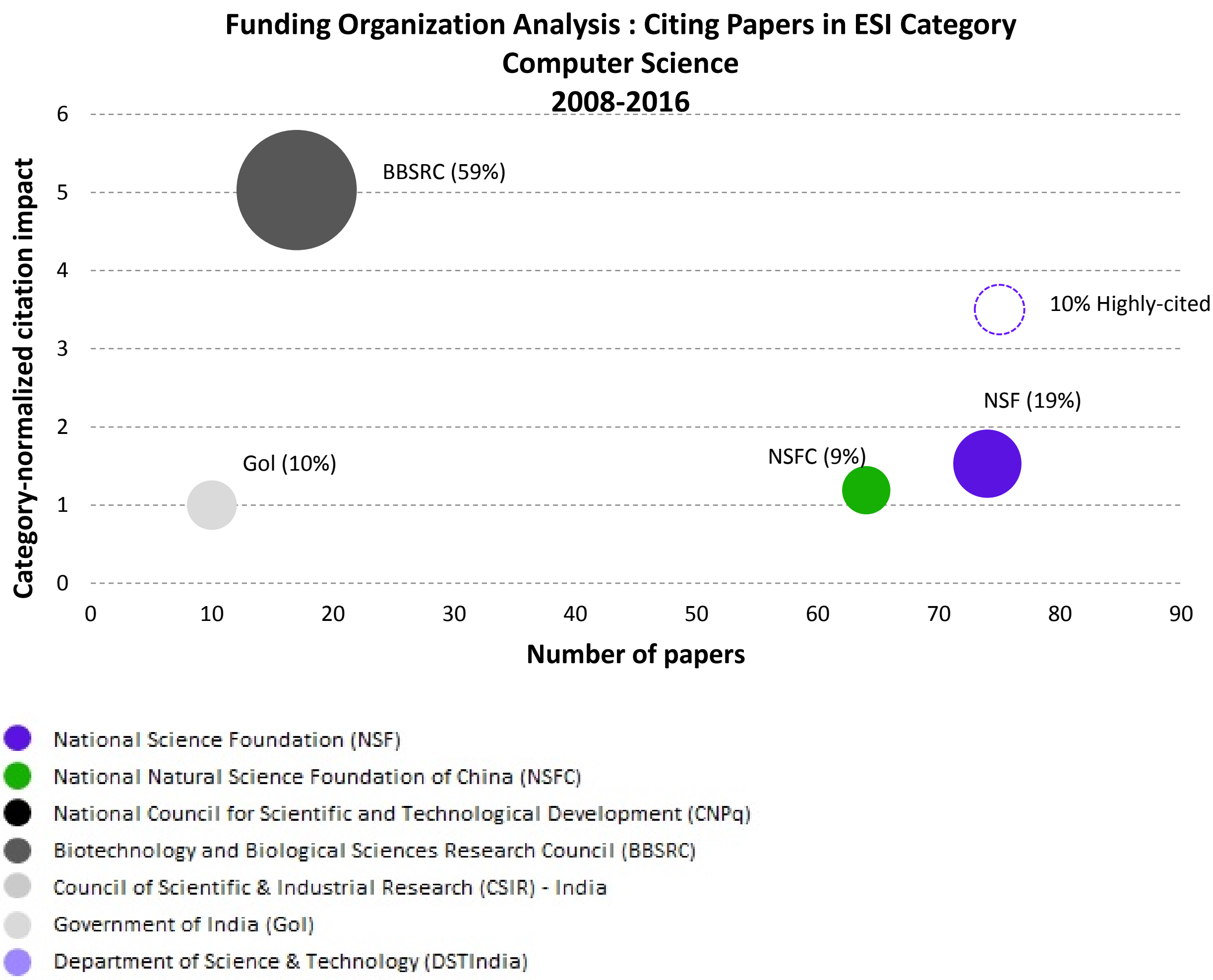


Table 14 Performance of citing papers in ESI category Computer Science by funding organization (2008-2016). The four funding organizations among the Top 25 funding organizations that supported research in Computer Science. This analysis focused on the Top 3 ESI of the Top 25 funding organizations.

Funding Organization	Number of Papers	Category-normalized citation impact	% Highly-cited Papers
Biotechnology and Biological Sciences Research Council (BBSRC)	17	5.03	59
National Science Foundation (NSF)	74	1.53	19
Government of India (GoI)	10	1	10
National Natural Science Foundation of China (NSFC)	64	1.19	9
National Council for Scientific and Technological Development (CNPq)	21	0.56	0
Council of Scientific & Industrial Research (CSIR) - India	15	0.77	0
Department of Science & Technology (DSTIndia)	8	0.72	0

6.2.5 Pharmacology & Toxicology and Clinical Medicine

The count of papers and field-normalized bibliometrics for the funding organizations supporting research in Pharmacology & Toxicology and Clinical Medicine are presented in Table 15. The citing papers in Pharmacology & Toxicology funding by MIUR performed at the world average while the other citing papers in this category and Clinical Medicine performed below the world average. About 8% of the citing papers in Pharmacology & Toxicology were supported by the two organizations among the Top 25 funding organizations and only about 4% of the citing papers in Clinical Medicine were supported by one of the Top 25 funding organizations[5].

Table 15 Performance of citing papers in ESI categories Pharmacology & Toxicology and Clinical Medicine by funding organization (2008-2016). The funding organizations among the Top 25 funding organizations that supported research in Pharmacology & Toxicology and Clinical Medicine. This analysis focused on the Top 3 ESI of the Top 25 funding organizations.

Funding Organization	ESI Category	Number of Papers	Category-normalized citation impact	% Highly-cited Papers
CAPES	Pharmacology & Toxicology	12	0.81	0
MIUR	Pharmacology & Toxicology	9	0.96	11
University Grants Commission, India	Clinical Medicine	3	0.28	0

6.3 Performance of citing papers supported by NIH, NSF and DOE by ESI category

The count of distinct papers and the Category-normalized citation impact of citing papers supported by NIH, NSF and DOE for all ESI categories are presented in Figure 11, Figure 12, and Table 16. The relative ranking of the number of citing papers by ESI category is similar to the general set of citing papers as presented in Table 4. Although the quantity of citing papers was relatively low for many categories (Figure 12), the quality was high. For ESI categories, except Engineering and Neuroscience &

Behavior, the citing papers supported by these USA federal agencies out-performed the world average (Figure 13). These results can encourage funders that future funding support to the lower volume categories (e.g., Computer Science, Physics, Plant & Animal Sciences and Clinical Medicine) could result in high-impact research.

The percentage of papers that were supported by NIH, NSF, and DOE in each ESI category is also presented Table 16. The NIH, NSF, and DOE supported a wide-breadth of research citing the Protein Data Bank. On average, these USA federal agencies supported 33% of the research across 17 research areas.

Figure 11 Number of citing papers supported by NIH, NSF and DOE by ESI category.

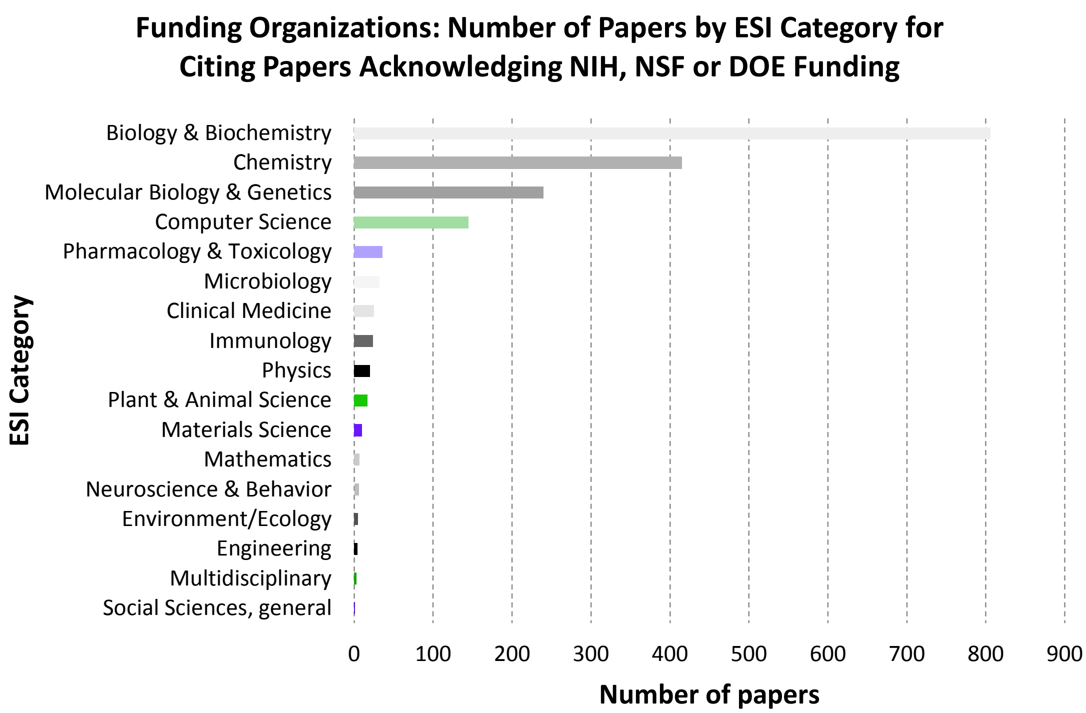
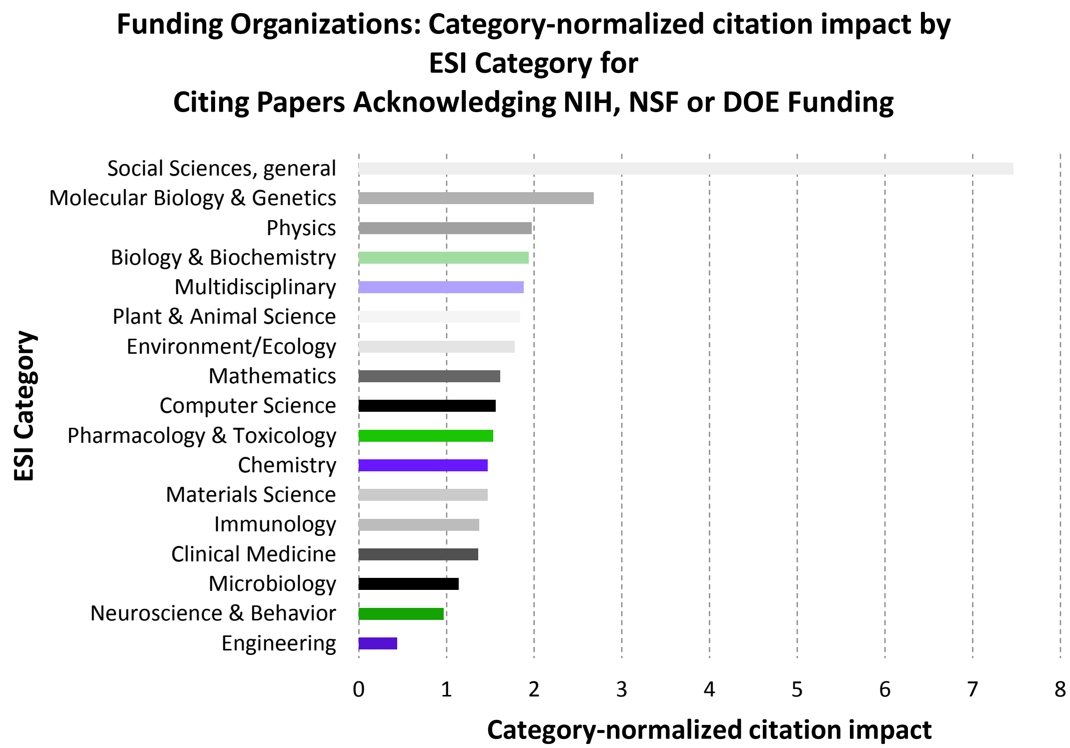


Figure 12 Category-normalized citation impact for citing papers supported by NIH, NSF and DOE by ESI category.



The NIH, NSF and DOE were not acknowledged by any citing papers in the ESI categories Agricultural Sciences, Economics & Business, Psychiatry/Psychology, or Geosciences.

Table 16 Performance of citing papers supported by NIH, NSF, and DOE funding by ESI category.

ESI Category	Number of Papers	Category-normalized citation impact	Percentage of All Citing Papers
Biology & Biochemistry	806	1.94	38%
Chemistry	415	1.47	31%
Molecular Biology & Genetics	240	2.68	44%
Computer Science	145	1.56	31%
Pharmacology & Toxicology	36	1.53	14%
Microbiology	32	1.14	30%
Clinical Medicine	25	1.36	32%
Immunology	24	1.37	39%
Physics	20	1.97	36%
Plant & Animal Science	17	1.84	26%
Materials Science	10	1.47	43%
Mathematics	7	1.61	39%
Neuroscience & Behavior	6	0.97	32%
Environment/Ecology	5	1.78	50%
Engineering	4	0.44	16%

ESI Category	Number of Papers	Category-normalized citation impact	Percentage of All Citing Papers
Multidisciplinary	3	1.88	18%
Social Sciences, general	1	7.47	50%

Table 17 provides the non-distinct count as well as field-normalized bibliometrics for those citing papers that acknowledged funding support from NIH, NSF, DOE and HHMI. The performance of citing papers supported by these four organizations was well above the world-average.

Table 17 Performance of citing papers for four US funding organizations.

Funding Organization	Number of Papers	Category-normalized citation impact	% Highly-cited papers
National Institutes of Health (NIH) - USA	1,457	1.99	19
National Science Foundation (NSF)	673	1.62	15
United States Department of Energy (DOE)	234	2.63	17
Howard Hughes Medical Institute (HHMI)	48	2.66	17

6.4 Funding Organization Review

Research citing the Berman paper acknowledged funding from government agencies throughout the world. All of the Top 25 funding organizations had Biology & Biochemistry and Chemistry as one of their Top 3 supported research areas. Fourteen of the Top 15 supported research papers that performed about the world-average as indicated by the Category-normalized citation impact.

Thirty-four percent of the citing research acknowledged funding from the USA Federal Agencies NIH, NSF, and DOE. These USA federal agencies were among the top 10 highest impact organizations and supported citing research in 17 out of the 22 ESI subject categories.

Appendix A Bibliometrics and Citation Analysis

Bibliometrics are about publications and their citations. The academic field emerged from 'information science' and now usually refers to the methods used to study and index texts and information.

Publications cite other publications. These citation links grow into networks, and their numbers are likely to be related to the significance or impact of the publication. The meaning of the publication is determined from keywords and content. Citation analysis and content analysis have therefore become a common part of bibliometric methodology. Historically, bibliometric methods were used to trace relationships amongst academic journal citations. Now, bibliometrics are important in indexing research performance.

Bibliometric data have particular characteristics of which the user should be aware, and these are considered here.

Journal papers (publications, sources) report research work. Papers refer to or 'cite' earlier work relevant to the material being reported. New papers are cited in their turn. Papers that accumulate more citations are thought of as having greater 'impact', which is interpreted as significance or influence on their field. Citation counts are therefore recognized as a measure of impact, which can be used to index the excellence of the research from a particular group, institution or country.

The origins of citation analysis as a tool that could be applied to research performance can be traced to the mid-1950s, when Eugene Garfield proposed the concept of citation indexing and introduced the Science Citation Index, the Social Sciences Citation Index and the Arts & Humanities Citation Index, produced by the Institute of Scientific Information.

We can count citations, but they are only 'indicators' of impact or quality – not metrics. Most impact indicators use average citation counts from groups of papers, because some individual papers may have unusual or misleading citation profiles. These outliers are diluted in larger samples.

Appendix B Data Source

The data used in this report came from the Clarivate Analytics Web of Science databases which give access not only to journals but also to conference proceedings, books, patents, websites, and chemical structures, compounds and reactions. Web of Science has a unified structure that integrates all data and search terms together and therefore provides a level of comparability not found in other databases. It is widely acknowledged to be the world's leading source of citation and bibliometric data. The Web of Science focuses on research published in journals, conferences and books in science, medicine, arts, humanities and social sciences.

The Web of Science was originally created as an awareness and information retrieval tool but it has acquired an important primary use as a tool for research evaluation, using citation analysis and bibliometrics. Data coverage is both current and retrospective in the sciences, social sciences, arts and humanities, in some cases back to 1900. Within the research community this data source was previously referred to by the acronym 'ISI'.

Unlike other databases, the Web of Science and underlying databases are selective, that is: the journals abstracted are selected using rigorous editorial and quality criteria. The authoritative, multidisciplinary content covers over 12,000 of the highest impact journals worldwide, including Open Access journals, and over 150,000 conference proceedings. The abstracted journals encompass the majority of significant, frequently cited scientific reports and, more importantly, an even greater proportion of the scientific research output which is cited. This selective process ensures that the citation counts remain relatively stable in given research fields and do not fluctuate unduly from year to year, which increases the usability of such data for performance evaluation.

Clarivate Analytics has extensive experience with databases on research inputs, activities and outputs and has developed innovative analytical approaches for benchmarking and interpreting international, national and institutional research impact.

Appendix C Database Categories

The source data can be grouped in various classification systems. Most of these are based on groups of journals that have a relatively high cross-citation linkage and naturally cluster together. Custom classifications use subject maps in third-party data such as the OECD categories set out in the Frascati manual.

Clarivate Analytics frequently uses the broader field categories in the Essential Science Indicators system and the finer journal categories in the Web of Science. There are 22 fields in Essential Science Indicators and 254 fields in Web of Science. In either case, our bibliometric analyses draw on the full range of data available in the underlying database, so analyses in our reports will differ slightly from anything created ‘on the fly’ from data in the web interface.

Most analyses start with an overall view across the data, then move to a view across broad categories and only then focus in at a finer level in the areas of greatest interest to policy, program or institutional purpose.

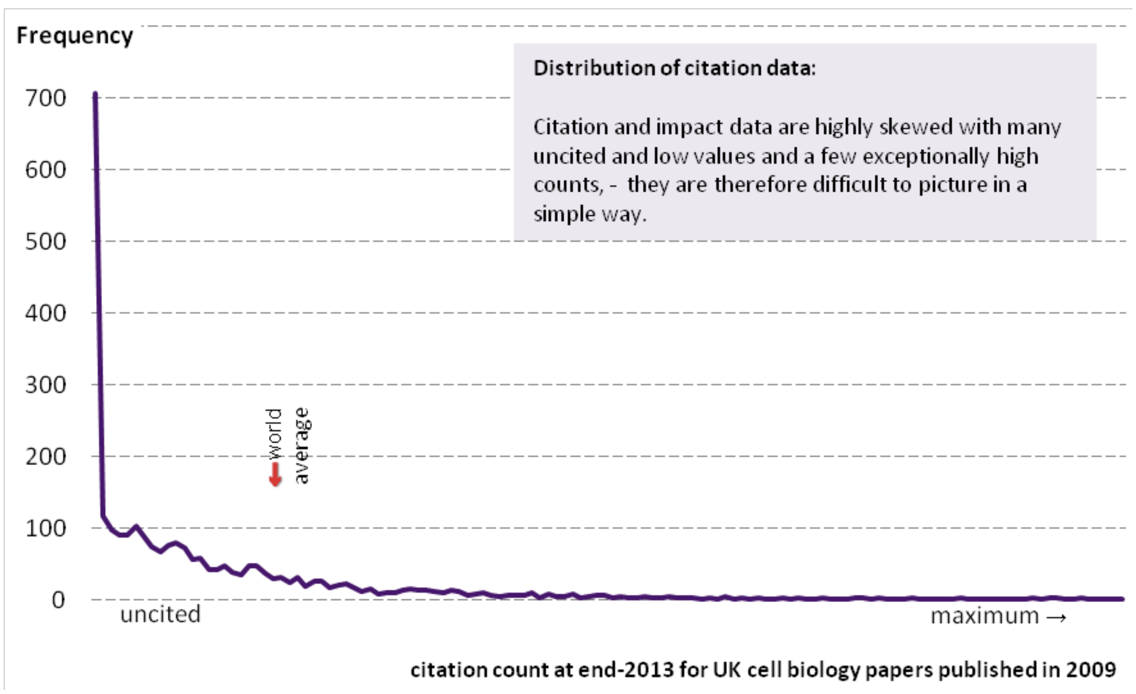
Appendix D Citation Counts

A publication accumulates citation counts when it is referred to by more recent publications. Some papers get cited frequently and many get cited rarely or never, so the distribution of citations is highly skewed.

Why are many papers never cited? Certainly some papers remain uncited because their content is of little or no impact, but that is not the only reason. It might be because they have been published in a journal not read by researchers to whom the paper might be interesting. It might be that they represent important but ‘negative’ work reporting a blind alley to be avoided by others. The publication may be a commentary in an editorial, rather than a normal journal article and thus of general rather than research interest, or it might be that the work is a ‘sleeping beauty’ that has yet to be recognized for its significance.

Other papers can be very highly cited: hundreds, even thousands of times. Again, there are multiple reasons for this. Most frequently cited work is being recognized for its innovative significance and impact on the research field of which it speaks. Impact here is a good reflection of quality: it is an indicator of excellence. But there are other papers which are frequently cited because their significance is slightly different: they describe key methodology; they are a thoughtful and wide-ranging review of a field; or they represent contentious views which others seek to refute.

Citation analysis cannot make value judgments about why an article is uncited nor about why it is highly cited. The analysis can only report the citation impact that the publication has achieved. We normally assume, based on many other studies linking bibliometric and peer judgments, that high citation counts correlate on average with the quality of the research.



The figure shows the skewed distribution of more or less frequently cited papers from a sample of UK authored publications in cell biology. The skew in the distribution varies from field to field. It is to compensate for such factors that actual citation counts must be normalized, or rebased, against a world baseline.

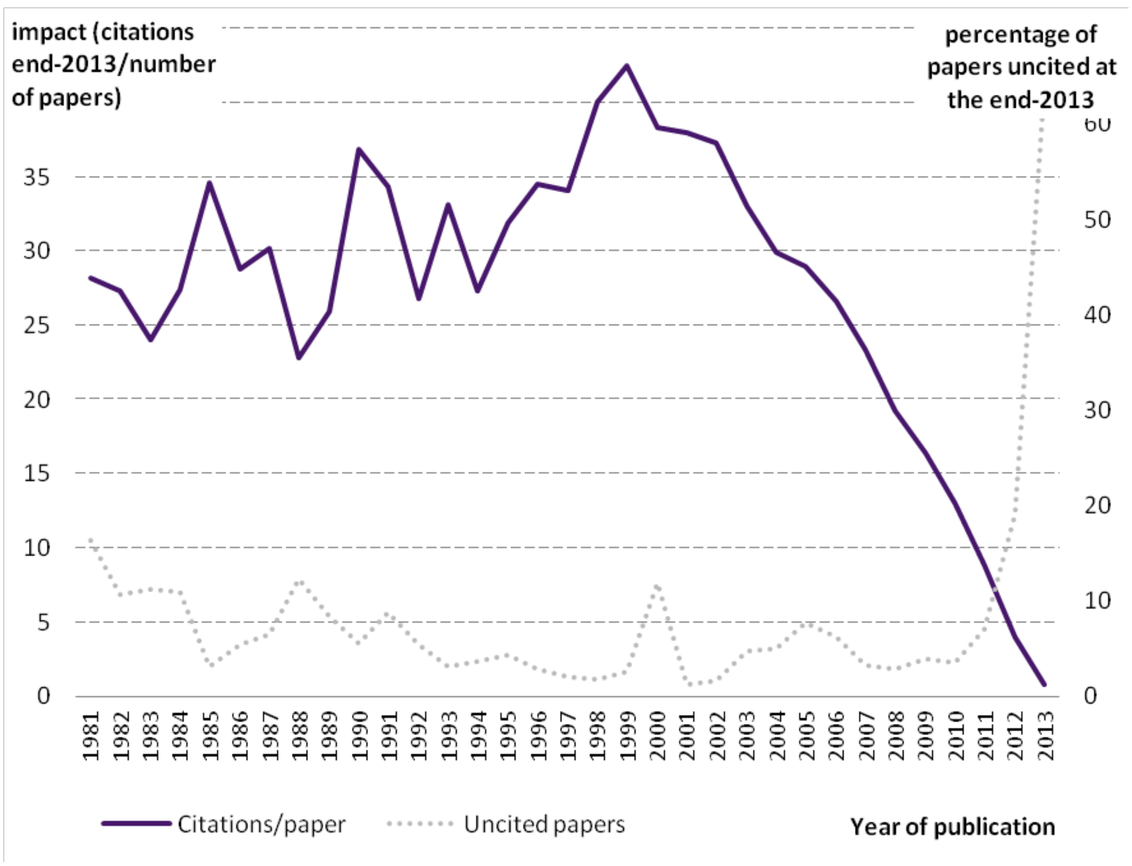
We do not seek to account separately for the effect of self-citation. If the citation count is significantly affected by self-citation then the paper is likely to have been infrequently cited. This is therefore only of consequence for low impact activity. Studies show that for large samples at national and institutional level, the effect of self-citation has little or no effect on the analytical outcomes and would not alter interpretation of the results.

Appendix E Time Factors

Citations accumulate over time. Older papers therefore have, on average, more citations than more recent work. The graph below shows the pattern of citation accumulation for a set of 33 journals in the journal category ***Materials Science, Biomaterials***. Papers less than eight years old are, on average, still accumulating additional citations. The citation count goes on to reach a plateau for older sources.

The graph shows that the percentage of papers that have never been cited drops over about five years. Beyond five years, between 5% and 10% or more of papers remain uncited.

Account must be taken of these time factors in comparing current research with historical patterns. For these reasons, it is sometimes more appropriate to use a fixed five-year window of papers and citations to compare two periods than to look at the longer term profile of citations and of uncitedness for a recent year and an historical year.



Appendix F Discipline Factors

Citation rates vary between disciplines and fields. For the UK science base as a whole, ten years produces a general plateau beyond which few additional citations would be expected. On the whole, citations accumulate more rapidly and plateau at a higher level in biological sciences than physical sciences, and natural sciences generally cite at a higher rate than social sciences.

Papers are assigned to disciplines (journal categories or research fields) by Clarivate Analytics, bringing cognate research areas together. Before 2007, journals were assigned to the older, well established Current Contents categories which were informed by extensive work by Thomson and with the research community since the early 1960s. This scheme has been superseded by the 252 Web of Science journal categories which allow for greater disaggregation for the growing volume of research which is published and abstracted.

Papers are allocated according to the journal in which the paper is published. Some journals may be considered to be part of the publication record for more than one research field. As the example below illustrates, the journal *Acta Biomaterialia* is assigned to two journal categories: ***Materials Science, Biomaterials and Engineering, Biomedical***.

Very few papers are not assigned to any research field and as such will not be included in specific analyses using normalized citation impact data. The journals included in the Clarivate Analytics databases and how they are selected are detailed here http://scientific.thomsonreuters.com/mjl/.

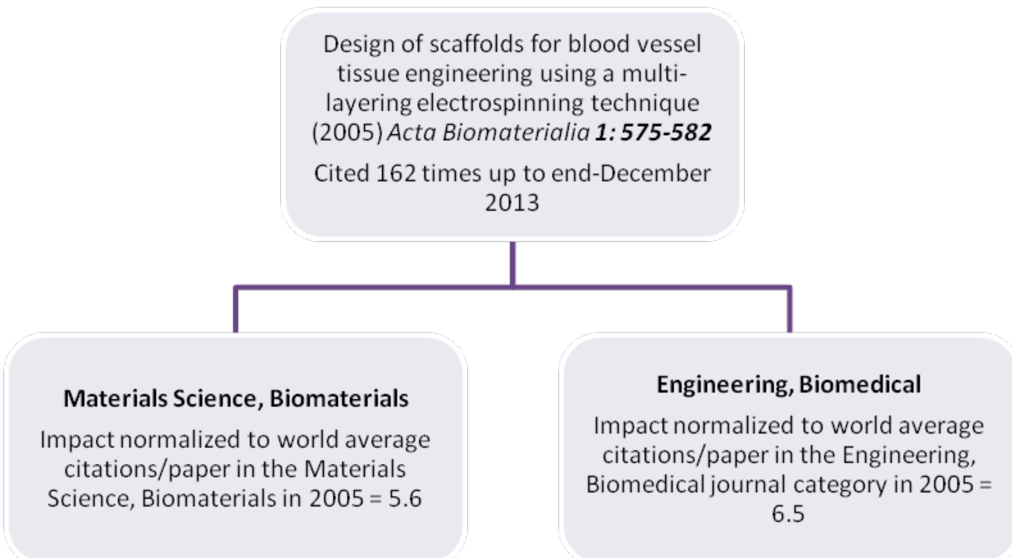
Some journals with a very diverse content, including the prestigious journals *Nature* and *Science* were classified as ***Multidisciplinary*** in databases created prior to 2007. The papers from these ***Multidisciplinary*** journals are now re-assigned to more specific research fields using an algorithm based on the research area(s) of the references cited by the article.

Appendix G Normalized Citation Impact

Because citations accumulate over time at a rate that is dependent upon the field of research, all analyses must take both field and year into account. In other words, because the absolute citation count for a specific article is influenced by its field and by the year it was published, we can only make comparisons of indexed data after normalizing with reference to these two variables.

We only use citation counts for reviews and articles in calculations of impact, because document type influences the citation count. For example, a review will often be cited more frequently than an article in the same field, but editorials and meeting abstracts are rarely cited and citation rates for conference proceedings are extremely variable. The most common normalization factors are the average citations per paper for (1) the year and (2) either the field or the journal in which the paper was published. This normalization is also referred to as 'rebasing' the citation count.

Impact is therefore most commonly analyzed in terms of 'normalized citation impact', or nci. The following schematic illustrates how the nci is calculated at the paper level and journal category level.



This article in the journal *Acta Biomaterialia* is assigned to two journal categories: ***Materials Science, Biomaterials and Engineering, Biomedical***. The world average baselines for, as an example, ***Materials science, Biomaterials*** are calculated by summing the citations to all the articles and reviews published worldwide in the journal *Acta Biomaterialia* and the other 32 journals assigned to this category for each year, and dividing this by the total number of articles and reviews published in the journal category. This gives the category-specific or nci_F (in the above example the category-specific nci_F for ***Materials Science, Biomaterials*** is 5.6 and the category-specific nci_F for Engineering, Biomedical is higher at 6.5). Most papers (nearly two-thirds) are assigned to a single journal category while a minority of them are assigned to more than five.

World average impact data are sourced from the Clarivate Analytics National Science Indicators baseline data for 2015.

Appendix H Mean Normalized Citation Impact

Research performance has historically been indexed by using average citation impact, usually compared to a world average that accounts for time and discipline. As noted, however, the distribution of citations amongst papers is highly skewed because many papers are never cited while a few papers accumulate very large citation counts. That means that an average may be misleading if assumptions are made about the distribution of the underlying data.

In fact, almost all research activity metrics are skewed: for research income, PhD numbers and publications there are many low activity values and a few exceptionally high values. In reality, therefore, the skewed distribution means that average impact tends to be greater than and often significantly different from either the median or mode in the distribution. This should be borne in mind when reviewing analytical outcomes.

Appendix I What are Uncited Papers?

It may be a surprise that some journal papers are never subsequently cited after publication, even by their authors. This accounts for about half the total global output for a typical, recent 10-year period.

There is variation in non-citation between countries and between fields. For example, relatively more engineering papers tend to remain uncited than papers in other sciences, indicative of a disciplinary factor but not a quality factor. While there is also an obvious increase in the likelihood of citation over time, most papers that are going to be cited will be cited within a few years of publication.

Appendix J What is the Threshold for "Highly Cited"?

Clarivate Analytics has traditionally used the term "Highly Cited Paper" to refer to the world's 1% of most frequently cited papers, taking into account year of publication and field. In rough terms, UK papers cited more than eight times as often as relevant world average would fall into the Clarivate Analytics Highly cited category. About 1-2% of papers (all papers, cited or uncited) typically pass this hurdle. Such a threshold certainly delimits exceptional papers for international comparisons but, in practice, is an onerous marker for more general management purposes.

After reviewing the outcomes of a number of analyses, we have chosen a more relaxed definition for our descriptive and analytical work. We deem papers that are in the world's top 10% of most frequently cited papers, taking into account year of publication and field, to be relatively highly cited for national comparisons.