Scientometrics
https://doi.org/10.1007/s11192-021-04057-z

Authorship and citation cultural nature in Density Functional Theory from solid state computational packages

Marie Dumaz[1,2] · Reese Boucher[1] · Miguel A. L. Marques[3] · Aldo H. Romero[1]

Received: 11 October 2020 / Accepted: 19 May 2021

Abstract
Density functional theory is the most used methodology in the characterization of the electronic structure of materials. Its applications have spread out to almost every STEM field and it is recognized as one of the most successful theories in materials science. In this paper we measure the specific impact of this theory by means of the citation record of the most important solid-state first principle *ab initio* packages. We report the exponential growth of publications and how the different electronic structure packages are supporting different scientific communities. Analysis of the growing community, relations between different communities, network strength, relation between citations and number of publications with respect to country of origin of the authors, number of authors per paper, words per title and publication journal is performed. We make several interesting observations, e.g., regarding the connection between the countries where the packages are developed and used, or concerning the collaboration networks. We also find bibliometrical evidence for the specialization of the software packages, even if they include similar capabilities.

Keywords Density Functional Theory · Country collaboration network · Citation impact analysis · Citation analysis

Introduction

The field of bibliometric analysis has been invigorated in the last few years due to the creation of many different bibliographic databases. Based on those databases, authorship analysis of research articles has been performed by many different scholars, in a large set of different research fields. In particular, as a metric to quantify the research output and achievements of the scientific community (Peters Van Raan 1991; Chow

This work was supported by NSF SI2-SSE Grant 1740112, DMREF-NSF 1434897, DOE DE-SC0016176 and DE-SC0019491 grants.

✉ Marie Dumaz
mcd0029@mix.wvu.edu

[1] Department of Physics, West Virginia University, Morgantown WV 26506, USA

[2] Lane Department of Computer Science and Electrical Engineering, Morgantown WV 26505, USA

[3] Institut für Physik, Martin-Luther-Universität Halle-Wittenberg, 06120 Halle (Saale), Germany

et al. 2015). Examples of this approach include the assignment of authorship credit (Hagen 2008), research performance (Moed 2006), the classification of journal impact factors (Garfield 1999; Amin and Mabe 2004), the definition of trends in collaborative research (Arya and Sharma 2011), the evaluation of international scientific collaboration (Glänzel et al. 1999), national and university trends in publication output and the existence of scientific networks (Barabási et al. 2002).

The bibliometric analysis also helps in identifying scientific growth and work recognition. With different bibliographic indices now used to classify research output, analysis of authorship and citations, it is now becoming more frequent to assess the impact, quality, or the development of a scientist, a university, a field, or a subset of researchers.

In this work, we analyze the research efforts in the field of electronic structure calculations and in particular, the use of density functional theory (DFT) (Hohenberg and Kohn 1964; Kohn and Sham 1965) to characterize solid-state materials at the atomic scale. DFT is the workhorse of electronic structure, and is used in the vast majority of works that use computational methods to characterize materials. While the theory is very mature, we expect to have a deeper understanding of the author's dynamics in this field by studying the related publication record.

While twelve DFT related papers are ranked in the one hundred most cited papers of all time (Van Noorden et al. 2014), the original papers, where the theoretical framework was developed, are not ranked near the top. The most cited papers reported by Van Noorden *et al.*, and connected to DFT, are from the Canadian chemist Axel Becke and the US-based theoretical chemists Chengteh Lee, Weitao Yang, and Robert Parr, in positions 8 and 7 respectively (Van Noorden et al. 2014). However, we should point out that the classification presented in reference (Van Noorden et al. 2014) is outdated. By May 2020, the development of the so-called PBE exchange correlation functional (Perdew 1996) has more citations than the theoretical works included in the one hundred most cited papers (Van Noorden et al. 2014).

Hence, the two DFT-related papers written by Becke & Lee et al, as well as the PBE exchange correlation functional by Perdew et al all have more citations than the original papers presenting DFT. This is an indication of the "obliteration by incorporation" (McCain 2014) concept: the theory of DFT has become so universally accepted and used that the original papers are not always required to be cited anymore.

To circumvent the lack of citations of the original papers where DFT has been introduced, we use the following premise: instead of studying the citations of the original DFT papers, we analyze the citations of the 31 most used solid state and electronic structure computational packages which have DFT as the underlying theory. The complete list of selected software is available in the Methodology section. This premise overcomes the problem that the citations record can be reduced in time when the basic idea is so spread out that citations to the original work are not required or used in research publications anymore.

While it is true that the number of publications citing the original papers is still exponentially growing, minimizing the case for obliteration by incorporation, it is to a lesser extent than software citations. Figure 1 displays the percentage difference between the number of publications citing the original theoretical work of DFT or an electronic structure package. We can see a clear growth in percentage difference, obtaining even 375% in 2019, meaning that even though the original papers were mainly cited before the 2000s, and that some authors are still citing them nowadays, the majority of authors do not anymore. Most of the electronic structure codes started to be developed after the

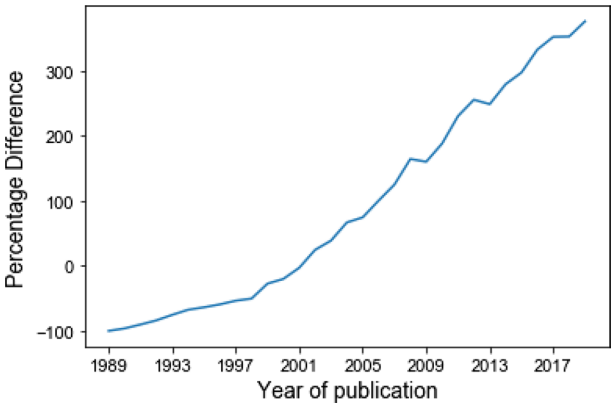

Fig. 1 Percentage difference per year between the number of publications citing at least one software package and publications citing the original DFT papers

2000s, which seems to indicate that authors have now the tendency to cite the packages without acknowledging the theoretical work behind DFT.

This is a clear indication that using the original papers does not provide a general idea about how the scientific electronic structure community works. In that respect, software users who cite the original computational implementations define clearer citation networks on which to perform an analysis in the field of electronic structure methods. As a publication accompanies any computational package, and in most cases, users are advised to cite these papers when the package is used, we expect that users' citations are more representative of the field.

Bibliometric impact of DFT has already been approached by other research groups, for example by classifying the different topics from a historic perspective (Haunschild et al. 2019), a materials focused analysis (Haunschild et al. 2016), from the marker paper approach or field classification (Haunschild and 2020). These publications all have in common the use of subclassifications created by databases or the identification of common topics. Our approach uses a different methodology as we base our analysis on the computational packages that are used to study the materials, independent of the field of application.

Methodology

We obtained our citation database from the Institute for Scientific Information (ISI), which is a part of Clarivate Analytics, and that has the graphical interface "Web of Science" (WOS). This database is very accessible and with a simple graphical interface, but it has also been criticized as its entries mostly come from journal titles in English (though it is the predominant language in scientific literature) and covers very few citations in books, web links and conference proceedings. This is a difference with respect to other databases such as Scopus, Citeseer, or Google Scholar. Though this lack of references could be a handicap for the proposed analysis, the fact that the citation list is more controlled with respect to other databases allows us to extract more general conclusions on trends and author behavior. The bibliographic search was performed since the first citation up to December 2019. In particular, we study the clustering of authors, and the impact of publications per author and per country of origin in the most popular journals. We also perform network analysis to describe the dynamical interaction between the code developers and

the users of the different packages. Here, we define the network such that vertices refer to a given country and are linked by edges to other affiliated countries present in one publication. This is also called the country collaboration network analysis.

As we focus on the impact of electronic structure calculations from the user perspective, we selected the most well-known density functional theory packages used to study crystalline systems (independently from their use of a free or commercial license). These computational codes are all mostly used to describe periodic (crystal/amorphous/nanostructure) systems. From the large variety of packages we decided to center our analysis on the six most cited packages: VASP, Castep, Quantum Espresso, Siesta, ADF and Turbomole. Though Turbomole and ADF are mostly used in the chemistry community, we still find that a large part of their application is related to periodic systems. On top of these six packages, we included all publications citing at least one of the following codes: Abinit, BigDFT, Casino, Conquest, CP2K, Crystal, Dacapo & ASE, Empire, EPW, Exciting, FHI-aims, FreeON, GPAW, JDFTx, NWChem, Octopus, Onetep, PySCF, QBOX, QMCPack, QuantumATK, RMG, TransSiesta, Wien2K and Yambo.

The citation metrics were acquired by searching the titles of all of the papers associated with each software package. After locating each paper, the option "Times Cited" was selected in order to see all of the occasions other people had cited the specific paper. The whole list of citations was downloaded with all possible entries created by WOS. The exact references considered for any of the codes are available in table 1 of the Supplementary. Some references, such as books or conference proceedings, could not be added as we only downloaded references through the "Basic Search" option of WOS. An example of this limitation would be the Wien2K manual (Blaha et al. 2001), that is recommended for reference by the Wien2K developers but can only be found through the "Cited Reference Search" option.

Only a few computational libraries exist that can read those WOS files. Due to the flexibility this package allows for the analysis, we selected Bibliometrix (Aria and Cuccurullo 2017), which works on R. Although it is an exceptional package when using already existing functions on the whole data set, some more precise analyses were easier to make and visualize with python tools. Hence, the dataframes created with the function readFiles from the Bibliometrix (Aria and Cuccurullo 2017) library were converted into Comma Separated Values (CSV) files that were used with our in-house Python script. Based on this script, a python package was developed to create the figures used in this paper and it is available to download (Dumaz 2020). However, before exporting the dataframes to files, some pre-processing steps were taken. First, duplicates, identified using the unique tag of each publication, were removed. Then, countries were extracted from the "C1" column using the method MetaTagExtraction in the Bibliometrix (Aria and Cuccurullo 2017) library.

The CSV files were opened in Python and manipulated through Pandas (McKinney 2010), Numpy (Stéfan van der Walt et al. 2011; Oliphant 2006) and Matplotlib (Hunter 2007) to create some of the figures. Before each computation, rows were removed from the original dataframe if they had null values in the columns of interest. This option was chosen as we had enough data and removing those rows would not substantially affect results. In the end, very few entries were removed: for example, out of 61 640 VASP citations, only 87 did not have any affiliated countries.

For any text analysis, the titles were pre-processed as follows: all symbols, punctuation and numbers were stripped, as well as stop words given by default in the nltk (Bird et al. 2009) package in Python. Then, the titles are tokenized to form a list of words. To compute the number of words per title, the normalization stage stops here. However, to find the most

common words, lemmatization was applied to each word. This technique finds the root of a word. For example, both the words "studies" and "studying" become "study".

Country collaboration networks were created and visualized using Bibliometrix (Aria and Cuccurullo 2017). Networks were generated from datasets where duplicates and observations with no address (no affiliated country) were removed. The method biblioNetwork creates a sparse matrix B where $B[i, j]$ is the number of publications for country i and j. Each node represents a country and there exists an edge between two countries if and only if they both collaborated on the same publication. The method networkPlot provides a visualization of the network where nodes are colored by clusters and distanced by the number of collaborations.

For the network analysis, we used different observables to characterize the network properties, in particular:

- *Density* This is simply the ratio of actual edges in the network to all possible edges in the network. In the undirected network defined for the citation bibliographic record for each DFT package, there could be a single edge between any two nodes, but seen in the visualization, only a few of those possible edges are actually depicted. Network density gives a quick sense of how closely knit a network is. This value ranges between 0 and 1.
- *Diameter* This is the longest of all shortest paths. This number provides a sense of the network's overall size, the distance from one end of the network to another
- *Transitivity* like density, expresses how interconnected a graph is as a ratio of existing triad connections with all possible triad connections. The idea is that if A is connected to B and B is connected to C, what is the probability that A is connected to C. Transitivity is scaled to the range from 0 to 1. When a graph is not very dense, there are fewer possible triangles (or triads) to begin with, which may result in slightly higher transitivity. That is, nodes that already have many connections are likely to be part of these enclosed triangles.

Results

Figure 2 shows the evolution of the research productivity in the field of density functional electronic structure by using the citations obtained from the most important computational packages, specialized mostly on periodic systems. The top panel shows a strong growth in the number of publications with respect to publication year, with more than 14 000 publications in 2019. This demonstrates that computational DFT is a healthy methodology and an active research field. The best fit for this growth is a powerlaw with an exponent of 2.82 (an exponential fit gives a coefficient of 0.13 but with lower r^2). Interestingly, the total number of citations per year (only until 2015) also follows a powerlaw dependence with an exponent of 1.41 (see Supplementary Fig. 1).

By looking at the number of different affiliated countries for each citing article, we find that the vast majority of publications are from a single country. Moreover, papers produced in three countries or more are uncommon and represent only about 9.4% of the dataset. Therefore, this field seems to be well geographically localized, while international endeavors probably come from large collaborations such as experimental/theoretical groups or code development teams. This is supported by an exponential fit with a strong decay (with an exponent of 0.64). Similar to the analysis of the number of publications, citation records

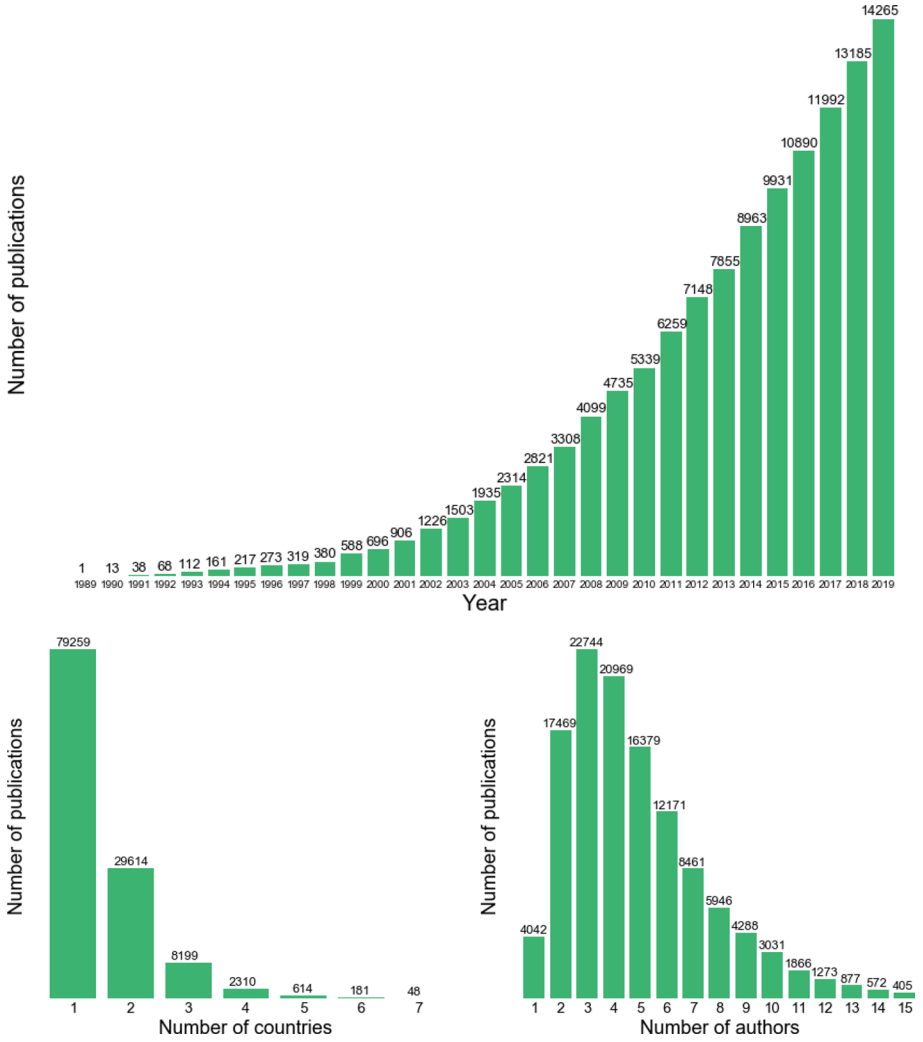

Fig. 2 Total publications (repeated entries removed) citing at least one entry of the selected group of DFT electronic structure packages. Top: total number of publications per year. Bottom left: Total number of publications as function of the number of different countries listed in the authors' institutions. Bottom right: Total number of publications as function of the number of different authors present in the paper

are larger for papers performed in the same country, although the exponential decay is slower than for the publications case.

The bottom right panel in Fig. 2 shows the number of publications with respect to the number of authors in the publication. A skewed Gaussian like behavior is observed with a maximum at three authors but with a very slow powerlaw decay (with a mean value of 1.62 and standard deviation of 3.83). This slow dependence demonstrates the efforts of the community to produce publications with a large number of authors. In other fields, such as in medical sciences, social sciences and other natural sciences, the number of authors has a maximum between 4 and 10 authors (Larivière et al. 2015); however, this is not the case in

electronic structure. With this analysis, we can conclude that there is no evidence of hyper-authorship. It is further supported by the fact that only 13 out of the 121 532 papers in our database have 50 or more authors. There is also a good number of single author papers (around 3%), which also supports that performing computational analysis can still be lone wolf research.

A very similar skewed Gaussian behavior is observed for the number of citations per number of authors. Moreover, 49.14% of papers with less than 5 citations were published in 2018 and 2019 whereas those numbers drop for older papers. Hence, even though around 38.82% of all papers have less than 5 citations it seems that a lot of them are just too recent and will gain more momentum after two years.

About 45% of all authors (identified by their ORCID number) in our database have only one publication. This percentage is very high and probably comes from large collaborations or publications made by graduate students or young faculty who are still very new to the domain. A figure of author productivity can be found in the Supplementary (Fig. 3). It shows the average number of publications per year, for researchers with at least two publications in our dataset. The highest peaks, in order, are at two, one, and three papers a year.

Figure 3 shows the ratio between citations and publications as a function of the number of authors per paper. This plot displays similarities with the number of publications depending on the number of authors, represented in the bottom right part of Fig. 2. They both exhibit a high peak at a low number of authors, followed by a rapid decrease. However, for the absolute count of publications, as well as citations, the peak happens at three authors per paper, whereas Fig. 3 reaches its maximum for two authors. It is interesting to see that papers with two and five authors have similar publication counts, but drastically different ratios of number of citations over publications. We also notice a slight growth from five to seven authors, where it then stabilize around 23. This could mean that even though large collaborations can be complicated, long or hard to maintain, they will pay off as most will have a significant effect on the community. We should note that this figure is sensitive to outliers and a few highly cited papers are significantly raising the peak for two authors per paper. However, the general trend is still valid without them, and papers written by two authors, as well as bigger collaborations, with seven or more authors, are often getting a bigger ratio of citations.

Left panel in Fig. 4 shows the number of publications as a function of the countries where the authors are from. Though the vast majority of DFT software packages are

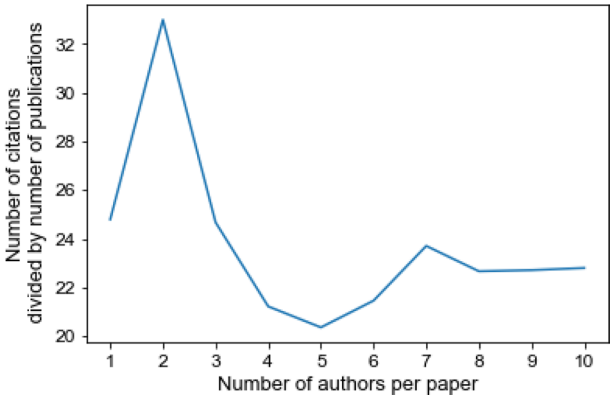

Fig. 3 Number of citations over number of publications as function of the number of authors per paper

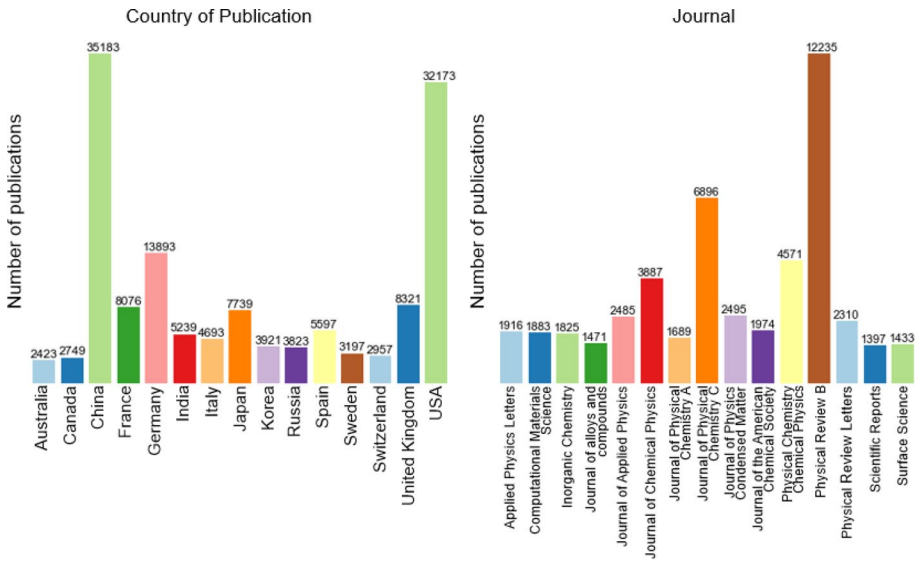

Fig. 4 Total number of publications per country and per journal (Only the top 15 are shown). Left: Total number of publications as function of the authors' countries. Right: Total number of publications with respect to the publishing journal

developed in Europe (at least those which are highly cited), the larger number of papers come from the USA and China by twice as much as Germany, which is ranked as the third country by publication number. Figure 5 shows a more detailed visualization of each country's productivity for each of the main DFT packages. Castep and Turbomole are both interesting cases, where one country really distinguishes itself from the others. China cited Castep the most, and by over 3 times as the next countries in list: the USA and Germany. Turbomole is the only code that was the most cited in a different country than the USA or China: Germany, where it originated. ADF and Quantum Espresso are both more popular in the USA and benefit from good support in Europe while VASP and Siesta display a very split number of citations between the USA and China. We need to point out that the records in this plot do not add to the total number of publications considered. Since a publication can have contributions from multiple countries, the publication is counted independently for each country. It is also clear that the USA is the country where the majority of the codes are used and the USA happens to be the country with the largest number of publications. VASP is by far the most used code overall, even if it is not a free package and has similar capabilities to other free licensed packages such as Quantum Espresso.

Studying the fields that are impacted the most by the use of DFT can offer an analysis over the strong effect of this methodology in the characterization and prediction of materials. This can be obtained by looking at the publication journal. This is what is reported on the right hand side of Fig. 4. From all journals used for publication, Physical Review B is the preferred journal to report analysis and discoveries of density functional theory, followed by Journal of Physical Chemistry C. However, the diversity of journals is very large, demonstrating the use of DFT in a large variety of problems. Almost all editorial companies have journals where DFT calculations are being reported. Although the full analysis of all publications provides a clear picture on the behavior of the electronic structure community, different applications and therefore packages are used in different journals, as

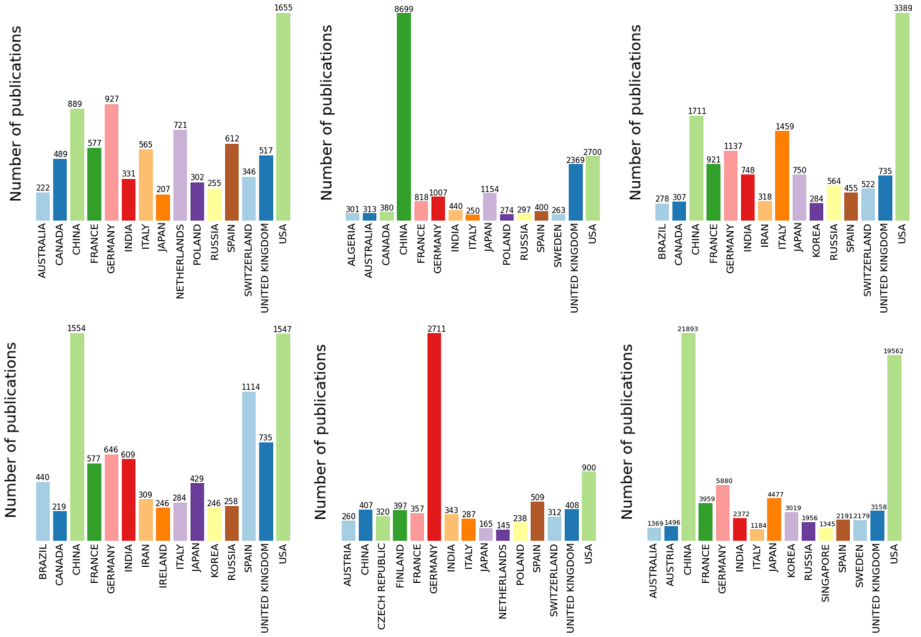

Fig. 5 Total number of publications as function of authors' country (Only top 10 are shown). Top row from left to right: ADF, Castep, Quantum Espresso. Bottom row from left to right: Siesta, Turbomole, VASP

can be concluded from Fig. 6. Though Physical Review B is one of the most well known journals in condensed matter physics, ADF citations prefer to publish in solid state chemical journals, as this particular software was created for solid state chemistry applications. Around 20% of published papers from Quantum Espresso are published in journals of the

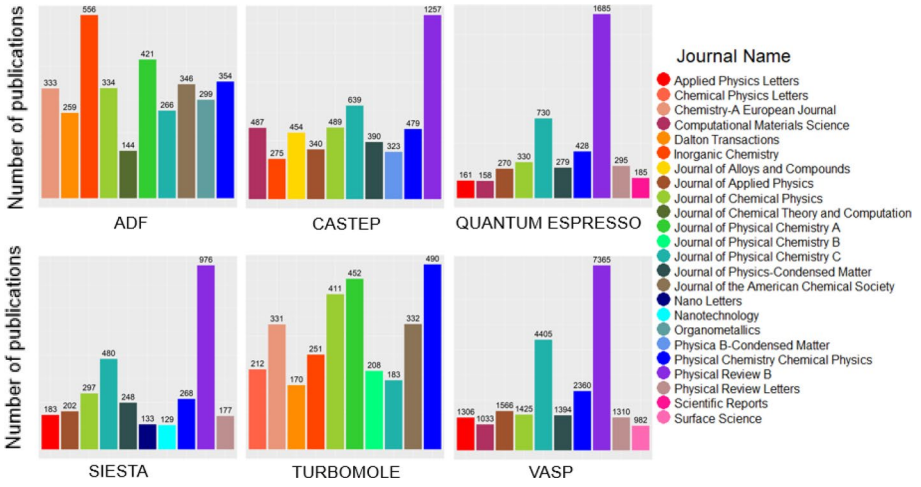

Fig. 6 Total number of publications per journal (Only top 10 are shown). Top row from left to right: ADF, Castep, Quantum Espresso. Bottom row from left to right: Siesta, Turbomole, VASP

American Physical Society while 18%, 15% and 35% from VASP, Siesta and ADF are published in journals of the American Chemical Society. It is also interesting to notice that the Nature Publishing Group is in the top ten of publishers with the highest number of publications, but when computing the number of citations, moves all the way up to the top 6.

Figure 11 in the Supplementary shows the normalized total number of citations, with respect to the total number of title words present in each paper. The words were selected following the procedure described in the methodology section. Papers for all 6 main codes reach their maximum number of citations between five and nine words, indicating that shorter titles may be the most successful. A possible explanation is that longer titles do not catch the attention of readers, therefore they are less cited. However, following findings from previous research (Deng 2015), it is important to note that this figure does not demonstrate causality between the length of the title and the number of times a paper is cited. Moreover, the figure displays big disparities between the distributions of each main code, making it harder to derive any trend.

An analysis of the twenty most common words in titles for each DFT code shows a common trend in the presence of theoretical-like words such as "First-Principles", "Studies", "Ab Initio" or "Properties" (see Supplementary Fig. 12). These are very general words and characterize many different theoretical publications. It is also clear that some codes have specific applications where users are more focused, for example "Graphene" in Siesta, "Optical properties" in Castep, "Center Dot" in Turbomole, "Ligand effects" in ADF, "surface effects" in Quantum Espresso and "Transition metal" in VASP. This plot also stresses the idea that communities created through specific topics tend to use one package.

Another source of analysis is how likely journals are to publish work on electronic structure calculations. We computed the number of citations normalized by the total number of citations for each one of the considered DFT codes with respect to the journals with the larger number of papers in this field. The normalized numbers are then divided by the impact factor reported in the SJR — SCImago Journal & Country Rank(SCImago 2007) for each journal (see Supplementary Fig. 4). The idea is to recover the impact of the paper based on the journal recognition. Castep is the code which has the highest impact in Physical Review B, closely followed by VASP. This can be explained from the fact that these two codes are mostly focused on condensed matter materials, which are very important topics in these journals. On the other hand, Turbomole and ADF have a larger impact in chemical journals such as Physical Chemistry Chemical Physics and Journal of Chemical Physics. VASP, which is the most used package, is very diverse and the citations are evenly distributed over different journals, which shows the diversity in users of this package. Therefore, this analysis shows that every one of the packages has created a localized community of users with VASP being the most general one, as it is the code with the largest journal diversity.

Figure 7 reports a log-log plot of the number of papers with respect to the number of citations for a selected set of countries to study the relationship between the two. Each plot shows the 10 countries with the highest H-index (Hirsch 2005), for a given code. If N is the total number of papers affiliated to a country, the H-index of that country is defined such that h papers out of N have at least h citations each and the other $(N - h)$ papers have $\leq h$ citations each. The H-index is an overall good estimate of a country's impact and corrects for the disproportionate weight of highly cited publications or publications that have not yet been cited. The exact H-index values for the 10 selected countries of each software are available in Figure 5 in the Supplementary.

The behavior displayed in Fig. 7 is very similar independently of the DFT package used. Evidently, countries with a large number of publications end up having a larger number

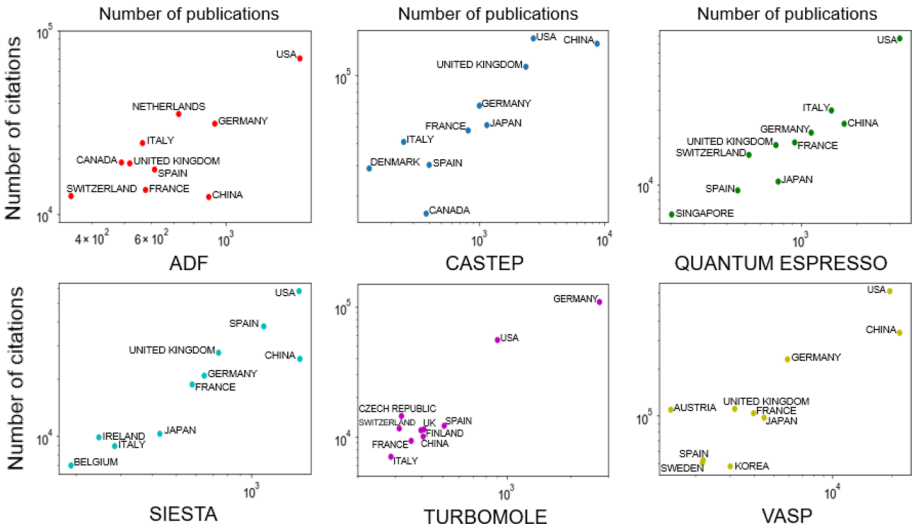

Fig. 7 Total number of publications as function of the total number of citations, on the logarithmic scale. The figures show the 10 countries with the highest H-index for a given code. Top row from left to right: ADF, Castep, Quantum Espresso. Bottom row from left to right: Siesta, Turbomole, VASP

of citations. Therefore an almost linear dependence is observed for almost all codes. The USA, China and Germany are often part of the countries with the highest number of both publications and citations. Electronic structure is therefore dominated by countries with large computational facilities and scientific diversity. The cases of Castep, Siesta and VASP are interesting and similar. For Castep, the USA and China both have a similar number of citations, but China achieves it with more than three times the number of American publications. However, Siesta and VASP display the same situation where the USA obtain a lot more citations than China for a similar number of publications. Moreover, China often has a smaller H-index. For example, while China is the third country with the highest number of publications citing ADF, it ranks 10th in H-index. Those profound differences in ranking and number of citations between the two countries demonstrate a better performance for papers from the USA.

To further analyze the dynamics between different countries, we computed the number of publications with at least one country in South America and one in either the USA, Canada or Europe (considered as European Union members, Switzerland, United Kingdom and Norway). The same analysis was made for African countries. Although South America and Africa are both developing continents, records show that South America benefits from a lot more collaborations with every other group defined previously than Africa does. However, the highest number of publications are created within two countries in the same class, highlighting how challenging it can be for developing countries to collaborate on scientific research with more economically advanced countries. We also report the number of authors per paper, when at least one of the affiliated countries is China or the USA (see Supplementary Fig. 8 and Fig. 9). While both countries tend to publish with authors from the same country, the USA are more open to cooperate, as its number of publications with several countries is higher.

In order to address the dynamics of the different collaborations, we now analyze the network properties of country collaboration networks. Basically, we create a network where

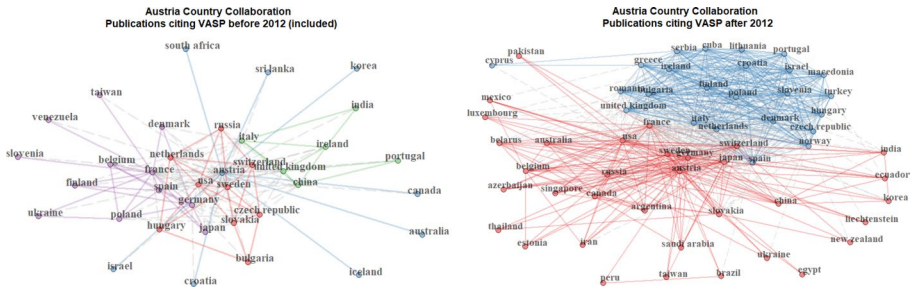

Fig. 8 Country collaboration network for publications citing the VASP code with at least one author in Austria. Left: publications published before 2012 (included). Right: publications between 2013 and 2019

Table 1 Country collaboration network properties, before and after 2012

	ADF	Castep	Quantum Espresso	Siesta	Turbomole	VASP
	Country Collaboration Network Before 2012 (included)					
Density	0.191	0.152	0.142	0.229	0.156	0.230
Diameter	2	2	2	2	2	2
Transitivity	0.361	0.333	0.312	0.390	0.342	0.449
	Country Collaboration Network After 2012					
Density	0.240	0.297	0.277	0.338	0.174	0.398
Diameter	2	2	2	2	2	2
Transitivity	0.556	0.735	0.705	0.760	0.411	0.742

every affiliated country in the same publication is connected. The analysis of the network properties is given in Table 1. We added the constraint that at least one of the authors come from the country where the code is developed. The idea is to study the internationalization of each code and what is the dynamic of the user base. We note that there is an important number of publications from the countries of the main developers of the code, in particular from free license packages (see Fig. 5). Some examples of high publication numbers in developers' countries include Turbomole (Germany), Castep (United Kingdom), Quantum Espresso (Italy) and Siesta (Spain). This also impacts the community around each one of the packages, as many publications come from personal networking of the code developers. In Fig. 8, we present a panel of two figures. The left hand side reports the network for publications up to 2012 and the right hand side refers to publications from 2013 to 2019. Colors represent clusters: groups of closely related countries. The same figures, for all 6 main packages are available in the Supplementary Figure 13.

A first thing to notice about the country collaboration networks after 2012 is the clear cluster created by European countries with smaller publication record like Sweden, Greece or Denmark. This could be explained by incentives given by the European Union for scientific collaboration, or simply by the fact that most of them are neighbors and can have more links, as researchers might have more chances to interact with each other. We can also observe, in this tight cluster, the presence of some non-European countries such as Cuba, Israel or Japan which could indicate a more general difficulty for smaller research communities to collaborate with more productive and higher cited countries.

The density of nodes is much higher for publications after 2012, which suggests that the number of publications increased, agreeing with Fig. 2. Indeed, Castep, Quantum Espresso, Siesta and VASP all have at least a 47% surge in density and transitivity after 2012. The nodes involved before 2012 and after also represent the dynamics of the collaborations. VASP is a code where the connectivity with its country of development is more diffuse and where links between other countries are much stronger, as it has a larger number of publications.

It is interesting to notice that the density and transitivity double only for Castep and Quantum Espresso. While all other main DFT codes have records before 2003, the first publication citing Quantum Espresso was in 2009. Hence, only about 14% of the citations for Quantum Espresso existed in 2012 and before. This difference in starting year might explain the sudden surge for Quantum Espresso, but it does not apply to Castep. The transitivity increase in Castep may be due to the license change of this package which has happened in the last few years, as the package is now freely available for academic use. On the other hand, Turbomole did not have as much of a surge in density and transitivity, in fact it is very small (about 12% and 20%), making it the network with the smallest density and transitivity after 2012.

Discussion

Bibliometric analysis is used to narrow the scholarship community of electronic structure calculations. Citations of computational packages implementing density functional theory show how influential and transcending this theoretical development in the understanding of material properties has been. The development and the impact is revealed by the large number of publications per year, which is the result of the efforts of a wide range of applications from many different material science related fields. At the same time, computational electronic structure is a field where a small number of groups work on developing ideas, as the number of participant countries and number of authors per paper is small in comparison with other research fields. In the 500 most cited papers, we also noted that the number of pages is not letter like, with an average of 15.25 pages per paper (with a standard deviation of 23.36). Therefore, long papers with important and new field concepts are very well cited by the community.

To give weight on the novel ideas, we find that only 1.6% of the whole dataset are defined as review papers (tag “DT” in the Web of Science database), but they represent about 17% of the 500 most cited papers, and 22% of the one hundred most cited papers. This implies that even though most published papers contain new ideas, representing more than 85% of the 500 most highly cited papers, the proportion of reviews increases in the top 100. It could be explained by the fact that reviews offer a great general panorama of a certain field, useful to introduce and explain the added value of a paper.

There is a very interesting correlation between the number of citations and the number of publications per country. Although the USA, China and Germany are among the countries with the highest numbers, there is correlation between the country where the package is mainly developed and the average number of citations per paper. This demonstrates that there is a payoff between the software initiative and the impact on the science done in those countries. Basically, the code development lies in the interest of the developers and that is why some of the codes considered here have a very unique set of tools not present in other implementations. The USA is an exception of code users, where scientists use many of the

most important DFT packages. On the other hand, China users use mostly VASP, which is a licensed package.

Developing an electronic structure package takes an important effort from developers and it needs a strong relationship between the different developer groups. Network analysis for each DFT code shows that during the first years of code development, the interaction happens mostly between the country where most of the development is performed and some other countries, which are directly related to the code developers. After some years and with the code reaching maturity (enough debugging by users and developers), the networks become much more dense and connected. It is interesting to see the degree of interaction between different continents and even geographical zones. Independently of the code, east European countries tend to collaborate more with other countries that have a smaller interaction density. We also notice that larger user communities surround codes which have simple interfaces, which increase number of publications and citation number. Therefore, efforts along the lines of making the codes more user friendly and with simpler interfaces are valued by the user community.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-021-04057-z.

Author Contributions Marie Dumaz and Reese Boucher downloaded the data from the Web of Science database and created the corresponding packages for data cleaning and data analysis. Marie Dumaz, Miguel A.L. Marques and Aldo H. Romero did the data analysis, created the data figures and wrote the paper.

Data availability The dataset that supports the findings of this study is available in Figshare with the identifier https://figshare.com/articles/dataset/bibliometric_DFT/12494654/4.

Declarations

Conflict of interests The authors declare that they have no conflict of interest.

References

Amin, M., & Mabe, M. (2004). Impact factors use and abuse. International Journal of Environmental Science and Technology (IJEST).

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007.

Arya, C., & Sharma, S. (2011). Authorship trends and collaborative research in veterinary sciences: A bibliometric study. *Chinese Librarianship: An International Electronic Journal, 34,* 1–9.

Barabási, A.-L., et al. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications, 311*(3–4), 590–614.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. Sebastopol, CA, USA: O'Reilly Media.

Blaha, P., Schwarz, K., Madsen, G.K.H., Kvasnicka, D., Luitz, J., (2001). "wien2k". In: An augmented plane wave+ local orbitals program for calculating crystal properties.

Chow, D. S., Ha, R., & Filippi, C. G. (2015). Increased rates of authorship in radiology publications: A bibliometric analysis of 142,576 articles published worldwide by radiologists between 1991 and 2012. *American Journal of Roentgenology, 204*(1), W52–W57.

Deng, B. (2015). "Papers with shorter titles get more citations". In: Nature News.

Dumaz, M. (2020). pyBilio. https://github.com/romerogroup/pyBiblio.

Garfield, E. (1999). Journal impact factor: a brief review.

Glänzel, W., Schubert, A., & Czerwon, H.-J. (1999). A bibliometric analysis of international scientific cooperation of the European Union (1985–1995). *Scientometrics, 45*(2), 185–202.

Hagen, N. T. (2008). Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis. *PLoS One, 3*(12), e4021.

Haunschild, R., Barth, A., & French, B. (2019). A comprehensive analysis of the history of DFT based on the bibliometric method RPYS. *Journal of Cheminformatics, 11*(1), 72.
Haunschild, R., Barth, A., & Marx, W. (2016). Evolution of DFT studies in view of a scientometric perspective. *Journal of cheminformatics, 8*(1), 52.
Haunschild, R. & Marx, W. (2020). Discovering seminal works with marker papers. In: Scientometrics, pp. 1–15.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences, 102*(46), 16569–16572.
Hohenberg, P., & Kohn, W. (1964). Inhomogeneous electron gas. *Physical Review, 136*(3B), B864.
Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering, 9*(3), 90–95.
Kohn, W., & Jeu, S. L. (1965). Self-consistent equations including exchange and correlation effects. *Physical Review, 140*(4A), A1133.
Larivière, V., et al. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology, 66*(7), 1323–1332.
McCain, K. W., (2014). Obliteration by incorporation. In: Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact, pp. 129–149
McKinney, W., (2010). Data structures for statistical computing in Python. In: ed. by Stéfan van der Walt and Jarrod Millman, pp. 51–56. https://doi.org/10.25080/Majora-92bf1922-00a.
Moed, H. F. (2006). *Citation analysis in research evaluation* (Vol. 9). Berlin, Germany: Springer Science & Business Media.
Oliphant, T. E. (2006). *A guide to NumPy*. USA: Trelgol Publishing.
Perdew, J. P., Burke, K., & Ernzerhof, M. (1996). Generalized gradient approximation made simple. *Physical review letters, 77*(18), 3865.
Peters, H., & Van Raan, A. (1991). Structuring scientific activities by co-author analysis: An exercise on a university faculty level. *Scientometrics, 20*(1), 235–255.
SCImago, . (2007). *SJR-SCImago Journal & Country Rank*.
Stéfan van der Walt, S., Colbert, C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering, 13*(2), 22–30.
Noorden, V., Richard, B. M., & Nuzzo, R. (2014). The top 100 papers. *Nature News, 514*(7524), 550.