The top 100 papers

The discovery of high-temperature superconductors, the determination of DNA's double-helix structure, the first observations that the expansion of the Universe is accelerating — all of these breakthroughs won Nobel prizes and international acclaim. Yet none of the papers that announced them comes anywhere close to ranking among the 100 most highly cited papers of all time.

Citations, in which one paper refers to earlier works, are the standard means by which authors acknowledge the source of their methods, ideas and findings, and are often used as a rough measure of a paper's importance. Fifty years ago, Eugene Garfield published the Science Citation Index (SCI), the first systematic effort to track citations in the scientific literature. To mark the anniversary, *Nature* asked Thomson Reuters, which now owns the SCI, to list the 100 most highly cited papers of all time. (See the full list at Web of Science Top 100.xls or the interactive graphic, below.) The search covered all of Thomson Reuter's Web of Science, an online version of the SCI that also includes databases covering the social sciences, arts and humanities, conference proceedings and some books. It lists papers published from 1900 to the present day.

The exercise revealed some surprises, not least that it takes a staggering 12,119 citations to rank in the top 100 — and that many of the world's most famous papers do not make the cut. A few that do, such as the first observation of carbon nanotubes (number 36) are indeed classic discoveries. But the vast majority describe experimental methods or software that have become essential in their fields.

The most cited work in history, for example, is a 1951 paper describing an assay to determine the amount of protein in a solution. It has now gathered more than 305,000 citations — a recognition that always puzzled its lead author, the late US biochemist Oliver Lowry. "Although I really know it is not a great paper … I secretly get a kick out of the response," he wrote in 1977.

The colossal size of the scholarly literature means that the top-100 papers are extreme outliers. Thomson Reuter's Web of Science holds some 58 million items. If that corpus were scaled to Mount Kilimanjaro, then the 100 most-cited papers would represent just 1 centimetre at the peak. Only 14,499 papers — roughly a metre and a half's worth — have more than 1,000 citations (see 'The paper mountain'). Meanwhile, the foothills comprise works that have been cited only once, if at all — a group that encompasses roughly half of the items.

Nobody fully understands what distinguishes the sliver at the top from papers that are merely very well known — but researchers' customs explain some of it. Paul Wouters, director of the Centre for Science and Technology Studies in Leiden, the Netherlands, says that many methods papers "become a standard reference that one cites in order to make clear to other scientists what kind of work one is doing". Another common practice in science ensures that truly foundational discoveries — Einstein's special theory of relativity, for instance — get fewer citations than they might deserve: they are so important that they quickly enter the textbooks or are incorporated into the main text of papers as

terms deemed so familiar that they do not need a citation.

Citation counts are riddled with other confounding factors. The volume of citations has increased, for example — yet older papers have had more time to accrue citations. Biologists tend to cite one another's work more frequently than, say, physicists. And not all fields produce the same number of publications. Modern bibliometricians therefore recoil from methods as crude as simply counting citations when they want to measure a paper's value: instead, they prefer to compare counts for papers of similar age, and in comparable fields.

Nor is Thomson Reuters' list the only ranking system available. Google Scholar compiled its own top-100 list for *Nature* (see 'An alternative ranking'). It is based on many more citations because the search engine culls references from a much greater (although poorly characterized) literature base, including from a large range of books. In that list, available at Google Scholar Top 100.xls, economics papers have more prominence. Google Scholar's list also features books, which Thomson Reuters did not analyse. But among the science papers, many of the same titles show up.

Yet even with all the caveats, the old-fashioned hall of fame still has value. If nothing else, it serves as a reminder of the nature of scientific knowledge. To make exciting advances, researchers rely on relatively unsung papers to describe experimental methods, databases and software.

Here *Nature* tours some of the key methods that tens of thousands of citations have hoisted to the top of science's Kilimanjaro — essential, but rarely thrust into the limelight.

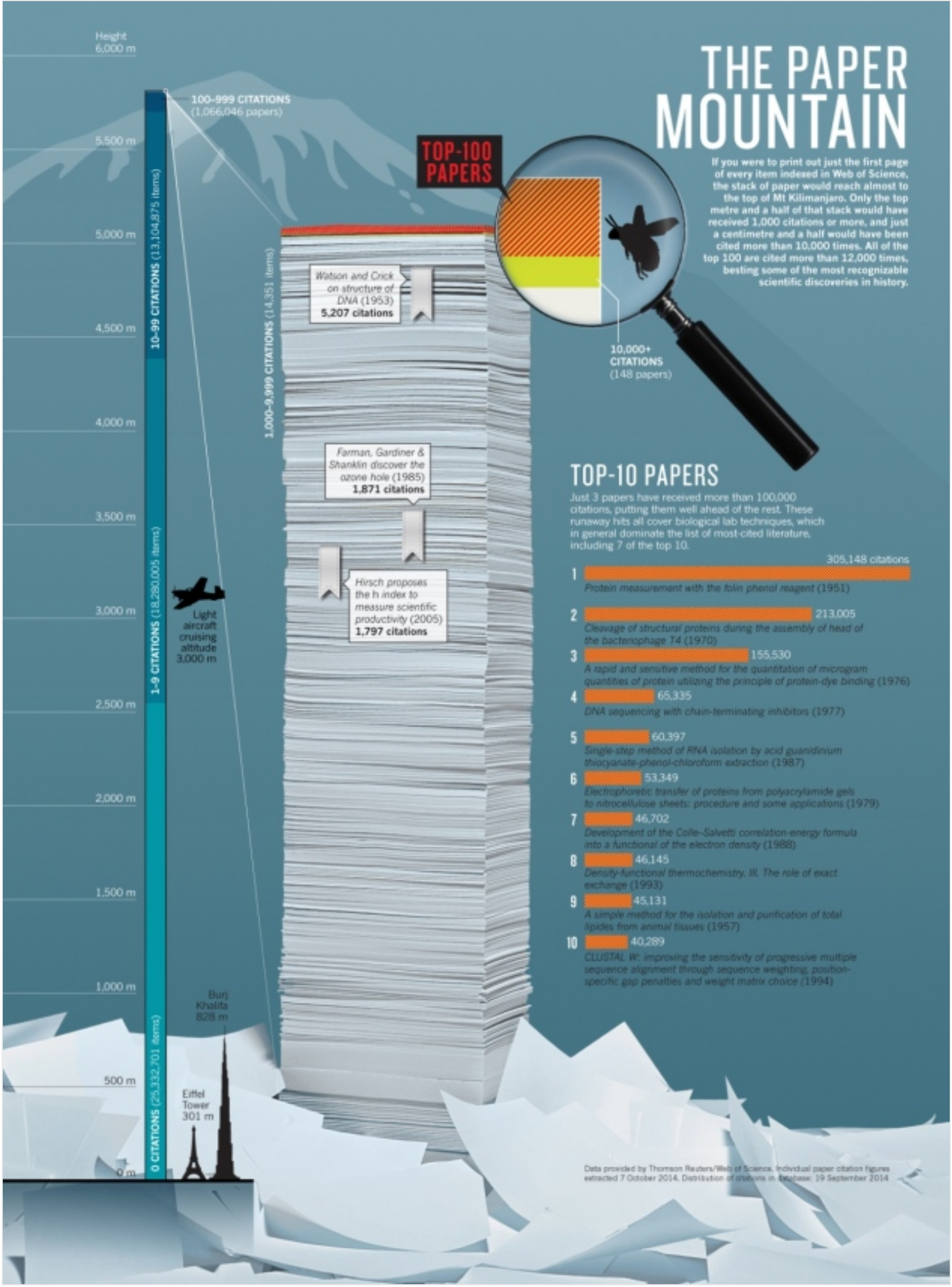

Photo by Kyle Bean; Design by Wesley Fernandes/nature

Biological techniques

For decades, the top-100 list has been dominated by protein biochemistry. The 1951 paper[2] describing the Lowry method for quantifying protein remains practically unreachable at number 1, even though many biochemists say that it and the competing Bradford assay — described by paper number 3 on the list — are a tad outdated. In between, at number 2, is Laemmli buffer, which is used in a different kind of protein analysis. The dominance of these techniques is attributable to the high volume of citations in cell and molecular biology, where they remain indispensable tools.

At least two of the biological techniques described by top-100 papers have resulted in Nobel prizes. Number 4 on the list describes the DNA-sequencing method that earned the late Frederick Sanger his share of the 1980 Nobel Prize in Chemistry. Number 63 describes polymerase chain reaction (PCR), a method for copying segments of DNA that earned US biochemist Kary Mullis the prize in 1993. By helping scientists to explore and manipulate DNA, both methods have helped to drive a revolution in genetic research that continues to this day.

Other methods have received less public acclaim, but are not without their rewards. In the 1980s, the Italian cancer geneticist Nicoletta Sacchi linked up with Polish molecular biologist Piotr Chomczynski in the United States to publish a fast, inexpensive way to extract RNA from a biological sample. As it became wildly popular — currently, it is number 5 on the list — Chomczynski patented modifications on the technique and built a business out of selling the reagents. Now at the Roswell Park Cancer Institute in Buffalo, New York, Sacchi says that she received little in the way of monetary rewards, but takes satisfaction from seeing great discoveries built on her work. The technique played a part in the explosive growth in the study of short RNA molecules that do not code for protein, for example. “That is what I would consider, scientifically speaking, a great reward,” she says.

Bioinformatics

The rapid expansion of genetic sequencing since Sanger’s contribution has helped to boost the ranking of papers describing ways to analyse the sequences. A prime example is BLAST (Basic Local Alignment Search Tool), which for two decades has been a household name for biologists wanting to work out what genes and proteins do. Users simply have to open the program in a web browser and plug in a DNA, RNA or protein sequence. Within seconds, they will be shown related sequences from thousands of organisms — along with information about the function of those sequences and even links to relevant literature. So popular is BLAST that versions[8, 9] of the program feature twice on the list, at spots 12 and 14.

But owing to the vagaries of citation habits, BLAST has been bumped down the list by Clustal, a complementary programme for aligning multiple sequences at once. Clustal allows researchers to describe the evolutionary relationships between sequences from different organisms, to find matches among seemingly unrelated sequences and to predict how a change at a specific point in a gene or protein might affect its function. A 1994 paper describing ClustalW, a user-friendly version of the

software, is currently number 10 on the list. A 1997 paper on a later version called ClustalX is number 28.

The team that developed ClustalW, at the European Molecular Biology Laboratory in Heidelberg, Germany, had created the program to work on a personal computer, rather than a mainframe. But the software was transformed when Julie Thompson, a computer scientist from the private sector, joined the lab in 1991. “It was a program written by biologists; I’m trying to find a nice way to say that,” says Thompson, who is now at the Institute of Genetics and Molecular and Cellular Biology in Strasbourg, France. Thompson rewrote the program to ready it for the volume and complexity of the genome data being generated at the time, while also making it easier to use.

The teams behind BLAST and Clustal are competitive about the ranking of their papers. It is a friendly sort of competition, however, says Des Higgins, a biologist at University College Dublin, and a member of the Clustal team. “BLAST was a game-changer, and they’ve earned every citation that they get.”

Phylogenetics

Another field buoyed by the growth in genome sequencing is phylogenetics, the study of evolutionary relationships between species.

Number 20 on the list is a paper that introduced the “neighbor-joining” method, a fast, efficient way of placing a large number of organisms into a phylogenetic tree according to some measure of evolutionary distance between them, such as genetic variation. It links related organisms together one pair at a time until a tree is resolved. Physical anthropologist Naruya Saitou helped to devise the technique when he joined Masatoshi Nei’s lab at the University of Texas in Houston in the 1980s to work on human evolution and molecular genetics, two fields that were starting to burst at the seams with information.

“We physical anthropologists were facing kind of the big data of that time,” says Saitou, now at Japan’s National Institute of Genetics in Mishima. The technique made it possible to devise trees from large data sets without eating up computer resources. (And, in a nice cross-fertilization within the top-100, Clustal’s algorithms use the same strategy.)

Number 41 on the list is a description of how to apply statistics to phylogenies. In 1984, evolutionary biologist Joe Felsenstein of the University of Washington in Seattle adapted a statistical tool known as the bootstrap to infer the accuracy of different parts of an evolutionary tree. The bootstrap involves resampling data from a set many times over, then using the variation in the resulting estimates to determine the confidence for individual branches. Although the paper was slow to amass citations, it rapidly grew in popularity in the 1990s and 2000s as molecular biologists recognized the need to attach such intervals to their predictions.

Felsenstein says that the concept of the bootstrap, devised in 1979 by Bradley Efron, a statistician at

Stanford University in California, was much more fundamental than his work. But applying the method to a biological problem means it is cited by a much larger pool of researchers. His high citation count is also a consequence of how busy he was at the time, he says: he crammed everything into one paper rather than publishing multiple papers on the topic, which might have diluted the number of citations each one received. “I was unable to go off and write four more papers on the same thing,” he says. “I was too swamped to do that, not too principled.”

Statistics

Although the top-100 list has a rich seam of papers on statistics, says Stephen Stigler, a statistician at the University of Chicago in Illinois and an expert on the history of the field, “these papers are not at all those that have been most important to us statisticians”. Rather, they are the ones that have proved to be most useful to the vastly larger population of practising scientists.

Much of this crossover success stems from the ever-expanding stream of data coming out of biomedical labs. For example, the most frequently cited statistics paper (number 11) is a 1958 publication by US statisticians Edward Kaplan and Paul Meier that helps researchers to find survival patterns for a population, such as participants in clinical trials. That introduced what is now known as the Kaplan–Meier estimate. The second (number 24) was British statistician David Cox’s 1972 paper that expanded these survival analyses to include factors such as gender and age.

The Kaplan–Meier paper was a sleeper hit, receiving almost no citations until computing power boomed in the 1970s, making the methods accessible to non-specialists. Simplicity and ease of use also boosted the popularity of papers in this field. British statisticians Martin Bland and Douglas Altman made the list (number 29) with a technique — now known as the Bland–Altman plot — for visualizing how well two measurement methods agree. The same idea had been introduced by another statistician 14 years earlier, but Bland and Altman presented it in an accessible way that has won citations ever since.

The oldest and youngest papers in the statistics group deal with the same problem — multiple comparisons of data — but from very different scientific milieux. US statistician David Duncan’s 1955 paper (number 64) is useful when a few groups need to be compared. But at number 59, Israeli statisticians Yoav Benjamini and Yosef Hochberg’s 1995 paper on controlling the false-discovery rate is ideally suited for data coming from fields such as genomics or neuroscience imaging, in which comparisons number in the hundreds of thousands — a scale that Duncan could hardly have imagined. As Efron observes: “The story is one of the computer slowly, then not so slowly, making its influence felt on statistical theory as well as on practice.”

An alternative ranking

The Web of Science is not the only index of citations available. Google Scholar has also generated a list

of the 'most-cited' articles of all time for *Nature* (Google Scholar Top 100.xls). Two-thirds of the entries are books, which Thomson Reuters did not include. "Folks have focused on journals, but there is this other world of books out there," says Anurag Acharya, a software engineer who leads the Google Scholar team in Mountain View, California. At number 4, the most-cited book is the manual *Molecular Cloning*, a mainstay of molecular-biology laboratories. But the list shows that research articles can be just as influential as books, notes Acharya. And at the top of both Google's and Thomson Reuters' rankings are the same three research articles — albeit in different order.

A separate Google Scholar top 100 showing only the top-cited research articles (Google Scholar Top 100 articles only.xls) throws up many similar papers to the Web of Science ranking. Noticeably, however, just over one-third of the list is different, with economics and psychology articles making considerable inroads, perhaps because they gain more citations from books than do other fields. Number 21, for example — a 1976 article on managerial behaviour in firms (M. C. Jensen & W. H. Meckling *J. Financ. Econ.* **3,** 305–360; 1976) — received 45,119 citations in Google's list, but just 8,372 in Web of Science. (Google gives most documents a higher number of citations than Web of Science, but a 5-fold difference is unusual). Highest of the Google Scholar new entrants, at number 4 in the list, is Claude Shannon's 1948 paper that birthed modern information theory (C. E. Shannon *Bell Syst. Tech. J* .**27,** 379–423; 1948). Google Scholar credits that with 69,273 citations, while the Web of Science gives it 10,239 citations — so just misses out on the top 100.

The Google Scholar top-10 list of articles with books interspersed

Google Scholar ranking (overall)	Times cited	Citation	Web of Science ranking	Times cited
1	223,131	Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227,** 680–685 (1970).	2	213,005
2	192,710	Lowry, O. H., Rosebrough, N. J., Farr, A. L. & Randall, R. J. Protein measurement with the folin phenol reagent. *J. Biol. Chem.* **193,** 265–275 (1951).	1	305,148
3	190,309	Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *J. Anal. Biochem.* **72,** 248–254 (1976).	3	155,530
*	172,540	Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning* (1989).		
*	110,822	Press, W. H. *Numerical Recipes: The Art of Scientific Computing* (1992).		
*	91,237	Yin, R. K. *Case Study Research: Design and Methods* (1984).		
*	73,818	Kuhn, T. S. *The Structure of Scientific Revolutions* (1962).		
*	70,807	Zar, J. H. *Biostatistical Analysis* (1974).		
4	69,273	Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27,** 379–423 (1948).	In top 150	10,239
*	67,824	Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (1969).		
*	64,956	Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and*		

		Machine Learning (1989).		
*	64,761	Glaser, B. G. & Strauss, A. L. *The Discovery of Grounded Theory: Strategies for Qualitative Research* (1967).		
5	64,031	Sanger. F., Nicklen, S. & Couslon, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74,** 5463–5467 (1977).	4	65,335
6	62,344	Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate phenol chloroform extraction. *J. Anal. Biochem.* **162,** 156–159 (1987).	5	60,397
*	61,929	Maniatis, T., Fritsch, E. F. & Sambrook, J. *Molecular Cloning: A Laboratory Manual* (1982).		
*	60,957	Nunnally, J. C., Bernstein, I. H. & Berge, J. M. F. T. *Psychometric Theory* (1967).		
*	58,915	Rogers, E. M. *Diffusion of Innovations* (1962).		
7	56,923	Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98,** 5648–5652 (1993).	8	46,145
8	54,365	Lee. C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37,** 785–789 (1988).	7	46,702
*	54,067	Porter, M. E. *Competitive Advantage: Creating and Sustaining Superior Performance* (1985).		
9	53,696	Murashige, T. & Skoog, F. A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiol. Plant.* **15,** 473–497 (1962).	15	36,132
10	53,423	Folstein, M. F., Folstein, S. E. & McHugh, P. R. Mini-mental state — practical method for grading cognitive state of patients for clinician. *J. Psychiatr. Res.* **12,** 189–198 (1975).	17	34,532

Density functional theory

When theorists want to model a piece of matter — be it a drug molecule or a slab of metal — they often use software to calculate the behaviour of the material's electrons. From this knowledge flows an understanding of numerous other properties: a protein's reactivity, for instance, or how easily Earth's liquid iron outer core conducts heat.

Most of this software is built on density functional theory (DFT), easily the most heavily cited concept in the physical sciences. Twelve papers on the top-100 list relate to it, including 2 of the top 10. At its heart, DFT is an approximation that makes impossible mathematics easy, says Feliciano Giustino, a materials physicist at the University of Oxford, UK. To study electronic behaviour in a silicon crystal by taking account of how every electron and every nucleus interacts with every other electron and nucleus, a researcher would need to analyse one sextillion (10^{21}) terabytes of data, he says — far beyond the capacity of any conceivable computer. DFT reduces the data requirement to just a few hundred kilobytes, well within the capacity of a standard laptop.

Theoretical physicist Walter Kohn led the development of DFT half a century ago in papers[20, 21] that now rank as numbers 34 and 39. Kohn realized that he could calculate a system's properties, such as its lowest energy state, by assuming that each electron reacts to all the others not as individuals, but as a

smeared-out average. In principle, the mathematics are straightforward: the system behaves like a continuous fluid with a density that varies from point to point. Hence the theory's name.

But a few decades passed before researchers found ways to implement the idea for real materials, says Giustino. Two[22, 23] top-100 papers are technical recipes on which the most popular DFT methods and software packages are built. One (number 8) is by Axel Becke, a theoretical chemist at Dalhousie University in Halifax, Canada, and the other (number 7) is by US-based theoretical chemists Chengteh Lee, Weitao Yang and Robert Parr. In 1992, computational chemist John Pople (who would share the 1998 Nobel prize with Kohn) included a form of DFT in his popular Gaussian software package.

Software users probably cite the original theoretical papers even if they do not fully understand the theory, says Becke. "The theory, mathematics and computer software are specialized and are the concern of quantum physicists and chemists," he says. "But the applications are endless. At a fundamental level, DFT can be used to describe all of chemistry, biochemistry, biology, nanosystems and materials. Everything in our terrestrial world depends on the motions of electrons — therefore, DFT literally underlies everything."

Crystallography

George Sheldrick, a chemist at the University of Göttingen in Germany, began to write software to help solve crystal structures in the 1970s. In those days, he says, "you couldn't get grant money for that kind of project. My job was to teach chemistry, and I wrote the programs as a hobby in my spare time." But over 40 years, his work gave rise to the regularly updated SHELX suite of computer programs, which has become one of the most popular tools for analysing the scattering patterns of X-rays that are shot through a crystal — thereby revealing the atomic structure.

The extent of that popularity became apparent after 2008, when Sheldrick published a review paper about the history of the system, and noted that it might serve as a general literature citation whenever any of the SHELX programs were used. Readers followed his advice. In the past 6 years, that review paper has amassed almost 38,000 citations, catapulting it to number 13 and making it the highest-ranked paper published in the past two decades.

The top-100 list is scattered with other tools essential to crystallography and structural biology. These include papers describing the HKL suite (number 23) for analysing X-ray diffraction data; the PROCHECK programs (number 71) used to analyse whether a proposed protein structure seems geometrically normal or outlandish; and two programs[27, 28] used to sketch molecular structures (numbers 82 and 95). These tools are the "bricks and mortar" for determining crystal structures, says Philip Bourne, associate director for data science at the US National Institutes of Health in Bethesda, Maryland.

An unusual entry, appearing at number 22, is a 1976 paper from Robert Shannon — a researcher at the

giant chemical firm DuPont in Wilmington, Delaware, who compiled a comprehensive list of the radii of ions in a series of different materials. Robin Grimes, a materials scientist at Imperial College London, says that physicists, chemists and theorists still cite this paper when they look up values of ionic size, which often correlate neatly with other properties of a substance. This has made it the highest formally-cited database of all time.

“We often cite these kinds of papers almost without thinking about it,” says Paul Fossati, one of Grimes’s research colleagues. The same could be said for many of the methods and databases in the top 100. The list reveals just how powerfully research has been affected by computation and the analysis of large data sets. But it also serves as a reminder that the position of any particular methods paper or database at the top of the citation charts is also down to luck and circumstance.

Still, there is one powerful lesson for researchers, notes Peter Moore, a chemist at Yale University in New Haven, Connecticut. “If citations are what you want,” he says, “devising a method that makes it possible for people to do the experiments they want at all, or more easily, will get you a lot further than, say, discovering the secret of the Universe”.