ILLUSTRATION BY DAVID PARKINS

# Predicting scientific success

**Daniel E. Acuna, Stefano Allesina** and **Konrad P. Kording** present a formula to estimate the future *h*-index of life scientists.

We research scientists often worry about the future of our careers. Is our research an exciting path or a dead end that will end our careers prematurely? Predicting scientific trajectories is a daily task for hiring committees, funding agencies and department heads who probe CVs searching for signs of scientific potential.

One popular measure of success is physicist Jorge Hirsch’s *h*-index[1], which captures the quality (citations) and quantity (number) of papers, thus representing scientific achievements better than either factor alone. A scientist has an *h*-index of *n* if he or she has published *n* articles receiving at least *n* citations each[2]. Einstein, Darwin and Feynman, for example, have impressive *h*-indices of 96, 63 and 53, respectively. According to Hirsch, an *h*-index of 12 for a physicist — meaning 12 papers with at least 12 citations each — could qualify him or her for tenure at a major university.

However, the *h*-index[3] and similar metrics[4] can capture only past accomplishments, not future achievements[5]. Here we attempt to predict the future *h*-index of scientists on the basis of features found in most CVs.

We maintain that the best way of predicting a scientist’s future success is for peers to evaluate scientific contributions and research depth, but think that our methods could be valuable complementary tools.

The typical research CV contains information on the number of publications, those in high-profile journals, the *h*-index and collaborators. One can also infer interdisciplinary breadth, the length and quality of training, the amount of funding received and even the standing of the scientist’s PhD adviser. Such factors are taken into account for hiring decisions, but how should they be weighted? Fortunately, obtaining data on the scientific activities of individual researchers has never been easier. Using all of these features, we can begin to probe the scientific enterprise statistically.

## VITAL STATISTICS

To construct a formula to predict future *h*-index, we assembled a large data set and analysed it using machine-learning techniques. Our initial sample from academictree.org — a crowd-sourced website listing scientists’ mentors, trainees and collaborators — contains the names and institutions of about 34,800 neuroscientists, 2,000 scientists studying the fruitfly *Drosophila* and 1,300 evolutionary researchers. We matched these authors to records in Scopus, an online database of academic papers and citation data. We restricted our analysis to authors who had accrued an *h*-index greater than 4 (to exclude inactive scientists); to publications after 1995 (because electronic records are sparse before then); to authors who had published their first manuscript in the past 5–12 years; and to authors who were identifiable in Scopus.

That left us with 3,085 neuroscientists, 57 *Drosophila* researchers and 151 evolutionary scientists for whom we constructed a history of publication, citation and funding.

For each year since the first article published by a given scientist, we used the features that were available at the time to forecast their *h*-index a number of years into the future. For example, we reconstructed how the CV features of a scientist looked five years after publishing his or her first article, and found a relationship between those features and the reconstructed *h*-index five years on.

Starting with neuroscientists, we attempted to predict the *h*-index of each scientist 5 years ahead — a timescale relevant for tenure decisions — using a linear regression with elastic net regularization[6] (see Supplementary Information at go.nature.com/mtvuzr). The ▶

➲ **NATURE.COM**
For more on science metrics, see:
go.nature.com/nj2xqk

## METRICS

### *Predict your future* h*-index*

These are approximate equations for predicting the *h*-index of neuroscientists in the future. They are probably reasonably precise for life scientists, but likely to be less meaningful for the other sciences. Try it for yourself online at go.nature.com/z4rroc.

- **Predicting next year** ($R^2=0.92$):

$$h_{+1}=0.76+0.37\sqrt{n}+0.97h-0.07y+0.02j+0.03q$$

- **Predicting 5 years into the future** ($R^2=0.67$):

$$h_{+5}=4+1.58\sqrt{n}+0.86h-0.35y+0.06j+0.2q$$

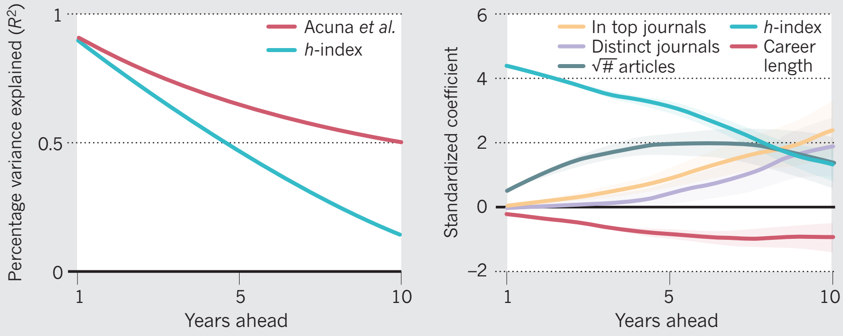
- **Predicting 10 years into the future** ($R^2=0.48$):

$$h_{+10}=8.73+1.33\sqrt{n}+0.48h-0.41y+0.52j+0.82q$$

Key: *n*, number of articles written; *h*, current *h*-index; *y*, years since publishing first article; *j*, number of distinct journals published in; *q*, number of articles in *Nature*, *Science*, *Nature Neuroscience*, *Proceedings of the National Academy of Sciences* and *Neuron.*

## PATHS TO SUCCESS

The accuracy of future *h*-index prediction decreases over time, but the Acuna *et al.* formula predicts future *h*-index better than does current *h*-index alone (left). The contribution of each factor to the formula accuracy also changes over time (right). Shading indicates 95% confidence error bars.



model predicted the future *h*-index accurately, yielding a respectable $R^2=0.67$, cross-validated across scientists (an $R^2$ of 1 would imply that the model predicts the data perfectly). A simplified model containing only the number of published articles, the *h*-index, years since first publication, number of publications in prestigious neuroscience journals (*Nature*, *Science*, *Nature Neuroscience*, *Neuron* and the *Proceedings of the National Academy of Sciences*) and the number of distinct journals still performed nearly equally well ($R^2=0.66$; see 'Predict your future *h*-index').

Predicting the future careers of *Drosophila* and evolutionary scientists leads to somewhat worse predictions ($R^2=0.54$ and $R^2=0.61$, respectively, based on scientists 3–15 years into their careers) but still better than predictions based on the *h*-index alone ($R^2=0.38$ and $R^2=0.39$, respectively). This indicates that generalizations to other fields within and outside of life science may be limited[1]. But for neuroscientists, at least, the predictions extend well to longer periods of time, such as ten years into the future ($R^2=0.52$). Over time, using just the *h*-index performs much worse than taking all features into account (see 'Paths to success', left panel).

The main five predictive features change in importance for predicting *h*-indices over increasingly longer periods (see 'Paths to success', right panel). The power of the *h*-index declines. The number of articles written, the diversity of publication in distinct journals and the number of articles published in five prestigious journals all become increasingly influential over time.

### FUTURE FORTUNES

It is risky to make any causal interpretations of these results. However, we will briefly speculate on why these features might be important predictors of future success. Some features directly affect the potential for a high *h*-index, such as the number of articles written. These features can also indirectly affect a scientist's future success, because scientists who are productive and publish many papers tend to remain productive. Publishing in many different journals may lead to fewer overlapping populations of scientists who cite the work, and hence higher growth potential for articles. A scientist who has published in several distinct journals is also likely to be someone with broad training who contributes in many ways. The number of publications in leading journals can increase the visibility of a scientist's other papers, past and future.

If promotion, hiring or funding were largely based on indices (*h*-index, the model used here or any other measure), then some scientists would adapt their behaviour to maximize their chances of success. Models such as ours that take into account several dimensions of scientific careers should be more difficult for researchers to game than those that focus on a single measure.

Our formula is particularly useful for funding agencies, peer reviewers and hiring committees who have to deal with vast numbers of applications and can give each only a cursory examination. Statistical techniques have the advantage of returning results instantaneously and in an unbiased way. Building and analysing massive data sets to track scientific careers could also help to identify potential gender, racial and other biases[7–9] and advance our understanding of how science develops.

Although our findings and predictions may not alleviate scientists' angst over their careers, the results offer some comfort by showing that the future is not so random. The occasional rejection of a paper may feel unjust and indiscriminate, but in the long run, such factors seem to average out, rendering *h*-index trajectories relatively predictable. ■

**Daniel E. Acuna** *is a research associate at the Rehabilitation Institute of Chicago, Illinois 60611, USA, and a research affiliate in biomedical engineering at Northwestern University, Evanston, Illinois 60208, USA.*
**Stefano Allesina** *is assistant professor in ecology and evolution and at the Computation Institute at the University of Chicago, Illinois 60637, USA.*
**Konrad P. Kording** *is associate professor of physical medicine and rehabilitation, physiology, and applied mathematics at Northwestern University, and at the Rehabilitation Institute of Chicago.*
*e-mail: daniel.acuna@northwestern.edu*

1. Hirsch, J. E. *Proc. Natl Acad. Sci. USA* **102,** 16569–16572 (2005).
2. Redner, S. *J. Stat. Mech. Theory Exp.* **3,** L03005 (2010).
3. Peterson, I. *ScienceNews* 2 December 2005; available at http://go.nature.com/iawd5o.
4. Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E. & Herrera, F. *J. Informetr.* **3,** 273–289 (2009).
5. Hirsch, J. E. *Proc. Natl Acad. Sci. USA* **104,** 19193–19198 (2007).
6. Zou, H. & Hastie, T. *J. Roy. Stat. Soc. B* **67,** 301–320 (2005).
7. Dwan, K. *et al. PLoS ONE* **3,** e3081 (2008).
8. Ginther, D. K. *et al. Science* **333,** 1015–1019 (2011).
9. Allesina, S. *PLoS ONE* **6,** e21160 (2011).