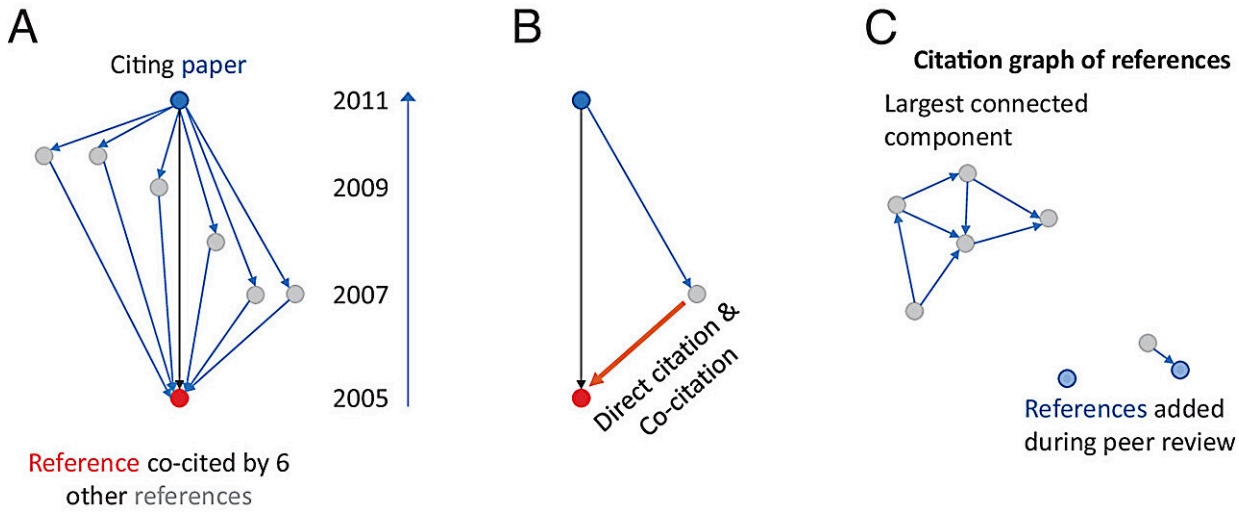
Machine learning analysis of research citations highlights importance of federal funding for basic scientific research

September 19 2023, by Chris Barncard



Local citation network features that may hold predictive power. (A) Illustration of references (gray) of a citing paper (blue) that all cocited the target referenced paper (red) found in the same citing paper's reference list. In this case, six other papers from this reference list (gray) cocited the referenced article (red), so the count is six. (B) Illustration of a direct citation that is also a cocitation. The blue paper cites both the other citing (gray) and referenced (red) articles, making this orange direct citation link also a cocitation. (C) The local citation network of the papers found in the reference list. Many referenced papers appear in the largest connected component, but in this illustration the two papers (blue) that were added to the reference list during peer review are not part of this component. Credit: *Proceedings of the National Academy of Sciences* (2023). DOI: 10.1073/pnas.2213697120

Biomedical research aimed at improving human health is particularly reliant on publicly funded basic science, according to a new analysis boosted by artificial intelligence.

"What we found is that even though research funded by the National Institutes of Health makes up 10% of published scientific literature, those published papers account for about 30% of the substantive research—the important contributions supporting even more new scientific findings—cited by further clinical research in the same field," says B. Ian Hutchins, a professor in the University of Wisconsin–Madison's Information School, part of the School of Computer, Data & Information Sciences. "That's a pretty big over-representation."

Hutchins and co-authors Travis Hoppe, now a data scientist at the Centers for Disease Control and Prevention, and UW–Madison graduate student Salsabil Arabi, published their findings recently in the *Proceedings of the National Academy of Sciences*.

Published research papers typically include lengthy sections citing all the previous work supporting or referenced within the study. "Predicting substantive biomedical citations without full text," the paper by Hutchins and Hoppe that you are reading about right now, cited no fewer than 64 other studies and sources in its "References" section.

Citations represent the transfer of knowledge from one scientist (or group of scientists) to another. Citations are extensively catalogued and tracked to measure the significance of individual studies and of the individuals conducting them, but not all citations included in any given paper make equally important contributions to the research they describe.

"We're taught that as scientists, when we make a factual claim, we're

supposed to back it up with some kind of empirical evidence," Hutchins says. "Like in Wikipedia entries, you can't have the little 'citation needed here' flag. You have to add that citation. But if that fact you're citing isn't actually describing key prior work that you built upon, then it doesn't really support the interpretation that the citation represents a necessary earlier step toward your results."

Hutchins and his collaborators figured citations added later in the publication process, like those that appear at the behest of peer reviewers—the subject-matter experts that evaluate scientific papers submitted to journals—are less likely to have been truly important to the authors' research.

"If you're building on other people's work, you probably identify that work earlier on in the research process," Hutchins says. "That doesn't mean all the references that are in an early version of the manuscript are important ones, but the important ones are probably more concentrated in that earlier version."

To make the early-late distinction, the researchers trained a machine learning algorithm to judge citations on their importance by feeding it citation information from a pool of more than 38,000 scholarly papers. Each paper's citation data came in two versions: a preprint version, posted publicly before peer review, and the eventual published version that had undergone peer review.

The algorithm found patterns to help identify the citations that were more likely to be important to each piece of published science. Those results revealed NIH-funded basic biological science appearing in the weightier citations at a rate three times the size of its share of all published research.

"Federal funding for basic research is under constant scrutiny from

members of the public and congressional leadership," Hutchins says. "This gives us some evidence, not just anecdotes, that this kind of basic research funding is really important for stimulating the kind of clinical research—treatments and cures for people—that Congress tends to be more receptive to funding."

More information: Travis A. Hoppe et al, Predicting substantive biomedical citations without full text, *Proceedings of the National Academy of Sciences* (2023). DOI: 10.1073/pnas.2213697120. www.pnas.org/doi/10.1073/pnas.2213697120

Provided by University of Wisconsin-Madison