Reproducible and Attributable Materials Science Workflows

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

Abstract

While small labs produce much of the fundamental experimental research in material science and engineering. However, little is known about their data curation practices and the extent to which they promote trust in, and transparency of, the published research. In this research, we conduct a case study of a leading MSE research lab to characterize the limits of current curation practice concerning replicability, data-sharing, and attribution. We systematically reconstruct the workflows underpinning four research projects by combining interviews, document review, and digital forensics. We then apply information graph analysis and computer-assisted retrospective auditing to identify where critical research information is unavailable or at risk. We find that while curation practices in this leading lab protect against computer and disk failure, they are insufficient to ensure reproducibility or correct attribution of work – especially when a group member withdraws before project completion. We conclude with recommendations for adjustments to MSE data curation practices to promote trustworthiness and transparency by adding lightweight automated file-level auditing and automated data transfer processes.

[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>

Mechanical

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 assess reproducibility and improve scientific practice systematically (*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.

Science stakeholders increasingly realize that a scientific discipline's reproducibility needs to be empirically evaluated, 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 conducted in the natural, clinical, and social sciences have yielded replication failure rates ranging from somewhat lower than 20% to higher than 80%. Further, the report found 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 & 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 as a particular 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. Few published studies 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 signals problems, most non-replicable research is generally not retracted, so a low retraction rate does not strongly suggest replicability. Another recent study examining data-sharing practices in small MSE labs (Wilson et al., 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 et al., 2019) 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 et al., 2017).

Experimental materials science typically does not generate large quantities of data through coordinated or collective studies compared to geology, genomics, and some disciplines within economics. In MSE, experimentalists generate materials property data in their 'small labs' individually and have not developed a shared practice of data sharing as in many other 'big data' disciplines. Moreover, gaps in experimental data availability have been identified as a barrier to computational materials science since the early 1980s (Westbrook & Rumble, 1983) and remain 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 also be productively applied to materials science ("Technology," 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 be the basis for valuable and insightful inferences (Raccuglia et al., 2016).

In the past decade, the renewed promise of machine learning and its applications in materials science has made the need for FAIR experimental data more urgent (Blaiszik et al., 2016). Further, applying machine learning and artificial intelligence to materials science at scale has been identified as a grand challenge for the discipline dependent on robust tools and practices for data sharing and replicable workflows (Stein & Gregoire, 2019). Last year, the NSF Division of Materials Research underscored the importance of transparent access to data by issuing specific policy guidance for the field (Foundation, n.d.).

Data resources can grow through open-science practices such as sharing data generated across the research lifecycle. Still, 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) (Himanen et al., 2019). For example, the 4Ceed project (Nguyen et al., 2017) developed a cloud framework 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 combine the data management infrastructure and AI-based tools facilitating materials design provide customizable solutions for corporate labs, which have 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.), a 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 been developing data management and sharing support. Their recent FAIRmat hands-on Tutorial Series (FAIR-DI, 2022) is designed to provide connections between the existing infrastructure and researchers' daily practices.

More recently, as a paradigm shift rooted in the exponential growth of computing power, integrated systems of Artificial intelligence (AI) based predictions and experimental automation via robotics are explored and examined to accelerate materials discovery with the promise of replacing the manual and human-intensive material discovery process (Pyzer-Knapp et al., 2022). For example, a technology roadmap was outlined to articulate the hardware and software infrastructure requirements and demonstrate a re-imagined role of humans as ensuring data is appropriately managed, aggregated, standardized, and shared (Delgado-Licona & Abolhasani, 2023). The analysis of the potential to apply accelerated materials discovery in clean energy highlighted insufficient experimental datasets for AI model training as a limitation for clean energy as a relatively new technology (Maleki et al., 2022).

Notwithstanding these particular efforts and the overall progress in 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. Instead, these labs use informal sociotechnical workflows that combine documented procedures, undocumented conventions, semi-automated tools, and manual processes. In this research, we elicit the informal workflows operating within a top material science lab, document and describe these using a formal workflow graph notation, and analyze these workflows using qualitative and mathematical graph analysis.

Research Questions

This study aims to identify potential gaps and challenges for small-lab MSE research replicability (trustworthiness), data availability (transparency), 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 probing the trustworthiness and transparency of MSE data curation:

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 another person's replication of a result 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?

Data and Methods

Overview

This study focuses on practices within the research group for several 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 of sufficient commercial value for an enterprise to attempt them in production may fail and be discarded without any subsequent reporting. Third, similarly, MSE relies almost entirely on internal processes to ensure appropriate attribution of work.

Our approach employs a purposive case-study design. We select a leading MSE research lab and interview selected lab members in detail about their most important research project. As illustrated in Figure 1, we code the collected interview data to create a standardized description of each task conducted for the project step and use this to construct a formal workflow process graph. By tracing across and within these graphs, we characterize overall collaboration and information flow patterns – and evaluate the extent to which information shared with the group is sufficient for replication and attribution.

We supplement these workflow graphs with information gained through a review of lab documentation and an audit of the lab's digital repository. Through manual coding of the documentation, we can extract rules -- which can be tested using extracted file-level metadata (digital forensics) collected through the audit.

Case and 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 a 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 exists. However, past research into research group size in selected disciplines and countries (e.g. (Brandt et al., 2021; Cook et al., 2015; Qurashi, 1984; Seglen & Aksnes, 2000)) suggest that 'small' research groups are a common or the predominant form of organization within the natural and applied sciences.

Concerning 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 worldwide exist – and a substantial proportion of these likely 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 the 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

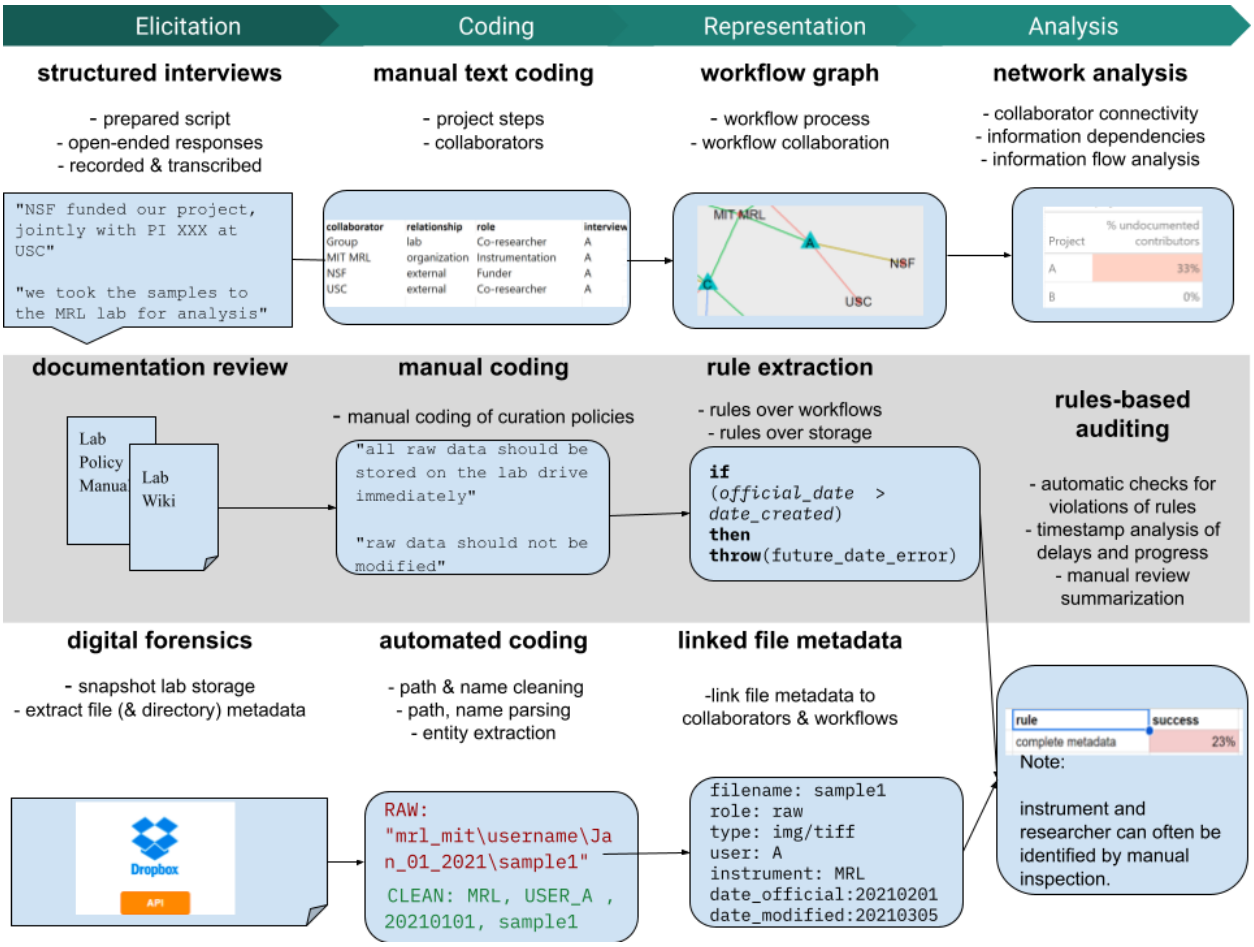


Figure 1: Fig 1: Overview of data sources and methods

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 FTEs, 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 several reasons.

External rankings of MIT's materials science department place it in the top five schools worldwide. MIT is a well-resourced institution, and MIT faculty are typically well-supported. MIT faculty, and Professor Jaramillo specifically, are generally successful in obtaining external research support. Professor Jaramillo is interested in reproducible research and open science: he has published in this area, group has developed related software prototypes and grant proposals, and 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 labs – 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 group members, including graduate students and postdoctoral fellows, allows for continuous monitoring lab equipment. This collaboration is facilitated by shared information repositories, including a group Electronic Lab Notebook (ELN) in LabArchives, a group Dropbox account, and Google Drive. Access 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 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 four 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 be exempt from further review by the Committee on the Use of Humans as Experimental Subjects (COUHES) at MIT on June 2, 2020.

Two researchers interviewed each graduate student: one served as the interviewer and the other as the transcriptionist. The interview audio was recorded and reviewed compared 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 was a guideline for the interviewer to construct the most complete narrative of each student's workflow. Each question was explicitly asked or indirectly answered through the student's response to a different question.

For the last section, Top Priority Project Workflow, it became evident that the most natural interview process was one in which the student first described the overall workflow for the selected project and then was prompted to recall each operational step. The interview would then ask follow-up questions to fill gaps and probe for additional detail.

Interview Coding

The interview coding process aimed to describe each step of the workflow in a systematic structure database. There is a wide range of existing formal models for provenance and workflow (see, for example, (Jandre et al., 2020)). However, most of these are 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 were labeled with their objective, task, and subtask, and the sequence, actor, input, output, data source, data target, level of automation, type of action, equipment, and methods used were recorded using standardized codes. (See appendix.) 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 an interpretation: the intended meaning of each statement was derived by 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 it requires its removal from storage. The first and second phases were completed for all interviewees before 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 interviewee's workflow can be derived from what is known from another's workflow. This was used mainly 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.

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.

Workflow Representation

We represent workflows as formal graphs and then apply social network analysis methods. (This follows a common approach to interpretation of workflows, first documented by (Tan et al., 2010).) The graph systematically describes all process, informational, and collaboration dependencies elicited through the interview process. By analyzing these graphs, we can visually and analytically identify workflow gaps, evaluate processes concerning stated policy, and probe potential interventions.

We augment this core workflow graph in several ways. First, we create a collaborator network graph by coding the interviews directly for any mentions of collaborations. Second, we derive separate dependency, collaboration, and information flow graphs directly from the workflow process graph. Finally, we apply graph methods from network and social network analysis (Carrington et al., 2005; Horwitz & Reps, 1992; Sharir, 1981) to probe questions related to collaboration (through analysis of connectivity and centrality), attribution (through comparing the explicitly elicited collaboration graph with its workflow-induced counterpart) and replicability (dependency, and subcomponent analysis of the information flow graph).

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

- A node on the graph represents each atomic action ("step") in the workflow. The node documents all of the characteristics of that single action.
- Process dependencies are represented through sequences and sub-sequences linked by "process" edges:

  - Actions performed by the same person, in a required sequence, for a single goal, and over a continuous period are represented by "sequence" nodes. Edges link each sequence to one or more child sub-sequences.
  - Actions performed within a sequence (and thus by the same person) and practically simultaneous (they have no natural order and occur during a brief period) are represented by sub-sequences. Edges link sub-sequences 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 holds.
  - 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 a single person conducts nodes during a 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.

Digital Forensics

```
library(tsibble)
library(lubridate)

dates.tsbl <-
  recent_files.tbl %>%
  transmute(date = as_date(client_modified)) %>%
  count(date) %>%
  filter(!is.na(date)) %>%
  as_tsibble(key = NULL, index = date) %>%
  fill_gaps()

library(feasts)
tmodel <-
  dates.tsbl %>%
  filter(year(date) > 2015) %>%
  index_by(period = ~ yearweek(.)) %>%
  summarise(count = sum(n, na.rm = TRUE)) %>%
  model(STL(count ~ season("1 year")))

fa1.plot <-
 components(tmodel)%>%
  autoplot()
```

Through the interview process, we determined that the lab used a shared folder in Dropbox as its official repository for collected data and documentation. With permission from the PI, we cloned a snapshot of the

repository contents, an collected all of the file-system metadata for use in a digital forensic analysis. For consistency with lab policies and with the timeline of the projects we evaluated, we restricted our analysis to files deposited between January of 2019 and January of 2022.

Overall, deposit patterns demonstrated that the repository was actively used – with deposit rates varying seasonally. (See Appendix). All information was collected for 31929 files – including file names, paths, content hashes, client-side-modification time, and deposit time. For image files, which are a common set of raw data formats used in this lab, we obtained additional internal timestamps and content metadata. Finally, many files used a naming convention to embed additional information such as creation date, creating user, and creating instrument – we used regular expression-based cleaning and parsing to extract this information where possible. This information set was then used to check for inconsistencies with documented information organization practices – as described below in the results section.

Post Analysis Validation

After the interviews were completed and their data was coded and analyzed, we validated the results with follow-up interviews and a documentation review. In the follow-up 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 omitted 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 data storage patterns per project and compare these to the patterns implied by the workflow analysis. In addition, we use content analysis to compare information organization naming practice with the documented lab policies.

Results

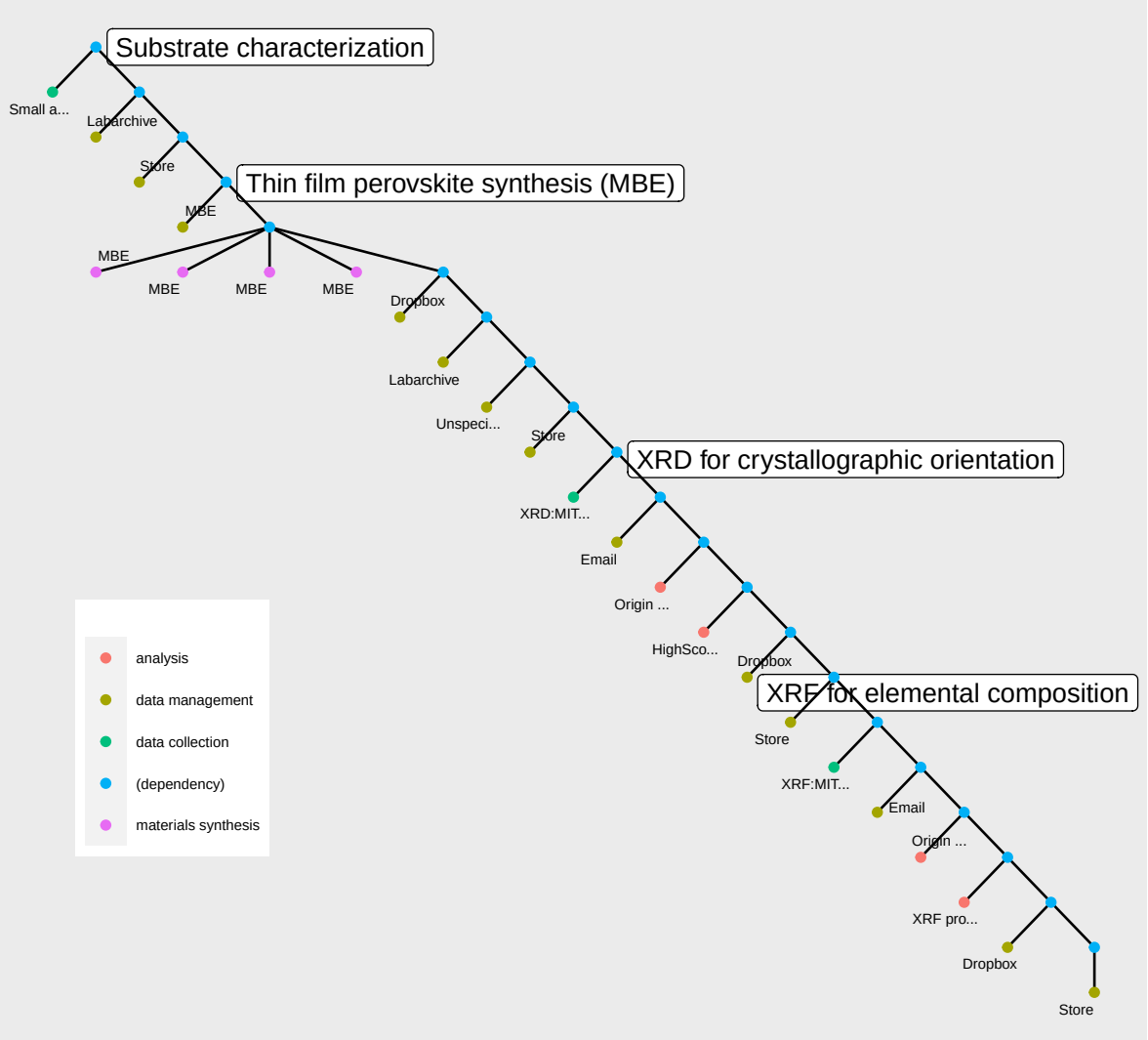
Workflow Overviews

For context, we describe the workflows for each of the four projects below:

Interviewee ‘A’ is involved in generating each output through the workflow sequence: sample preparation, synthesis, characterization, and analysis. The workflow sequence is iterative. Therefore, the analysis phase results inform how the next iteration’s synthesis process will be tuned. They use 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.

Figure 2A: workflow overview −− project A

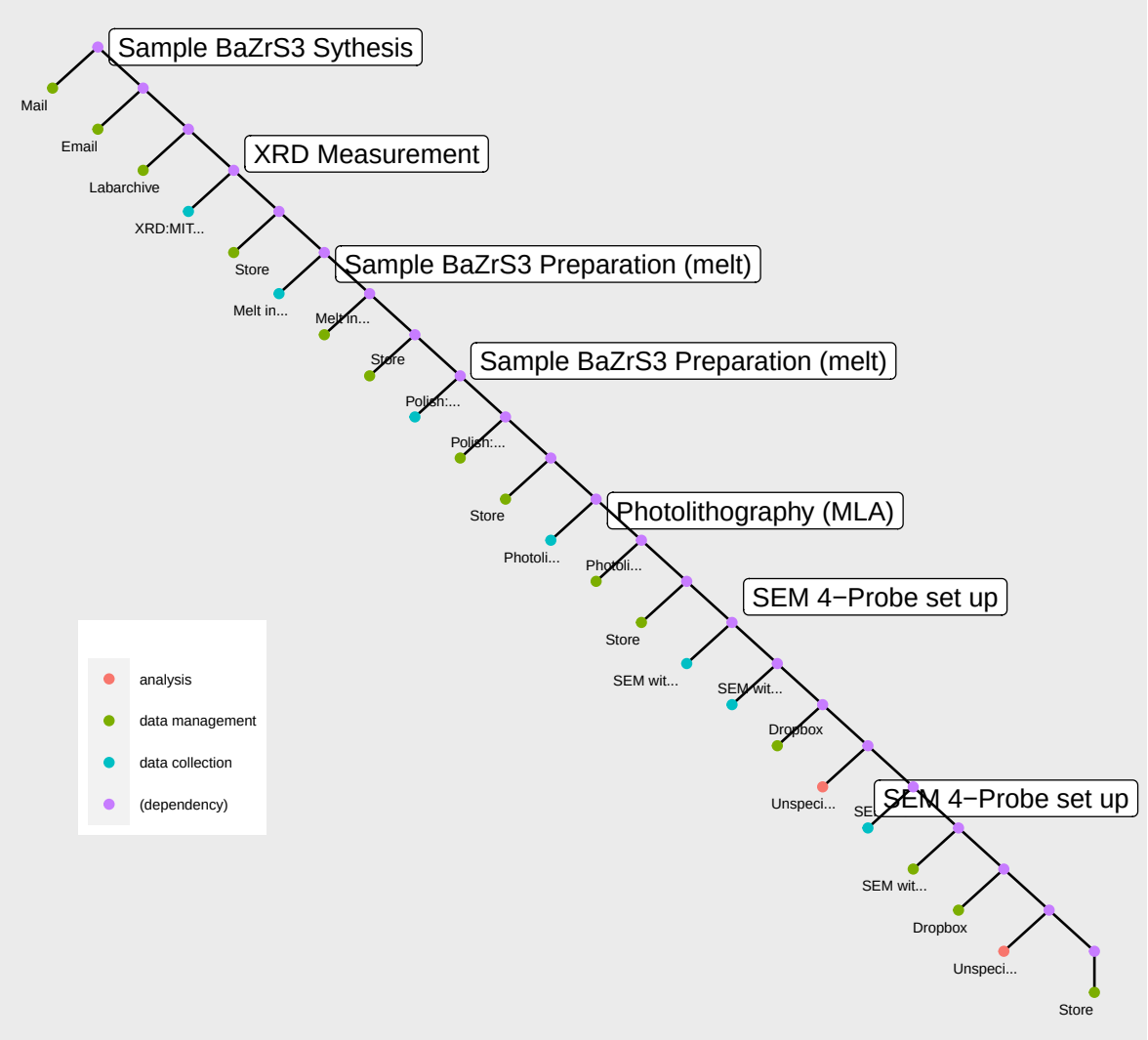
Workflow steps by phases, description, and type.



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 directly from instruments. Post-processed data is saved to a personal Dropbox.

Figure 2B: workflow overview −− project B

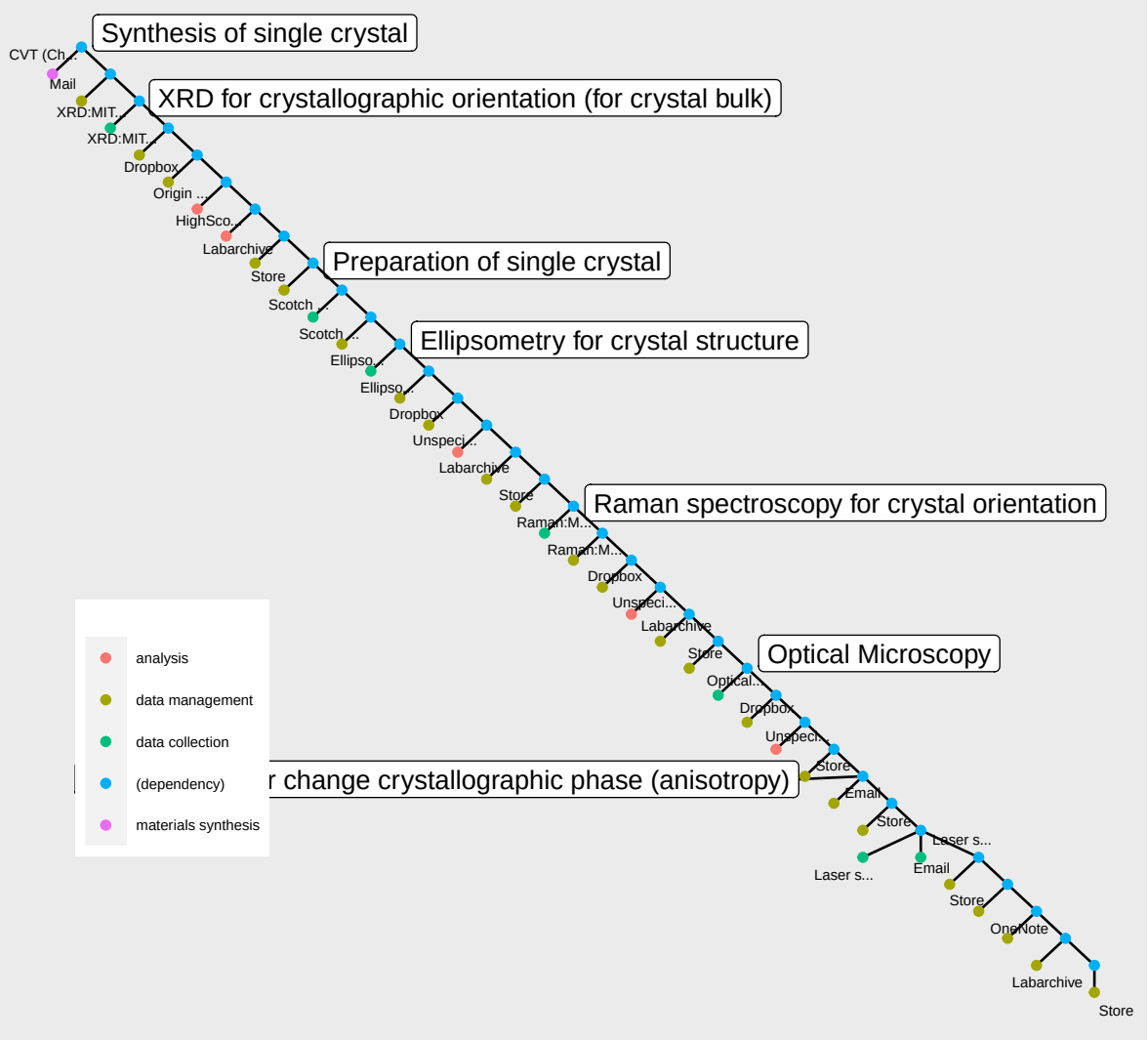
Workflow steps by phases, description, and type.



Interviewee 'C' received the synthesized materials from a collaborator. They prepare the acquired sample for analysis, characterize it, and analyze 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.

Figure 2C: workflow overview −− project C

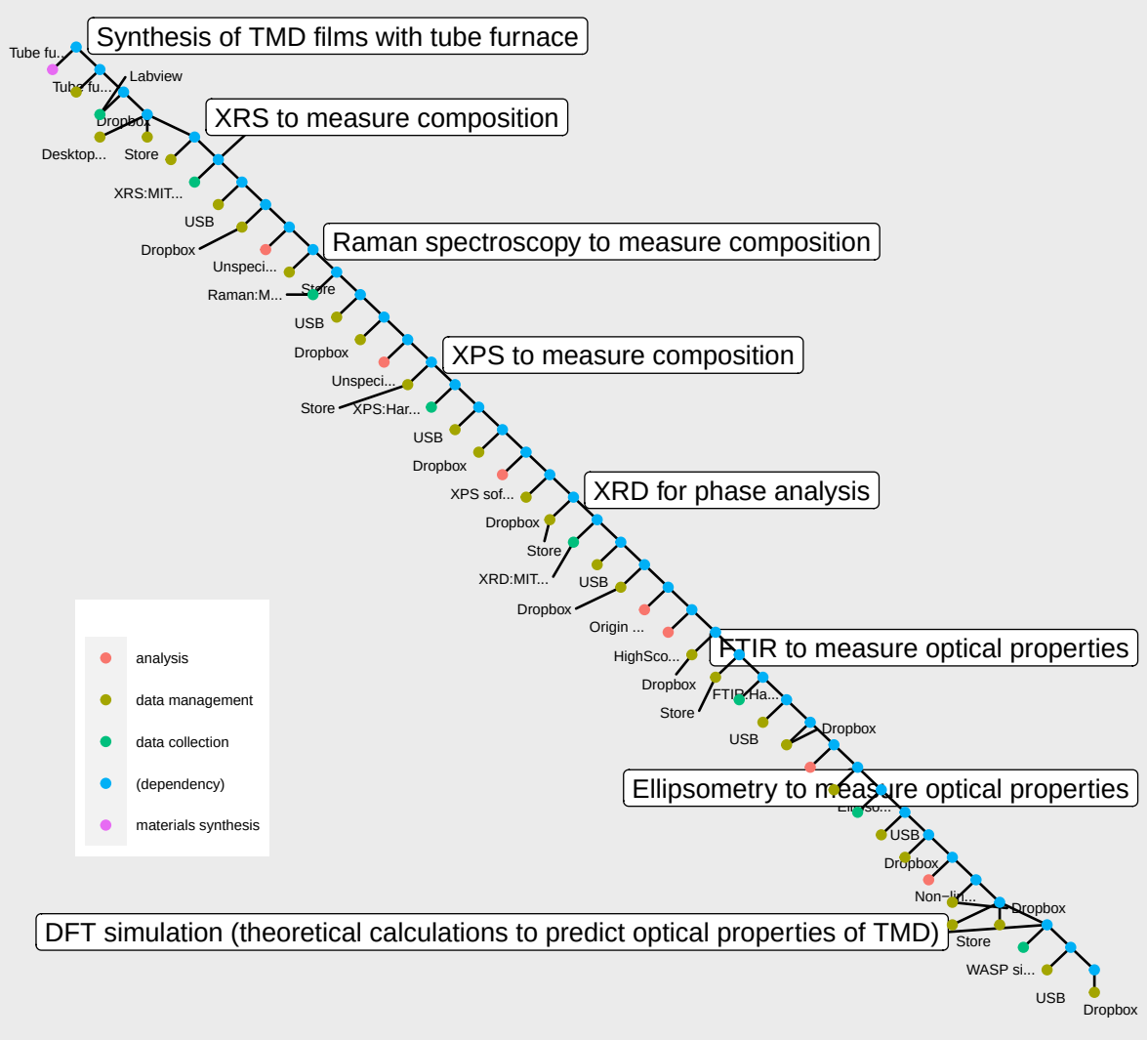
Workflow steps by phases, description, and type.



Interviewee 'D' is directly involved in each sequence step, 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.

Figure 2D: workflow overview –– project D

Workflow steps by phases, description, and type.



Note that each workflow is hierarchical – each project does 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.

All four interviewees used some instruments or equipment outside their 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 creates additional data transfer and documentation challenges. 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 acknowledgment of

equipment use.

Workflow Automation

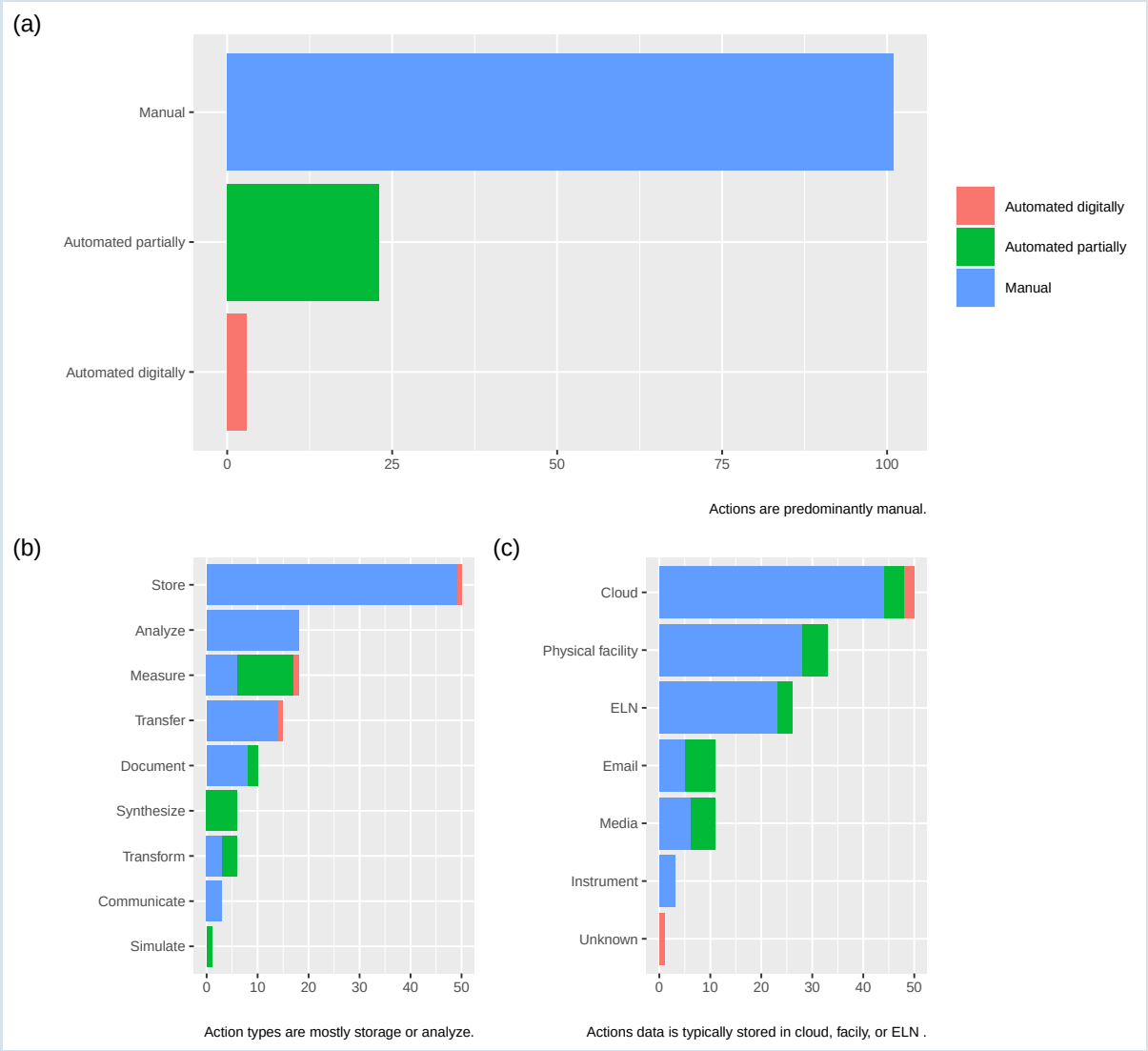
While there is very little systematic research on the rate and frequency of human errors in scientific research generally (or MSE specifically), a long history of research in the fields of human performance and reliability engineering suggests that the human error rates are substantial in the absence 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 the adoption of automated recording systems, and the use of explicit checklists and logs for manual proceduresRodziewicz & Houseman (2021).

During the interviews, we collected information about the types of automation associated with each instrument, storage facility, and analysis method. As described in the methodology section, each workflow step was coded for the type of action performed and level of automation used: as 'automated' if the step is initiated automatically following the prior step; 'partially' automated if the operation was launched manually, but was entirely described by digital metadata (e.g., configuration files), or as 'manual' if the step depended on manual initiation and manually configuration for correct operation.

Figure 3 summarizes selected characteristics of the workflow process. This figure reveals that, overall, workflow is dominated by manual activities.

Figure 3: Selected characteristics of the workflow steps

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



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?

Automation is not a panacea and can increase system complexity or decrease local transparency in ways that increase errors across a broader system. Notwithstanding, automation is often recommended for tasks that do not involve complex judgment (e.g., file transfers) and are not otherwise associated with specific performance, audit, and quality assurance procedures. 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 significant opportunity for human error in data management and organization.

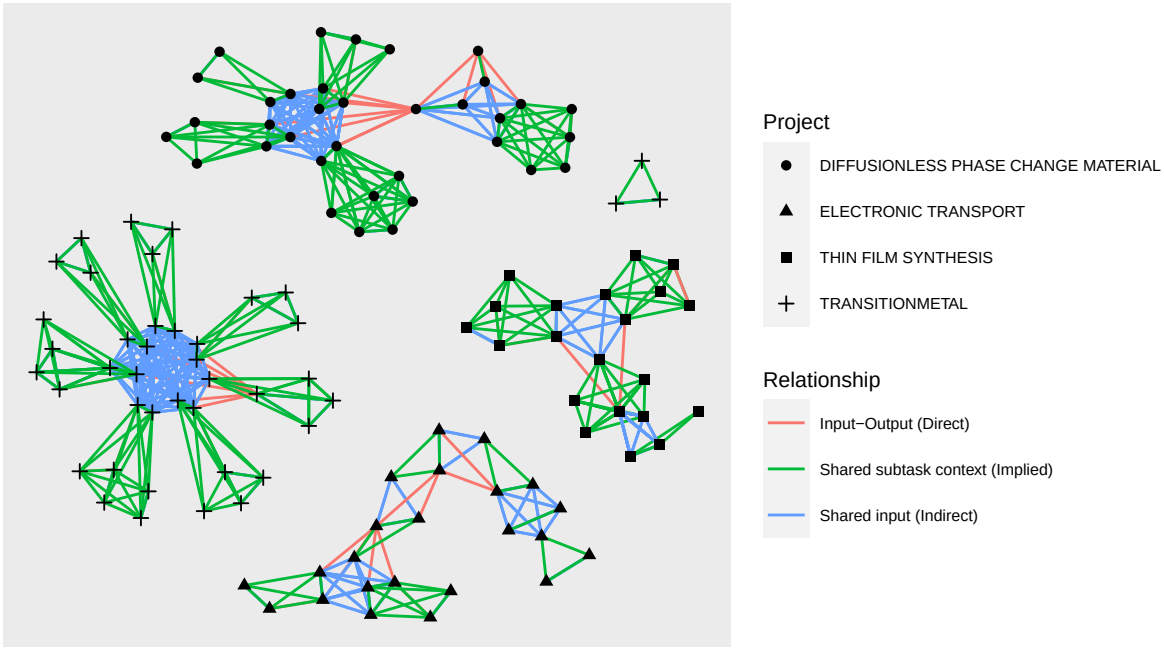
Collaboration and Information Flow

In the section above, we used the workflows to show the dependencies between steps in the research process. We use the same workflow graphs to derive the dependency graph for each analysis and in conjunction with interview data to derive the collaboration networks.

The information connections within and across projects (Figure 3) are densely interconnected within projects – a contrast with the linearity of the process used to produce the information. Further, most information flow is implicit – through shared context. Information rarely flows through direct input-output. There is no information flow between projects.

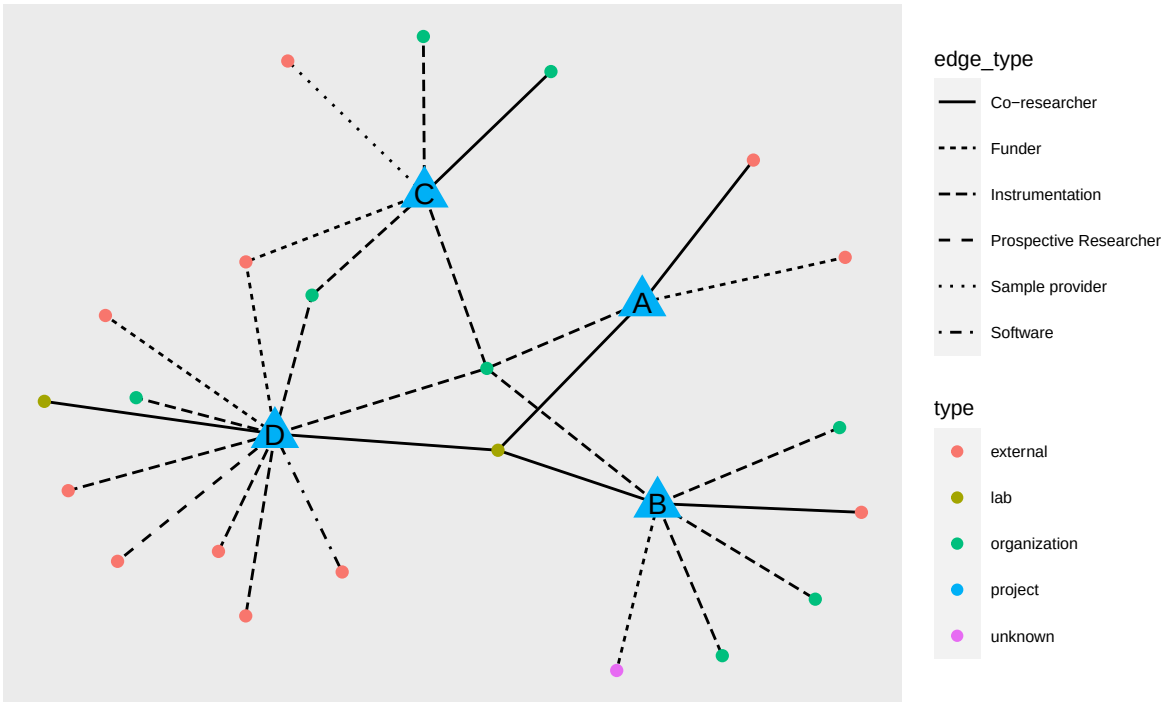
Figure 4: Project Information Exchange.

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 grid varies substantially across projects.

Figure 5: Project Collaboration.



Lab Practices

Characterizing Documented Practices

Our first research question concerns documented practices:

> 2(a) What data and research workflow management processes 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 processes from the subjects to characterize these.

During the interviews, one interviewee, who 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 and 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 individual practices elicited through interviews. We reviewed these documents both for relevance to the research question above and to inform the interpretation of the workflow networks below.

We identified 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, a 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") must be stored in the group Dropbox folder and should never be modified. All internal lab computers are configured to save data to the group Dropbox folder automatically. Data collected outside the group lab must be manually transferred to the group Dropbox folder. Examples of raw data are JPG from a microscope, TXT from a probe station, or files in a proprietary format such as RAW from XRD.
- P3. Group members can store their analysis results wherever is most convenient.

Information Security

- P4. Group members must 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\sam`
- P6. Samples must be named consistently with a given scheme, including `YYMMDD` and a serial number.

Consistency of Documented vs. Recalled and Observed Practices

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

1. In general, we identified all instances where subjects explicitly referred to documented or established practices during the interview, either during the description of their project or separately.
2. Concerning 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. Concerning 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. Concerning 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 Recalled 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 most significant 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. Furthermore, where collection dates were assigned, they often (53.5 %) pre-dated – and in rare cases, post-dated – the modification time stamps provided by the researcher's computers when they were delivered to Dropbox (or by the image creation software, where applicable). In the absence of more systematic processes for maintaining the provenance and authenticity of digital records, his discrepancy raises the possibility that data files could have been modified after collection.

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

Process Robustness Assessment

To address the remaining research questions we measuring and compared 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 another person's replication of a result 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 separately stored. As noted in the previous subsection, the documented practice in this lab does not include active replication of results before publication, nor does it 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.

The group exhibits documentation practices during the data collection and analysis process to 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 interpreted. When this capability is used, we describe the documentation 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. The researcher can manually add this separate documentation–e.g., a lab notebook entry or notes file. Alternatively, documentation may be implied by a previous step – e.g., when a measurement process is controlled by a configuration file already recorded.

We use the workflow information graph to identify when 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 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 format 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 processes 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 ut is not sufficient: We did not evaluate the completeness of the documentation, if it existed – only its presence. Notwithstanding, analysis documentation was missing in over half of the cases examined. This obstructs future replication of results and publications – which must rely on communication with the researcher who conducted this analysis (and upon their memory) and trial-and-error.

Robustness of Storage Practices

The following 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?

We use the workflow information graph to probe this question to identify all collected data (digital and physical samples created as part of each scientific workflow), metadata, and analysis results. We then use the process of the graph to trace the flow of these objects across tasks and into storage location. From this set of traces, we can infer the content of the 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 process's end. 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 insufficient 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 depend on manual information transfer rather than being automated. Given the high frequency of manual operations documented in the previous section, it is unsurprising that 100% of the analyses relied on manual information management at an earlier 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. We can measure the delay between the data creation and deposit by comparing the manually recorded date in the path with the automatically recorded date in the shared filesystem. For half of these files, the delay is relatively 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 transferred by hand, errors, interruptions, or forgetfulness can contribute to the delayed deposit. Second, during the validation interviews, we identified 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 impossible to determine the proportion of lag attributable to each mechanism because of the inconsistent use of documented and naming practices and the variation of undocumented practices.)

The two years of files examined it 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 can compute the elapsed time between creation and deposit. The delay is quite small for most of these files – 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 relationships. This list includes active collaborators (e.g., actors who supply material, perform an analysis, or contribute 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 standard process or policy concerning 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 principal investigator. Where this practice was followed with a particular 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 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 omitted from workflow documentation or action altogether. The table below summarizes these omissions.

Table 4: Undocumented collaborators
Types of collaborations that were recalled, but not documented in project work.

Project	Undocumented Types	% undocumented contributors
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 work process, the information used in it, or the analysis produced by it.

Analysis Validation

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

Further, a number of participants reflected that since the initial interviews, they had noted some of these gaps 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, depositing 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 in automating data collection from instruments outside the lab.

Discussion - Toward More Reproducible and Attributed Practices

In sum, the workflow, documentation, and digital forensics analyses revealed both strengths and limitations of the current practice. Practices in the lab are sufficient to mitigate the risk of data loss resulting from the failure of an individual's computer and to ensure access to the raw data collected for the lab research.

Preservation of the data is necessary but insufficient for trustworthiness and transparency. We find that lab data curation practices often deviate from the stated policy and vary across projects, especially concerning the metadata and documentation needed to contextualize and analyze the collected data. Moreover, neither policy nor practice are sufficient for the attribution, replication, or verification of the labs' published results.

As a result, the integrity and continuity of lab research are threatened if an individual fails to keep private records of attribution and data provenance or withdraws from the research 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 or storage of critical information, preferably in standardized formats with necessary metadata.
- Changes to the lab policies regarding requirements for those activities are done manually.

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

- Recommendation 1: Concerning the documented practices, minimal automated audits – in support of sanity checks – could verify that documented naming conventions are being followed and that systems are running backup software. Concerning 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 flag unusual activity patterns 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 the '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, is attractive because they do not require people to change behavior– which is often costly, difficult to assess, error-prone, and requires consistent focus to maintain. While a fully automated infrastructure for materials science remains currently too expensive and immature for many labs, more minor 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 reported workflows involved the extensive use of personal portable storage to transfer data from experimental instruments manually. Further, the file forensic analysis shows that the delays between file creation and deposit can be quite significant. Moreover, no systems or 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.

The portable storage is typically a simple offline USB "flash drive." Alternatives USB-compatible mobile storage devices, including built-in wireless networking and data synchronization capabilities, are now readily available. Although researchers would still need to transport these storage devices with a network connection to the instruments and plug them in, the manual data transfer to cloud storage could then be automated, reducing the risk to reproducibility. Using this type of portable storage device 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 or facilities owners 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 and a more systematized practice of folder organization for work products kept in personal storage would enable more reliable and robust data lab practices without sacrificing the convenience of personal cloud-storage accounts.

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

- Recommendation 5: Develop explicit practices around collaborator attribution. Practices are needed to identify the contributions of collaborators systematically. This might include (a) enhancing existing workflow project documentation (e.g., in the lab notebook) to identify when the researcher uses externally contributed resources clearly, borrowed equipment, or 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 allow 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 to be conducted in a framework that builds 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, is needed to reliably measure how better practices can improve reproducibility and research attribution. (Altman & Cohen, 2021;

NASEM, 2018) Moreover, these practices are embedded in and responsive to a much broader system of scientific incentives, institutional and organizational collaboration, and professional training (Altman & 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 their own study could further motivate individual researchers and their teams. Hiring data curators or research workflow facilitators to provide discipline-specific support for particular groups and departments could further enable researchers to 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 lowering individual groups' barriers to fully adapting 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 creativity.

Appendix

Interview Protocol & Coding Dictionary

The interview protocol and coding dictionary used in the study are shown in Appendix I.

Tables

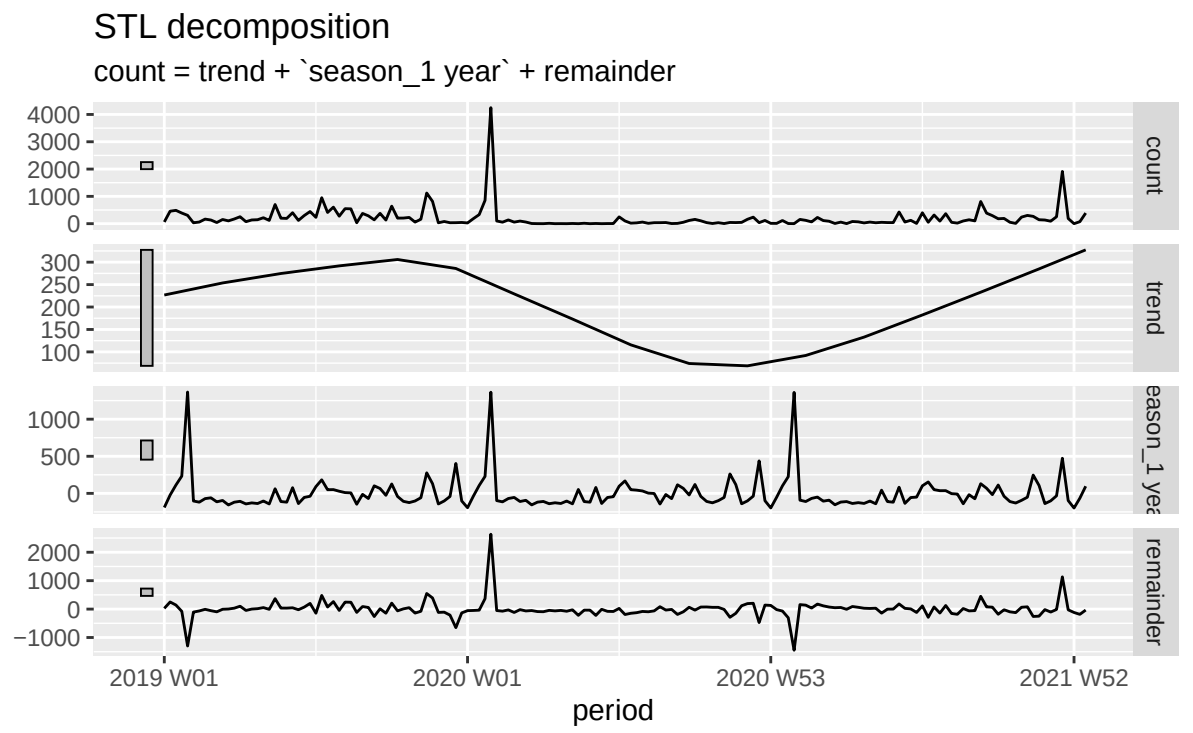
Table A1: Information Exchange Graph Statistics

Direct Connections	Graph Diameter	Mean Distance Between Steps
1660	5	0.2172682

Table A2: Collaboration Network

Direct Connections	Graph Diameter	Mean Distance Between Steps
29	1	1

Figure A1. Trends in file creation in group storage.



Data Availability

The deidentified datasets generated during or analyzed during the current study, as well as the R scripts used for analysis and generating the research report, are available in the Zenodo repository under CC BY 4.0 license, https://doi.org/10.5281/zenodo.7158715

Acknowledgments

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 and 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 Professor Rafael Jaramillo at MIT for his commentary and 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, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature*, *508*(7496), 312–313. https://doi.org/10.1038/508312a

Altman, M., & Bourg, C. (2018). A Grand Challenges-Based Research Agenda for Scholarly Communication and Information Science. *MIT Grand Challenge Participation Platform*. https://doi.org/10.21428/62b3421f

Altman, M., & Cohen, P. N. (2021). *The scholarly knowledge ecosystem: Challenges and opportunities for the field of information*. http://dx.doi.org/10.31235/osf.io/ctdb9

Blaiszik, B., Chard, K., Pruyne, J., Ananthakrishnan, R., Tuecke, S., & Foster, I. (2016). The Materials Data Facility: Data Services to Advance Materials Science Research. *JOM*, *68*(8), 2045–2052. https://doi.org/10.1007/s11837-016-2001-3

Brandt, N., Griem, L., Herrmann, C., Schoof, E., Tosato, G., Zhao, Y., Zschumme, P., & Selzer, M. (2021). Kadi4Mat: A Research Data Infrastructure for Materials Science. *Data Science Journal*, *20*. https://doi.org/10.5334/dsj-2021-008

Carrington, P. J., Scott, J., & Wasserman, S. (2005). *Models and methods in social network analysis* (Vol. 28). Cambridge university press.

Cook, I., Grange, S., & Eyre-Walker, A. (2015). Research groups: How big should they be? *PeerJ*, *3*, e989. https://doi.org/10.7717/peerj.989

Coudert, F.-X. (2019). Correcting the Scientific Record: Retraction Practices in Chemistry and Materials Science. *Chemistry of Materials*, *31*(10), 3593–3598. https://doi.org/10.1021/acs.chemmater.9b00897

Delgado-Licona, F., & Abolhasani, M. (2023). Research Acceleration in Self-Driving Labs: Technological Roadmap toward Accelerated Materials and Molecular Discovery. *Advanced Intelligent Systems*, *5*(4), 2200331. https://doi.org/10.1002/aisy.202200331

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, N. S. (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, R., Walton, K. S., & Sholl, D. S. (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, J., Mannodi-Kanakkithodi, A., Ramprasad, R., & Meredig, B. (2018). *Materials Data Infrastructure and Materials Informatics* (D. Shin & J. Saal, Eds.; pp. 193–225). Springer International Publishing. https://doi.org/10.1007/978-3-319-68280-8_9

Himanen, L., Geurts, A., Foster, A. S., & Rinke, P. (2019). Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science*, *6*(21), 1900808. https://doi.org/10.1002/advs.201900808

Horwitz, S., & Reps, T. (1992). The use of program dependence graphs in software engineering. *Proceedings of the 14th International Conference on Software Engineering - ICSE '92*. https://doi.org/10.1145/143062.143156

Informatics, C. (2022). What is the Citrine Platform? In *Citrine Informatics*. https://citrine.io/product/what-is-the-citrine-platform/.

Jacobs, P. (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, E., Diirr, B., & Braganholo, V. (2020). Provenance in Collaborative in Silico Scientific Research. *ACM SIGMOD Record*, *49*(2), 36–51. https://doi.org/10.1145/3442322.3442329

Maleki, R., Asadnia, M., & Razmjou, A. (2022). Artificial Intelligence-Based Material Discovery for Clean Energy Future. *Advanced Intelligent Systems*, *4*(10), 2200073. https://doi.org/10.1002/aisy.202200073

McNutt, M. K., Bradford, M., Drazen, J. M., Hanson, B., Howard, B., Jamieson, K. H., Kiermer, V., Marcus, E., Pope, B. K., Schekman, R., Swaminathan, S., Stang, P. J., & Verma, I. M. (2018). Transparency in

authors' contributions and responsibilities to promote integrity in scientific publication. *Proceedings of the National Academy of Sciences*, *115*(11), 2557–2560. https://doi.org/10.1073/pnas.1715374115

Medina-Smith, A., Becker, C. A., Plante, R. L., Bartolo, L. M., Dima, A., Warren, J. A., & Hanisch, R. J. (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, P., Konstanty, S., Nicholson, T., O'Brien, T., Schwartz-Duval, A., Spila, T., Nahrstedt, K., Campbell, R. H., Gupta, I., Chan, M., Mchenry, K., & Paquin, N. (2017). *2017 17th IEEE/ACM international symposium on cluster, cloud and grid computing (CCGRID)*. 11–20. https://doi.org/10.1109/CCGRID.2017.51

Park, J., Howe, J. D., & Sholl, D. S. (2017). How Reproducible Are Isotherm Measurements in Metal–Organic Frameworks? *Chemistry of Materials*, *29*(24), 10487–10495. https://doi.org/10.1021/acs.chemmater.7b04287

Plante, R. L., Becker, C. A., Medina-Smith, A., Brady, K., Dima, A., Long, B., Bartolo, L. M., Warren, J. A., & Hanisch, R. J. (2021). Implementing a Registry Federation for Materials Science Data Discovery. *Data Science Journal*, *20*. https://doi.org/10.5334/dsj-2021-015

Pyzer-Knapp, E. O., Pitera, J. W., Staar, P. W. J., Takeda, S., Laino, T., Sanders, D. P., Sexton, J., Smith, J. R., & Curioni, A. (2022). Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *Npj Computational Materials*, *8*(1), 1–9. https://doi.org/10.1038/s41524-022-00765-z

Qurashi, M. M. (1984). Publication rate as a function of the laboratory/group size. *Scientometrics*, *6*(1), 19–26. https://doi.org/10.1007/bf02020110

Raccuglia, P., Elbert, K. C., Adler, P. D. F., Falk, C., Wenny, M. B., Mollo, A., Zeller, M., Friedler, S. A., Schrier, J., & Norquist, A. J. (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 (A Consensus Study Report). (2019). The National Academies of Science, Engineering, and Medicine. https://doi.org/10.17226/25303

Rodziewicz, T. L., & Houseman, J. E., Benjamin Hipskind. (2021). *Medical error reduction and prevention*. StatPearls. https://www.ncbi.nlm.nih.gov/books/NBK499956/

Savage, C. J., & Vickers, A. J. (2009). Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. *PLoS ONE*, *4*(9), e7078. https://doi.org/10.1371/journal.pone.0007078

Seglen, P. O., & Aksnes, D. W. (2000). *Scientometrics*, *49*(1), 125–143. https://doi.org/10.1023/a:1005665309719

Sharir, M. (1981). A strong-connectivity algorithm and its applications in data flow analysis. *Computers & Mathematics with Applications*, *7*(1), 67–72. https://doi.org/10.1016/0898-1221(81)90008-0

Stein, H. S., & Gregoire, J. M. (2019). Progress and prospects for accelerating materials science with automated and autonomous workflows. *Chemical Science*, *10*(42), 9640–9649. https://doi.org/10.1039/c9sc03766g

Tan, W., Zhang, J., & Foster, I. (2010). Network analysis of scientific workflows: A gateway to reuse. *Computer*, *43*(9), 5461.

Technology: Sharing data in materials science. (2013). *Nature*, *503*(7477), 463–464. https://doi.org/10.1038/503463a

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Timmer, J. (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). https://www.ideals.illinois.edu/handle/2142/94738

Westbrook, J. H., & Rumble, J. (1983, January 1). *Computerized materials data systems*. https://ieeexplore.ieee.org/stamp/st
https://www.osti.gov/biblio/6969565-computerized-materials-data-systems

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

Wilson, S. L., Altman, M., & Jaramillo, R. (2019). *Methods for open and reproducible materials science*. https://doi.org/10.31235/osf.io/ag8zu