Examining gender imbalance in chemistry authorship

Adam D. Cotton and Ian B. Seiple

Abstract

The gender gap in chemistry has been the topic of much debate, with many perspectives stating that the field has improved, and outdated sexist views are behind us. While these views are common, we wanted to assess the accuracy of these comments from a data-driven perspective. In this study, we use Pubmed to obtain the names of all first and last authors for every paper published since 2005 across 15 journals. Each name was assigned a gender using a name-based gender predictor. We show that historically there has been an extensive imbalance of gender for both first and corresponding authors, and that there is no strong trend towards parity since 2005. We demonstrate that female corresponding authors are more likely to be equitable in the gender of their first authors. Finally, we find that there is significant variability among journals in the gender make-up of their editorial boards. We hope this analysis spurs creative discussions on how we can improve female representation in chemistry publications.

Introduction

In recent weeks there has been extensive conversation regarding gender diversity in chemistry, spawned in part by an article published (and then removed) in *Angewandte Chemie* entitled "'Organic synthesis–Where now?' is 30 years old. A reflection on the current state of affairs". While numerous commentaries since its publications have stated our field has improved and outdated sexist views are behind us, we wanted to assess how accurate these perspectives are using a data-driven approach. Our goal is to provide data to the community to assist in constructive conversations about how we can progress. There have been numerous validated studies demonstrating how diversity in teams improves creativity and productivity, so we will not labor this point[1,2]. We simply ask that people read our work without preconceived opinions and to remind themselves that there is an equal number of males and females in the general population, and any deviation from that norm is caused by systemic biases.

We wrote a python-based web scraper to extract the first (given) name of the first and last (corresponding) authors from papers published on PubMed each year since 2005. 2005 was chosen due to the lack of quality data in prior years. We focused our efforts on organic chemistry journals, the closely related chemical biology field, and Nature, Cell, Science (NCS). In addition to authors' names, we obtained the country of the corresponding author and assumed this would be the same for the first author. With data lists of names and countries, we utilized Gender API[4], the leading name-based gender predictor to assign a gender to each author in the list. Due to the binary nature of the API used, we could only account for two genders: male and female. We desired to include other genders in the analysis; however, we could not find a method to include them. There were occasions where the API couldn't interpret the name, so these instances were removed from future analyses. Gender API outputs an accuracy score and sampling size, we set strict data quality requirements that were used to calculate population parameters. Any years for a journal with a high-quality data set of less than 25 were removed from any further analyses. For all graphical figures, individual journal graphs can be found in the supporting-information. All python scripts used are available on github[5].

Results

The first question we assessed was if there has been a gender disparity for first authors and for corresponding authors, cumulatively over the past 16 years, and if there were any dissonance between different journals. It quickly became clear that there has historically been exceedingly low representation of females in Chemistry (Figure 1). In *Journal of the American Chemical Society* (JACS*)*, a leading chemistry journal, females have been represented 9.5% of corresponding authors and 21.4% of first authors (Figure 1 a,b). These values were consistent across the different chemistry journals. Interestingly medicinal chemistry journals such as *Journal of Medicinal Chemistry* (J Med Chem) had a higher female representation, while better, they still only reached 15.7% for corresponding authors and 29.9% for first authors. These values may arise due to the many different contributing factors to authorship in academia versus industry. Due to this difference, medicinal chemistry journals were separated from the other traditional chemistry journals. Chemical biology is a closely related field to Chemistry, so much so that many departments and graduate programs have fully integrated the two fields. ACS Chemical Biology had the highest female representation with 18.4% of corresponding authors and 34.6% of first authors. It is important to note that these values still are strikingly far from the 50/50 ratio seen in the general population. Interestingly the NCS journals separated in two groups throughout our analyses, with Nature and Science closely correlating with each other and Cell showing male to female ratios that were closer to parity. Finally, it is apparent there is a large disparity between first authors and corresponding authors, which tracks with faculty and student gender percentages; 39% of chemistry graduate students and only 12% of faculty are female. This aligns with the documented decline of females in science between PhD and faculty jobs[3].

Figure 1: Rank-ordered gender breakdown of first and corresponding authors for all papers published since 2005. A) Gender breakdown of first authors for papers published in chemistry, chemical biology and NCS journals since 2005. B) Gender breakdown of corresponding authors for papers published in chemistry, chemical biology and NCS journals since 2005. *Name-predicted gender–see methods section for details.

Next, we plotted the percentage of male and female corresponding authors for each journal since 2005. Given the results from the collated data we separated the journals into chemistry (figure 2a), medicinal chemistry (figure 2b), chemical biology (figure 2c) and NCS (figure 2d). Dashed lines on the graphs represent the US male and female academic faculty percentages in synthetic chemistry in 2020 (n = 518). There is a clear lack of trend across all journals towards parity over

the past 16 years, suggesting gender bias in chemistry is not a thing of the past. *Journal of Medicinal Chemistry* had the highest representation of females of all journals, with a 23.7% female corresponding author percentage so far for 2020.

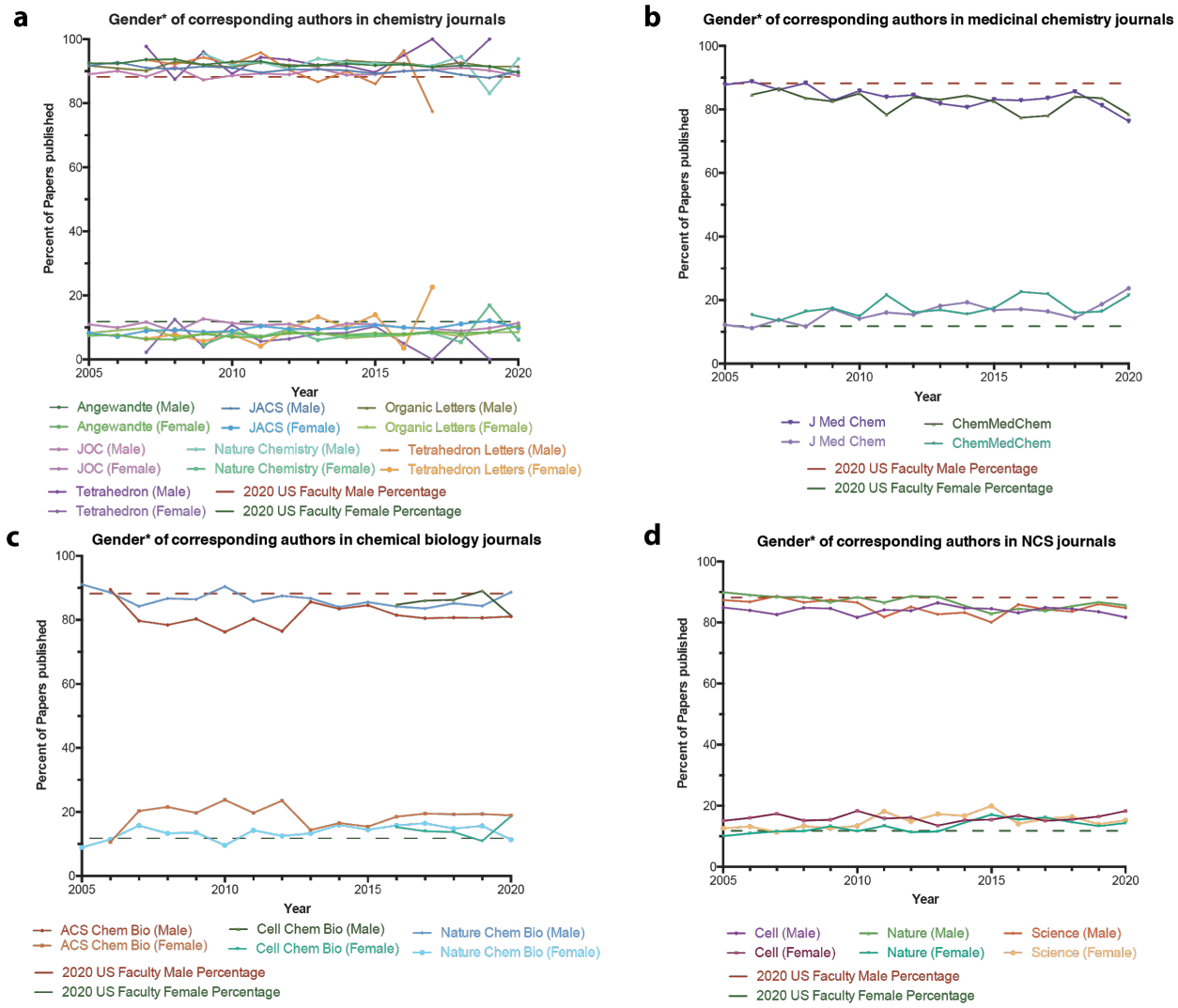

Figure 2: Gender percentages of corresponding authors for different journal types per year since 2005. Dashed lines represent current (2020) faculty percentages in US chemistry departments. A) Chemistry Journals. B) Medicinal Chemistry Journals. C) Nature, Cell, Science. D) Chemical Biology Journals. *Name-predicted gender–see methods section for details.

We completed the same analysis as figure 2, but for first authors. The dashed lines here represent NSF self-reported gender percentages for US graduate student and postdoctoral fellows combined. The gender percentages have stayed steady over the past 16 years (Figure 3). Similar to corresponding authors, chemistry journals perform the worst in this metric (Figure 3a). Both medicinal chemistry and chemical biology first author percentages have tracked closely with the percentages seen in graduate schools, suggesting male and females are equally likely to receive first authorship in these fields.

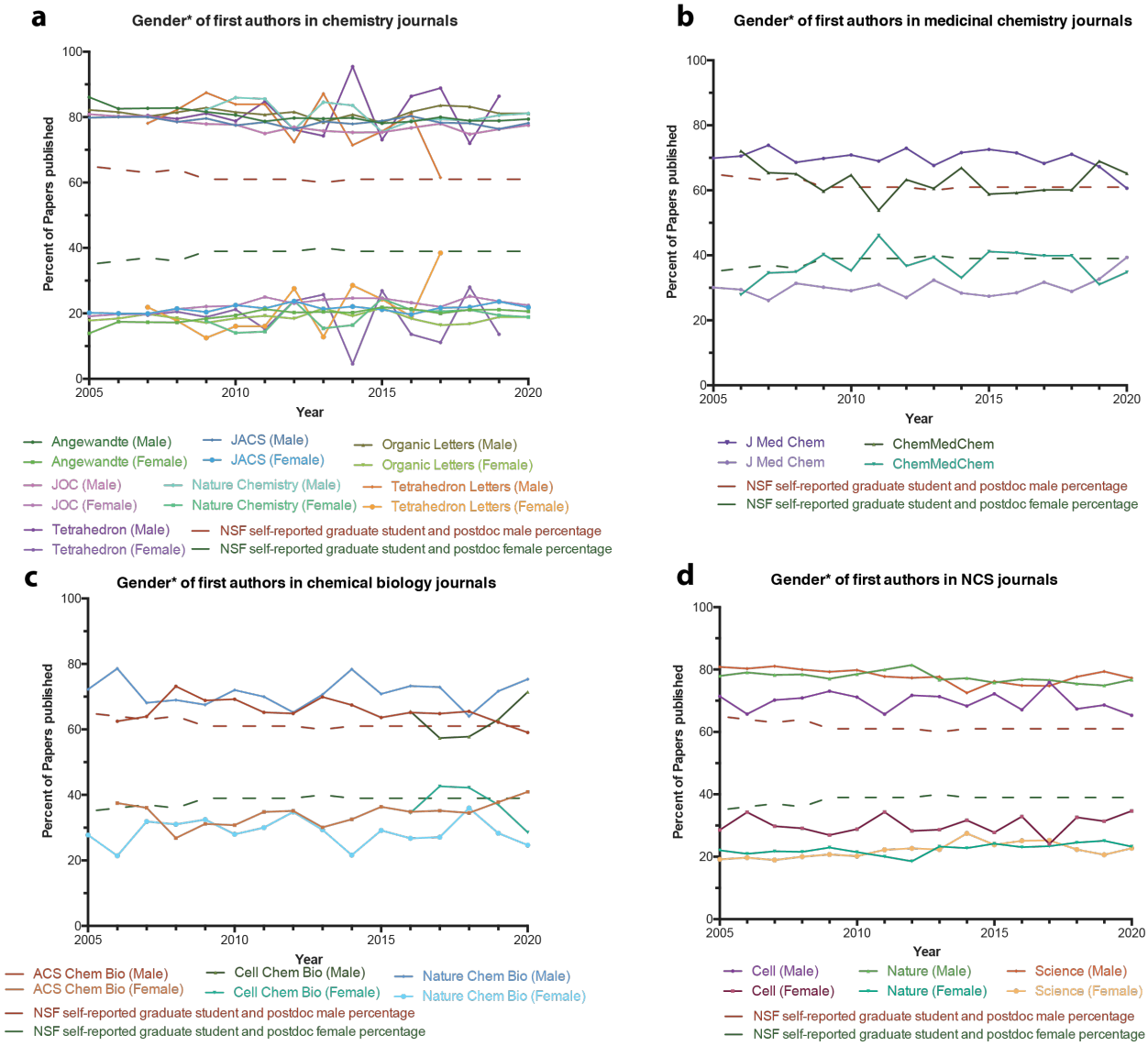

Figure 3: Gender percentages of first authors for different journal types per year since 2005. Dashed lines represent the NSF self-reported gender percentages of US graduate students and postdoctoral fellows. A) Chemistry Journals. B) Medicinal Chemistry Journals. C) Nature, Cell, Science. D) Chemical Biology Journals. *Name-predicted gender–see methods section for details.

The final analysis we conducted on our data sets, was to assess if first authors were more likely to be male or female depending on the gender of the corresponding author. We show four representative journals in Figure 4, and the remaining data can be accessed in the Supporting Information. Across all journals first authors were more likely to be female if the corresponding author was female versus male. In numerous cases including all chemical biology journals female corresponding authors had multiple years at 50:50 first author ratios. Neither gender shows an improvement in their first author percentages over time, suggesting any progress towards parity in total first author ratios (Figure 3) will be due to the gradually increasing number of female faculty.

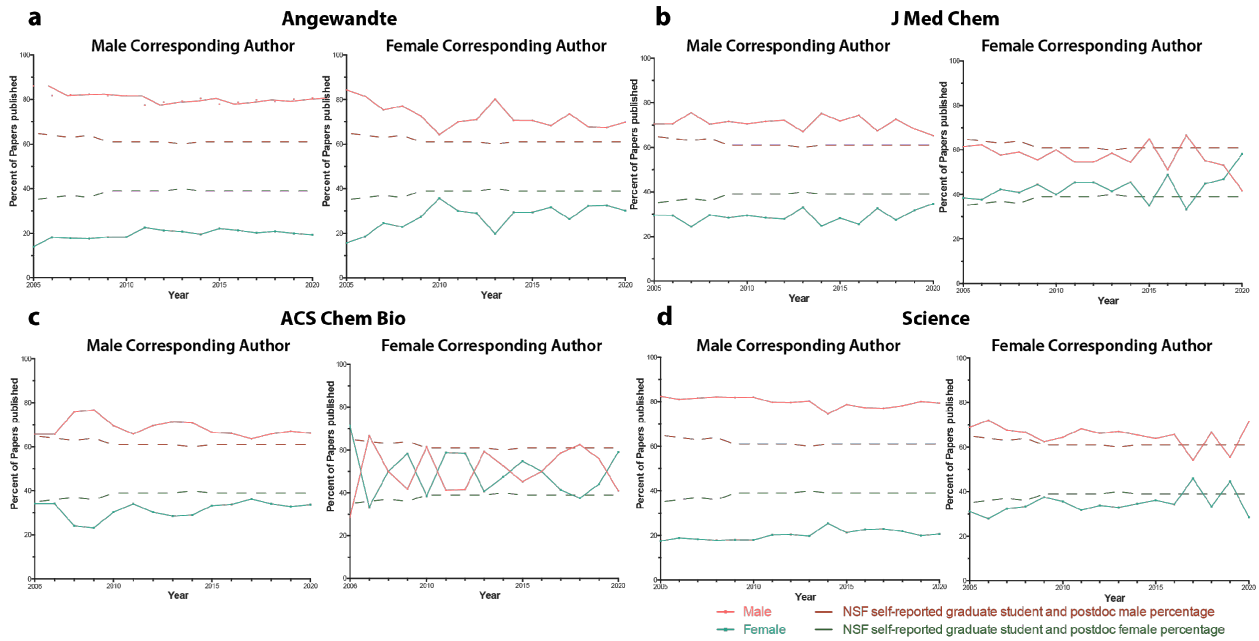

Figure 4: Gender* percentages of first authors for both male and female corresponding authors. Dashed lines represent the NSF self-reported gender representation of US graduate students and postdoctoral fellows. a) First author gender percentage in Angewandte with either a male (left) or female (right) corresponding author. B) First author gender percentage in JMedChem with either a male (left) or female (right) corresponding author. C) First author gender percentage in ACS Chemical Biology with either a male (left) or female (right) corresponding author. D) First author gender percentage in Science with either a male (left) or female (right) corresponding author. *Name-predicted gender–see methods section for details.

In addition to authorship data, we were interested in gender representation on journal editorial boards. We manually curated journal websites to assign gender percentages. Due to the wide breadth of topics covered in Nature and Science, the editorial boards for NCS were not analyzed. As expected, the editorial gender ratios are far from equitable (Figure 5). Tetrahedron and Tetrahedron Letters have journal advisory boards of over sixty members in which there is less than 10% female representation. Nature Chemistry and Nature Chemical Biology both have parity in their boards, but it should be noted that these journals differ from the others in that they have a small team of full-time editors.

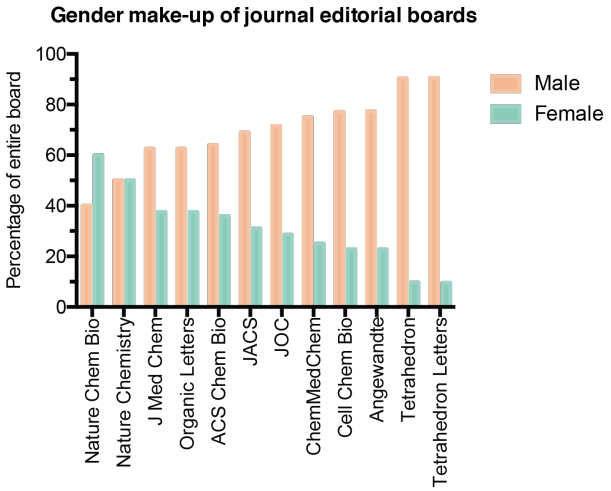

Figure 5: Rank-ordered gender ratios of journal editorial boards in 2020. Data collected from respective journal websites.

Conclusion

Some salient trends emerged in our dataset:

1. Chemistry journals have a large disparity in gender ratios, especially for corresponding authors.
2. Even the best performing journal with respect to gender parity ACS Chemical Biology only achieved male to female ratios of 65:35 and 82:18 for first and corresponding authors respectively.
3. In the past 16 years there has been no increase in the representation of females as either corresponding or first authors.
4. There is a large drop off in female representation between graduate school and faculty positions.
5. First author representation is more likely to be equitable when the corresponding author is female.
6. With few exceptions, the gender ratios of editorial boards heavily favor men.

It is clear that, despite recent publications on the topic, there is still a large problem in our field that needs to be addressed. The goal of our analyses is to provide data on authorship which we hope will foster further discussion about how to improve gender representation in chemistry. Open discussion on the topic of diversity yields creative ideas and we hope our data allows informed conversations about how we as a community can improve female representation in chemistry and the wider sciences.

Methods

On June 10th 2020 a web-scraper was used to search PubMed for all papers published in a journal each year since 2005. For journals that were created post-2005, their date of creation was used as the first year. The scraper pulled the first name of both the first and corresponding authors. The country of the corresponding author was extracted, and this was assumed to be the same for the first author. The country information improves the accuracy of the gender predicting algorithm. Any papers that have a solo author were discarded from the data. This resulted in 196,157 articles. The data sets containing a first name and country were uploaded to Gender API to predict the gender of the author based on their first name. Gender API outputs a predicted gender, an accuracy score out of 100 and data size sampled. The accuracy score is a probability of the gender being correctly assigned, thus any data points below 95% confidence or small sample size (below 100, with confidence below 100) were removed from the data set. This resulted with a set of high-quality data points for 296,245 authors. A histogram demonstrating the quality of the data set is found in supporting information. Any years for a journal with less than 25 quality data points was removed from any time-series analysis. Each author was then assigned a binary gender and population totals calculated.

NSF self-reported data were used to calculate the US graduate student and postdoc gender percentages. These data sets are available for 2005-2014; post-2014 values were extrapolated from the publicly available data.

ACS Division of Organic Chemistry has a curated list of 518 current US-based faculty who work on synthetic organic chemistry. Similar to above, a web-scraper was used to extract first-names, and Gender API was used to assign a name-predicted gender.

Journal websites were used to obtain their editorial boards. These lists were manually extracted, and board members' genders manually assigned based on lab websites and publicly available data.

Our approach has some inherent limitations. The use of a gender API is the best high throughput method to address the disconnect between an extensive list of authors and their gender at the scale required for this study. Due to the limitations of first name-based gender APIs, we could only account for two genders: male and female. We desired to include other genders in the analysis; however, we could not find a method to include them. Additionally, we used the affiliated country of the last (corresponding) author to inform gender assignment by the API, which uses country to improve gender prediction. There are likely cases in which that affiliated country will not reflect the origins of the names of the first or last authors. Finally, there are many cases where there are multiple first or multiple corresponding authors, and, more rarely, cases where the last author is not a corresponding author. Our approach does not account for these cases. Due to the data quality standards we set, gender neutral names were eliminated from the data sets.

Acknowledgements

We thank Etan Schwartz for help with analyzing the data sets, James Fraser and Seemay Chou for helpful discussion.

References:

1. Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N. & Malone, T. W. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* **330**, 686–688 (2010).
2. Joshi, A. By Whom and When Is Women's Expertise Recognized? The Interactive Effects of Gender and Education in Science and Engineering Teams. *Administrative Science Quarterly* **59**, 202–239 (2014).
3. Finkelstein, M. J., Conley, V. M. & Schuster, J. H. Taking the measure of faculty diversity. *Advancing Higher Education*, TIAA Institute (2016).
4. https://gender-api.com/
5. https://github.com/adamcotton/gender-scraper