Ten recommendations for reducing the carbon footprint of research computing in human neuroimaging

Souter, Nicholas E.; Lannelongue, Loïc; Samuel, Gabrielle; Racey, Chris; Colling, Lincoln J.; Bhagwat, Nikhil; Selvan, Raghavendra; Rae, Charlotte L.

Published in:
Imaging Neuroscience

DOI:
10.1162/imag_a_00043

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Ten recommendations for reducing the carbon footprint of research computing in human neuroimaging

Nicholas E. Souter\textsuperscript{a}, Loïc Lannelongue\textsuperscript{b,c,d,e}, Gabrielle Samuel\textsuperscript{f}, Chris Racey\textsuperscript{a}, Lincoln J. Colling\textsuperscript{a}, Nikhil Bhagwat\textsuperscript{g}, Raghavendra Selvan\textsuperscript{h,i}, Charlotte L. Rae\textsuperscript{a}

\textsuperscript{a}School of Psychology, University of Sussex, Brighton, United Kingdom
\textsuperscript{b}Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom
\textsuperscript{c}British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom
\textsuperscript{d}Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge, United Kingdom
\textsuperscript{e}Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, United Kingdom
\textsuperscript{f}Department of Global Health and Social Medicine, King’s College London, London, United Kingdom
\textsuperscript{g}McConnell Brain Imaging Centre, The Neuro (Montreal Neurological Institute - Hospital), McGill University, Montreal, Canada
\textsuperscript{h}Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
\textsuperscript{i}Department of Neuroscience, University of Copenhagen, Copenhagen, Denmark

Corresponding Author: Nicholas E. Souter (n.souter@sussex.ac.uk)

ABSTRACT
Given that scientific practices contribute to the climate crisis, scientists should reflect on the planetary impact of their work. Research computing can have a substantial carbon footprint in cases where researchers employ computationally expensive processes with large amounts of data. Analysis of human neuroimaging data, such as Magnetic Resonance Imaging brain scans, is one such case. Here, we consider ten ways in which those who conduct human neuroimaging research can reduce the carbon footprint of their research computing, by making adjustments to the ways in which studies are planned, executed, and analysed; as well as where and how data are stored.

Keywords: neuroimaging, neuroscience, carbon footprint, computing, sustainability, green

1. INTRODUCTION
We are in the midst of a climate crisis, with exponentially increasing rates of carbon emissions leading to increases in global temperatures. This, in turn, leads to an increased incidence of natural disasters, including floods, fires, and droughts, as well as a loss of biodiversity (IPCC, 2022). Given that the technological path to removing carbon from the atmosphere remains unclear (Carton et al., 2020), and the benefits of carbon offsetting schemes are dubious (Watt, 2021), real-term reductions in emissions will be needed to combat this crisis.

As data-literate individuals with power over the design of research paradigms and the dissemination of knowledge, researchers should reflect on the carbon footprint of their work. One may ask where responsibilities should lie. All actors in the research ecosystem, including governments, institutions, journals, funders, researchers, and data hosts, have a role in incentivising and supporting research practices that promote reductions in carbon emissions (Farley, 2022; Lannelongue et al., 2023; Urai & Kelly, 2023)\textsuperscript{1}. While systemic changes are undoubtedly critical, climate conscious researchers should also take initiative through collective action for two important

\textsuperscript{1} Indeed, recent reports from research funders demonstrate interest in this topic, Wellcome; https://wellcome.org/reports/advancing-environmentally-sustainable-health-research, UKRI; https://zenodo.org/record/8199984

Received: 21 September 2023 Revision: 16 November 2023 Accepted: 16 November 2023 Available Online: 21 November 2023
reasons. First, as practising scientists, we have a much deeper understanding of our own research processes than governing bodies, and are therefore well placed to address their impacts. Second, by doing so, we create a social mandate for change with governing bodies. Given there are often barriers to individuals making meaningful change in the face of institutional incentive structures, acting in the domains where we are empowered to act can help in pushing our institutions further. One such domain is research computing.

In many fields, researchers rely heavily on computing. The information communication technology (ICT) sector accounts for an estimated 1.8-3.9% of global CO2 emissions, largely accounted for by electricity production (Freitag et al., 2021). Although a small percentage in absolute terms, it is likely to continue growing as we process and store increasing amounts of data. This has become more pertinent in recent years with increasing adoption of large models trained through artificial intelligence (Selvan et al., 2022). Belkhir and Elmeligi (2018) estimated data centre energy use to account for 45% of the greenhouse gas emissions produced within the ICT sector in 2020, up from 33% in 2010. Despite this, the amount of data being collected and processed is relatively neglected in climate policies and initiatives concerning data-driven health research (Samuel & Lucassen, 2022). We therefore urgently need to establish and use best practices for greener computing moving forward. Such initiatives are already being taken in some compute-heavy fields, including bioinformatics (Grealey et al., 2022), machine learning (Selvan et al., 2022), and astronomy (Portegies Zwart, 2020).

Despite initiatives in other disciplines, there has been little attention so far to the computing carbon footprint of human neuroimaging research. This is another compute-heavy field which frequently relies on computationally expensive data processing and analysis. Fortunately, there is scope to reduce this footprint by “computing carefully” (Rae et al., 2022)—reducing a project’s required computing power and, therefore, energy production. Here, we outline several factors that contribute to the energy required for computing in human neuroimaging, and provide ten recommendations for how researchers can reduce these costs (summarised in Box 1). There are a number of other ways in which neuroimaging research contributes to the climate crisis, including through the procurement of specialist equipment, extraction of experimental resources (such as liquid helium for MRI scanning), and frequent flights to international conferences (Aron et al., 2020; Zak et al., 2020). Methods for reducing the carbon footprint of these aspects are beyond the scope of this paper (but see Rae et al., 2022). In the coming years, advances in artificial intelligence may supplement neuroimaging data processing in ways that modulate its carbon footprint—this is also beyond the scope of this present paper.

2. RECOMMENDATIONS

2.1. Plan and preregister analysis

Unnecessary repetitions of data analysis represent a waste of energy consumption and should be avoided. Here, we are not referring to replication studies of existing paradigms—these are important in increasing the credibility of science (Open Science Collaboration, 2015). Instead, we are referring to repetitions that occur as the result of unforeseen obstacles. For example, neuroimagers may run analysis for multiple participants, only to discover that results are unusable due to fundamental issues

---

**BOX 1. SUMMARY OF TEN RECOMMENDATIONS FOR REDUCING THE CARBON FOOTPRINT OF NEUROIMAGING COMPUTING**

1. Preregister a study analysis plan in order to avoid repetitions
2. Quantify and report the carbon footprint of your computing using available carbon tracking tools
3. Only run the preprocessing and analysis steps that you need
4. Run your computing at lower carbon intensity times and in lower carbon intensity locations
5. Regularly remove files that you do not need
6. Plan where, and for how long, you will store files, aided by research technicians
7. Advocate for non-commercial and centralised data storage solutions
8. Publicly share sufficient data to ensure it is FAIR (Findable, Accessible, Interoperable, Reusable), but consider the extent of what others will actually need or use
9. Make use of existing preprocessed data when possible, instead of acquiring and processing new data
10. Discuss the importance of greener computing with other neuroimagers and advocate for systemic change
with event timing files (e.g., for fMRI) or missing data. Historically, neuroimagers may have also tweaked analysis pipelines to identify the settings that produced the “best” (i.e., most statistically significant) results. Such repetitions not only contribute to increased carbon emissions, but can also be inconsistent with good research practices. Both issues can be addressed through preregistration of your plan for data collection, preprocessing, and analysis. Preregistration involves uploading a detailed study plan to an online repository (e.g., Open Science Framework (OSF); https://osf.io registries, AsPredicted; https://aspredicted.org) before data have been collected and/or analysed. Doing so can increase the credibility of your research by clearly delineating between confirmatory and exploratory analyses and providing evidence against suspicions of having “p-hacked” significant results (Gorgolewski & Poldrack, 2016). While writing a preregistration can be initially time consuming, engaging with this process has downstream benefits such as increasing confidence in methods used and improving the efficiency of the analysis stage, thereby reducing the need for unnecessary repetitions and computing. In cases where further exploratory analysis is needed, we recommend that one designs and tests analysis pipelines on a single subject before applying them to the entire sample. This will help with the elimination of code bugs, reducing the amount of unnecessary repetitions and therefore energy use.  

**Suggested Action:** Preregister a study analysis plan in order to avoid repetitions

### 2.2. Track your emissions

In recent years, several tools have been developed to systematically track and quantify carbon emissions associated with computational processes. For example, Green Algorithms (Lannelongue et al., 2021; https://www.green-algorithms.org) is an online calculator that allows users to input parameters for a given job, including runtime, number of cores, and available memory, in order to generate estimates of resulting carbon emissions before a job has started running. This calculator also takes the location of computing into account, given that carbon intensity of energy use will vary by country (see Recommendation 4). It also provides the server-side tool “GA4HPC,” which uses log information to estimate carbon emissions for jobs utilising high performance computing (HPC). Other packages, such as CodeCarbon (Goyal-Kamal et al., 2021; https://codecarbon.io) and Carbontracker (Anthony et al., 2020; https://github.com/lfwa/carbontracker), can be embedded directly into existing tools, allowing researchers to estimate carbon emissions without manually inputting parameters. Again, both packages consider the location of computing, and Carbontracker even makes use of real-time carbon intensity data for a given country, when possible. Recent experiments have shown that these tools provide sensible estimates of energy usage and carbon footprints (Jay et al., 2023).

As of version 22.1.0 (December 12th, 2022), the fMRI preprocessing pipeline fMRIPrep (Esteban et al., 2019) has had CodeCarbon integrated into its code (see https://fmriprep.org/en/stable/changes.html#december-12-2022). Simply by toggling on a “track-carbon” flag and providing a relevant “country-code” (e.g., GBR for United Kingdom) in the command line, fMRIPrep users are provided estimates of carbon emissions for the preprocessing of a given participant. There are strengths and weaknesses to online calculators, server-side tools, and embedded packages, and the ideal solution for a given neuroimager will depend on these factors (Lannelongue & Inouye, 2023a). For example, embedded packages allow automatic collection of computing metrics but are not necessarily compatible with all programming languages, while the reverse is true for online calculators. Whichever approach you use, estimating the carbon footprint of analysis or preprocessing is a good first step to understand the carbon emissions associated with your research computing (Henderson et al., 2020; Lannelongue et al., 2023). Small-scale experimentation, including manipulations of analysis or preprocessing parameters in conjunction with carbon tracking (see Recommendation 3), may further allow researchers to understand which elements of their research computing particularly tax energy usage.

Beyond this, we recommend neuroimagers (and other compute-heavy researchers) provide an “Environmental impact statement” in published papers—openly reporting the carbon footprint of their project in kilograms of carbon dioxide equivalent emissions (CO₂eq; Rae et al., 2022). For particularly compute-heavy projects, this will involve using carbon trackers to provide estimates of the carbon footprint of data processing, whenever available. Box 2 provides an example of the form such a statement could take, based on one pipeline of an ongoing preregistered study focusing on the carbon footprint of fMRI preprocessing (see Recommendation 3).

This approach has already been taken by researchers in compute-heavy fields (e.g., Lannelongue & Inouye, 2023b; Xu et al., 2023), and a comprehensive framework for reporting these figures is provided by the Scientific CO₂nduct initiative (Sweke et al., 2022; https://scientific
We encourage researchers to be transparent and pragmatic in reporting this figure. Accurate and representative estimates will allow for synthesis across studies, facilitating a better understanding of which elements of neuroimaging data processing may have a particularly large footprint, and what can be done to reduce this footprint. The adoption of this practice could mirror that of the “Data availability statement,” a relatively recent open science initiative that is now culturally accepted within the life sciences, and expected by many journals for the publication of papers.

**Suggested Action:** Quantify and report the carbon footprint of your computing using available carbon tracking tools

### 2.3. Preprocess conservatively

When working with raw neuroimaging data, preprocessing is a necessary but computationally expensive process. fMRI preprocessing steps include brain extraction, registration, smoothing, and denoising (Caballero-Gaudes & Reynolds, 2017). EEG steps include noise and artefact removal, elimination of bad channels, and re-referencing (Kim, 2018). Following previous lab procedures using existing scripts can be a reliable way to produce good-quality data through a pipeline that runs without error. However, doing so often means that redundant steps are performed which have little or no impact on the final product. From our own experience, existing lab scripts for fMRIPrep have included registration of BOLD data to multiple output spaces and the creation of CIFTI files (storing connectivity data), despite the fact that these files are frequently not used in subsequent analyses. While it may feel useful to store such files “just in case,” it is possible to reduce compute and runtime for your preprocessing by carefully planning which files you will need prior to starting a project. Aspects of analysis may similarly use unnecessary compute. For instance, independent component analysis (ICA) denoising of fMRI data is a computationally lengthy process. While often beneficial in producing higher sensitivity to statistical results, it can operate with varying degrees of success (Scheel et al., 2022). In the absence of good theoretical motivations to conduct steps such as ICA denoising, one should consider which aspects of analysis are necessary. The scope for meaningful reductions in emissions during job execution will also be impacted by the baseline energy consumption of HPC cluster nodes when idling (not in use). For example, when examining energy costs of running a large language model, Luccioni et al. (2022) found only 54.5% of energy use to be attributable to running code, 13.5% to infrastructure including storage and cooling, and 32% to idling costs needed to keep nodes on regardless of whether code was running on them. Reduction of these costs will likely rely on advances in hardware.

Even within energy costs associated with running a job, it can be challenging for end users to know in advance which preprocessing and analysis steps have meaningful versus negligible effects on compute, as there is a lack of systematic investigation into this question. In the absence of empirical data, researchers may rely on the assumption that individual preprocessing steps that take a particularly long time to complete require particularly intensive compute power. To provide a more systematic estimate, in an ongoing preregistered study (https://osf.io/839pa), we are evaluating the effect of different fMRIPrep parameters on both performance (in analytical sensitivity) and the carbon footprint of preprocessing. Similar investigations using carbon trackers in conjunction with a wider range of packages will allow for a better understanding of how compute power and data quality can be teased apart in order to identify the carbon footprint-optimised set of parameters that balance climate costs and scientific gains. We encourage these future investigations.

**Suggested Action:** Only run the preprocessing and analysis steps that you need

### 2.4. Time and location matter

Periods of peak energy use put strain on a national grid’s available renewable energy and increase reliance on carbon-intensive sources in order to meet demand. Data from the UK National Grid ESO carbon intensity API (https://carbonintensity.org.uk), for example, reveal characteristic
patterns of carbon intensity, with as much as a 36.8% average decrease from peak to lowest point within a given day of the week (using available data from 2023). Available data from 2017 to 2023 can be seen in Figure 1—with the average carbon intensity presented for each 30-minute period of each day of the week within a given year. During the working week (Monday-Friday), predictable peaks occur at approximately 7:30-8 am and 6:30-7:30 pm—preceding the start and following the end of the working day, when domestic energy use spikes. Carbon intensity is considerably lower overnight and on the weekend. This implies that carbon savings can be made by running analyses at times of lower carbon intensity.

Unfortunately, live carbon intensity data are not publicly available for many countries (see https://app.electricitymaps.com for available sources). However, in cultures following a common 9-5 Monday-Friday working week, peaks and troughs of carbon intensity should approximately resemble those of the UK in Figure 1. When available, live data facilitate the creation of automated job schedulers, which can schedule jobs to run at forecasted periods of low carbon intensity. One recent example is the Climate-Aware Task Scheduler (CATS; https://github.com/GreenScheduler/cats), which can be implemented in the UK for any HPC task. Institutions could take this initiative further by imposing user-specific “carbon budgets” for research computing, in conjunction with task schedulers. While such a move may be controversial among HPC users, this could incentivise researchers to be more mindful in their computing, including the use of scheduling when possible.

The data in Figure 1 should also provide cause for optimism. The transition from 2017 to 2023 reflects an overall reduction in mean carbon intensity in the UK energy mix of 46.4%. Continuing adoption of renewable energy should see this trend continue within the UK. However, given there is a limit to renewable infrastructure that can be created, and that many other aspects of society need to be electrified (e.g., transport), the overall amount of energy available for research computing in a renewable powered world will still be limited.

Carbon intensity also depends on the location in which computing occurs, because energy grids of countries and regions differentially employ carbon-intensive and renewable energy sources. As seen in Figure 2, there is considerable variation in carbon intensity both between and within countries—Iceland’s carbon intensity is 0.01% that of South Africa, due to greater reliance on renewable energy sources.

---

2 Changes in carbon intensity by month are not entirely consistent across years, see Supplementary Figure 1.

3 At the time of writing, data were not available for all of 2017 or 2023. The equivalent reduction from 2018 to 2022 is 26.3%.
Fig. 2. The carbon intensity of energy supply for (a) countries of the world, (b) US states, (c) Canadian provinces, and (d) Australian states. Data taken from the 2022 v1.0 release of Country Specific Electricity Factors (2022) from www.carbonfootprint.com. All available data from this report are plotted. Dotted line reflects the average of countries/states/provinces within each graph. gCO2e/kWh = grams of carbon dioxide equivalent per kilowatt hour. US = United States, DKIS = Darwin Katherine Interconnected System, NWIS = North Western Interconnected System, SWIS = South West Interconnected System.4

4 Note that exact estimates of carbon intensity vary by source; there is disparity between numbers presented here and those provided by Our World in Data (https://ourworldindata.org/grapher/carbon-intensity-electricity), although the rank order of countries by intensity across sources is comparable.

adjust the location in which computing is done, as researchers may be tied to the physical location of institutional servers. This becomes more tractable when considering cloud computing services, which may house servers in more or less carbon-intensive areas. When storing and sharing data, researchers may choose to
prioritise data repositories supported by servers in areas with low carbon intensity. However, researchers should remain cognisant of the implications of their choices in terms of, for example, any socio-political issues associated with where data centres are constructed, and any ramifications on local areas and/or communities. Well thought-through and equitable international inter-institutional collaborations may be critical in providing researchers in low- and middle-income countries with access to low carbon intensity computing opportunities (Lannelongue et al., 2023). Overall, by scheduling pre-processing or analysis to run at periods or in locations of low carbon intensity, you could emit considerably less carbon while using the same amount of energy.

_Suggested Action:_ Run your computing at lower carbon intensity times and in lower carbon intensity locations

### 2.5. Tidy up “junk” files

Energy is required not only to process data, but also to store it. Increasing the amount of data stored on a server can impact workload, by providing a larger amount of material to backup, for instance. The mere storage of data also incurs energy consumption due to the requirement for powering hard drives and air conditioning in server rooms. The more we store, the greater the energy consumption. Additionally, as institutions run out of space to store data, it becomes necessary to acquire additional hardware. Even before servers are in use, the production of computing hardware contributes a substantial portion of the carbon impact for this sector, 15-40% for data centre servers⁵, and 70-90% for consumer devices (Clément et al., 2020). Overall, 10 kg of CO₂-equivalent is the order of magnitude of the carbon footprint of each terabyte of data stored on a hard drive (Lannelongue et al., 2021).

It is common for neuroimaging pipelines to produce large amounts of intermediary files that will never be used by the researcher. This includes files generated both in working directories and for the final output. For the aforementioned preregistered study (https://osf.io/839pa), we have been processing data for 257 subjects in fMRIPrep. For a single pipeline, fMRIPrep generated a total average of 5.55 GB per subject (across output files, working directories, and logs). Only 0.23 GB, 4.0% of the total size, corresponded to files intended for use in subsequent statistical analysis (see Fig. 3). To address this unnecessary output, we provide an open source tool, fMRIPrepCleanup, available for download on GitHub (https://github.com/NickESouter/fMRIPrepCleanup), designed to delete unnecessary fMRIPrep files within a given directory. Using such automated scripts to cleanup junk files can place less stress on existing storage infrastructures and reduce the need for additional server purchases and manufacturing. However, extreme care should be taken when writing and executing such scripts, including the one linked here. They will need to be customised based on the research needs of the user and the output file structure they have created. If the overarching directory and file paths to be saved are not correctly specified, you risk irreversibly deleting important data. We recommend executing such a script on a copy of one participant’s dataset first, and using the Brain Imaging Data Structure (BIDS) file organizational structure (Gorgolewski et al., 2016) to make it easier to index files that are to be kept and deleted as appropriate. It is typically more energy-intensive to regenerate files than it is to store them. As such, files should only be deleted following an evaluation of exactly what will need to be retained, in order to avoid unnecessary repetition.

_Suggested Action:_ Regularly remove files that you do not need

---

2.6. Plan your long-term storage

As well as making efforts to delete unneeded files, researchers should consider the method and duration of storage of files that do have value. As covered in Recommendation 5, the long-term storage of files has a carbon footprint. Researchers in computationally expensive fields such as neuroimaging can explicitly consider the carbon footprint implications of data storage in data management plans prior to beginning a study, as outlined by the Digital Humanities Climate Coalition (DHCC Information, Measurement and Practice Action Group, 2022; https://sas-dhrh.github.io/dhcc-toolkit). Many institutions have policies requiring researchers to retain data on local storage for a minimum time period; these are sometimes specific in scope (e.g., 10 years) or are sometimes vague and unspecified (Briney et al., 2015; e.g., that data should be retained for an “appropriate” period of time, as judged by the researcher). It is unclear what typically happens to data after such a retention period. Although there may be an implicit assumption that researchers delete data after this period, they may not do so without explicit encouragement. This can contribute to the accumulation of “dark data” that is poorly indexed or simply unneeded, and therefore becomes functionally invisible and unused while taking up space (Schembera & Durán, 2020). Here, institutional research technicians or administrators could play a vital role in prompting researchers to regularly remove files that are no longer needed or within the retention period. Having recommended options for data at the end of a project’s life cycle would help avoid the accumulation of dark data. For neuroimagers concerned about permanently losing access to data, transitions from digital disk storage to long-term tape storage may provide substantial savings in both storage costs and carbon emissions (Johns, 2020), as tape storage does not require energy to air condition servers or cover baseload. Researchers could consider utilising offline or solid storage for their data unless they have specific reasons not to do so.

**Suggested Action:** Plan where, and for how long, you will store files, aided by research technicians

2.7. Push for publicly owned centralised data storage

Centralised data storage avoids the duplication of datasets in each research group and benefits the research community as well as the environment. General open science repositories, such as the OSF (https://osf.io), and neuroimaging-specific platforms, such as Neurovault (https://neurovault.org) and OpenNeuro (https://open-neuro.org), allow neuroimagers to provide public access to raw and processed neuroimaging data. This practice has helped make neuroimaging research more credible, reproducible, and accessible (Gorgolewski & Poldrack, 2016). Although larger data centres tend to be more energy efficient than decentralised small data storage infrastructures (Masanet et al., 2020), it is important to assess the carbon footprint of these facilities when building such centralised storage resources. Moreover, these repositories rely on commercial cloud computing services. For instance, data for OSF are hosted by Google, and data for OpenNeuro are hosted by Amazon Web Services. Some researchers may perceive ethical issues with entrusting public medical data to such commercial platforms—including issues surrounding privacy and data security, and ownership and control (Chiruvella & Guddati, 2021). BigTech is also associated with producing and amplifying a range of social injustices and inequalities (Couldry & Mejias, 2018). Besides, claims made by large cloud providers around renewable energy and sustainability can be difficult to verify—transparency on this matter is not always evident.

Such concerns may be addressed through the use of publicly owned, non-commercial “trusted research environments” (TRE), which allow secure access to large amounts of medical data (Graham et al., 2022), and are run for, and by, the research community. When operating at a large scale, such services can also have power use effectiveness comparable to that of commercial services. The European High Performance Computing Joint Undertaking (EuroHPC JU) provides a promising example of a non-commercial approach (https://eurohpc-ju.europa.eu). For example, one EuroHPC JU initiative, LUMI in Finland, reportedly ranks as one of the most energy-efficient supercomputers in the world while also being fully powered by renewable hydroelectric energy. When considering data storage, non-commercial solutions such as this may provide an optimal balance between computing power, energy efficiency, and transparency. While the power of individual neuroimagers in this sense is limited, researchers can help here by advocating for and supporting the use of non-commercial computing initiatives and TREs.

**Suggested Action:** Advocate for non-commercial and centralised data storage solutions

---

2.8. Reflect on what needs to be shared

In a recent survey of neuroimaging researchers, 54% indicated that they were likely to share all raw imaging data in online repositories for their next project (Paret et al., 2022). This demonstrates an impressive commitment from much of the neuroimaging community to transparent and accountable science. When doing so, researchers should ensure that the data they share are FAIR—findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). These principles are often not adhered to (Crüwell et al., 2023), and sharing of unFAIR data may in some cases be worse than sharing nothing. Access to all data for a project can significantly enhance the utility of a dataset. However, as the demands on the ICT sector continue to grow, data centres will require more energy and space to operate (Freitag et al., 2021; although see Masanet et al., 2020 for an opposing view). This will also manifest as increased costs to data sharing platforms7. Even technological improvements in cloud computing may elicit a rebound effect, whereby increasingly efficient data storage actually leads to net increases in demand and therefore storage space and energy used (Widdicks et al., 2023). For their part, researchers should consider which aspects of their data will realistically be necessary or helpful to share. In most cases, it may be sufficient to upload preprocessed data only—enough to replicate analysis and test novel hypotheses, while placing minimal strain on cloud computing. Additionally, sharing too much can make it harder for users to navigate and correctly use datasets. We are not advocating against the sharing of data. However, exponential increases in the amount of existing data may make it necessary to ask difficult questions. Ultimately, it may be wise for researchers, institutions, or data-sharing platforms to place expiration dates on datasets, with removal after a set period. For now, researchers can exercise their own best judgement on the balance between the usefulness and size of their public data. Neuroimagers may benefit from the publication of a consensus paper, providing guidelines for sharing imaging data with sufficient rigour while also considering greener computing.

**Suggested Action:** Publicly share sufficient data to ensure it is FAIR (Findable, Accessible, Interoperable, Reusable), but consider the extent of what others will actually need or use

2.9. Use existing data

In neuroimaging investigations, the default approach is to design a novel paradigm and collect raw data from a novel sample. Making use of pre-existing, and often preprocessed data which are already suitable for statistical analysis, allows one to save time and resources, and importantly to avoid the energy use associated with processing novel data (a computationally expensive process, see Recommendation 3; Rae et al., 2022). This will only apply when the data necessary to answer a research question already exist, and are publicly available (see Recommendation 8). Large public datasets for research use include the Human Connectome Project (https://www.humanconnectome.org; Van Essen et al., 2013), which incorporates multimodal datasets across young adult, developmental, ageing, and clinical samples. These contain task-based fMRI, resting-state fMRI, MEG, and PET data. Similarly, UK Biobank is a mass-scale study comprising diverse phenotypic and genotypic data, including structural, diffusion, and functional MRI (https://www.ukbiobank.ac.uk; Miller et al., 2016). As of October 2022, over 60,000 volunteers have been imaged, with a target of 100,000 individuals in the final sample8. These data are used to generate over 4,000 imaging-derived phenotypes—metrics such as structure volume and connectivity that can be used as predictors of disease risk factor (Alfaro-Almagro et al., 2018). Note that this platform does not offer free access to its data, meaning it is not an entirely accessible public resource. Alternatively, the platform OpenNeuro provides free access to over 800 public datasets spanning MRI, fMRI, PET, MEG, EEG, and iEEG data (https://openneuro.org). Other open data repositories are listed in Table 19. When possible, the re-use of existing data provides a good example of how open science practices can intersect with opportunities to reduce one’s personal compute emissions.

**Suggested Action:** Make use of existing preprocessed data when possible, instead of acquiring and processing new data

---

7 [https://www.cos.io/blog/shared-investment-in-osf-sustainability](https://www.cos.io/blog/shared-investment-in-osf-sustainability)

8 [https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/worlds-most-ambitious-imaging-study-scans-60-000th-participant](https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/worlds-most-ambitious-imaging-study-scans-60-000th-participant)

9 As presented in this table, there are issues here with the diversity of data—all projects found are based in the US or Europe. All sources appear free to access, with the exception of UK Biobank.
Table 1. Overview of open access neuroimaging projects/data repositories, with available links, modalities, countries of origin, dataset sizes, and brief descriptions.

<table>
<thead>
<tr>
<th>Project</th>
<th>Link</th>
<th>Imaging modalities</th>
<th>Based in</th>
<th>Dataset size (participants)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A large and rich EEG dataset for modeling human visual object recognition</td>
<td><a href="https://figshare.com/articles/dataset/A_large_and_rich_EEG_dataset_for_modeling_human_visual_object_recognition/18470912">https://figshare.com/articles/dataset/A_large_and_rich_EEG_dataset_for_modeling_human_visual_object_recognition/18470912</a></td>
<td>EEG</td>
<td>Germany</td>
<td>10</td>
<td>Contains EEG data for responses to images of objects on a natural background. Ten participants each with a large number of trials.</td>
</tr>
<tr>
<td>Adolescent Brain Cognitive Development (ABCD) study</td>
<td><a href="https://abcdstudy.org">https://abcdstudy.org</a></td>
<td>MRI, fMRI, dMRI</td>
<td>US</td>
<td>11,880</td>
<td>A large long-term study of brain development and child health.</td>
</tr>
<tr>
<td>Alzheimer’s Disease Neuroimaging Initiative (ADNI)</td>
<td><a href="https://adni.loni.usc.edu">https://adni.loni.usc.edu</a></td>
<td>MRI, PET</td>
<td>US</td>
<td>800 (ADNI 1) 507 (ADNI 2)</td>
<td>A longitudinal multicenter study designed to develop biomarkers for the early detection of Alzheimer’s disease.</td>
</tr>
<tr>
<td>Amsterdam Open MRI Collection (AOMIC)</td>
<td><a href="https://openneuro.org/datasets/ds002785/versions/2.0.0">https://openneuro.org/datasets/ds002785/versions/2.0.0</a></td>
<td>MRI, fMRI, dMRI</td>
<td>Netherlands</td>
<td>216</td>
<td>Multiple large datasets containing data for various task-based fMRI paradigms, psychometrics, and demographics.</td>
</tr>
<tr>
<td>Autism Brain Imaging Data Exchange (ABIDE)</td>
<td><a href="https://fcon_1000.projects.nitrc.org/indi/abide">https://fcon_1000.projects.nitrc.org/indi/abide</a></td>
<td>MRI, fMRI, DTI</td>
<td>US/Europe</td>
<td>1,112 (ABIDE 1) 1,000+ (ABIDE 2)</td>
<td>Aggregates data from institutions around the world to further our understanding of the neural bases of autism.</td>
</tr>
<tr>
<td>Cambridge Centre for Ageing and Neuroscience (Cam-CAN)</td>
<td><a href="https://www.cam-can.org">https://www.cam-can.org</a></td>
<td>MRI, DTI, DKI, MEG, fMRI</td>
<td>UK</td>
<td>623-653 (varies by imaging modality)</td>
<td>A large collaborative project, focused on how individuals can retain cognitive abilities into old age.</td>
</tr>
<tr>
<td>Enhanced Nathan Kline Institute - Rockland Sample (NKI-RS)</td>
<td><a href="https://fcon_1000.projects.nitrc.org/indi/enhanced">https://fcon_1000.projects.nitrc.org/indi/enhanced</a></td>
<td>MRI, fMRI</td>
<td>US</td>
<td>1,000+</td>
<td>A large community sample of participants across the lifespan, contains diverse data types.</td>
</tr>
<tr>
<td>ERP CORE</td>
<td><a href="https://erpinfo.org/erp-core">https://erpinfo.org/erp-core</a></td>
<td>EEG</td>
<td>US</td>
<td>40</td>
<td>Contains event-related potential data for six paradigms relating to different components.</td>
</tr>
<tr>
<td>Human Connectome Project (HCP)</td>
<td><a href="https://www.humanconnectome.org">https://www.humanconnectome.org</a></td>
<td>MRI, fMRI, MEG, PET</td>
<td>US</td>
<td>1,200 (Young Adult) 1,200 (Aging) 1,350 (Development) 500 (Lifespan Baby) 1,500 (Lifespan Developing)</td>
<td>Contains multiple large datasets spanning different age groups across diverse tasks.</td>
</tr>
<tr>
<td>Imaging and Data Archive (IDA)</td>
<td><a href="https://ida.loni.usc.edu">https://ida.loni.usc.edu</a></td>
<td>MRI, CT, SPECT, PET, EEG</td>
<td>US</td>
<td>Signposts 150 studies; 95,100 participants</td>
<td>A resource for archiving and signposting neuroscience data repositories, including some of those listed here.</td>
</tr>
<tr>
<td>International Neuroimaging Data-Sharing Initiative (INDI)</td>
<td><a href="http://fcon_1000.projects.nitrc.org">http://fcon_1000.projects.nitrc.org</a></td>
<td>MRI, fMRI</td>
<td>US</td>
<td>1,200+</td>
<td>Includes a public release of resting-state fMRI datasets from 33 sites.</td>
</tr>
<tr>
<td>MEG UK</td>
<td><a href="https://meguk.ac.uk/database">https://meguk.ac.uk/database</a></td>
<td>MEG, MRI</td>
<td>UK</td>
<td>~500 (prospective)</td>
<td>A partnership between eight UK labs, adding towards a single shared repository of MEG data.</td>
</tr>
<tr>
<td>Project</td>
<td>Link</td>
<td>Imaging modalities</td>
<td>Dataset size (participants)</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>-------------------</td>
<td>-----------------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>Multimodal data, includes a language task and resting-state data. in BIDS format</td>
<td><a href="https://data.donders.ru.nl/collections/dccn/DSC_3011020.09_236?0">https://data.donders.ru.nl/collections/dccn/DSC_3011020.09_236?0</a></td>
<td>MRI, fMRI, MEG</td>
<td>Netherlands</td>
<td>204</td>
<td>A multimodal dataset focused on the Moos, including language and resting-state data. Includes 30 participants over two sessions in BIDS format.</td>
</tr>
<tr>
<td>Moos, Multisubject, multimodal face processing</td>
<td><a href="https://openneuro.org/datasets/ds000117/versions/1.0.3">https://openneuro.org/datasets/ds000117/versions/1.0.3</a></td>
<td>MRI, fMRI, MEG, EEG</td>
<td>UK</td>
<td>16</td>
<td>A project aimed at making neuroimaging datasets freely available for download. Contains five existing datasets and dementia.</td>
</tr>
<tr>
<td>NSD, Multisubject, multimodal face processing</td>
<td><a href="http://naturalscenesdataset.org">http://naturalscenesdataset.org</a></td>
<td>Ultra high field MRI, fMRI, MEG</td>
<td>US</td>
<td>8</td>
<td>A public repository of unthresholded statistical maps derived from neuroimaging studies.</td>
</tr>
<tr>
<td>Neurosynth</td>
<td><a href="https://neurosynth.org">https://neurosynth.org</a></td>
<td>MEG, PET</td>
<td>US</td>
<td>Many, separate datasets</td>
<td>Aggregates existing results</td>
</tr>
<tr>
<td>NeuroVault</td>
<td><a href="http://neurovault.org">http://neurovault.org</a></td>
<td>MRI, fMRI, PET</td>
<td>US</td>
<td>416 (OASIS-1) 1,150 (OASIS-2), 370 (OASIS-3), 451 (OASIS-3; TAUS), 663 (OASIS-4)</td>
<td>A free platform that provides access to over 800 public datasets, all BIDS compliant. Formerly “OpenfMRI.”</td>
</tr>
<tr>
<td>Open Access Series of Imaging Studies (OASIS)</td>
<td><a href="https://www.oasis-brains.org">https://www.oasis-brains.org</a></td>
<td>MRI, fMRI, PET, MEG, EEG, iEEG</td>
<td>US</td>
<td>416 (OASIS-1) 1,150 (OASIS-2), 370 (OASIS-3), 451 (OASIS-3; TAUS), 663 (OASIS-4)</td>
<td>A project aimed at making neuroimaging datasets freely available for download. Contains five existing datasets and dementia.</td>
</tr>
<tr>
<td>OpenNeuro</td>
<td><a href="https://openneuro.org">https://openneuro.org</a></td>
<td>MRI, fMRI, dMRI, PET, MEG, EEG, iEEG</td>
<td>US</td>
<td>800+ datasets</td>
<td>A free platform that provides access to over 800 public datasets, all BIDS compliant. Formerly “OpenfMRI.”</td>
</tr>
<tr>
<td>OpenForest</td>
<td><a href="http://www.studyforest.org">http://www.studyforest.org</a></td>
<td>MRI, fMRI, dMRI</td>
<td>UK</td>
<td>217</td>
<td>A project centering around the movie Forrest Gump, providing highly reproducible scanning of rich contexts.</td>
</tr>
<tr>
<td>UK Biobank</td>
<td><a href="https://www.ukbiobank.ac.uk">https://www.ukbiobank.ac.uk</a></td>
<td>MRI, fMRI, dMRI, IDP</td>
<td>UK</td>
<td>60,000+</td>
<td>A biomedical database containing data from UK participants. The world's largest imaging study.</td>
</tr>
</tbody>
</table>

Note: This list is not exhaustive. MRI = magnetic resonance imaging, fMRI = functional MRI, dMRI = diffusion MRI, PET = positron emission tomography, MEG = magnetoencephalography, IDP = imaging-derived phenotypes, EEG = electroencephalography, TAC = tissue and angiography, SPECT = single-photon emission CT, DTI = diffusion tensor imaging, DKI = diffusion kurtosis imaging, US = United States, UK = United Kingdom, BIDS = brain imaging data structure. Other lists of neuroimaging databases are available (e.g., https://en.wikipedia.org/wiki/List_of_neuroscience_databases; https://imaging.mrc-cbu.cam.ac.uk/methods/OpenDatasets; https://sccn.ucsd.edu/~arno/fam2data/publicly_available_EEG_data.html).
2.10. Talk about greener computing

When we talk to neuroimaging colleagues about greener computing, many say they have not previously considered the carbon emissions associated with this aspect of the research process. This carbon footprint is perhaps challenging to intuitively conceptualise, compared to more visible sources of carbon emissions such as aviation. However, by reflecting on the above recommendations and actively discussing them with your neuroimaging community, you can raise awareness about this issue. In recent years, we have been inspired by increasing engagement on environmental sustainability issues within the neuroscience community, from conference attendance and participation in green neuroscience- and computing-themed sessions (e.g., British Neuroscience Association 2021, “Environmental impacts of computing in health & life sciences research” workshop 2023), to environmental chapters within neuroscience societies (e.g., the Sustainability and Environment Action Special Interest group in the Organization for Human Brain Mapping). As with recent drives for open science practices (Gorgolewski & Poldrack, 2016) and equitable publication costs (Sanderson, 2023) within this community, there is an appetite amongst neuroimagers to be more environmentally conscious in their work. Neuroimagers can—and we believe should—exercise advocacy in their own work, challenging established norms that exacerbate the footprint of the field.

Individual researchers may feel that their actions have a negligible impact on net emissions. However, fostering a culture of thinking seriously about these issues will contribute to the implementation of ideas into standard practice, as we have seen with the open neuroimaging movement. Systemic change from governments, institutions, and funders will be critical in facilitating this change. Alongside this, however, researchers should take small steps in their own work when possible. This can include the implementation of novel practices, such as “Environmental impact statements” (see Recommendation 2), HPC task scheduling, and carbon budgets (see Recommendation 4). We call on all neuroimagers to actively consider the carbon footprint of their work, especially that derived from computing, where there are already meaningful steps that can be taken.

Suggested Action: Discuss the importance of greener computing with other neuroimagers and advocate for systemic change

3. CONCLUSION

We have discussed 10 ways in which neuroimagers can reduce the carbon footprint of their research computing.
ETHICS STATEMENT

No novel data were collected for this study, and as such it was not necessary to obtain ethical approval. All figures were created using existing open access datasets.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available with the online version here: https://doi.org/10.1162/imag_a_00043.

REFERENCES


