Reproducible and Attributable Materials Science Workflows

Ye Li[1,*] Sara Wilson[2] Micah Altman[3,*]

Abstract

In this research we apply network analysis to data gathered through structured interviews and file forensics to evaluate the information flows and research workflows in materials science research projects. We discuss the gaps in reproducibility highlighted by this analysis, and suggest practical steps to making these scientific workflows more robust and sharable.

[1] MIT Libraries, Massachusetts Institute of Technology
[2] Department of Mechnical Engineering, Massachusetts Institute of Technology
[3] Center for Research in Equitable and Open Scholarship, Massachusetts Institute of Technology

[*] Correspondence: Ye Li <yel@mit.edu>, Micah Altman <escience@mit.edu>

Introduction

Background

Increasing Attention to Reproducibility, Openness and Attribution in Science

Reproducibility is a foundation of science. Over the last two and half decades however, mounting evidence has called into question the reproducibility of findings in a continually expanding set of fields – leading to regular calls to systematically assess reproducibility and improve scientific practice ("Reproducibility and Replicability in Science" 2019). And more recently, there have been high-profile calls and initiatives by research societies, funders, and publishers to make scientific practice and data more open and transparent NASEM (2018)and to develop systematic attribution standards and (McNutt et al. 2018) practices for contributors to scientific publications and outputs.

Stakeholders in science are increasingly coming to the realization that the reproducibility of a scientific discipline needs to be empirically evaluated, and not simply assumed. A hallmark study by the National Academies ("Reproducibility and Replicability in Science" 2019) reviewing the state of knowledge on scientific transparency finds that the evidence base of non-replicability across all science and engineering research is incomplete: Scientific practices of replication are neither sufficiently consistent nor sufficiently enough to make confident statements about the rate of replicability in most fields. However, the major empirical studies of replication failure that have been conducted in the natural, clinical, and social sciences have yielded rates of replication failure that range from somewhat lower than 20% to higher than 80%. Further, the report found that there is an uneven awareness of issues related to replicability practices and awareness across fields and within fields of science and engineering.

Similarly, although many fields have widespread norms or even stated policies on research transparency (e.g. making data available after publication) and appropriate attribution of contributors, these policies are unreliable predictors of practice. See for example Savage and Vickers (2009) Empirical evaluation is needed to understand how and where these practices are followed, and what effects they yield.

Studies of Practices in Experimental MSE

Schechtman's Nobel-winning discovery of quasi-crystals stands a particularly occurrence (and eventual resolution) of the classic "file-drawer" problem (Timmer 2011) that is highlighted by open-science advocates — but this is one illustration, with a happy ending, and cannot establish a pattern. There are few published studies that describe or evaluate practices related to replication, transparency, and attribution in Materials Science and Engineering (MSE).

A more recent study suggests a rosier picture – an analysis of retractions in MSE publications finds a relatively low rate (0.03%) (Coudert 2019). However, while a high retraction rate is a signal of problems, generaly most non-replicable research is not retracted – so a low retraction rate does not strongly suggest replicability. Another recent study, examining data sharing practices in small MSE labs (Wilson, Altman, and Jaramillo 2019) revealed that while many researchers in materials science embrace the idea of open science, reproducible research, and data sharing, they are frustrated with the inadequate infrastructure, tools, and practice guidelines. This finding suggests the potential for gaps between aspiration (for reproducibility, openness, etc.) and practice. Perhaps most concerning, however, is a recent set of case studies (Han, Walton, and Sholl 2019) in published in the *Annual Review of Chemical and Molecular Engineering* that found a high (20%) rate of reproducibility failure in the two research areas, the properties of metal-organic frameworks (MOFs) and synthesis of crystalline nanoporous materials, were targeted for study. A 2017 study on isotherm measurements in MOFs also revealed a similar level of irreproducible rate (Park, Howe, and Sholl 2017).

Experimental materials science typically does not generate large quantities of data through coordinated or collective studies compared to fields such as geology, genomics, and some disciplines within economics. In MSE, experimentalists generate materials property data in their 'small labs' individually and has not developed a shared practice of data sharing as in many other 'big data' disciplines. Moreover, gap of experimental data availability has been identified as a barrier for computational materials science since the early 1980's (Westbrook and Rumble 1983) and remains a significant obstacle to progress.

Rapid progress in data science and the ever-increasing number of demonstrated applications of data science approaches in data-rich fields produce optimism that data science can be productively applied to materials science as well ("Technology: Sharing Data in Materials Science" 2013). Significant progress in this direction requires significant data resources. Pioneering studies highlight the difficulty in assembling large quantities of experimental materials science data that can then be the basis for useful and insightful inferences (Raccuglia et al. 2016). In the past decade, the renewed promise of machine learning and their applications in materials science has made the need of FAIR experimental data more urgent (Blaiszik et al. 2016). Further, the application of machine learning and artificial intelligence to materials science at scale has been identified as a grand challenge for the discipline that is dependent on robust tools and practices for data sharing and replicable workflows (Stein and Gregoire 2019). In March 2022, National Science Foundation (NSF) issued NSF 22-055 *Dear Colleague Letter: Effective Practices for Making Research Data Discoverable and Citable (Data Sharing)* (Foundation, n.d.). NSF 22-055 emphasized that the Division of Matierals Research (DMR) expected its awardees to embrace NSF's policy on data sharing (Foundation 2022) and provided their guidance on preparing Data Management Plan (DMPs) (https://www.nsf.gov/bfa/dias/policy/dmpdocs/dmr.pdf)

Data resources can grow through open-science practices such as sharing data generated across the research lifecycle, but experimental materials science lacks the norms, standards, and tools to make this widespread, especially for academic labs. There have been notable efforts to develop infrastructure, standards, and tools to enable experimental reproducible workflow management and data sharing in materials science (Hill et al. 2018). For example, the 4Ceed project (Nguyen et al. 2017) developed a cloud frameworks and associated curation services for real-time capturing of materials data from instruments based on a survey they carried out among experimentalists ("User Study and Survey on Material-Related Experiments" 2016). The Materials Data Facility (MDF) service launched in 2016 (Blaiszik et al. 2016) was designed to provide an interconnection point for data sharing, discovery, access, and analysis. The MDF (Material Science Data Facility), sponsored by the National Institute of Standards (NIST) and the Center for Hierarchical Materials Design (CHIMaD), now hosts about 578 datasets (116 experimental datasets) and indexes over 970,000 records of Materials data from other repositories as of December 2021. Other recent efforts include infrastructure for a federated registry of information resources for materials science (Plante et al. 2021) ; a proposed controlled vocabulary and metadata schema for materials discover (Medina-Smith et al. 2021); and a new experimental infrastructure under development for the integration of Electronic Lab Notebooks and data archiving systems with materials science workflows (Brandt et al. 2021). In industry, software platforms (e.g. Citrine Platform (Informatics 2022)) that combines the data management infrastructure and AI-based tools facilitating materials design provide customizable solutions for corporate labs, which has more consistent pipeline workflows and can afford the resource-intensive infrastructure. FAIR-DI (FAIR Data Infrastructure for Physics, Chemistry, Materials Science, and Astronomy e.V.), an European originated effort, aims at building a reliable infrastructure for data from materials science, engineering, and astronomy that follows FAIR principles (FAIR-DI 2022). FAIR-DI launched the NOMAD repository (https://nomad-lab.eu/) in 2014 and has since been developing support to data managment and sharing. Their recent FAIRmat hands-n Tutorial Series (FAIR-DI 2022) is designed to provide connections between the existing infrastructure and researchers' daily practices.

Notwithstanding these particular efforts and the overall progress made in the area of developing tools standards and practices, the adoption of these infrastructure and tools by individual "small" labs remains limited. No direct solutions have been provided for individual labs to streamline their workflows and efficiently prepare their data for sharing throughout the research lifecycle.

Research Questions

In this study, we aim to identify potential gaps and challenges in small-lab MSE replicability, data availability, and attribution, through an in-depth analysis of the practices supporting workflow and data management at a leading lab.

To identify the gaps and opportunities in the current research practice for such improvements, we designed our study to answer the following research questions.

1. To what extent does research depend on manual processes for information management?
2. Explicit processes:
   (a) What processes concerning data and research workflow management are documented?
   (b) To what extent are documented processes consistent with practice?
3. To what extent are documentation processes complete enough to support replication of a result by another person within the lab (without further communication with the original researcher)?
4. To what extent are data management processes robust enough to survive the departure of a project member or the loss of an individual's personal computer or storage?
5. To what extent are workflow data, outputs and documentation sufficient to describe responsibility (or support attribution) for published results?

This study focus on practices within the research group, for a number of reasons. First, internal data management is a prerequisite for external data sharing and transparency. If research information created by one research becomes unavailable, uninterpretable, or irreproducible for a close team member, there is little hope it can be made meaningfully available for external reuse and review. Second, MSE relies in large part on internal processes to guarantee replicability – there are no formal processes for external validation, systematic studies of replicability conducted across the field, nor systematic reporting guidelines for reporting failures. Further, null results and those deemed uninteresting may end up in the file-drawer, and thus not made available for any external examination. Moreover, even published results that are of sufficient commercial value for an enterprise to attempt them in production may fail , and be discared without any subsequent reporting. Third, similarly, MSE relies almost entirely on internal processes to ensure appropriate attribution of work.

Data and Methods

Interviewee Selection

The use of 'small lab' is common in the literature, but often used unaccompanied by a precise definition. Within this paper we use the term 'small lab' to refer to a set of researchers that: (a) self-identified as research collective (b) aims to conduct research and produce scholarly communications, (c) is substantially responsible for identifying its research agenda, design, and methods (d) contains under twenty people, and (e) conducts experiments.

Although it is not possible to precisely determine the number of 'small labs' in science generally because no comprehensive survey of research groups exist. However past research into research groups size in selected disciplines and countries (e.g [@cook2015,@brandt2012,@qurashi1984,@seglen2000]) suggest that 'small' research groups are a common or the predominant form of organization within the natural and applied sciences.

With respect to MSE, the total number of research groups is unknown. However, public rankings of universities establish that at least seven-hundred and fifty academic materials science programs world-wide exist – and it is likely that a substantial proportion of these include small MSE labs

Professor Rafael Jaramillo's group conducts experimental materials science within the Department of Materials Science and Engineering at the Massachusetts Institute of Technology. Their research focuses on synthesis, properties, and application of electronic materials. Each research project in the group generates many experimental datasets and can be supplemented by computational studies for revealing mechanisms or analyzing structures. This typical type of workflow bonds the four elements of MSE research, structure/composition, synthesis/processing, properties, and performance (Flemings 1999).

The Jaramillo lab meets all of the criteria of a small MSE lab – it has a scientific aim, collects its own data from local experiments, comprises less than 20 FTE, provides its own scientific direction, and oversees its own methods and infrastructure. However, the lab is unlikely to be statistically representative of small materials science labs, for a number of reasons.

External rankings of MIT's materials science department place it in the top five schools worldwide. MIT is a well-resources institution, and MIT faculty are typically well-supported. MIT faculty in general, and Professor Jaramillo specifically, are generally successful in obtaining external research support. Professor Jaramillo himself is interested in reproducible research and open science: he has published in this area; group has developed related software prototypes and grant proposals; and he has advocated for reproducible and open science practice within his institution and discipline. Thus this lab should be considered a near-best-case for FAIR data workflows in small materials science lab – it is implausible that many other small experimental materials science groups have the resources, experience or interest to do substantially better in this area.

Synergistic collaboration between members of the group, including graduate students and postdoctoral fellows, allows for the continuous monitoring of lab equipment. This collaboration is facilitated by shared repositories of information, including a group Electronic Lab Notebook (ELN) in LabArchives, a group Dropbox account, and Google Drive. Accesses to all the cloud-based storage and services are provided to the group via MIT campus-wide site licenses. Protocol for saving and sharing information is specified in a group manual, which all researchers in the group are encouraged to follow for both the group repositories and their personal data storage systems. In this way, this lab is representative of good practices for data sharing, as individual data from one researcher is, ideally, stored in a format that is comprehensible to and a location that is accessible by all members of the lab. Consequently, reproducibility of research is possible in the absence of the originator of the research.

Investigating the workflow of four researchers within the Jaramillo group highlights which practices are most essential to open and reproducible research - these practices appear to be standardized across the researchers in the lab despite idiosyncrasies due to personal preference. Identifying these practices allows for other "small academic labs" to formulate and adopt the most effective structure for their data storage framework.

Data Collection Methods

We conducted structured interviews with 4 graduate students in Jaramillo's group to obtain the specifications of their workflow, data profile, and challenges in daily practices. This study (Exempt ID: E-2317) has been determined to exempt from further review by the Committee on the Use of Human as Experimental Subjects (COUHES) at MIT on June 2, 2020.

Each graduate student was interviewed by two researchers: one served as the interviewer and the other as the transcriptionist. The interview audio was recorded and reviewed in comparison to the transcribed notes post-interview for completeness and accuracy.

The interview protocol (see Appendix I) consisted of three sections: Interviewee Background, Top Priority Project Background, and Top Priority Project Workflow. The protocol served as a guideline for the interviewer to construct the most complete narrative of each student's workflow. Each question was either specifically asked or indirectly answered through the student's response to a different question.

For the first section, Top Priority Project Workflow, it became evident that the most natural interview process was one in which the student first described his or her workflow, and then the interviewer asked follow-up questions for details which were incomplete or missing.

Interview Coding

The aim of the interview coding process was to describe the each step of the workflow in a systematic structure database. There are a wide range of existing formal models for provenance and workflow (see for example (Jandre, Diirr, and Braganholo 2020)). However, most of these designed for automated execution, and contain much more detail than is feasible to elicit during a standard interview. We thus used a simplified coding approach in which the actions each described action was labeled with its objective, task, and subtask, and, the sequence, actor, input, output, data source, data target, equipment, and methods used were recorded using standardized codes. This tabular data was then used to impute collaboration and data-dependency graphs (see the *Results* section below).

The coding of the research workflow was conducted in four phases.

The first phase involved the direct translation of each interviewee's narration. During this phase, only the steps in the sequence that were explicitly stated were recorded.

The second phase was interpretation: the intended meaning of each statement was derived through assessing what the researcher implied but did not explicitly state. Each step of the workflow sequence has a series of subsequences that occur before and after the main objective. For example, when a physical material is placed in storage, it is implied that the next step involving said material requires its removal from storage. The first and second phases were completed for all interviewees prior to progressing.

The third phase was inference. The same synthesis, characterization, and analysis techniques were often used across interviewees, and each lab member was subject to the same regulations to achieve each objective. Therefore, knowledge of one interviewees workflow can be derived from what is known from another's workflow. This was used particularly for details such as the names of analysis software and data output formats.

The fourth phase was extrapolation. The primary coder of this data is a materials scientist who conducted research in the same facilities as those used by the interviewees. This familiarity allows for inferring implied steps from the workflow narrative that may not have been uncovered during the interview process.

No additional assumptions were made during the coding process. Any gaps in information that could not be acquired through these four steps were left blank.

Data Validation

After the interviews were completed and data from them was coded and analyzed, we validated the results with follow-up interviews and a documentation review. In the followup interviews, we reviewed with each subject the gaps presented by the preliminary analysis, confirmed whether either the subject believed the gap to exist, and addressed the gap in the workflow − through some action not noted in the original interview, the gap was addressed in some other manner. Where these discussions pointed to workflow steps that had been elided during the initial interview, we updated the workflow graphs to include these additions.

We also reviewed the content of the existing group storage systems (specifically, names, directories, and file types) to characterize patterns of data storage per project and compare these to the patterns implied by the workflow analysis. In addition, we use analysis of content to compare information organization naming practice with the documented lab policies.

Results

Interview Overviews

To summarize each interviewee's workflow, which is visualized from the interview codings:

Interviewee 'A' is involved in the generation of each output through the workflow sequence: sample preparation, synthesis, characterization, and analysis. The workflow sequence is iterative. Therefore, the results from the analysis phase inform how the synthesis process for the next iteration will be tuned. They uss a personal LabArchives notebook to record observations. Metadata from equipment is recorded in the group Dropbox. Pre- and post-processed data are saved to the group Dropbox.

Interviewee 'B' received the synthesized sample from a collaborator. They are responsible for preparing the sample for analysis, characterizing it, and analyzing the data. They use a personal LabArchives notebook as an electronic lab notebook, so any conditions needed to interpret and replicate a process are recorded. The group LabArchives notebook is used for recording measurements on lab tools that are shared, as to maintain a consistent tool log (required by the Professor). The group Dropbox is used for saving raw data that is directly output from instruments. Post-processed data is saved to a personal Dropbox.

Interviewee 'C' received the synthesized materials from a collaborator. They prepare the acquired sample for analysis, characterizes it, and analyzes the results. Finally,they transforms the sample via a laser set-up; this process is iterative, as the transformed sample is characterized. A personal OneNote notebook is used for experimental notes. OneNote is manually synchronized to the group LabArchives notebook. In the group Dropbox, they record all sample notes, raw data, and analyzed data.

Interviewee 'D' is directly involved in each step of the sequence, which includes sample preparation, characterization, analysis, and simulation. A personal LabArchives notebook is used to write details of each experiment and record measurements from equipment from the synthesis process. The group Dropbox is used to save equipment metadata and only raw or lightly processed data from characterization. A personal Dropbox is used for processed data.

All four interviewees used some instruments or equipment outside their own lab, either at a shared facility or in a collaborator's lab. Each interviewee saved a copy of raw data from those instruments in the group Dropbox but had different practices with transferring data. Each in-house instrument in the lab is overseen by an unofficially designated group member for its maintenance. Regular maintenance notes for each in-house instrument are recorded in the shared LabArchives notebook folder.

Group members regularly use equipment outside of the lab and outside of MIT − which interviewees indicate create additional challenges for data transfer and documentation. Interviewees noted that equipment within the MSE Department is locatable through an internal wiki − but there is no other central documentation or standardization around equipment configuration, data transfer, network access, or acknowledgement of equipmnet use.

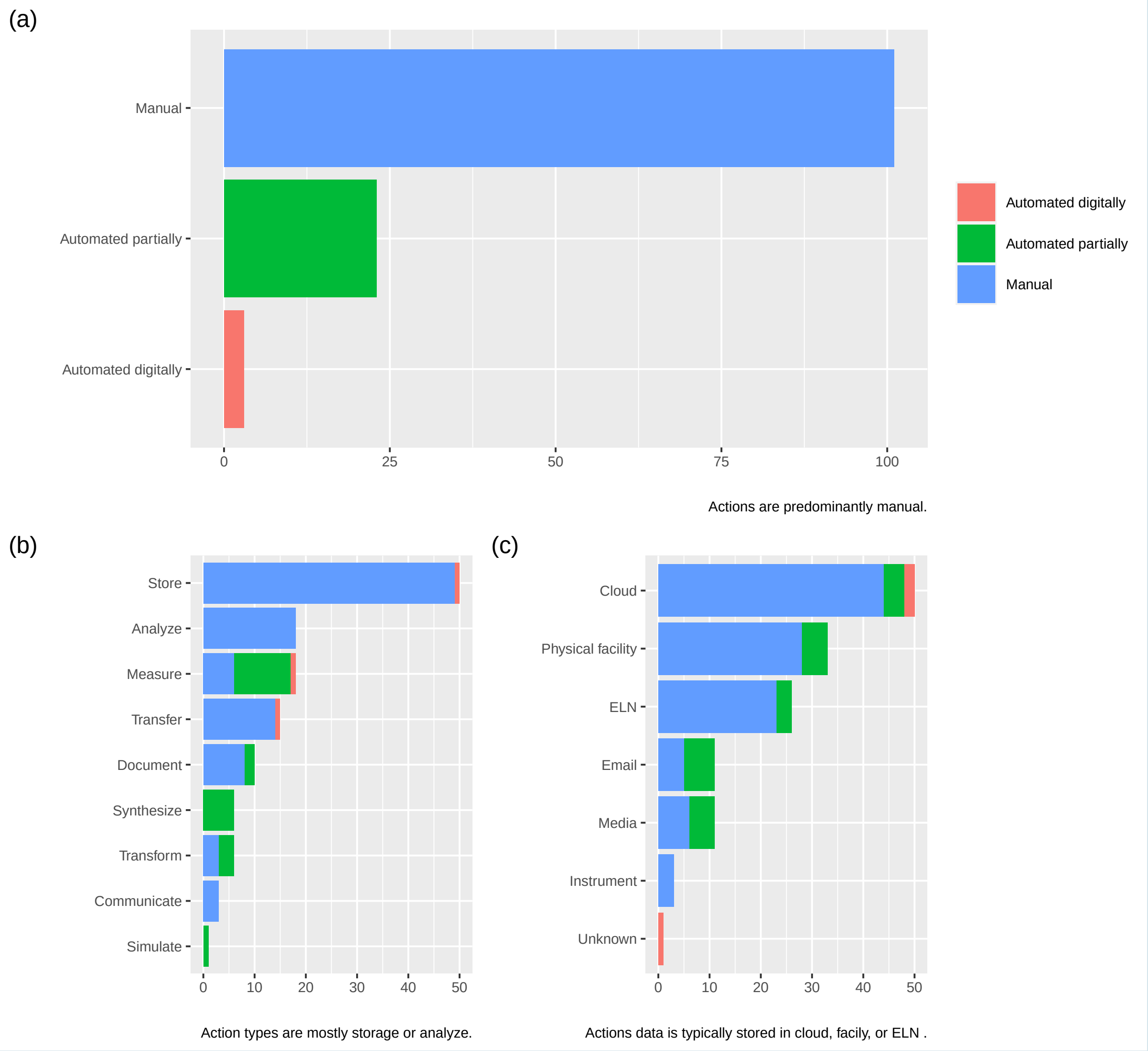
Workflow Automation

While there is very little systematic research on the rate and frequency of human errors in scientific research generally (or MSE specifically) a long history of research in the fields of human performance and reliability engineering that suggests that in the human error rates are substantial in the absence of well-engineered monitoring and error-mitigation regimes Jacobs (1995). For example, over the last fifteen years, human error in medicine has been a focus of study – and systematic reviews demonstrate both the high level of harmful and avoidable human error, and the efficacy of error-reduction processes such as adoption of automated recording systems, and the use of explicit checklists and logs for manual proceduresRodziewicz and Houseman (2021).

It is not possible to determine from the data observed whether or not these workflows contained error – instead, this analysis aims to identify areas of potential risk. Figure 1 summarizes selected characteristics of the workflow process.

Figure 1: Selected characteristics of the workflow steps

Summarizes type of action described in each step, proximate data source, and level of automation.



Actions are predominantly manual.

Action types are mostly storage or analyze.

Actions data is typically stored in cloud, facily, or ELN .

(N =127 workflow steps)

This figure provides an answer to the first research question, concerning automation:

1. To what extent does research depend on manual processes for information management?

Each workflow step was coded as 'automated' if it was performed without any manual intervention; as 'partially' automated if the operation was launched manually, but was completely described by digital metadata (e.g. configuration files), or as 'manual' if it depended on manually entered parameters for correct operation. These figure shows that workflow steps are predominantly manual, and that storage and analysis steps are almost entirely manual.

Automation is not a panacea, and can increase systems complexity or decrease local transparency in ways that make increase errors across a broader system. Notwithstanding, automation is often recommended for tasks that do not involve complex judgment (e.g. file transfers), and that are not otherwise associated with specific procedures for performance, audit, and quality assurance. Further, targeted automation enables people to shift their efforts to tasks where judgment is required, and reduces the cost and effort of logging and auditing – so where errors do occur they are more readily detected.

Further, explicit communication and documentation are relatively infrequent; there is a high level of reliance on manual transmission of information (e.g. for instrument setup, or for contextualization of the analysis); and a substantial incidence of e-mail and portable media for information storage. Together this suggest that there is substantial opportunity for human error in data management and organization.

Workflow Dynamics

We represent these workflows as formal graphs, and then apply social network analysis methods. (This follows a common approach to interpretation of workflows, first documented by (Tan, Zhang, and Foster 2010).) The graph systematically describes all process, informational, and collaboration dependencies elicited through the interview process. Through analysis of this graph we can identify workflow gaps, evaluate processes with respect to stated policy, and probe potential interventions. Further, the graph structure provides a natural way of visualizing these workflows.

The process of creating the graph is summarized below. (For replication purpose, we have placed in a public archive all of the de-identified and coded interview data, and the software code necessary to construct the graph in detail, as well as all of the code necessary to reproduce all figures and table.)

- Each atomic action ("step") in the workflow is represented by a node on the graph. The node documents all of the characteristics of that single actions.
- Process dependencies are represented through sequences and sub-sequences. linked by "process" edges:
    - Actions that are all performed by the same person, in a required sequence, fore a single goal, and over a continuous period of time are represented by "sequence" nodes. Each sequence is linked by edges to one or more child sub-sequences.
    - Actions that are all performed within a sequence (and thus by the same person) and are practically simultaneous (they have no natural order, and occur during a brief period) are represented by sub-sequences. Sub-sequences are linked by edges to one or more child steps.
- Informational dependencies are represented by augmenting the graph with "informational" edges. An edge is created whenever one of the following conditions hold.
    - When nodes share common data inputs – this represents passive information sharing.
    - When the output of one node is the input for another – this represents active information sharing.
    - When nodes are conducted by the a single person during an continuous time (i.e. they are part of the same sequence) – this represents implicit information sharing.

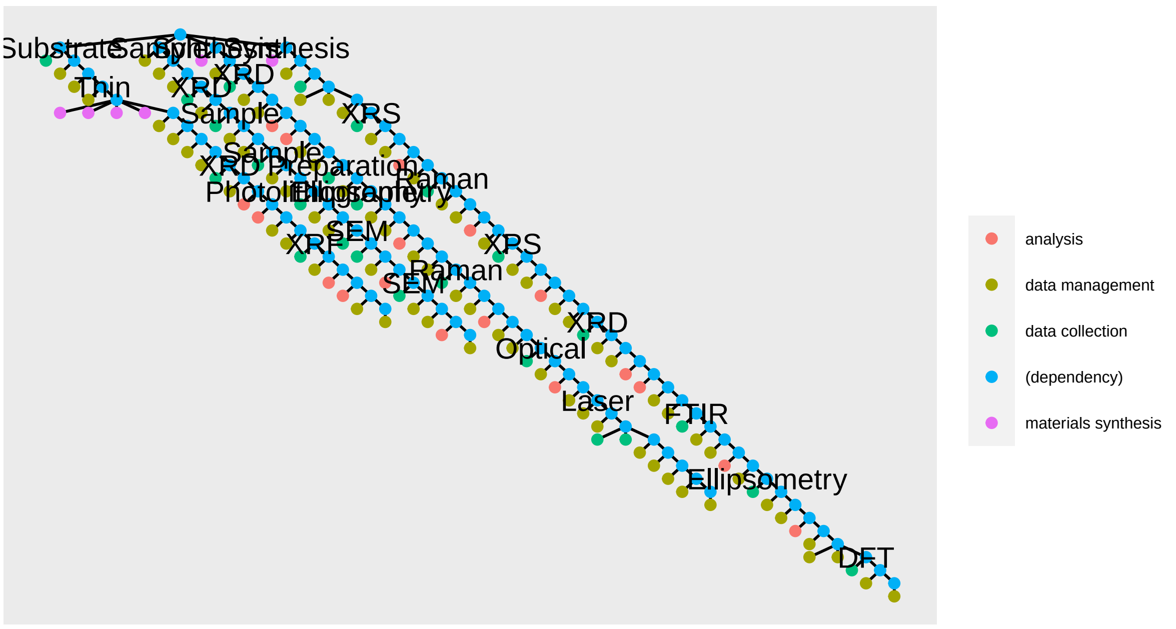
- Collaboration (attribution) dependencies are represented by augmenting the graph with typed nodes and edges.
  - Collaborator nodes represent individual or organizational collaborators.
  - Edges are created from workflows to collaborators when either the collaborator is explicitly referenced in the action (e.g. sending results to a collaborator, receiving samples from a collaborator) or by implication – when the action involves some instrument (or other tool) provided by a collaborator.

The three figures below present the entire set of workflows through the lens process dependency, information dependence, and attributional dependence (respectively).

Figure 2 illustrates the workflow processes over time. The workflow is hierarchical – each of the projects do not interact (there are no connecting branches), and the work can be represented as a set of independent self-contained tasks. (Summary graph statistics are shown in the Appendix, Table A1.) Most of the tasks contain only one atomic action. Further, there is a rhythm across each workflow in which the type of task at each step alternates.

Figure 2: Workflow steps and dependencies.

Four workflows visualized by phases and individual step. Each project is independent of the others, proceeds in a roughly linear fashion.

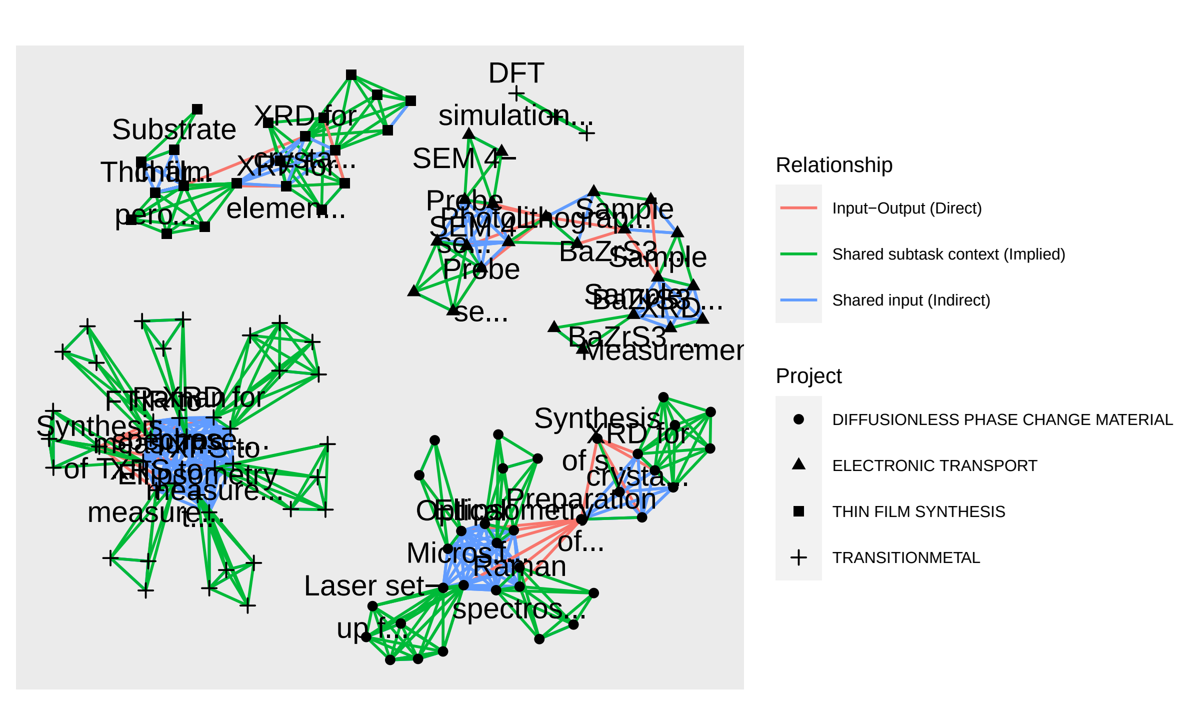


(Step types have been grouped into categories for visual clarity.)

In contrast to process flow, as shown in Figure 3, the information flow is connected in dense clusters within projects. Most information flow is implicit – through shared context. Information rarely flows through direct input-output. There is no information flow between projects.

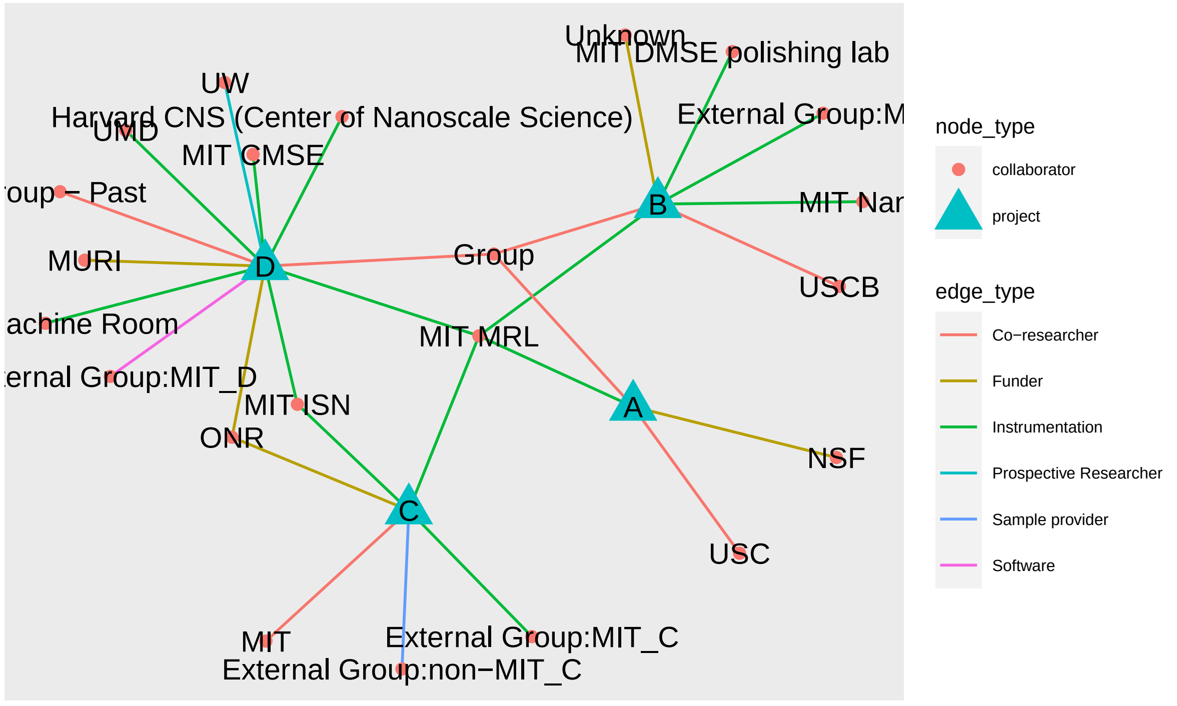
Figure 3: Project Information Exchange.

Implicit and indirect information exchange occurs frequently within projects, but does not connect projects.



Collaboration networks (Figure 4) are also partitioned by project/workflow. The size of the network varies substantially across projects.

Figure 4: Project Collaboration.



Lab Practices

Characterizing Documented Practices

Our first research question concerns documented practices:

> 2(a) What processes concerning data and research workflow management are documented?
> 2(b) To what extent are documented processes consistent with practice?

We identified the practices through direct interview questions administered during face-to-face interviews of the PI and group members. We then obtained copies of the documented processed from the subjects in order to characterize these.

During the interviews, one interviewee, who was self-identified as a founding member of the group, mentioned a document titled “Jaramillo group new member checklist”, which described shared computing resources, lab safety and access, as well as data management practices. The interviewee shared the document afterwards, as a part of their Group Handbook. We reviewed the document along with the Computing Resources page on the group Wiki site to summarize the essential practices required for each group member. We also compared the documented group practice with the self-reported individual practices we summarized from interviews.We reviewed these documents both for relevance to the research question above, and for to inform the interpretation of the workflow networks below.

We identified the the practices most relevant to data-management and scientific workflow management, and grouped them into three categories – information sharing, information security, and information organization:

Information Sharing.

- P1. Shared data storage and management resources include a shared group account in Dropbox, Group wiki, a shared group lab notebook in LabArchives, and a group Zotero account for sharing literature references.
- P2. All raw data (defined as "data as-recorded by the measurement instruments") are required to be stored in the group Dropbox folder and should never be modified. All internal lab computers are configured to automatically save data to the group Dropbox folder. Data collected outside the group lab must be manually transferred to the group Dropbox folder. Examples of raw data given are JPG from microscope, TXT from a probe station, or files in proprietary format such as RAW from XRD.
- P3. Group members can store their analysis results wherever is most convenient.

Information Security

- P4. Group members are required to use MIT Enterprise version of CrashPlan to keep group-owned and individual computers backed up, especially the directories containing data or codes.
- P5. Group members are requested not to store raw data outside of group-managed storage.

Information Organization

- P5. The group Dropbox folder should be kept organized using the folder structure: `instruments\username\YYMMDD\samp`
- P6. Samples are required to be named consistently with a given scheme including `YYMMDD` and a serial number.

Consistency of Documented vs. Observed Practices

We employed four strategies to evaluate the consistency of observed practices with documented practices:

1. In general, we identified all instances where subjects explicitly referred to documented and/or established practices during interview, either during the description of their project, or separately.
2. With respect to information sharing practices, we used network analysis of the workflows, to identify where each information object was stored, and compare this against documented policy.
3. With respect to information security processes, whenever the network analysis identified information as stored only in a non-group location, we verified with the subject whether the location was backed up using CrashPlan or an equivalent MIT service.
4. With respect to information organization, we reviewed the group dropbox file listing to confirm practices.

The results are summarized below

Table 1: Comparison of Documented Group Practice with Self-reported Individual Practices

Documented Procedures	Inconsistencies with Practice
Information sharing	Practices are predominantly consistent with documentation, although occasional lapses occur.
Information security	Practices are consistent with documentation.
Information organization	Practices are frequently inconsistent with documentation, however the instrument, username, data and sample can often be identified by human inspection of the file and directory name.

The largest deviation with formal documented practice is in the area of information organization. Of the 31929 deposited over two years of proximity, less than a quarter (23%) provided could be readily assigned a collection date, researcher, and instrument.

The **Assessment** section below, provides more detail on information sharing.

Process Robustness Assessment

To address the remaining research questions we measuring and comparing the mathematical graphs describing the workflow process, information, and collaboration.

Internal replicability

The next research question concerns internal replicability:

> 3. To what extent are documentation processes complete enough to support replication of a result by another person within the lab (without further communication with the original researcher)?

Generally a documentation process may be implicit or explicit, and the documentation may be integrated with analytic outputs or separated stored. As noted in the previous subsection, the documented practice in this lab does not include active replication of results prior to publication, nor require that materials and instructions sufficient to replicate published articles be made available. Follow-up interviews (discussed at the end of this section) revealed that some projects have since adopted an informal local practice of depositing replication materials to the group drive after publication.

There group does exhibit practices of documentation during the process of data collection and analysis that would aid in future replication. The interviews and workflow analysis demonstrate the use of multiple documentation strategies. For example, some data (and analysis) formats and systems provide the capability to store information about how the data (or analysis) was produced, and how it is to be interpret. When this capability is used, we describe the documentation as as integrated into the data (equivalently, one could refer to the data as "self-documenting".)

Much of the time, however, documentation is stored separately from the outputs produced by measurement, experiment and analysis. This separate documentation can be manually added by the researcher – e.g. adding a lab notebook entry, or notes file. Alternatively, documentation may be implied by a previous step – e.g. when a measurement process or is controlled by a configuration file already recorded.

We use the workflow information graph to identify each time data or analysis is produced. We then analyze the graph to match each output to potential documentation based on the following

- Outputs were coded as having “manual” documentation based on an analysis of the workflow graph to determine that both data and documentation objects were produced during the same substage, or supplementary statements in the interviews that a specific output was manually documented.
- Outputs were coded as having “integrated” documentation when the output format matched a specific formats that was confirmed through the interviews to be part of a general self documentation process.
- Outputs were coded as having “implicit” documentation when they were derived from process that were (semi-)automated and where either log files or generating scripts were also stored. The table below summarizes these categories of documentation:

Table 2: Documentation of outputs.
Missing documentation obstructs reproducibility.

	integrated	manual	implicit	missing
processed data	0 (0%)	8 (88.89%)	1 (11.11%)	0 (0%)
analysis	0 (0%)	8 (44.44%)	0 (0%)	10 (55.56%)
raw data	7 (50.00%)	5 (35.71%)	2 (14.29%)	0 (0%)

Note that the existence of documentation is necessary for unassisted replication, but is not sufficient: We did not evaluate the completeness of the documentation, if it existed – only its presence. Notwithstanding, in over half of the cases examined, documentation of analysis was missing. This obstructs future replication of results and publications – which will need to rely on communication with the researcher who conducted this analysis (and upon their memory), and trial-and-error.

Robustness of Storage Practices

The next research question concerns robustness of storage practices:

4. To what extent are data management processes robust enough to survive the departure of a project member or the loss of an individual’s personal computer or storage?

To probe this question we use the workflow information graph to identify all of collected data (both digital and physical samples that were created as part of the each scientific workflow), metadata, and analysis results. We than use the process the graph to trace the flow of these objects across tasks, and into storage location. And from this set of traces we can infer the content of designated group storage location post-analysis. The results are summarized in the table below.

Table 3: Proportion of output in group managed storage, by type.
A substantial portion of highlighted outputs are at risk.

metadata	44%
analysis	44%
raw data	79%
processed data	100%

Note: processed data includes derived, linked and cleaned data; metadata includes configuration files, output logs, and manual documentation

On the positive side, almost all data objects (with exceptions) are deposited into institutionally-managed shared-group storage by the end the process. This is consistent with documented lab policy, and is necessary for the work to support future data sharing, and for the workflow to be robust to the loss of an individual computer.

However, over half of the metadata/documentation and half of the analysis produced is never copied or transmitted to a group location but remains accessible solely from individually-owned media, computers, or accounts. This will decrease the utility of data sharing – as most of the data is not self-documenting; and threatens the replicability of analysis: If a group member were to depart, there is not sufficient information available to ensure that the work can be replicated or re-validated, even internally. Further, in a small number of cases (2) raw data was stored outside of group storage, contrary to documented policy.

We used the same approach as above to identify when analyses are dependent on manual information transfer, rather than being automated. Given the high frequency of manual operations documented in the previous section, it is not surprising that 100% of the analyses relied on manual information management at a previous step in the experiment and measurement process.

Serendipitously, file-forensics data collected from the lab shared storage system provides a glimpse of the reliability of manual transfer processes. By comparing the manually recorded date in the path with the automatically recorded date in the shared filesystem we can measure the delay between the data creation and deposit. For half of these files the delay is quite small (40% of these files were deposited within 1 days). However, a substantial percentage were considerably delayed (25% of files were deposited only after a delay exceeding 95 days). (Note that two mechanisms could produce significant delays. First, where raw data is collected and transfered by hand, errors, interruptions, or forgetfulness can contribute to the delaying deposit. Second, we identified during the validation interviews that some projects adopted an informal process of adding files associated with a publication – after that publication had been accepted. Those added files can include processed data files and descriptions of data collection and analysis processes. It is not possible to determine the proportion of lag attributable to each mechanisms, because of the inconsistent use of documented and naming practices, and the variation of undocumented practices.)

The two years of files examined also included a substantial number (863) of image files that contained internal creation-time metadata produced by the original creating software. By comparing this time-stamp with the shared file-system time-stamp we are able to compute the elapsed time between creation and deposit. For most of these files the delay is quite small – less than a workday (75% of these files were deposited within 5 hours). However, the distribution of deposit latencies has a long tail, with some files not deposited until months (3008 hours) after creation.

The final research question concerns attribution:

5. To what extent are workflow data, outputs and documentation sufficient to describe responsibility (or support attribution) for published results?

To examine the final question, we interviewed respondents to elicit lists of all the collaborators on the project and their general collaborative relationship. This list includes both active collaborators (e.g., actors who supplies material, performs an analysis, or contributes to writing for publication) and passive collaborators (actors who provide access to equipment or software). Through the interview we confirmed that there was no written or common process or policy with respect to recording or acknowledging collaborators. In assigning attribution, interviewees reported relying primarily on memory rather than written documentation and outputs.

A partial exception to the reliance on memory is an informal practice discovered during the file forensics analysis. A common practice was to structure the directory trees such that data produced by a specific instrument was contained under a folder named for the Principle Investigator. Where this practice was followed with a specific instrument we code this as documentation of the collaboration.

Workflows may document collaborations explicitly (e.g. through entries in a lab notebook, or an author line in an analysis document) or indirectly (through an e-mail correspondence history). To quantify the degree to which attribution relies on memory, we compared the the list of collaborators stated by interviewers to a list of collaborators that could be detected through workflow outputs and documentation. To do this we extracted direct and implied collaborators from each workflow step – e.g. when another person was recorded as doing the analysis, when the interviewee sent someone an analysis by e-mail or an analysis when an external instrument was used.

As expected from the interviews, many collaborators are altogether omitted from workflow documentation or action. The table below summarizes these omissions.

Table 4: Undocumented collaborators

Types of collaborations that were recalled, but not documented implicitly or explicitly.

Project	Undocumented Contributor	% of total collaborators
A	Co-researcher	33%
B	[None]	0%
C	Co-researcher, Sample provider	40%
D	Instrumentation, Prospective Researcher, Co-researcher, Software	40%

As shown in the table, a significant proportion of the collaborations could not be associated with either the process of the work, the information used in it, or the analysis produced by it.

Analysis Validation

We conducted semi-structured interviews with all participants to assess strength of (dis)agreement with the analysis described above and its main conclusions, and with the with the recommendations below; and to probe for additional comments, reflections and recommendations. Participants consistently expressed agreement with the analysis, and confirmed existence of the gaps we note.

Further a number of participants reflected that since the initial interviews, they had noted some of these gaps themselves and adopted informal practices within their project to address them. For example, one project had a local, undocumented, but intentional practice of on the occasion of formally publishing an article, to deposit into lab storage all analysis scripts necessary to reproduce the analysis in the article.

Moreover, participants agreed with all areas of recommendations. One caveat – most participants noted that they faced institutional challenges to automating data collection from instruments outside of the lab.

Discussion - Toward More Reproducible and Attributed Practices

The results of the workflow analysis above reveal the strengths and limitations of the current practice. Actual practices in the lab is, for the most part, observed to be consistent with the documented group processes, and is sufficient to mitigate against risk of data loss resulting from the failure of an individual computer and storage system.

However, the documented policies nor the actual practice is sufficient to generally ensure reproducibility or complete attribution, nor to mitigate the risk of loss of critical information if an individual's withdraws suddenly from from the group. We conclude that improvements are needed.

Several general strategies can be employed to address workflow gaps generally, and should be considered as an approach to the gaps discussed above:

- The addition of processes to regularly audit/validate ongoing projects for reproducibility and attribution.
- Changes to research infrastructure (defined broadly) to automate the capture, transfer and/or storage of critical information, preferably in standardized formats with necessary metadata.
- Changes to the lab policies with respect to requirements for those activities done manually.

Auditing. It is a truism that in order to be effective manual processes and policies must be regularly auditing and verified. Auditing and verification should evaluate both the use of documented practice and the achievement of desired outcomes

- Recommendation 1: With respect to the documented practices, minimal automated audits – in support sanity checks – could verify that documented naming conventions are being followed, and that systems are running backup software. With respect to outcomes, less frequent (e.g. semi-annual) manual audits could be used to validate that the current analytic results from each project can be reproduced from (or at least traced back to) data and metadata curated in the group storage.
- Recommendation 2: Even automated processes sometimes fail or are misconfigured. Automated validation can be used to detect system failures, and to flag unusual patterns of activity for further investigations. For example, automated analysis of group storage can be used to flag the absence of data collection and processing for purportedly active projects. Automated analysis of deposits could provide evidence of 'liveness' of projects and individuals. Automated analysis could also correlate the timing of lab notebook updates with the timing of data deposits into the group storage system – substantial data changes/updates without corresponding lab notebook signal a potential threat to reproducibility.

Upgrading infrastructure, where feasible, are attractive because they do not require people to make changes in behavior – which often is costly, difficult to assess, error prone, and requires consistent focus to maintain. While a fully automated infrastructure for materials science remains is currently too expensive and immature for many labs, smaller changes in infrastructure and tooling have the potential to mitigate a number of the gaps identified by the workflow analysis:

- Recommendation 3: All of the observed workflows involved the extensive use of personal portable storage to manually transfer data from experimental instruments. Further, the file-forensic analysis shows that there are the delays between file creation and deposit can be quite large. Moreover no systems are processes are in place that would detect common categories of human errors that occur at this stage, such as erasing or overwriting local files, loss or replacement of the storage device, failure to delete files after the transfer is complete, or transfer of the files to an incorrect destination (such as the user's personal computer or cloud) – should these occur. This suggests that reducing manual data transfer and operations will increase errors.

  Typically the portable storage used is a simple offline USB "flash drive". Alternatives USB-compatible portable storage devices are now readily available that include built-in wireless networking and data synchronization capabilities. Although researchers would still need to transport these storage devices with network connection to the instruments and plug them in, the manual data transfer to cloud storage could then be automated, reducing risk to reproducibility. Using this type of portable storage devices will not introduce more security risks for instruments in shared facilities than an offline USB “flash drive” would. (During the analysis validation interviews participants noted that enacting this recommendation will require agreement and action from the equipment and/or facilities owners in order to align information security policies.)

- Recommendation 4: Similarly, most workflows involved a significant amount of regular transfer from personal cloud storage (such as dropbox) to a group cloud storage. When having multiple independently managed locations is not necessary for data processing, analysis, and backup, eliminating the use of multiple storage locations will further lower the risk of introducing inconsistency. When multiple independently managed services are necessary, services are now readily available that can monitor target folders in one storage system and replicate or synchronize them with another. Using these tools, along with a more systematized practice of folder organization for work products kept in personal storage, would enable a more reliable and robust data lab practices, without sacrificing the convenience of a personal cloud-storage accounts.

Refining practices. Although infrastructure and audition of current practices can be expected to ameliorate the workflow gaps identified in this analysis, additional refinement to lab practices are also likely to be necessary in two areas:

- Recommendation 5: Develop explicit practices around collaborator attribution. Practices are needed to systematically identify the contributions of collaborators. This might include (a) enhancing existing workflow project documentation (e.g. in the lab notebook) to clearly identify when the researcher uses externally contributed resources, borrowed equipment, of information received from collaborator, (b) explicitly saving contributed data, analyses and comments from collaborators in the group storage, rather than in personal e-mail, and (c) defining contributor roles according to taxonomy, such as the CRediT (Contributor Roles Taxonomy, https://credit.niso.org/) in group documentations.
- Recommendation 6: Develop explicit practices around reproducibility beyond the stage of raw data.
    - Make documentation for standard practices at commonly used equipment in external locations (i.e. other MIT facilities, like MRL, CMSE). Consistent practices at these facilities would all for the comprehensible transfer of data between researchers within the lab
    - Establish a group shared location for metadata (especially equipment parameters) since this is essential to reproducibility. Monitor the progress in open data standards in the field and start to adopt them.
    - Encourage analyses be conducted in a framework that build-in reproduciblity – e.g. using executable scripts or notebooks stored in cloud storage, rather than spreadsheets transmitted by email.

Future Research

In this article we have identified gaps in an exemplar set of materials science workflow process, and characterized approaches to address those gaps. However, the effectiveness of specific practices and approaches is an open question: Empirical evidence, preferably from designed interventions are needed to reliably measure how better practices, can in general improve reproducibility and research attribution. (NASEM 2018; Altman and Cohen 2021) Moreover, these practices are embedded in and responsive to a much broader system of scientific incentives, institutional and organizational collaboration, and professional training (Altman and Bourg 2018) – research is needed in how effective practices can be aligned with incentives, training, institutional coordination, and infrastructure improvement. Intrinsically, recognizing the value of FAIR data sharing and computational use of experimental data for the research community in general and for their own study could further motivate individual researchers and their teams. Hiring data curators or research workflow facilitators to provide discipline-specific support for individual groups and departments could further enable researchers and overcome the barriers of starting new practices. Improvement of interfaces for human-computer interaction, accessibility and security of cloud-based systems could also be the key to lower the barriers for individual groups to fully adapt digital workflows recommended, especially when shared instrument facilities are often inseparable components of the infrastructure. With the improvement of research infrastructure for MSE that can integrate experimental data management and sharing as well as AI-based materials design, it will become critical to study how "small academic labs" could adapt to such infrastructure cost-effectively for open and reproducible research while maximizing the creativity.

Appendix

Interview Protocol

The interview protocol used in the study is shown in Appendix I.

Tables

Table A1: Workflow Process Statistics

Direct Connections	Graph Diameter	Mean Distance Between Steps
244	42	12.21399

Table A2: Information Exchange Graph Statistics

Direct Connections	Graph Diameter	Mean Distance Between Steps
1660	5	2.172682

Table A3: Collaboration Network

Direct Connections	Graph Diameter	Mean Distance Between Steps
29	1	1

Acknowledgements

The authors describe contributions to this article using a standard taxonomy. (Allen et al. 2014) All authors equally shared the core formulation of the research goals and aims. All authors co-developed the research design as well as the interview instrument and plan. M.A. lead the formal analysis and visualization. Y.L. lead administration, funding acquisition, and supervision. S.W. and Y.L. led the data curation (including collect), and investigation. All authors shared in writing of the original manuscript and in further refinement through commentary, review, editing, and revision.

The authors thank the Professor Rafael Jaramillo at MIT for his commentary and for enabling access to lab records, and the members of the Jaramillo research group for participation in interviews.

The authors thank MIT Libraries for the special fund and support to the project.

References

Allen, Liz, Jo Scott, Amy Brand, Marjorie Hlava, and Micah Altman. 2014. "Publishing: Credit Where Credit Is Due." *Nature* 508 (7496): 312–13. https://doi.org/10.1038/508312a.

Altman, Micah, and Chris Bourg. 2018. "A Grand Challenges-Based Research Agenda for Scholarly Communication and Information Science." *MIT Grand Challenge Participation Platform*, December. https://doi.org/10.21428/62b3421f.

Altman, Micah, and Philip N. Cohen. 2021. "The Scholarly Knowledge Ecosystem: Challenges and Opportunities for the Field of Information." http://dx.doi.org/10.31235/osf.io/ctdb9.

Blaiszik, B., K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, and I. Foster. 2016. "The Materials Data Facility: Data Services to Advance Materials Science Research." *JOM* 68 (8): 2045–52. https://doi.org/10.1007/s11837-016-2001-3.

Brandt, Nico, Lars Griem, Christoph Herrmann, Ephraim Schoof, Giovanna Tosato, Yinghan Zhao, Philipp Zschumme, and Michael Selzer. 2021. "Kadi4Mat: A Research Data Infrastructure for Materials Science." *Data Science Journal* 20. https://doi.org/10.5334/dsj-2021-008.

Coudert, François-Xavier. 2019. "Correcting the Scientific Record: Retraction Practices in Chemistry and Materials Science." *Chemistry of Materials* 31 (10): 3593–98. https://doi.org/10.1021/acs.chemmater.9b00897.

FAIR-DI. 2022. "FAIR-DI e.V. - FAIR Data Infrastructure for Physics, Chemistry, Materials Science, and Astronomy." https://www.fair-di.eu/about/info.

Flemings, M. C. 1999. "WHAT NEXT FOR DEPARTMENTS OF MATERIALS SCIENCE AND ENGINEERING?" *Annual Review of Materials Science* 29 (1): 1–23. https://doi.org/10.1146/annurev.matsci.29.1.1.

Foundation, National Science. 2022. "Dissemination and Sharing of Research Results- NSF Data Management Plan Requirements." https://www.nsf.gov/bfa/dias/policy/dmp.jsp.

———. n.d. "Dear Colleague Letter: Effective Practices for Making Research Data Discoverable and Citable (Data Sharing)." https://www.nsf.gov/pubs/2022/nsf22055/nsf22055.jsp.

Han, Rebecca, Krista S. Walton, and David S. Sholl. 2019. "Does Chemical Engineering Research Have a Reproducibility Problem?" *Annual Review of Chemical and Biomolecular Engineering* 10 (1): 43–57. https://doi.org/10.1146/annurev-chembioeng-060718-030323.

Hill, Joanne, Arun Mannodi-Kanakkithodi, Ramamurthy Ramprasad, and Bryce Meredig. 2018. "Materials Data Infrastructure and Materials Informatics." In, edited by Dongwon Shin and James Saal, 193–225. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-68280-8_9.

Informatics, Citrine. 2022. "What Is the Citrine Platform?" *Citrine Informatics*. https://citrine.io/product/what-is-the-citrine-platform/.

Jacobs, Philip. 1995. "*Human Reliability and Safety Analysis Data Handbook* by David I. Gertman & Harold S. Blackman 1994, 448 Pages, $69.95 New York: John Wiley & Sons ISBN 0-471-59110-6." *Ergonomics in Design: The Quarterly of Human Factors Applications* 3 (2): 33–34. https://doi.org/10.1177/106480469500300209.

Jandre, Eduardo, Bruna Diirr, and Vanessa Braganholo. 2020. "Provenance in Collaborative in Silico Scientific Research." *ACM SIGMOD Record* 49 (2): 36–51. https://doi.org/10.1145/3442322.3442329.

McNutt, Marcia K., Monica Bradford, Jeffrey M. Drazen, Brooks Hanson, Bob Howard, Kathleen Hall Jamieson, Véronique Kiermer, et al. 2018. "Transparency in Authors' Contributions and Responsibilities to Promote Integrity in Scientific Publication." *Proceedings of the National Academy of Sciences* 115 (11): 2557–60. https://doi.org/10.1073/pnas.1715374115.

Medina-Smith, Andrea, Chandler A. Becker, Raymond L. Plante, Laura M. Bartolo, Alden Dima, James A. Warren, and Robert J. Hanisch. 2021. "A Controlled Vocabulary and Metadata Schema for Materials Science Data Discovery." *Data Science Journal* 20. https://doi.org/10.5334/dsj-2021-018.

NASEM. 2018. *Open Science by Design*. National Academies Press. https://doi.org/10.17226/25116.

Nguyen, Phuong, Steven Konstanty, Todd Nicholson, Thomas O'Brien, Aaron Schwartz-Duval, Timothy Spila, Klara Nahrstedt, et al. 2017. "2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)." In, 11–20. https://doi.org/10.1109/CCGRID.2017.51.

Park, Jongwoo, Joshua D. Howe, and David S. Sholl. 2017. "How Reproducible Are Isotherm Measurements in Metal–Organic Frameworks?" *Chemistry of Materials* 29 (24): 10487–95. https://doi.org/10.1021/acs.chemmater.7b04287.

Plante, Raymond L., Chandler A. Becker, Andrea Medina-Smith, Kevin Brady, Alden Dima, Benjamin Long, Laura M. Bartolo, James A. Warren, and Robert J. Hanisch. 2021. "Implementing a Registry Federation for Materials Science Data Discovery." *Data Science Journal* 20. https://doi.org/10.5334/dsj-2021-015.

Raccuglia, Paul, Katherine C. Elbert, Philip D. F. Adler, Casey Falk, Malia B. Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A. Friedler, Joshua Schrier, and Alexander J. Norquist. 2016. "Machine-Learning-Assisted Materials Discovery Using Failed Experiments." *Nature* 533 (7601): 73–76. https://doi.org/10.1038/nature17439.

Reason, J. 1995. "Understanding Adverse Events: Human Factors." *Quality and Safety in Health Care* 4 (2): 80–89. https://doi.org/10.1136/qshc.4.2.80.

"Reproducibility and Replicability in Science." 2019. A Consensus Study Report. The National Academies of Science, Engineering, and Medicine. https://doi.org/10.17226/25303.

Rodziewicz, Thomas L., and John E. Houseman Benjamin Hipskind. 2021. *Medical Error Reduction and Prevention*. StatPearls. https://www.ncbi.nlm.nih.gov/books/NBK499956/.

Savage, Caroline J., and Andrew J. Vickers. 2009. "Empirical Study of Data Sharing by Authors Publishing in PLoS Journals." Edited by Chris Mavergames. *PLoS ONE* 4 (9): e7078. https://doi.org/10.1371/journal.pone.0007078.

Stein, Helge S., and John M. Gregoire. 2019. "Progress and Prospects for Accelerating Materials Science with Automated and Autonomous Workflows." *Chemical Science* 10 (42): 9640–49. https://doi.org/10.1039/c9sc03766g.

Tan, Wei, Jia Zhang, and Ian Foster. 2010. "Network Analysis of Scientific Workflows: A Gateway to Reuse." *Computer* 43 (9): 5461.

"Technology: Sharing Data in Materials Science." 2013. *Nature* 503 (7477): 463–64. https://doi.org/10.1038/503463a.

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." Edited by Cameron Neylon. *PLoS ONE* 6 (6): e21101. https://doi.org/10.1371/journal.pone.0021101.

Timmer, John. 2011. "Symmetry Free Quasicrystals Given the Nobel Prize in Chemistyry." https://arstechnica.com/science/2011/10/symmetry-free-quasicrystals-given-the-nobel-prize-in-chemistry/.

To Err Is Human. 2000. National Academies Press. https://doi.org/10.17226/9728.

"User Study and Survey on Material-Related Experiments." 2016, November. https://www.ideals.illinois.edu/handle/2142/94738.

Westbrook, J. H., and Jr Rumble. 1983. "Computerized Materials Data Systems." In. Steering Committee of the Computerized Materials Data Workshop,Fairfield Glade, TN. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=11 https://www.osti.gov/biblio/6969565-computerized-materials-data-systems.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1). https://doi.org/10.1038/sdata.2016.18.

Wilson, Sara L, Micah Altman, and Rafael Jaramillo. 2019. "Methods for Open and Reproducible Materials Science." https://doi.org/10.31235/osf.io/ag8zu.