Atypical Combinations and Scientific Impact

Brian Uzzi [1, 2] Satyam Mukherjee [1, 2] Michael Stringer [2, 3] Ben Jones [1, 4, 5]

[1] Northwestern University
[2] Northwestern University Institute on Complex Systems (NICO)
[3] Datascope Analytics
[4] NBER
[5] Corresponding Author

August 26, 2013

Abstract

Novelty is an essential feature of creative ideas, yet building blocks of new ideas are often embodied in existing knowledge. From this perspective, balancing atypical knowledge with conventional knowledge may be critical to the link between innovativeness and impact. Our analysis of 17.9 million papers spanning all scientific fields suggests that science follows a nearly universal pattern: the highest-impact science is primarily grounded in exceptionally conventional combinations of prior work yet simultaneously features an intrusion of unusual combinations. Papers of this type were twice as likely to be highly cited works. Notably, novel combinations of prior work are rare, yet teams are 37.7% more likely than solo authors to insert novel combinations into familiar knowledge domains.

Scientific enterprises are increasingly concerned that research within narrow boundaries is unlikely to be the source of the most fruitful ideas (1). Models of creativity emphasize that innovation is spurred through original combinations that spark new insights (*2-10*). Current interest in team science and how scientists search for ideas is premised in part on the idea that teams can span scientific specialties, effectively combining knowledge that prompts scientific breakthroughs (*11-15*).

Yet the production and consumption of boundary-spanning ideas can also raise well-known challenges (*16-21*). If, as Einstein believed (*21*), individual scientists inevitably become narrower in their expertise as the body of scientific knowledge expands, then reaching effectively across boundaries may be increasingly challenging (*4*), especially given the difficulty of searching unfamiliar domains (*17-18*). Moreover, novel ideas can be difficult to absorb (*19*) and communicate, leading scientists to intentionally display conventionality. In his *Principia*, Newton presented his laws of gravitation using accepted geometry rather than his newly developed calculus, despite the latter's importance in developing his insights (*22*). Similarly, Darwin devoted the first part of the *Origin of Species* to conventional, well-accepted knowledge of the selective breeding of dogs, cattle, and birds. From this viewpoint, the balance between extending science with atypical combinations of knowledge while maintaining advantages of conventional domain-level thinking is critical to the link between innovativeness and impact. However, little is known about the composition of this balance or how scientists can achieve it.

In this study, we examined 17.9 million research articles in the Web of Science (WOS) to see how prior work is combined. We present facts that inform (i) the extent to which scientific papers reference novel versus conventional combinations of prior work, (ii) the relative impact of papers based on the combinations they draw upon, and (iii) how (i) and (ii) are associated with collaboration.

We considered pairwise combinations of references in the bibliography of each paper (*23*, *24*). We counted the frequency of each co-citation pair across all papers published that year in the WOS and compare these observed frequencies to those expected by chance, using randomized citation networks. In the randomized citation networks, all citation links between all papers in the WOS were switched by means of a Monte Carlo algorithm. The switching algorithm preserves the total citation counts to and from each paper and the distribution of these citation counts forward and backward in time to ensure that a paper (or journal) with n citations in the observed network will have n citations in the randomized network. For both the observed and the randomized paper-to-paper citation networks, we aggregated counts of paper pairs into their respective journal pairs to focus on domain-level combinations (*24-26*). In the data, there were over 122 million potential journal pairs created by the 15,613 journals indexed in the WOS.

Comparing the observed frequency with the frequency distribution created with the randomized citation networks, we generated a z-score for each journal pair. This normalized measure describes whether any given pair appeared novel or conventional. Z-scores above zero indicate pairs that appeared more often in the observed data than expected by chance, indicating relatively common or "conventional" pairings. Z-scores below zero indicate pairs that appear less often in the observed WOS than expected by chance, indicating relatively atypical or "novel" pairings. For example, in the year 1980, the pairing *Tetrahedron* and *Experientia* had a high z-score (21.55) indicating a conventional pairing, while *Tetrahedron* paired with *Life Sciences* had a negative z-score (-17.67) indicating a pairing more unusual than chance. The SOM details these computations, the null model, and an illustrative example (Tables S1, Fig S1-S3).

As a simple validation of the z-score measure, we found that journal pairs from the same WOS disciplinary designation had significantly higher z-scores than interdisciplinary journal pairs (Table S3, Fig S11). At the same time, only a minority (40.1%) of interdisciplinary journal pairs are novel, having z-scores below zero in the 1990s. This pattern indicates that observed journal pairings from the same WOS disciplines tend to be conventional and interdisciplinary WOS journal pairings are less substantially conventional but still not consistently novel.

The above method assigns each paper a distribution of journal pair z-scores based on the paper's reference list (Fig 1A). To characterize a paper's tendency to draw together conventional and novel combinations of prior work, we take two summary statistics. First, to characterize the central tendency of a paper's combinations, we consider the paper's median z-score. The median allows us to characterize conventionality in the paper's main mass of combinations. Second, we consider the paper's 10^{th} percentile z-score. The left tail allows us to characterize the paper's more unusual combinations where novelty may reside.

We find that papers typically relied on very high degrees of conventionality. Figure 1B presents the distribution of papers' median z-scores for the WOS in the indicated decades. Considering that a z-score below zero represents a journal pair that occurs less often than expected by chance, the analysis of median z-scores suggests very high degrees of conventionality. Half the papers have median z-scores exceeding 69.0 in the 1980s and 99.5 in the 1990s. Moreover, papers with a median z-score below zero are rare. In the 1980s only 3.54% of papers had this feature, while in the 1990s the percentage fell to 2.67%, indicating a persistent and prominent tendency for high conventionality.

Focusing on each paper's left tail combinations, we found that even among the paper's relatively unusual journal combinations, the majority of papers did not feature atypical journal pairs. Figure 1C shows that 40.8% of the papers in 1980s and 40.7% in the 1990s have a 10^{th} percentile z-score below zero. Overall, by these measures, science typically relies on highly conventional combinations and rarely incorporates journal pairs that are uncommon compared to chance. Analyses in the SOM (Fig S6) show that these empirical regularities for the WOS taken as a whole are largely replicated on a field-by-field basis and across time.

Our next finding indicates a powerful relationship between combinations of prior work and ensuing impact. Figure 2 presents the probability of a "hit" paper conditional on the combination of its referenced journal pairs. Hit papers are operationalized as those in the upper 5^{th} percentile of citations received across the whole dataset, as measured by total citations through 8 years after publication (the SOM considers alternative definitions of hit papers). The vertical axis shows the probability of a hit paper conditional on a 2x2 categorization indicating the paper's (i) "median conventionality" (an indicator for whether the paper's median z-score is in the upper or lower half of all median z-scores) and (ii) "tail novelty" (and indicator for whether the paper's 10^{th} percentile z-score is above or below zero).

Papers with “high median conventionality” and “high tail novelty” display a hit rate of 9.11 out of 100 papers, or nearly twice the background rate of 5 out of 100 papers. All other categories show significantly lower hit rates. Papers featuring high median conventionality but low tail novelty displayed hit rates of 5.82 out of 100 papers, while those featuring low median conventionality but high tail novelty display hit rates of 5.33 out of 100 papers. Finally, papers low on both dimensions have hit rates of just 2.05 out of 100.

Further analyses suggest a universality of these relationships for scientific work across time and fields. We considered the same relationships for different time periods (Fig S4), for different definitions of high impact papers (Fig S5), and for each of 243 fields of science (Fig S6 and Table S2). These analyses confirmed the findings above. Thus, novelty and conventionality are not opposing factors in the production of science; rather, papers with an injection of novelty into an otherwise exceptionally familiar mass of prior work are unusually likely to have high impact.

Collaboration is often claimed to produce more novel combinations of ideas (*10-14*), but the extent to which teams incorporate novel combinations across the universe of fields is unknown. Team-authored papers were more likely to show atypical combinations than single or pair-authored papers. Figure 3A shows that the distribution of 10^{th} percentile z-scores shifted significantly leftward as the number of authors increased (KS tests indicate solo vs. pair p=0.016, pair vs. team p=0.001, team vs. solo p<0.001). Papers written by one, two, or three or more authors showed high tail novelty in 36.1%, 39.8%, and 49.7% of cases respectively, indicating that papers with three or more authors showed an increased frequency of high tail novelty over the solo-author rate by 37.7 percent.

Teams were neither more nor less likely than single authors or pairs of authors to display high median conventionality. Figure 3B indicates no significant statistical difference in the median z-scores distributions (KS tests indicate solo vs. pair p=0.768, pair vs. team p=0.417, team vs. solo p=0.164). Teams thus achieve high tail novelty more often than solo authors. Yet, teams were not simply “more novel” but rather displayed a propensity to incorporate high tail novelty without giving up a central tendency for high conventionality.

In our final analysis, we examined the interplay between citation, combination, and collaboration using regression methods (Fig. 4). Papers were binned into eleven equally sized categories of median conventionality. A separate regression is run for each category of median conventionality and each team size, with field fixed effects. The SOM details the regression methodology and presents additional confirmatory tests (Fig S7-S10).

There were three primary findings. First, high tail novelty papers had higher impact than low tail novelty papers, an impact advantage that occurred at any level of conventionality and regardless of authorship structure. Second, peak impact occurs in the $85\text{-}95^{th}$ percentile of median conventionality, an exceptionally high level. This peak and its position appeared irrespective of tail novelty/no tail novelty or authorship structure. These generic features suggest fundamental underlying rules relating combinations of prior work to the highest impact science.

Finally, Figure 4 indicates that at virtually all mixes of tail novelty and median conventionality, larger teams were associated with higher impact. Thus, while teams incorporated the highest impact mixes more frequently (Fig. 3), teams also tended to obtain higher impact for any particular mix (Fig. 4). Nonetheless, despite teams' advantage in citations across virtually all fields of science (*12*), even teams had low impact at low levels of median conventionality and tail novelty.

Our analysis of 17.9 million papers across all scientific fields suggests that the highest-impact science draws on primarily highly conventional combinations of prior work with an intrusion of combinations unlikely to have been joined together before. These patterns suggest that novelty and conventionality are not factors in opposition; rather, papers that mix high tail novelty with high median conventionality have nearly twice the propensity to be unusually highly cited.

These findings have implications for theories about creativity and scientific progress. Combinations of existing material are centerpieces in theories of creativity, whether in the arts, the sciences, or commercial innovation (*2-4, 6-10, 16*). Across the sciences, the propensity for high impact work is sharply elevated when combinations of prior work are anchored in substantial conventionality, not novelty, while mixing in a left tail of combinations that are rarely seen together. In part, this pattern may reflect advantages to being within the mainstream of a research trajectory, where scientists are currently focused, while being distinctive in one's creativity. Combinations of prior work also relate to "burden of knowledge" theory, which emphasizes the growing knowledge demands upon scientists (*4, 17, 21*). New articles indexed by the WOS now exceed 1.4 million per year across 251 fields, encouraging specialization and challenging scientists' capacity to comprehend new thinking across domains. The finding that teams preserve high conventionality yet introduce tail novelty suggests that teams help meet the challenge of the burden of knowledge by balancing domain-level depth with a capacity for atypical combinations.

The methodology considered paper and journal pairings but can be applied at the level of disciplines, papers, or topics within papers, allowing the examination of combinations of prior work at different resolutions in future studies of creativity and scientific impact. Beyond science, links between novelty and conventionality in successful innovation also appear. E-books retain page-flipping graphics to remind the reader of physical books, and blue jeans were designed with a familiar watch pocket to look like conventional trousers. From this viewpoint, the balance between extending technology with atypical combinations of prior ideas while embedding them in conventional knowledge frames may be critical to human progress in many domains. Future research questions also arise from our findings. Science is dynamic, with research areas shifting and new fields arising. While we find that the regularities relating novelty, conventionality, and impact persist across time and fields, understanding how research trajectories shift and how new fields are born are questions that measures of novelty and convention may valuably inform. At root, our work suggests that creativity in science appears to be a nearly universal phenomenon of two extremes. At one extreme is conventionality and at the other is novelty. Curiously, notable advances in science appear most closely linked not with efforts along one boundary or the other but with efforts that reach toward both frontiers.

List of references:

1. "Facilitating interdisciplinary research" (Committee on Facilitating Interdisciplinary Research, National Academy of Sciences, National Academy of Engineering, Institute of Medicine, 2004).
2. H. S. Becker, Art Worlds (Univ. of California Press, California, 1982)
3. R. Guimera, B. Uzzi, J. Spiro, L. A. Nunes Amaral, Team assembly mechanisms determine collaboration network structure and team performance. Science. 308, 697-702 (2005).
4. B. Jones, The burden of knowledge and the death of the Renaissance Man: is innovation getting harder? Rev. Econ. Stud. (2009).
5. B. Jones, S. Wuchty, B. Uzzi, Multi-university research teams: shifting impact, geography, and stratification in science. Science. 322, 1259-1262 (2008).
6. J. Schumpeter, Business Cycles. (McGraw-Hill Book Company, Inc., New York, 1939).
7. A. P. Usher, A History of Mechanical Invention. (Cambridge, MA, 1954).
8. M. L. Weitzman, Recombinant growth, Q. J. Econ. 331-360 (1998).
9. M. Schilling. A 'small world' network model of creative insight. Creativity Res. J. 17, 131-154 (2005).
10. B. Uzzi, J. Spiro, Collaboration and creativity: the small world problem. Am. J. Sociology. 111, 447-504 (2005).
11. K. Börner, N. Contractor, H. Falk-Krzesinski, S. Fiore, K. Hall, J. Keyton, J. Thompson Klein, B. Spring, D. Stokols, W. Trochim, Mapping a research agenda for the science of team science. Res. Evaluat. 20(2) (2011).
12. S. Wuchty, B. Jones, B. Uzzi, The increasing dominance of teams in the production of knowledge. Science. 316, 1036-1039 (2007).
13. D. Stokols, K. L. Hall, B.K. Taylor, R.P. Moser, The science of team science: overview of the field and introduction to the supplement. Am. J. Prevent. Med. 35, S77-S89 (2008).
14. S. M. Fiore, Interdisciplinarity as teamwork: how the science of teams can inform team science. Small. Group. Res. 39, 251–277 (2008).
15. J. Evans, J. Foster, Metaknowledge. Science. 331(6018), 721-725 (2011).
16. R. Collins, The Sociology of Philosophies: A Global Theory of Intellectual Change. (Harvard Univ. Press, Cambridge MA, 1998).
17. L. Fleming, Recombinant uncertainty in technological search management. Science. 47, 117–132 (2001).
18. M. Schilling, E. Green. Recombinant search and breakthrough idea generation: an analysis of high impact papers in the social sciences. Res. Pol. 40, 1321-1331 (2011).
19. R.M. Henderson, K.B. Clark, Architectural Innovation: The Reconfiguration Of Existing, Administrative Science Quarterly. 35, 1; ABI/INFORM Global pg. 9-30 (1990).
20. P. Azoulay, J.G. Zivin, G. Manso, Incentives and Creativity: Evidence from the Howard Hughes Medical Investigator Program, The RAND Journal of Economics. 42, 527-554 (2011).
21. A. Einstein, The World as I See It. (Citadel Press, Secaucus NJ, 1949).
22. D. T. Whiteside, The mathematical principles underlying Newton's principia mathematica. J. Hist. Astron. 1, 116-138 (1970).
23. H. Small, Co-citation in scientific literature: a new measure of the relationship between two documents. J. Amer. Sot. Inform. Sci. 24, 265-9 (1973).

24. M. Stringer, M. Sales-Pardo, L. A. Nunes Amaral, Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. J. Am. Soc. Inf. Sci. Technol. 61(7), 1377–1385 (2010).
25. M. Stringer, M. Sales-Pardo, L. A. Nunes Amaral, Effectiveness of journal ranking schemes as a tool for locating information. PLoS ONE. 3(2), e1683. doi:10.1371/journal.pone.0001683 (2008).
26. Milo, R. N. Kashtan, S. Itzkovitz, M. E. J. Newman, U. Alon. Subgraphs in Networks. Phys. Rev. E. 70, 058102 (2004).
27. Bollen, J. Rodriguez, A. Marko and H. Van de Sompel. Journal status. Scientometrics. 69, (2006).
28. L. Rosenkopf and P. McGrath. Advancing the conceptualization and operationalization of novelty in organizational research. Organization Science. 22 1297 (2011).

Acknowledgements:

Sponsored by the Northwestern University Institute on Complex Systems (NICO), by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 and DARPA BAA-11-64, Social Media in Strategic Communication. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.

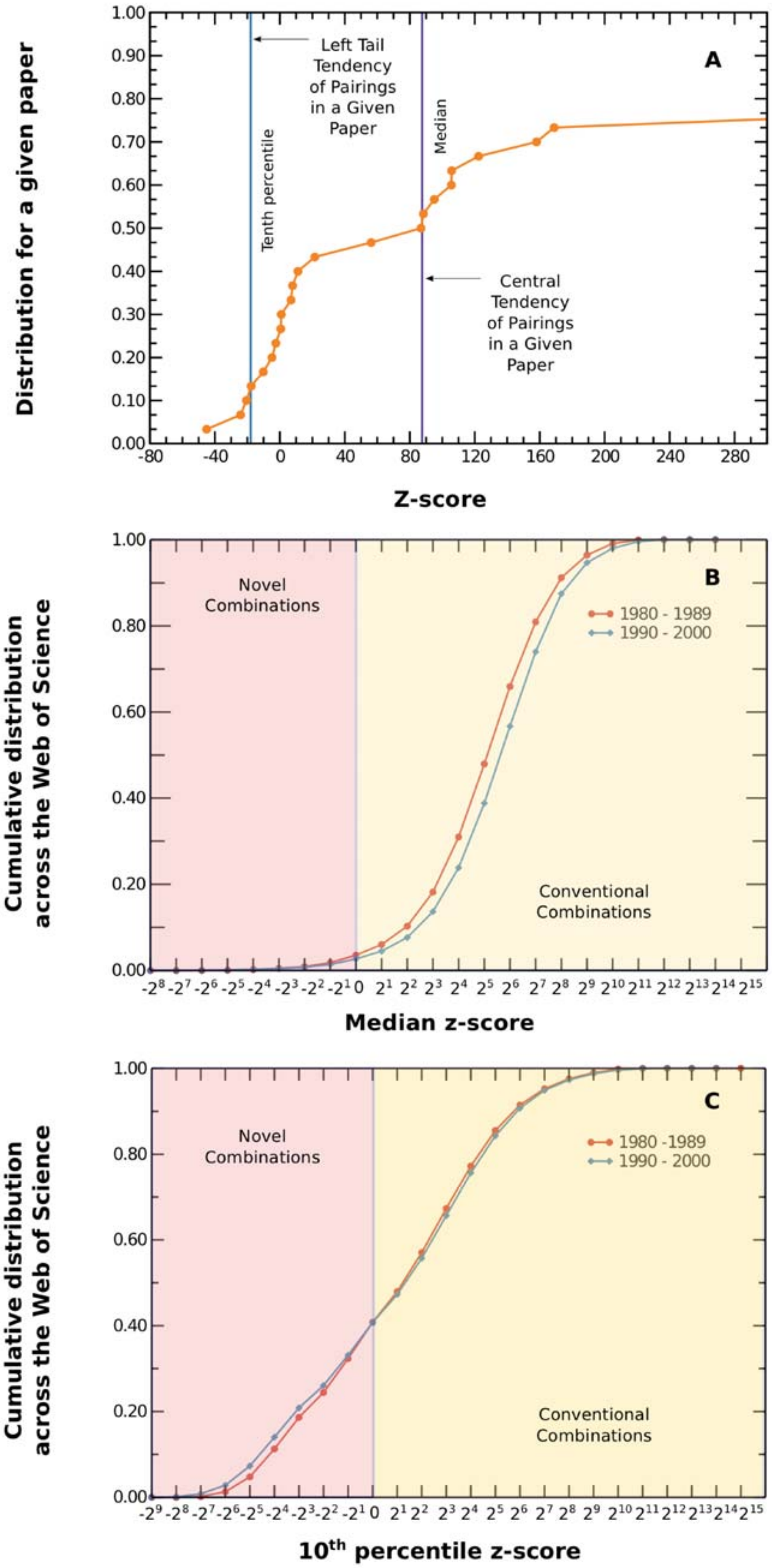

Figure 1: Novelty and Conventionality in Science. For a sample paper, Fig. 1A shows the distribution of z-scores for that paper's journal pairings. The z-score shows how common a journal pairing is compared to chance. For each paper we take two summary measures: its median z-score, capturing the paper's central tendency in combining prior work, and the 10th percentile z-score, capturing the paper's journal pairings that are relatively unusual. For the population of papers, we then consider these values across all papers in the WOS published in the 1980s or 1990s. Fig. 1B considers the median z-scores and shows that the vast majority of papers displays a high propensity for conventionality; in the 1980s and 1990s fewer than 4% of papers have median z-scores below 0 and more than 50% of papers have median z-scores above 64. Fig 1C considers the 10th percentile z-scores, which further suggest a propensity for conventionality; only 41% of papers in the 1980s and 1990s have a 10th percentile z-score below 0. Overall, by these measures, science rarely draws on atypical pairings of prior work.

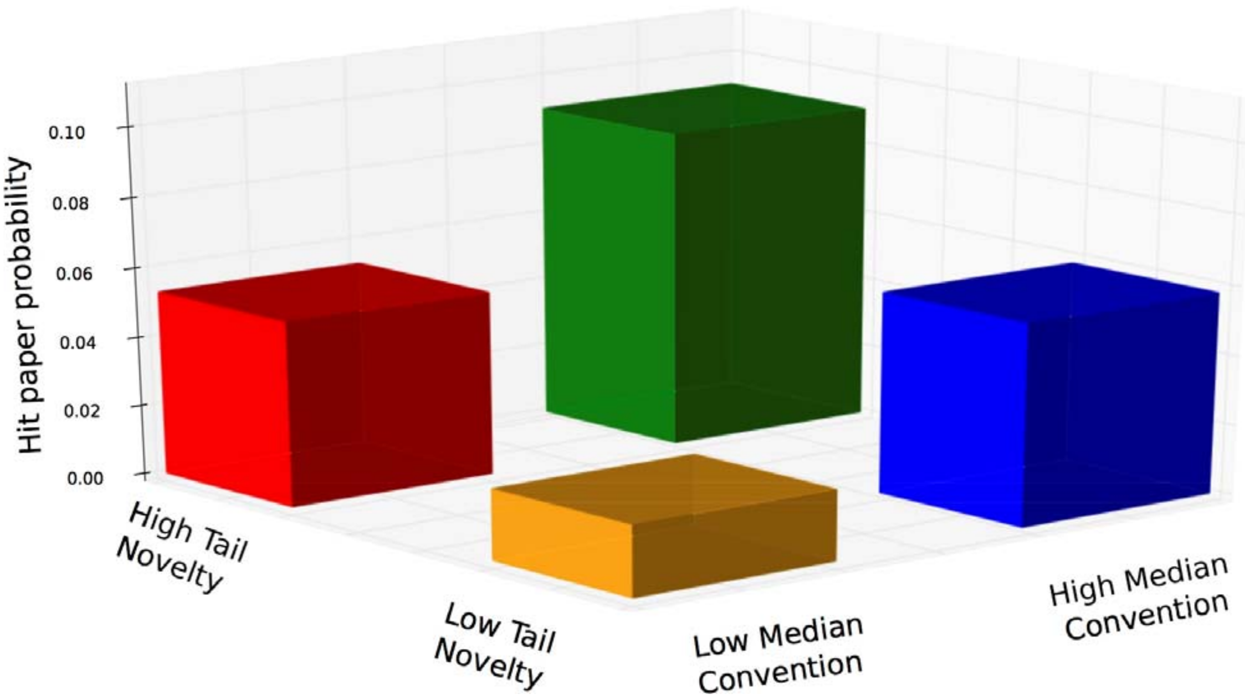

Figure 2: The Probability of a "Hit" Paper Conditional on Novelty and Conventionality
Figure 2 presents the probability of a paper being in the top 5% of the citation distribution conditional on two dimensions: whether a paper exhibits (1) high or low median conventionality and (2) high or low tail novelty, as defined in the text. Papers that combine high median conventionality and high tail novelty are hits in 9.11 out of 100 papers, a rate nearly double the background rate of 5%. Papers that are high on one dimension only – high median conventionality or high tail novelty but not both -- have hit rates about half as large. Papers with low median conventionality and low tail novelty have hit rates of only 2.05 out of 100 papers. The sample includes all papers published in the WOS from 1990-2000. The SOM shows similar findings when considering (i) all other decades from 1950-2000; (ii) "hit" papers defined as the top 1% or 10% by citations, and (iii) analyses controlling for field and other observable differences across papers, hinting at a universality of these relationships for scientific work. The difference in the hit probabilities for each category is statistically significant ($p < 0.00001$). The percentage of WOS papers in each category are: Green Bar (6.7%), Gold Bar (23%), Red Bar (26%), and Blue Bar (44%).

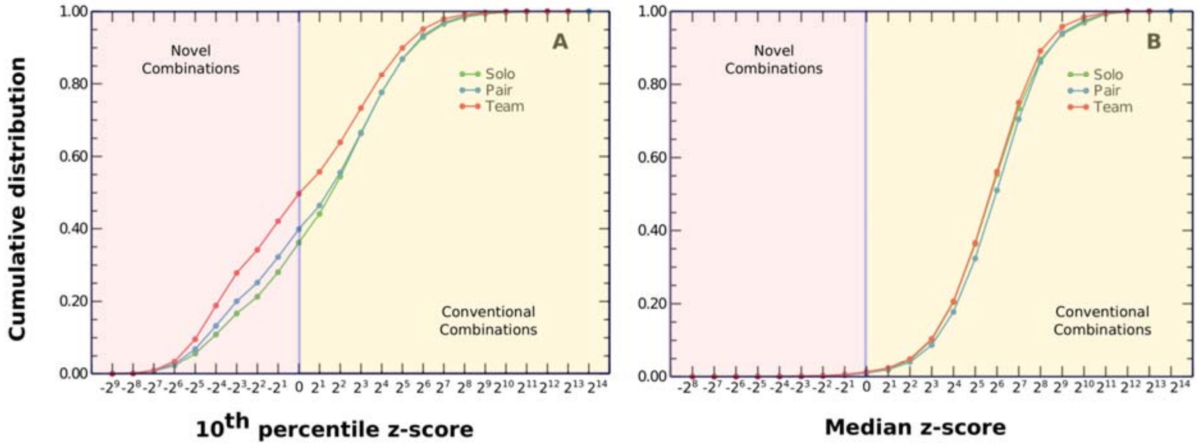

Figure 3: Authorship Structure, Novelty, and Conventionality
Team-authored papers are more likely to incorporate tail novelty but without sacrificing a central tendency for high conventionality. Papers introduce tail novelty (a 10^{th} percentile z-score less than 0) in 36.2%, 39.9%, and 49.7% of cases for solo authors, dual authors, and three or more authors respectively (Fig 3A). Kolmogorov-Smirnov tests confirm the distributions of tail novelty are distinct (solo vs. pair p=0.016, pair vs. team p=0.001, team vs. solo p<0.001). By contrast, each team size shows similar distributions for median conventionality (Fig 3B, K-S tests indicate no statistically significant differences). These findings suggest that a distinguishing feature of teamwork, and teams' exceptional impact, reflects a tendency to incorporate novelty.

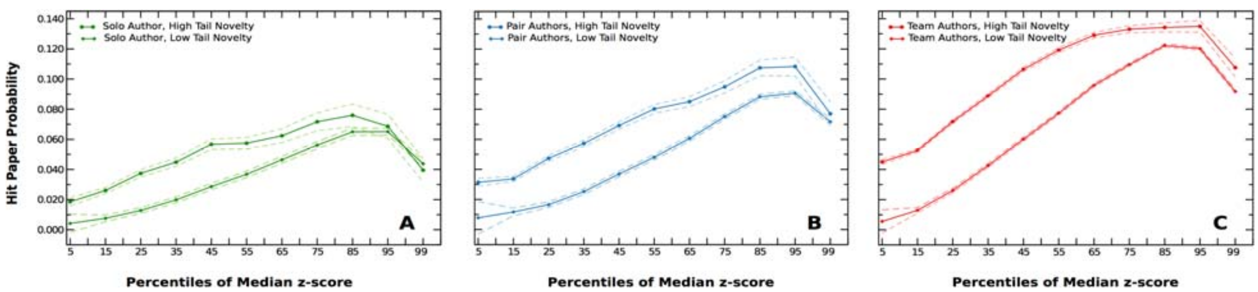

Figure 4: Novel and Conventional Combinations in the Production of Science
The interplay between tail novelty, median conventionality, and hit paper probabilities show remarkable empirical regularities (Fig 4A-C). First, high tail novelty papers have higher impact than low tail novelty papers at (i) any level of conventionality and (ii) regardless of authorship structure. Second, increasing median conventionality is associated with higher impact up to the 85-95th percentile of median conventionality after which the relationship reverses. Third, larger teams obtain higher impact given the right mix of tail novelty and median conventionality. Nonetheless, at low levels of median convention and tail novelty even teams have low impact, further emphasizing the fundamental relationship between novelty, conventionality, and impact in science.

Supplementary Materials for

Atypical Combinations and Scientific Impact

Brian Uzzi, Satyam Mukherjee, Michael Stringer, Benjamin Jones

correspondence to: bjones@kellogg.northwestern.edu

This PDF file includes:

Data and Methods

Data

We examined 17.9 million scientific publications across 15,613 journals, constituting all research articles indexed in the Thomson Reuters Web of Science (WOS) database that were published over the 1950-2000 period. According to each journal's subject area, the ISI currently defines three fields and constituent subfields: science and engineering (171 subfields), social sciences (54 subfields), and arts and humanities (27 subfields) with coverage for research publications in science and engineering since 1945, social sciences since 1956, and arts and humanities since 1975. The WOS records papers' citations, number of authors, and citation links to other papers in the database.

Methods

We measure the relative conventionality and novelty of the prior work that a paper combines by examining the papers referenced in a paper's bibliography (*23*, *24*). This section first provides an overview of our methodology, followed by an illustrative example and further details.

Overview

We look at pairwise combinations of prior work. Our basic measurement question is to assess how common or novel any pairwise combination of prior work is. To determine how conventional or novel prior combinations of referenced work are, we would like to know both the (i) observed frequency of any given pairing of references in the WOS and (ii) the frequency of that pairing that would have occurred by chance. Comparing the observed frequency to the frequency expected by chance creates a normalized z-score measure for whether any given pairing appears novel or conventional.

To measure the observed frequency of any given pairing in the WOS, we take the following five steps:

(1) Take the references listed in a given paper's bibliography.
(2) Consider all pairwise combinations of the papers referenced in the bibliography of the paper.
(3) For each pairwise combination, record the two journals that were paired.
(4) Repeat steps (1)-(3) for every paper in the WOS.
(5) Count the aggregate, population-wide frequency of each journal pairing for all referenced pairs from a given publication year.

Figure S1 presents a stylized example for Steps 1-3, showing for a given paper how pairs of references are counted from that paper's reference list. The algorithm repeats this counting process for every article in the WOS and aggregates the counts for each given publication year.

Our method counts specific journal pairings, using journals to proxy for different areas of knowledge. Journal-level analysis is well-positioned to distinguish domains of knowledge while having precedence in the literature for being relatively transparent, interpretable, and computationally feasible (*23*, *24*, *27*).[1]

Having determined the observed frequency of each journal pairing, we consider the frequency distribution for each journal pairing that would have occurred by chance. The null model randomly reassigns the citation links between papers. As further detailed below, the method uses a variation of the Markov Chain Monte Carlo (MCMC) algorithm to randomly switch co-citations between all 17.9 million papers into a synthetic network with 302 million citations (edges), the same number of papers and citations as the observed network. Note that this method preserves the detailed paper-level structure of the global citation network; specifically, the number of citations to and from each paper is preserved, as are the dynamics of citation timing.

Using this approach, we create 10 synthetic instances of the entire WOS, each with its own set of randomized citation links. For each instance of the WOS, we then repeat steps (1)-(5) above, calculating the frequency of each co-referenced journal pair. Looking across all ten randomized cases of the WOS, we generate a distribution of frequencies for each journal pair. We can then evaluate the z-score for each observed journal pair relative to what was expected by chance:

$$z = (obs - exp)/\sigma$$

Where obs is the observed frequency of the journal pair in the actual WOS while exp is the mean and σ is the standard deviation of the number of journal pairs obtained from the 10 randomized simulations of the paper-to-paper citation network.

Finally, returning to categorizing a paper's prior work in terms of novelty and conventionality, we can now assign a z-score to each of the journal pairs in that paper's reference list. Each paper thus has a distribution of journal pairings, where any given pairing may be more or less common compared to chance. To summarize the information in this distribution, we take two primary summary statistics:

(i) The median z-score for that paper
(ii) The 10^{th} percentile z-score for that paper

The first measure is a summary statistic for the central tendency of the combinations of journals that a paper cites. The larger the median z-score for a paper, the more common the main mass of journal combinations in that paper compared to chance. The second measure is a summary statistic for the left tail of combinations of journals that a paper cites – journal pairings that are relatively unusual, compared to chance, among the set of journal pairings in that paper's reference list.

Illustrative Example of Methodology and Further Detail
To illustrate these procedures, consider the follow example, based on a single paper.

[1] Other operationalizations might consider lower resolution pairings using the ISI's 252 subfield categories, text-based combinations, or conceptualizations for measuring novelty beyond combinatorial pairs (*28*).

- Step 1. Take the references in a bibliography in a given paper. Consider the paper "Synthesis of the 5 Natural Cannabis Spirans," which was published in *Tetrahedron Letters* in the year 1980. This paper references 11 papers and one thesis (Fig S2).

- Step 2. Consider all pairwise combinations of the papers referenced in the bibliography of that paper. As can be seen in Figure S2, pairwise paper combinations include, for example, (i) El-Feraly et al. 1976 with Boeren et al. 1977, (ii) El-Ferely et al. 1976 with Bull et al. 1975, and (iii) Boeren et al. 1977 with Bull et al. 1975. With 11 referenced papers, we have 55 (i.e. 11 choose 2) pairwise paper combinations.

- Step 3. Map the observed paper pairs into observed journal pairs. The 55 paper pairs are mapped into 55 journal pairs, where some journal pairs in this list appear multiple times. For example, *Tetrahedron* and *Experientia* are paired 6 times.

- Step 4. Repeat steps (1)-(3) for every paper in the WOS. The above steps, shown for a single article, are now repeated for every paper in the WOS. References to materials outside the WOS (for example, books) are not included.

- Step 5. Count the frequency of each observed journal pairing for a given publication year, using the referenced works of every paper published that year in the WOS. Information from the sample paper above would be counted as part of the year 1980. Hence, we allow journal pair frequencies to vary over time.

Having completed steps (1)-(5) for the observed papers in the WOS, we repeat them for each synthetic instance of the WOS, as created by the null model. Comparing the observed frequency of journal pairs under the real WOS with the frequency distribution that appears across instances of the null model, we compute a z-score for each journal pair. Continuing our illustrative example, the observed frequency, expected frequency, and z-score for several journal pairings that appear in the paper "Synthesis of the Five Natural Cannabis Spirans" are presented in Table S1. As Table S1 demonstrates (for a subsample of journal pairs), each published paper has a distribution of journal pairs, some of which are highly conventional (such as *Tetrahedron-Tetrahedron*) while others are unusual compared to chance (such as *Tetrahedron-Life Sciences*). Figure 1A (main text) presents the distribution of z-scores for this illustrative paper and indicates the median z-score and the 10^{th} percentile z-score in that paper's distribution.

Table S1 further shows the importance of normalizing the observed frequencies. For example, compare the pairings (1) *Tetrahedron* and *Experientia* and (2) *Journal of the American Chemical Society* and *Life Sciences*. Both have similar observed co-citation frequencies in the WOS: 454 and 469 respectively. However, compared to chance, the first pairing appears to have high conventionality while the second pairing appears to have high novelty. This result follows because *Tetrahedron* and *Experientia* receive fewer total citations in the database, so that co-citations are less likely by chance, averaging only 256 co-citations under the null model. The latter pairing, representing journals receiving many citations, averages 3,147 co-citations under the null model. Thus normalizing the observed counts given the underlying citations frequencies to each journal is essential to accurately describe the relative conventionality or novelty of any journal pair.

Null Model Detail

The null model creates random synthetic instances of the WOS while incorporating realistic aspects of the data and its network structure. In particular, the null model incorporates two basic empirical facts about citation patterns:

- Citation distributions are skewed. Some papers and journals are cited far more often than other papers and journals and consequently are referenced more frequently in papers' bibliographies.
- Citation counts are dynamic processes that vary by journal (*25*), so that the rate at which papers accumulate citations is journal dependent.

Keeping these facts in mind, the null model preserves for each paper in the WOS the same number of references to past work, the same number of citations from subsequent papers, and the same distribution of these citations over time (Fig S3, left panel and middle panel). The right panel of Figure S3 shows the distributions of observed frequency and expected frequency of journal papers for the example paper above.

Specifically, we use a variation of Markov Chain Monte Carlo (MCMC) algorithm to construct randomized citation networks for all papers in the WOS database. The switching of endpoints of citation links is constrained to randomly chosen endpoints within the same class (Fig S3), where the link classes are defined as having the same origin year and target year (*26*). One can think of each link class as a subgraph of the global citation network, which can then be randomized in the usual way by performing Q*E switches, where E is the number of links in the subgraph. There is no proof for when the Markov Chain converges, however it is suggested (*26*) to set Q at a safe value of 100. Since the citation network has 302 million edges, the scale of the computation is large, and we used a slightly less conservative value of Q =2log(E) in order to reduce computational burden. As can be noted in the original paper on the MCMC switching algorithm (*26*), this value of Q is well in the region where correlations with the original network cannot be detected.

Supplementary Text

Results over Time and by Definition of Hit Papers

In the text, we focused on the 1990s and defined hit papers as being in the upper 5th percentile by citations received. In Fig S4 we show that the results hold (a) over 5 decades of data recorded in the WOS from 1950-2000 and (a) using the upper 1^{st} or 10^{th} percentiles of citation impact.

Results using Alternative Definitions of Tail Novelty

In the main text, we defined tail novelty using the 10^{th} percentile z-score of a paper's citation pairings, where high (low) tail novelty indicates a 10^{th} percentile z-score below (above) zero. In Fig S5, we define the cutoff for high and low tail novelty at different percentiles of a paper's z-score: the 1^{st}, 5^{th}, 20^{th}, 30^{th}, and 40^{th}. Fig S5 shows that using the 1^{st}, 5^{th}, 10^{th}, or 20^{th} percentiles all capture significant positive associations between impact and tail novelty in the 1990s. Beyond the 30^{th} percentile the significant association between impact and tail novelty disappears. These patterns suggest that the concept of tail novelty is not sensitive to a single value and that beyond a precise focus on the 10^{th} percentile the construct is related to impact so long as one continues to consider the left tail of the distribution.

Results by Subfields

The following analysis shows that the results presented in the main text for the whole of the WOS continue to appear quite broadly when examining patterns within individual subfields. By subfield, we present (1) the tendency for tail novelty and median conventionality, and (2) the relationship between novelty, conventionality, and hit papers. We examine all 243 subfields that appear in the WOS over the 1990s.

To examine subfield-specific patterns with regard to tail novelty and median conventionality, we grouped all the papers in each subfield. We then examined the central tendency, by subfield, for the median and 10^{th} percentile z-scores for each paper. Consistent with our main result, Fig S6A indicates a strong subfield-specific tendency towards conventionality among papers' median z-scores. On a field-by-field basis, papers typically reference journal pairings that are much more likely than expected by chance. Moreover, Fig S6B indicates that few fields display a propensity for tail novelty. The subfield-specific central tendency of papers' 10^{th} percentile z-score is below zero for just 6.6% of subfields, indicating that combinations of journal pairs that are unusual compared to chance are rare.

To examine any field specific relationships between novelty, conventionality, and hit papers, we calculate the subfield-specific probabilities of a "hit" by the four categories used in Figure 2 and defined in the text. We then ranked these four categories in each subfield, where 1 indicates the highest probability of hit, 2 indicates the second highest probability of a hit and so on. Consistent with the main results, Table S2 shows that in 64.4% of fields, a paper's likelihood of being a hit paper is greatest when combining prior work characterized by high tail novelty and high median conventionality. This category (GREEN) is ranked first or second in 86.3% of subfields. Notably, to the extent that this category is not dominant within a subfield, the category featuring a more general shift toward novelty (RED) appears prominently, suggesting that tail novelty is an especially generic feature of the highest-impact papers. Conversely, the category (ORANGE) featuring low tail novelty and low median conventionality ranks lowest in 70.4% of subfields.

In summary, these subfield specific analyses indicate that the results presented in the main text for the whole of the WOS appear consistently on a field-by-field level.

Regression Methods and Results

Figure 4 in the main text uses regression methods to consider the relationships between median conventionality, tail novelty, and impact for each authorship category. We use logistic regression to predict the probability of hit papers in the 1990s and run these regressions in a flexible manner that avoids imposing functional forms on the data. In particular, we first divide papers into subsamples based on their median conventionality (11 categories, from least to greatest median conventionality, as defined in the main text) and the number of authors (3 categories, for solo authors, two-author pairs, and three or more authors). This creates 33 distinct subsamples. We then run a separate regression for each subsample. For a given subsample, a regression takes the form

$$\Pr(y_i) = f\,(\beta Tail\_Novelty_i + \sum_f \gamma_f\, Field_{fi})$$

where $y_{ij} \in \{0,1\}$ is an indicator variable for a "hit" paper, and $Tail\_Novelty_i \in \{0,1\}$ is an indicator variable for whether a paper's 10th percentile z-score is below zero. The regression includes a full set of fixed effects for each of 243 subfields indeed by the WOS in the 1990s, where the indicator variables $Field_{fi} \in \{0,1\}$ are equal to 1 if the paper *i* is in field *f*. Inclusion of these fixed effects accounts for any mean differences in hit probabilities and tail novelty across subfields. We further restrict the sample to papers with at least ten known references, which ensures that the each paper in the sample has many pairwise combinations of prior work.

Figure 4 establishes a large positive relationship between tail novelty and hit papers, which appears independently in each of the 33 subsamples. The regressions further establish that the probability of hit papers increases with median conventionality, peaking at approximately the 85th percentile of median conventionality.

These strong empirical regularities extend to alternative analyses. The main text defines hit papers as those in the top 5 percent of citations received. Figure S7 reconsiders these regressions defining hit papers to be in the top 1 percent of citations received. The results for this higher threshold for a "hit" paper look extremely similar. Second, Figure S8 reconsiders the regressions when controlling for the number of references made by the paper to other papers in the WOS. These regressions are of the form

$$\Pr(y_i) = f\left(\beta Tail\_Novelty_i + \sum_f \gamma_f \, Field_{fi} + \sum_r \rho_r \, Ref_{ri}\right)$$

which include fixed effects for each of 10 ranges of reference counts, where the indicator variables $Ref_{ri} \in \{0,1\}$ are equal to 1 if the paper *i* makes the number of references in category *r*. Fig S8 shows that controlling for the number of references presents similar patterns as reported in the main text and underscores the empirical regularity of these findings.

In our regression analyses presented in the main text and above, we condition on papers with at least 10 references to ensure that each paper analyzed has a rich distribution of underlying journal pairs. That said, in practice there is no substantive distinction when analyzing these papers. Fig S9 below confirms the results when looking at the all papers together, regardless of the number of references. Fig S10 below further confirms the results when looking at the subset of papers with less than 10 references, although the relationships for this restricted sample are somewhat noisier, given the smaller sample sizes.

Interdisciplinary Journal Pairings

As a validation exercise, we examined the relationships between our measure of novelty and conventionality and interdisciplinary journal pairs. (We thank an anonymous reviewer for suggesting this transparent analysis.) The broad expectation is that novel journal pairings encompass journals from different fields/disciplines and conventional journal pairings encompass journals from the same field. Specifically, Fig S11 shows the relationship between journal pair z-scores (our measure) and whether the journal pair shares a common WOS field designation (e.g., economics, ecology, or physics). We define a binary variable, "journal similarity," which is equal to 1 if two journals share a common WOS field and equal to 0 otherwise.

As shown in Fig S11A, journal pairs sharing the same WOS field have much higher average conventionality (z-scores) than those which do not. Fig S11B aggregates the same binary measure of journal pair similarity to the paper level and shows similar construct validity with our measure. Thus, our measures of novelty and conventionality are strongly associated with field-level dissimilarity and similarity respectively, providing face validity and further transparency to our approach.

At the same time, we observe that journal pairs from different WOS fields are an imprecise metric for assessing actual novelty because journals from different fields are commonly referenced together in papers. For example, consider the journals Human Genetics and Nucleic Acids Research, which are in distinct WOS disciplines but have a z-score of 3386, suggesting a remarkably conventional pairing. By contrast, consider the New England Journal of Medicine paired with Brain Research, which also sit in distinct WOS disciplines but in this case are novel, with a z-score of -121. Table S3, examines these tendencies for each year from 1990-2000. We see that journal pairs from different WOS fields tend to be conventional. The majority of journal pairs exhibit positive z-scores; that is, these journal pairs appear together in reference lists substantially more often than chance. At the same time, the truly novel journal pairings with z-scores less than zero that guide impact in our analyses are only a subset of interdisciplinary journals pairings. These results suggest the high quantitative precision and added information gleaned from our approach compared to simpler, heuristic measures.

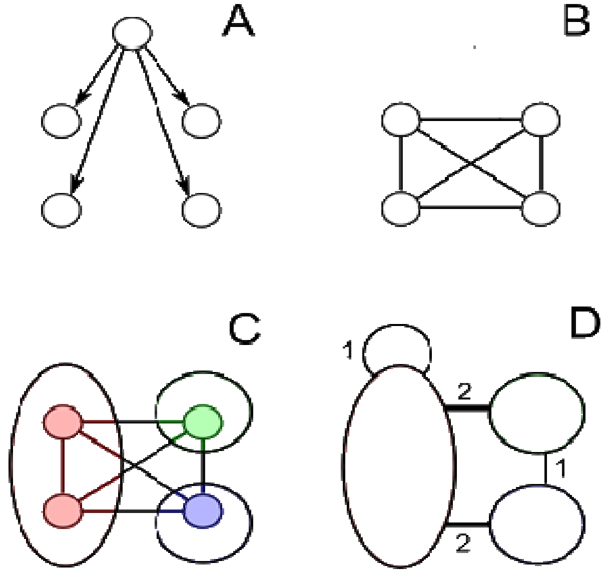

Fig. S1

Paper Pairs and Journal Pairs. This figure presents a stylized example of how paper pairs and journal pairs are drawn from the network structure of citations. In panel A, the circular nodes represent papers and the directed links exist when the top paper cites the bottom four papers. In panel B, the circular nodes represent papers and the undirected co-citation links between papers are shown in black. A co-citation exists between each pair of papers that occurs in the reference list of the focal paper. Here there are 4 references and therefore 6 (i.e. 4 choose 2) co-citation links. In panel C, paper nodes are grouped by journal; the shaded ovals represent the three journals in which each of the cited papers is published. Finally, in panel D, the co-citation links between papers are mapped to the journal level, and the black links represent journal co-citations. Note that the total number of paper-to-paper co-citation links (6) is preserved at the journal co-citation level.

Title: CANNABISPIRONE AND CANNABISPIRENONE 2 NATURALLY OCCURRING SPIRO-COMPOUNDS
Author(s): BERCHT, CAL; VANDONGEN, JPCM; HEERMA, W; et al.
Source: TETRAHEDRON Volume: 32 Issue: 23 Pages: 2939-2943 Published: 1976

Title: BETA-CANNABISPIRANOL - NEW NON-CANNABINOID PHENOL FROM CANNABIS-SATIVA L
Author(s): BOEREN, EG; ELSOHLY, MA; TURNER, CE; et al.
Source: EXPERIENTIA Volume: 33 Issue: 7 Pages: 848-848 Published: 1977

Title: SYNTHESIS AND TOXICITY EVALUATION OF AFLATOXIN-P1
Author(s): BUCHI, G; SPITZNER, D; PAGLIALU.S; et al.
Source: LIFE SCIENCES Volume: 13 Issue: 8 Pages: 1143-1149 Published: 1973

Title: EFFICIENT METHOD FOR CONVERTING 17-OXO-STEROIDS INTO 17-ACETYL STEROIDS
Author(s): BULL, JR; TUINMAN, A
Source: TETRAHEDRON Volume: 31 Issue: 17 Pages: 2151-2155 Published: 1975

Title: ISOLATION OF CANNABISPIRADIENONE AND CANNABIDIHYDROPHENANTHRENE - BIOSYNTHETIC RELATIONSHIPS BETWEEN THE SPIRANS AND DIHYDROSTILBENES OF THAILAND CANNABIS
Author(s): CROMBIE, L; CROMBIE, WML; JAMIESON, SV
Source: TETRAHEDRON LETTERS Issue: 7 Pages: 661-664 Published: 1979

Title: BIOMIMETIC SYNTHESIS OF CANNABISPIRAN
Author(s): ELFERALY, FS; CHAN, YM; ELSOHLY, MA; et al.
Source: EXPERIENTIA Volume: 35 Issue: 9 Pages: 1131-1132 Published: 1979

Title: CRYSTAL AND MOLECULAR-STRUCTURE OF CANNABISPIRAN AND ITS CORRELATION TO DEHYDROCANNABISPIRAN - 2 NOVEL CANNABIS CONSTITUENTS
Author(s): ELFERALY, FS; ELSOHLY, MA; BOEREN, EG; et al.
Source: TETRAHEDRON Volume: 33 Issue: 18 Pages: 2373-2378 Published: 1977

Title: CANNABIS .19. OXYGENATED 1,2-DIPHENYLETHANES FROM MARIHUANA
Author(s): KETTENESVANDENBOSCH, JJ; SALEMINK, CA
Source: RECUEIL DES TRAVAUX CHIMIQUES DES PAYS-BAS-JOURNAL OF THE ROYAL NETHERLANDS CHEMICAL SOCIETY Volume: 97 Issue: 7-8 Pages: 221-222 Published: 1978

Title: [not available]
Author(s): KETTENESVANDENB.JJ
Source: THESIS UTRECHT Published: 1978

Title: GENERAL ONE-STEP SYNTHESIS OF NITRILES FROM KETONES USING TOSYLMETHYL ISOCYANIDE - INTRODUCTION OF A ONE-CARBON UNIT
Author(s): OLDENZIEL, OH; VANLEUSEN, D; VANLEUSEN, AM
Source: JOURNAL OF ORGANIC CHEMISTRY Volume: 42 Issue: 19 Pages: 3114-3118 Published: 1977

Title: CANNABIS .13. 2 NEW SPIRO-COMPOUNDS, CANNABISPIROL AND ACETYL CANNABISPIROL
Author(s): SHOYAMA, Y; NISHIOKA, I
Source: CHEMICAL & PHARMACEUTICAL BULLETIN Volume: 26 Issue: 12 Pages: 3641-3646 Published: 1978

Title: METHODS IN ALKALOID SYNTHESIS - IMINO ETHERS AS DONORS IN MICHAEL REACTION
Author(s): TROST, BM; KUNZ, RA
Source: JOURNAL OF THE AMERICAN CHEMICAL SOCIETY Volume: 97 Issue: 24 Pages: 7152-7157 Published: 1975

Fig. S2
Reference list for example paper. The paper “Synthesis of the 5 Natural Cannabis Spirans” references 11 different papers and one thesis.

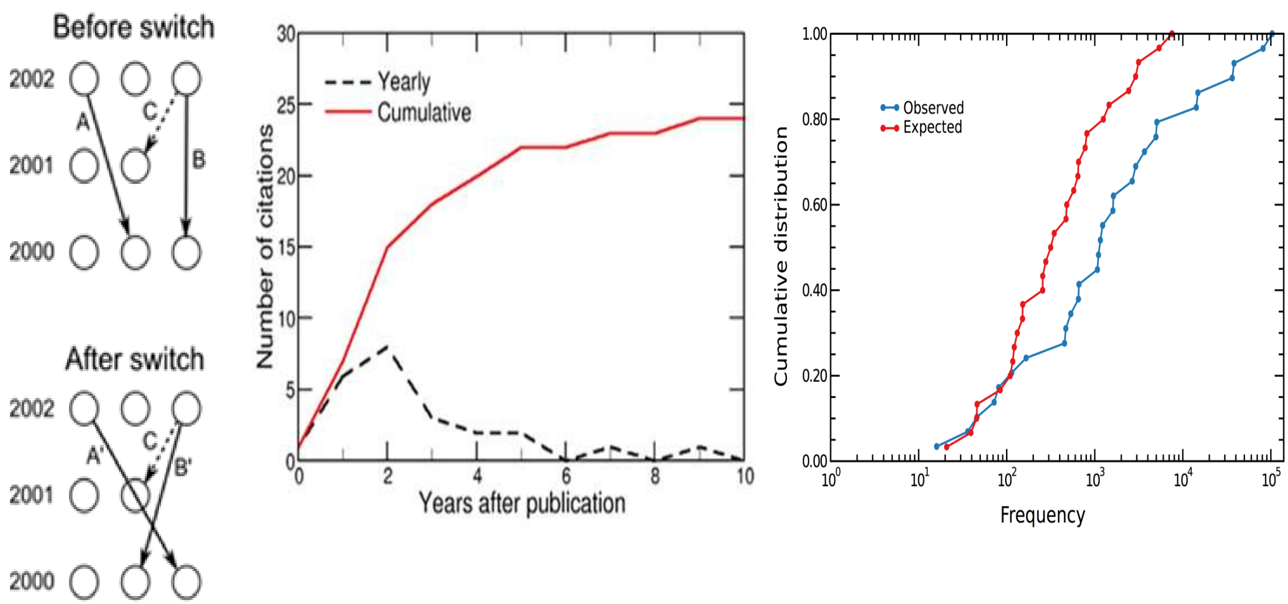

Fig. S3
Link switching in the null model and example distributions of observed and expected frequency of journal pairs. Citation links between papers are switched randomly but constrained to have the same origin year and target year. Thus in the left panel, switching links A and B is allowed, while switching links A and C is not allowed. The switching algorithm thus preserves for each paper its (i) number of references, (ii) citation count, (iii) citation accumulation dynamics, and (iv) the age distribution of referenced work. Performing QE switches converges to a random graph from the configuration model (*26*) where the number of and dynamics of citations are preserved but the origin of the citations is randomized. Since each node is equally likely to be the originating node of any citation, given the constraints, we know a priori that no disciplines exist in this randomized citation network. The middle panel above demonstrates the citation history of a paper -- the citation history of every paper is exactly preserved under our null model, ensuring that we control for both the variation in magnitude and dynamics of citation accumulation to papers. The right panel above further shows, for the example paper highlighted in Table S1, the frequency distribution for the observed journal pairings (blue line) and the frequency distribution for these journal pairings when averaged across instances of the null model (red line).

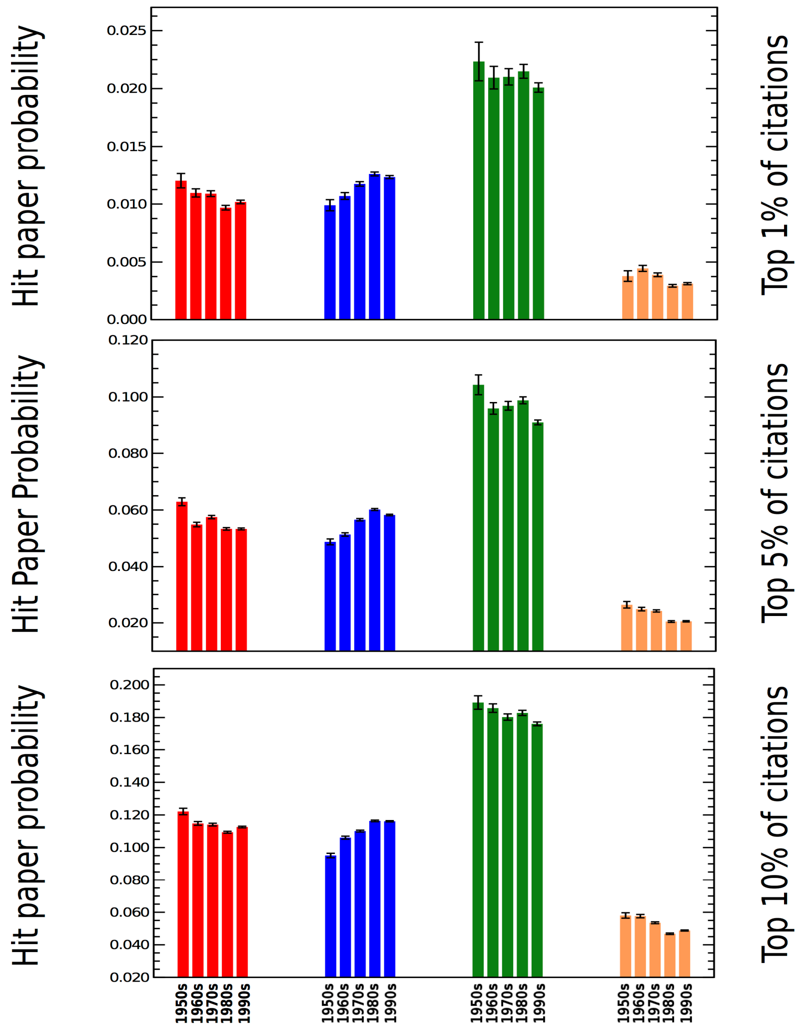

Fig. S4
Citation impact results generalize by decade and by definition of "hit" paper. This figure shows broadly consistent patterns both over time and by the definition of "hit" paper, suggesting a remarkably robust and strong empirical regularity between scientific impact and how prior work is combined. Specifically, the figure shows that high tail novelty combined with high median conventionality (GREEN bars) outperforms other categories in all decades from 1950-2000, and regardless of whether a "hit" paper is defined as a top 1%, 5%, or 10% by citations received, broadly showing hit rates that are approximately twice the background rate. By contrast, papers that feature neither high tail novelty nor high median conventionality (ORANGE bars) see hit rates at only half or less the background rate.

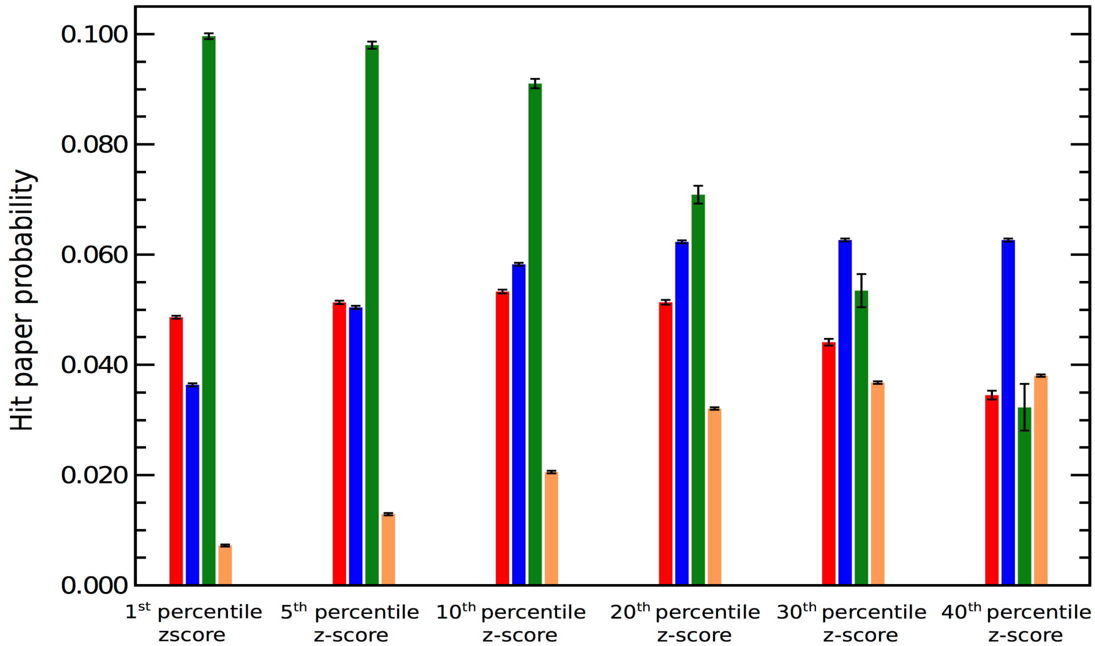

Fig. S5
Citation impact results generalize to broader definitions of left tail novelty. The figure presents the relationship between tail novelty and impact using alternative definitions of tail novelty. In each case, tail novelty is defined as an indicator for whether the p^{th} percentile of a paper's z-score distribution is less than zero. The x-axis indicates the value of p. It is seen that for p≤20, high tail novelty combined with high median convention (GREEN bars) outperforms other categories. The results in the main text, which use the 10^{th} percentile, thus extend broadly to other definitions of tail novelty so long as the measure emphasizes the paper's left tail of combinations.

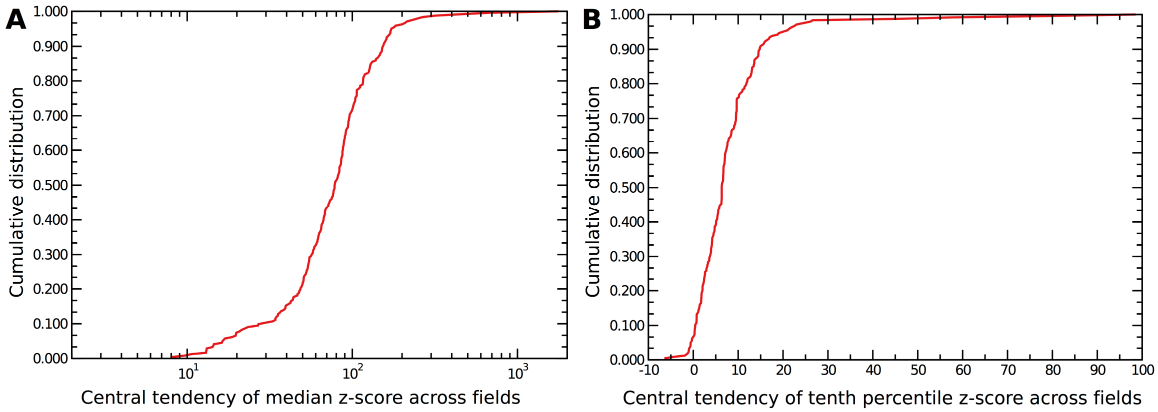

Fig. S6
High median conventionality and low tail novelty are common features across subfields. Grouping papers by each of the 243 subfields indexed by the WOS in the 1990s, we examine the median and 10th percentiles z-scores. Taking the central tendency (median) of each of these measures in each subfield, the plots indicate that no subfield displays a strong tendency for novel journal pairings. All subfields display a characteristic central tendency for drawing on highly conventional pairings of prior work (A) while just 6.6% of fields display 10th percentile z-scores that are typically less than zero (B)**.**

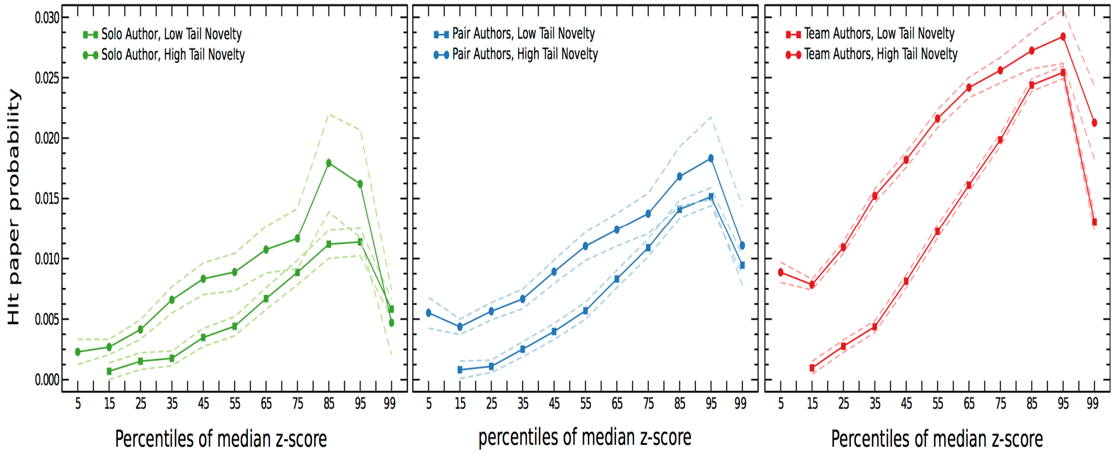

Fig. S7
Novelty, authorship and impact for top 1% papers. This Figure repeats Figure 4 in the main text but defines hit papers as those that receive citations within eight years of publication that are in the upper 5 percent of all papers published that year.

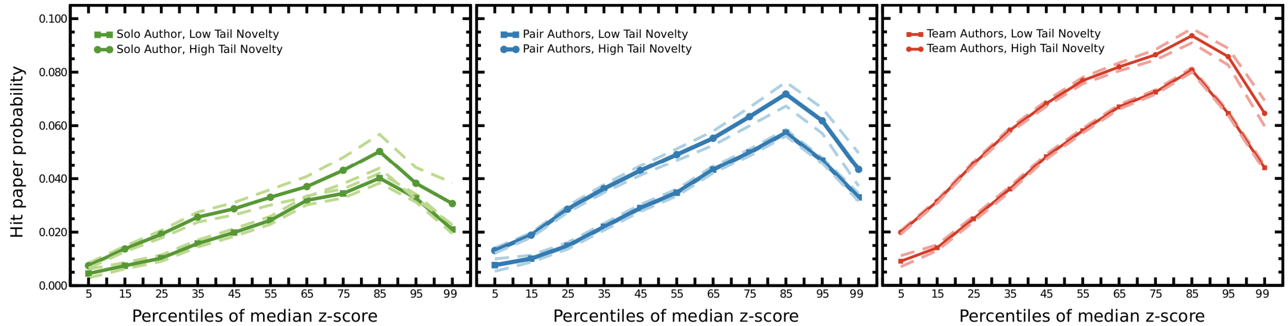

Fig. S8

Novelty, authorship and impact for top 5% papers with controls for referencing behavior. This Figure repeats Figure 4 in the main text but with fixed effects for number of WOS references, using ten categories of reference counts.

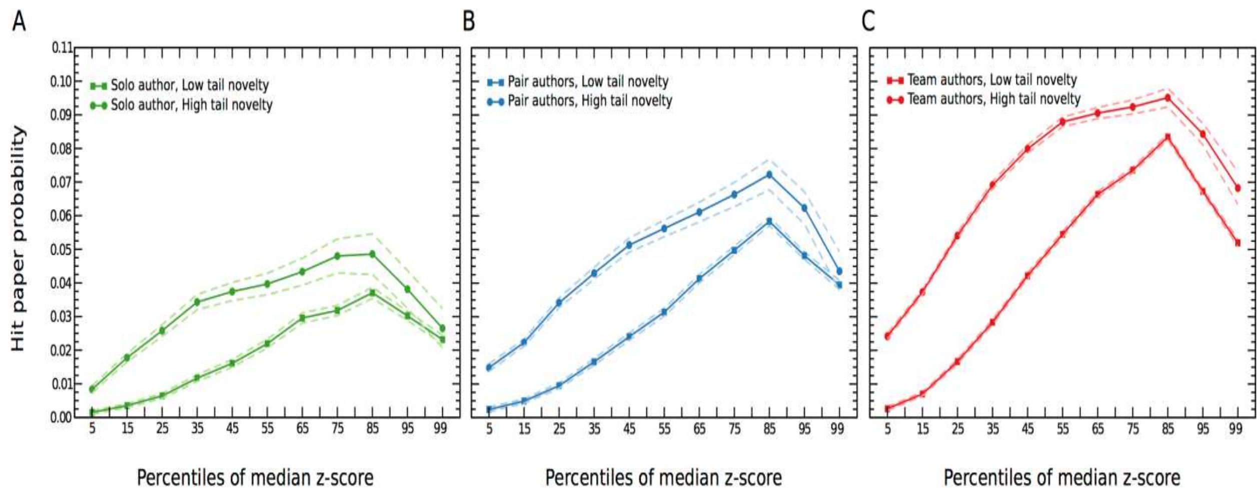

Fig. S9
Novel and conventional combinations, full sample. This figure shows that results for all WOS papers, regardless of the number of references they make, and shows that the results are similar to the main text results shown in Figure 4(A-C).

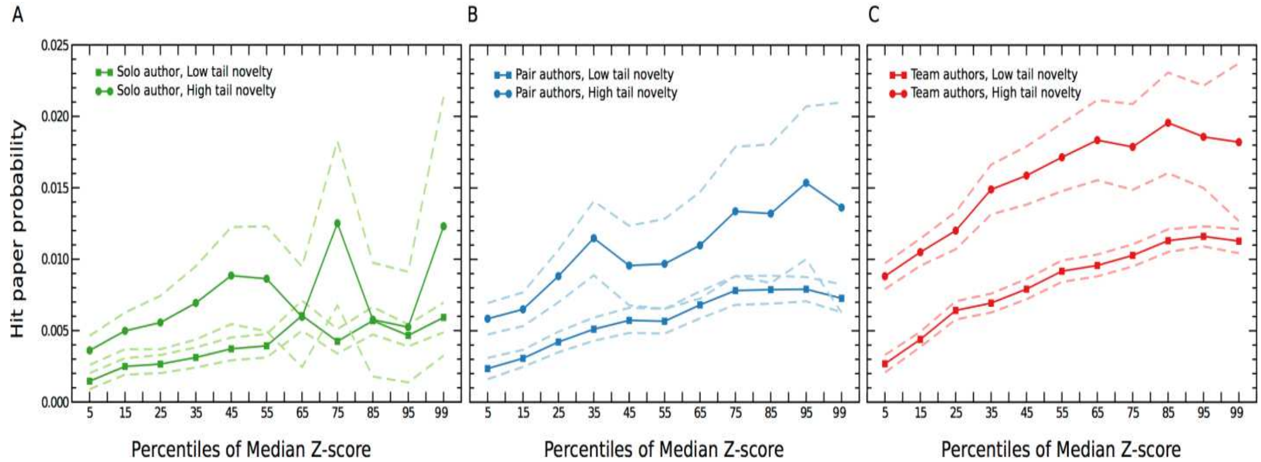

Fig. S10
Novel and conventional combinations, references < 10. This figure shows that results for the subsample of papers with fewer than ten references are broadly similar to the main text results shown in Figure 4(A-C). The noise for solo authors in cases with high tail novelty and high median z-scores reflects the small number of observations in those cases.

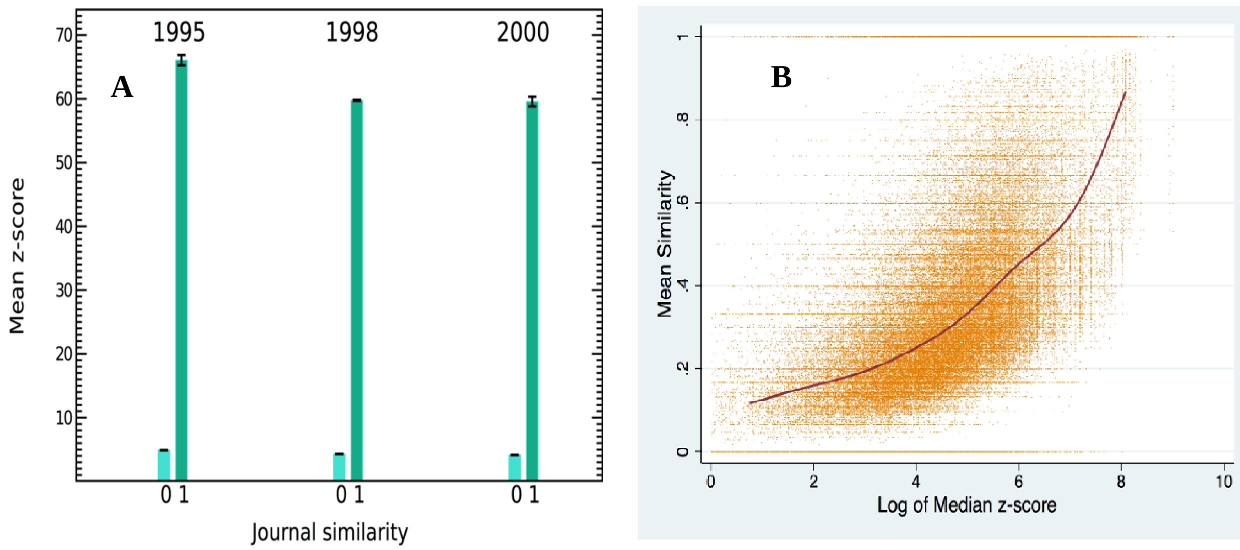

Fig. S11
Novelty, conventionality, and journal field similarity. In the left panel, the x-axis divides journal pairs into those that share a common WOS field designation (journal similarity = 1) and those that do not share a WOS field (journal similarity = 0). The y-axis shows the mean z-score (and indicated 95% confidence interval) within each set of journal pairs. We consider three different years (1995, 1998, and 2000). Journal pairs sharing a WOS field show high z-scores on average, indicating highly conventional combinations. Journal pairs that do not share a WOS field are on average much less conventional combinations; however, mean z-scores remain greater than zero, indicating that journal combinations from distinct WOS fields are on average actually not novel compared to chance. In the right panel, the x-axis presents the "median conventionality" of each paper, using the paper's median z-score. The y-axis indicates the mean journal similarity at the paper level, averaging the journal similarity variable across all the referenced journal pairs in a given paper. Data are for the year 1995. We again see the expected positive relationship between high conventionality and high field similarity.

Table S1. Examples of Journal Pair Frequencies for Illustrative Paper

Journal Pairs	Observed	Expected	Z-score
Tetrahedron - Tetrahedron	5071	151.89	637.77
Experientia - Experientia	1159	109.59	95.07
Tetrahedron - Experientia	454	256.06	21.55
Experientia - Tetrahedron Lett	661	481.07	6.88
Z-score of Zero means obs is as likely as chance			0.0
Chem Phar Bull - Life Sci	114	151.19	-2.4
Life Sci - R J Royal Neth C	16	45.45	-4.82
Life Sci – Tetrahedron	36	315.78	-17.67
Life Sci – J Organic Chemistry	166	813.72	-24.21
J Am Chem Soc - Life Sci	469	3147.65	-45.07

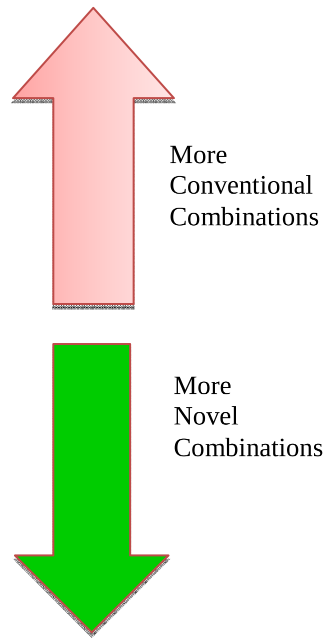

Table S2.

Novelty, convention, and citation impact by field. For each of 243 subfields indexed by the WOS in the 1990s, we rank the categories of papers according to their probability of producing hit papers. Hit papers are defined as those in the upper 5% of citations received in that subfield. We focus on all papers published across all subfields in the 1990s. This analysis reveals that high tail novelty and high median conventionality are the highest impact papers in 64.4% of subfields and either first or second in 86.3% of fields. By contrast, low tail novelty and low median conventionality rank lowest or second lowest in 87.4% of fields.

	Rank			
	1st	2nd	3rd	4th
High tail novelty and low median conventionality	20.3%	44.5%	28.7%	6.5%
Low tail novelty and high median conventionality	9.7%	26.7%	50.6%	13.0%
High tail novelty and high median conventionality	64.4%	21.9%	3.6%	10.1%
Low tail novelty and low median conventionality	5.7%	6.9%	17.0%	70.4%

Table S3.
Journal pairs from common and distinct WOS disciplinary designations. Using the 243 field categories in the WOS in the 1990s, this table divides journal pairs into those that share a common WOS field and those that do not. We then consider, by year, the mean z-score for each category of journal pairs and the percentage of journal pairs that are conventional (z-score > 0). We see that journal pairs from distinct WOS fields are much less conventional than journal pairs that share a WOS field. At the same time, observed journal pairs from different WOS fields are, in their majority, conventional combinations because some disciplines regularly publish together.

	Journal Pairs that do not share a common WOS field		Journal Pairs that share a common WOS field	
Year	**% Conventional**	**Mean z-score**	**% Conventional**	**Mean z-score**
1990	61.3%	4.89	92.3%	64.63
1991	61.3%	5.07	92.1%	66.34
1992	60.9%	4.95	91.9%	68.96
1993	61.3%	5.00	91.9%	68.42
1994	61.4%	5.05	91.7%	68.73
1995	60.7%	4.89	91.5%	66.07
1996	59.8%	4.63	91.3%	64.19
1997	59.4%	4.84	90.9%	62.13
1998	59.1%	4.30	90.8%	59.77
1999	58.5%	4.12	90.7%	59.33
2000	58.4%	4.14	90.6%	59.52