Automation and Autonomous Experimentation for Sol-Gel Nanomaterial Synthesis

Brenden Pelkie

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Lilo D. Pozzo, Chair

Zachary Sherman

David Beck

Program Authorized to Offer Degree:

Department of Chemical Engineering

University of Washington

Abstract

Automation and Autonomous Experimentation for Sol-Gel Nanomaterial Synthesis

Brenden Pelkie

Chair of the Supervisory Committee:

Lilo D. Pozzo

Department of Chemical Engineering

The nexus of laboratory automation and machine learning enables a new paradigm of autonomous experimental materials research. Autonomous experimentation integrates highly automated materials synthesis and characterization experiments with machine-learning guided experimental design strategies to adaptively execute experiments that advance specific research goals. These systems have the potential to accelerate materials development timelines compared to traditional manual processes by efficiently targeting experimental efforts. Sol-gel processes are used to synthesize a diverse array of metal oxide nanomaterials for many applications. In particular, mesoporous colloidal silicas are promising materials for use as catalysis support matrices,

chromatographic separation media, and drug delivery systems. Achieving retrosynthetic control over particle morphologies is critical for advancing use in these applications. In this work, automated and autonomous systems for the synthesis and optimization of colloidal silicas are developed. An open-source platform for flexible laboratory automation is developed to enable democratized access to autonomous experimentation. A system for the fully automated synthesis and characterization of colloidal mesoporous silicas is developed which integrates the open-hardware automation platform to perform automated sol-gel synthesis with synchrotron and laboratory X-ray scattering instruments for characterization. Synthesis campaigns executed with this system have produced colloidal silicas with a range of particle morphologies and mesopore phase structures. Finally, progress towards integrating a Bayesian optimization based experimental design strategy to optimize the morphology of silica nanoparticles is discussed. This work represents an important step towards accelerating the development of sol-gel materials with autonomous experimentation.

Contents

List of Figures

List of Tables

Acknowledgements

The work presented here would not have been possible without the contributions of many people. I'd like to thank my advisor, Lilo Pozzo, for helping me navigate the transition to experimental research and for her guidance throughout this process. Thank you also to Stephanie Valleau for introducing me to the world of computational chemistry and machine learning. The Science-Jubilee work discussed here was a particularly collaborative effort. Thank you to Maria Politi for leading the adoption and development of Jubilee in our research group. Machine Agency, especially Blair Subbaraman, Danli Luo, Wm Salt Hale, and Nadya Peek, were key to the developments made on the Science-Jubilee project. Thank you to Sonya Vasquez for making herself available for informative discussions and providing encouragement throughout my time working with the Jubilee platform. The support of Peter Beaucage and Tyler Martin was critical to the successful development of the NIST-AFL sample changer and its integration with a beamline described in Chapter 6. Zachery Wylie collected TEM images for this work. Illustrations generated by Maria Politi were used in this work. Thank you also to all the other Pozzo research group members and collaborators who helped make this work possible. Finally, I'd like to thank my wife Mary for her love and support along this journey.

I would also like to acknowledge financial support from the Department of Chemical Engineering John C. Berg Fellowship, the College of Engineering Dean's Fellowship, the UW Clean Energy Institute, and the DIRECT program. Work presented within was performed on APS beam time award https://doi.org/10.46936/APS-188896/60013388 from the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science user facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. The use of SEM and SAXS instruments in this work was supported by the University

of Washington Molecular Engineering and Materials Center, an NSF MRSEC supported under award number DMR-2308979. This work benefited from the use of the SasView application, originally developed under NSF award DMR-0520547. SasView contains code developed with funding from the European Union's Horizon 2020 research and innovation programme under the SINE2020 project, grant agreement No 654000. This work was supported by use of facilities and instrumentation supported by the Molecular Analysis Facility, which is supported in part by funds from the Molecular Engineering & Sciences Institute, the Clean Energy Institute, and the National Science Foundation (NNCI-2025489 and NNCI-1542101).

1 Introduction

Advanced new materials and chemicals are needed as part of a multi-faceted approach to adapt to and mitigate climate change and other emerging threats[1]. Autonomous experimentation is a powerful tool for accelerating the development of these materials. This research paradigm integrates automated laboratory experiment execution with data-driven experimental planning approaches to create systems that iteratively and autonomously optimize material synthesis procedures to achieve target properties[2]. These systems are especially effective at addressing optimization flavored problems that involve navigating complex experimental parameter spaces to optimize for properties of interest or understand how synthesis parameters impact property outcomes.

Autonomous experimentation builds from effective automated experimentation workflows, which use robotics and other mechanical automation hardware to perform laboratory processes that would traditionally be done manually by a researcher. Successfully implementing an automated experiment can be a challenging and complex undertaking[3]. It requires acquiring the capabilities to perform all the main steps in an experimental procedure with automation equipment, either by purchasing off-the-shelf hardware, building existing open-source solutions, or developing new systems entirely from scratch. These capabilities then need to be integrated to form a coherent sample synthesis and characterization workflow. Implemented workflows need to be validated, and potentially improved, to ensure that they perform the synthesis they are designed for repeatably and reliably. Implementing an automated experiment adds significant complexity and effort compared to simply running a few rounds of the experiment by hand. However, it can provide major benefits including improved reproducibility, reduced experiment time, and higher experimental throughput[4]. Stand-alone automated experimentation is a powerful tool for

advancing materials research that can be used to perform large-scale screenings of synthesis and composition parameter spaces to identify promising candidates for particular applications and provide experimental data to train machine learning predictors[5]. An effective automated experiment implementation also enables extension to autonomous experimentation by integrating data-driven experimental design approaches to actively adapt and select interesting experiments that most effectively further a particular research goal[6]. The canonical version of autonomous experimentation uses Bayesian optimization to select experimental parameter values that optimize some target material property. This approach trains a machine learning model embedded in the Bayesian optimization algorithm to learn how modifiable experiment parameters impact a property of interest, then uses this predictor to select maximally useful experiments to run. This can focus experiments on areas likely to perform well while minimizing experiments resulting in poorly performing materials, thus accelerating optimization by reducing the total number of experiments needed. Other implementations of autonomous experimentation can contribute to improving fundamental understanding of the materials system, such as autonomous phase mapping[7]. Despite the attention it has received recently, autonomous experimentation is not a panacea for all materials research problems and limitations to the applicability of these systems exist. Fundamental exploratory research may require flexibility that is not achievable with rigid automation workflows. Some experimental processes require a degree of 'human intuition' that is not presently replicable with automated systems. Current optimization-focused autonomous experiment approaches are not well suited to generating new insights and knowledge about the systems that they study. Despite these limitations, autonomous experimentation can play an important role in developing new materials and chemicals.

Extending an automated experiment into an autonomous one adds an additional layer of complexity. In addition to the execution of the experiment, an appropriate adaptive experimental design strategy needs to be selected and implemented. Adequate data and metadata management is needed to keep up with the high rate of experiments[8]. An orchestration infrastructure to integrate these components and keep the experiment running reliably is also needed[9,10]. Of course, an implementation of an autonomous experiment still needs to keep the original material or chemical problem in sight and incorporate domain knowledge needed to enable an improvement in performance. When it comes to autonomous experiments, it is truly turtles all the way down.

This dissertation presents contributions to several aspects of autonomous experimentation. Early work to predict reaction rate constants for heterogeneous catalyzed reactions with machine learning is presented in chapter 3. While not directly applicable to an autonomous experimentation workflow, machine learning property prediction can play an important role in selecting promising candidates for experimental study. Work on an alternative design space delineation approach using data-driven engineering criteria is described in appendix 2. The need for effective data management infrastructure to handle the increased data flux from automated experiments is discussed in chapter 4. A major component of this work is the development of an open-source flexible laboratory automation system known as Science-Jubilee. The development of tools, software, and documentation for this platform is discussed in chapter 5. Following the development of this lab automation tool, an automated experimentation platform was implemented that incorporates Science-Jubilee as well as other open-source automation equipment to perform sol-gel nanomaterial synthesis. The development of this platform and results from initial campaigns are described in chapter 6. Finally, significant progress was made toward executing autonomous optimization of sol-gel nanomaterials using the automated experiment platform

discussed in chapter 6. Results and insights from this development work are discussed in chapter 7.

A major theme throughout the work presented here is democratization of autonomous experimentation. 'Democratization' describes efforts to make autonomous experimentation broadly accessible to as many researchers as possible. Motivation for and action toward democratization can be understood by exploring the question, 'Who gets to do their science with self-driving labs?'. A brief survey of the current state of the field suggests that automated experimentation development is converging toward complex, expensive, task-specific implementations[11–13]. These systems can involve hundreds of thousands of dollars of automation equipment and require dedicated engineering staff to integrate and support. They take years to develop and implement and are often dedicated to a single materials system at a time. This level of infrastructure investment is out of reach for most researchers. Democratization efforts seek to develop pathways to laboratory automation that reduce the time and financial investment needed to get up and running with autonomous experimentation[14]. This can be done by developing accessible automation equipment like the Science-Jubilee platform, adapting existing commonly used equipment for new uses, and openly sharing development efforts to lower the barrier to entry for new lab automation practitioners. The work described here contributes to these efforts through the development of the open-hardware, accessible Science-Jubilee platform, and by demonstrating a successful automated experimentation campaign that uses this platform and other open-source hardware components. Making autonomous experimentation accessible to as many researchers as possible so that they are deployed widely to address a broad range of scientific questions will lead to the largest possible gains from these systems.

1.1 References

(1) Pörtner, H.-O.; Roberts, D. C.; Adams, H.; Adelekan, I.; Adler, C.; Adrian, R.; Aldunce, P.; Ali, E.; Begum, R. A.; Bednar-Friedl, B.; Bezner Kerr, R.; Biesbroek, R.; Birkmann, J.; Bowen, K.; Caretta, M. A.; Carnicer, J.; Castellanos, E.; Cheong, T. S.; Chow, W.; Cissé, G.; Clayton, S.; Constable, A.; Cooley, S.; Costello, M. J.; Craig, M.; Cramer, W.; Dawson, R.; Dodman, D.; Efitre, J.; Garschagen, M.; Gilmore, E.; Glavovic, B.; Gutzler, D.; Haasnoot, M.; Harper, S.; Hasegawa, T.; Hayward, B.; Hicke, J. A.; Hirabayashi, Y.; Huang, C.; Kalaba, K.; Kiessling, W.; Kitoh, A.; Lasco, R.; Lawrence, J.; Lemos, M. F.; Lempert, R.; Lennard, C.; Ley, D.; Lissner, T.; Liu, Q.; Liwenga, E.; Lluch-Cota, S.; Löschke, S.; Lucatello, S.; Luo, Y.; Mackey, B.; Mintenbeck, K.; Mirzabaev, A.; Möller, V.; Vale, M. M.; Morecroft, M. D.; Mortsch, L.; Mukherji, A.; Mustonen, T.; Mycoo, M.; Nalau, J.; New, M.; Okem, A.; Ometto, J. P.; O'Neill, B.; Pandey, R.; Parmesan, C.; Pelling, M.; Pinho, P. F.; Pinnegar, J.; Poloczanska, E. S.; Prakash, A.; Preston, B.; Racault, M.-F.; Reckien, D.; Revi, A.; Rose, S. K.; Schipper, E. L. F.; Schmidt, D. N.; Schoeman, D.; Shaw, R.; Simpson, N. P.; Singh, C.; Solecki, W.; Stringer, L.; Totin, E.; Trisos, C.; Trisurat, Y.; Aalst, M. van; Viner, D.; Wairu, M.; Warren, R.; Wester, P.; Wrathall, D.; Ibrahim, Z. Z. Technical Summary. In *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Pörtner, H.-O., Roberts, D. C., Tignor, M. M. B., Poloczanska, E. S., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., Rama, B., Eds.; Cambridge University Press, 2022.

(2) Tom, G.; Schmid, S. P.; Baird, S. G.; Cao, Y.; Darvish, K.; Hao, H.; Lo, S.; Pablo-García, S.; Rajaonson, E. M.; Skreta, M.; Yoshikawa, N.; Corapi, S.; Akkoc, G. D.; Strieth-Kalthoff, F.; Seifrid, M.; Aspuru-Guzik, A. Self-Driving Laboratories for Chemistry and Materials Science. *Chem. Rev.* **2024**, *124* (16), 9633–9732. https://doi.org/10.1021/acs.chemrev.4c00055.

(3) Christensen, M.; E. Yunker, L. P.; Shiri, P.; Zepel, T.; L. Prieto, P.; Grunert, S.; Bork, F.; E. Hein, J. Automation Isn't Automatic. *Chem. Sci.* **2021**, *12* (47), 15473–15490. https://doi.org/10.1039/D1SC04588A.

(4) Stach, E.; DeCost, B.; Kusne, A. G.; Hattrick-Simpers, J.; Brown, K. A.; Reyes, K. G.; Schrier, J.; Billinge, S.; Buonassisi, T.; Foster, I.; Gomes, C. P.; Gregoire, J. M.; Mehta, A.; Montoya, J.; Olivetti, E.; Park, C.; Rotenberg, E.; Saikin, S. K.; Smullin, S.; Stanev, V.; Maruyama, B. Autonomous Experimentation Systems for Materials Development: A Community Perspective. *Matter* **2021**, *4* (9), 2702–2726. https://doi.org/10.1016/j.matt.2021.06.036.

(5) Green, M. L.; Choi, C. L.; Hattrick-Simpers, J. R.; Joshi, A. M.; Takeuchi, I.; Barron, S. C.; Campo, E.; Chiang, T.; Empedocles, S.; Gregoire, J. M.; Kusne, A. G.; Martin, J.; Mehta, A.; Persson, K.; Trautt, Z.; Van Duren, J.; Zakutayev, A. Fulfilling the Promise of the Materials Genome Initiative with High-Throughput Experimental Methodologies. *Appl. Phys. Rev.* **2017**, *4* (1), 011105. https://doi.org/10.1063/1.4977487.

(6) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving Laboratories. *Trends Chem.* **2019**, *1* (3), 282–291. https://doi.org/10.1016/j.trechm.2019.02.007.

(7) Martin, T. B.; Sutherland, D. R.; McDannald, A.; Kusne, A. G.; Beaucage, P. A. Autonomous Small-Angle Scattering for Accelerated Soft Material Formulation Optimization. arXiv March 14, 2025. https://doi.org/10.48550/arXiv.2503.11859.
(8) G. Pelkie, B.; D. Pozzo, L. The Laboratory of Babel: Highlighting Community Needs for Integrated Materials Data Management. *Digit. Discov.* **2023**, *2* (3), 544–556. https://doi.org/10.1039/D3DD00022B.
(9) Fei, Y.; Rendy, B.; Kumar, R.; Dartsi, O.; Sahasrabuddhe, H. P.; McDermott, M. J.; Wang, Z.; Szymanski, N. J.; Walters, L. N.; Milsted, D.; Zeng, Y.; Jain, A.; Ceder, G. AlabOS: A Python-Based Reconfigurable Workflow Management Framework for Autonomous Laboratories. *Digit. Discov.* **2024**, *3* (11), 2275–2288. https://doi.org/10.1039/D4DD00129J.
(10) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: An Orchestration Software to Democratize Autonomous Discovery. *PLOS ONE* **2020**, *15* (4), e0229862. https://doi.org/10.1371/journal.pone.0229862.
(11) Rupnow, C. C.; MacLeod, B. P.; Mokhtari, M.; Ocean, K.; Dettelbach, K. E.; Lin, D.; Parlane, F. G. L.; Chiu, H. N.; Rooney, M. B.; Waizenegger, C. E. B.; Hoog, E. I. de; Soni, A.; Berlinguette, C. P. A Self-Driving Laboratory Optimizes a Scalable Process for Making Functional Coatings. *Cell Rep. Phys. Sci.* **2023**, *4* (5). https://doi.org/10.1016/j.xcrp.2023.101411.
(12) Szymanski, N. J.; Rendy, B.; Fei, Y.; Kumar, R. E.; He, T.; Milsted, D.; McDermott, M. J.; Gallant, M.; Cubuk, E. D.; Merchant, A.; Kim, H.; Jain, A.; Bartel, C. J.; Persson, K.; Zeng, Y.; Ceder, G. An Autonomous Laboratory for the Accelerated Synthesis of Novel Materials. *Nature* **2023**, *624* (7990), 86–91. https://doi.org/10.1038/s41586-023-06734-w.
(13) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A Mobile Robotic Chemist. *Nature* **2020**, *583* (7815), 237–241. https://doi.org/10.1038/s41586-020-2442-2.
(14) Pelkie, B.; Baird, S.; Aissi, E.; Aspuru-Takata, K.; Cao, Y.; Chang, J. H.; Gambhir, K.; Hale, W. S.; Hao, L.; Hattrick, C.; Hein, J.; Luo, D.; Melville, O.; Ngan, M.; Nyeland, L. L. B.; Peek, N.; Politi, M.; Rajkumar, E. E.; Siemenn, A.; Subbaraman, B.; Vasquez, S.; Watchorn, J.; Zhang, W.; Ziskason, R.; Pozzo, L.; Buonassisi, T.; Vegge, T. Democratizing Self-Driving Labs through User-Developed Automation Infrastructure. ChemRxiv February 12, 2025. https://doi.org/10.26434/chemrxiv-2025-zhkrf.

2 Methods and Background

2.1 Gaussian Process Regression

Gaussian process regression (GPR) is a non-parametric supervised machine learning method. Like other supervised machine learning regression models, GPR models learn to predict the value of a target variable for a given example of input variables by 'training' on several known examples of input-output pairs. In contrast with many other regression models, GPRs can provide reasonable predictive performance with relatively few training examples and provide an estimate of their uncertainty about predictions[1]. These properties make Gaussian process regression an attractive model method for machine learning problems related to computational and experimental materials research, where data availability can be sparse[2] and uncertainty estimates can be used to design experimental campaigns[3]. In the GPR context, a process is the extension of the concept of a probability distribution on a random variable to functions. A probability distribution describes random variables that are scalars or vectors. Processes describe distributions on functions.

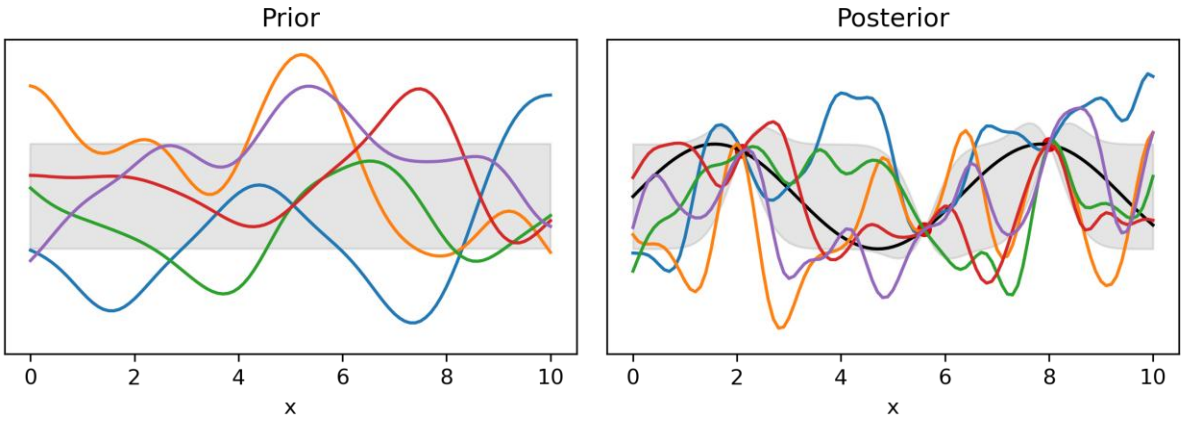


Figure 2.1: The left panel plots 5 functions drawn from a prior process defined by a radial basis function kernel. The gray shading plots the mean of the prior +/- one standard deviation. The right panel shows functions sampled from the posterior (colored curves) after the model has been fit to 3 points (red dots) observed from a target function (black curve). The range encompassed by one standard deviation is small near observed values but large away from them. Figure generation code inspired by reference 4.

Focusing on processes that are Gaussian results in tractable math. Figure 2.1 illustrates a somewhat naïve but descriptive explanation of how Gaussian process regression works[4]. Consider a modeling problem with a 1-dimensional input x. Before any observations are made, no information on how the target y depends on x is available. In defining a process with this information, we can make minimal assumptions about the distribution, perhaps only assuming that the mean of the distribution of functions is constant and zero. This results in the *prior* distribution, which in a Bayesian formalism, describes assumptions about the problem without having seen any information about it. Next, an experiment is run and three data points are collected. Fitting the model to these data could be understood as the process of running through every function in the (infinite) prior distribution and discarding any function that does not describe the observed data, perhaps to within a noise tolerance. This results in a distribution of functions like the one illustrated in the right panel of Figure 2.1. This posterior distribution can 'confidently' predict targets for inputs near the observed data, but the spread of the distribution gets much larger for values of the input that are far from observed data points. Calculating the mean and variance of this posterior distribution at a location of interest provides a means of making a prediction of both the target value and a measure of confidence in the predicted value. There are mathematical formulations for GPR that directly solve for a posterior and are more efficient than the iterative process above[1]. All supervised machine learning methods make an assumption that points which are 'near' each other will have similar target values. In Gaussian process regression, a definition of 'nearness' is critically important for obtaining reasonable results. A covariance function or kernel is used to measure distance between points in the input or feature space. This covariance function determines the properties of the functions in the prior and posterior, such as the characteristic length scale on which they vary. Kernel selection is critically important to effective implementation of GPR and

selecting kernel function hyperparameters is a main task in model selection for GPR. It is common to either use marginal likelihood optimization or cross validation to select model parameters. Marginal likelihood can be calculated directly from the model and theoretically incorporates a model fit vs. complexity tradeoff by definition, but cross validation can still be used to avoid overfitting.

2.2 Bayesian Optimization

Globally optimizing black box functions that are expensive to evaluate is a challenging task. Traditional gradient based optimization methods like gradient descent can't be used when a function's gradient is not available. Expensive function evaluations make exhaustive search or derivative approximation methods unfeasible. Materials researchers are often interested in finding the set of synthesis and processing parameters that optimize the performance of a system of interest. Often, the physics behind materials systems of interest is either not fully understood or too complex to model theoretically, making the desired performance metric a black box function of selected parameter values. Measuring an outcome of a set of parameter values requires running an entire experiment, likely involving many synthesis and characterization steps. These experiments are certainly expensive. Bayesian optimization is a global optimization technique that is well suited to working with expensive, black box functions, so it has gained significant interest for use in optimizing material properties[3,5,6]. Bayesian optimization works well for problems that can be formulated as a matter of maximizing or minimizing a measurable property over a finite parameter space. In Bayesian optimization, a probabilistic model of the response of the variable to be optimized to the values of variable parameters is developed using available information about the system. An acquisition function or utility function that evaluates the utility of evaluating potential points in the parameter space is then optimized using the understanding of the system

encoded in the probabilistic model. Traditional optimization is performed on this acquisition function, which is cheaper to evaluate as it only requires calls to the probabilistic model, not the main black box function of interest. The most optimal point selected from the acquisition function is then evaluated with the main function of interest. The probabilistic model is updated with the new result, and the acquisition process is repeated[7]. This process is 'Bayesian' in the sense that it incorporates available information about the system and updates with new information as it becomes available. This method requires a probabilistic model that can provide both predictions of the value of the target function at a point in the parameter space, as well as an estimate of the uncertainty in making that prediction. The model also must provide useful predictions in a low data regime. Gaussian process regression fits both requirements. There is an inherent explore vs. exploit tradeoff in Bayesian optimization. At each experiment, the decision-making algorithm must consider 'exploring' areas of parameter space that have not been sampled yet, or 'exploiting' existing information about the best parameter values discovered so far. Researchers can tune this tradeoff during their selection of an acquisition function for the problem. Three acquisition functions are commonly used. Each assigns a utility to any point in the parameter space by considering the model's prediction about the outcome of an experiment at that point along with the uncertainty associated with that prediction. Each also contains a parameter that allows for the balance between exploration and exploitation to be tuned. Probability of improvement assigns the highest utility to the point most likely to improve on the current 'best' point. This acquisition function only considers the probability of improving, and not the expected magnitude of the improvement. Expected improvement considers the magnitude of the predicted improvement over the current best point. Upper confidence bound is another formulation that considers both the predicted function value as well as the uncertainty. In all 3 acquisition functions, it is possible for

a point with very high uncertainty to have a high utility, even if the predicted value at that point is low.

2.3 Autonomous Experimentation

Autonomous experimentation combines active experimental design strategies like Bayesian optimization with automated experiment execution to independently achieve defined research objectives[8]. These systems are most commonly used to optimize a material property of interest over a prescribed experimental parameter space. Autonomous experimentation is also often referred to as closed-loop optimization because executed experiments feed back into the experiment selection process. Figure 2.2 illustrates the standard 'closed-loop' autonomous experimentation workflow for material optimization. First, an initial batch of pre-selected samples are synthesized and characterized to seed the Bayesian optimization algorithm. After these initial samples are finished, they are used to fit a GPR model. A Bayesian optimization acquisition function is then optimized over this model to select the next highest-utility sample(s) to measure. This sample is synthesized and characterized using the automated experiment workflow. The results of this experiment are used to update the GPR model, which is then used to again select the next best experiment. This process is iterated until a convergence criterion is satisfied or the experiment budget is exhausted.

This process requires several components to be useful and effective. A clearly defined, quantitative, measurable target property to optimize is needed. Multiple target properties can also be addressed through multi-objective optimization approaches. Characterization techniques that either directly provide or whose results can easily be processed into scalar values are well suited for autonomous experimentation. Techniques that require extensive human analysis, such as small-angle scattering or imaging methods, are harder to integrate into autonomous

experimentation loops. A well-defined parameter space over which to optimize is needed. This parameter space is generally selected through existing knowledge of the material system and initial experiments. An automated experiment execution workflow is generally considered to be necessary for autonomous experimentation. Fully automated experiments can provide greater reproducibility and reliability compared to manual experiments, may have higher experiment throughput, and free researchers to focus on other tasks. However, fully automating many experiments is challenging. Many simple, everyday tasks done in lab such as uncapping bottles, measuring solid powders, or opening doors can be devilishly difficult and expensive to reliably automate. The recent conversations around human-in-the-loop automated experiments tend to acknowledge that autonomous experimentation may have more impact in terms of research gains

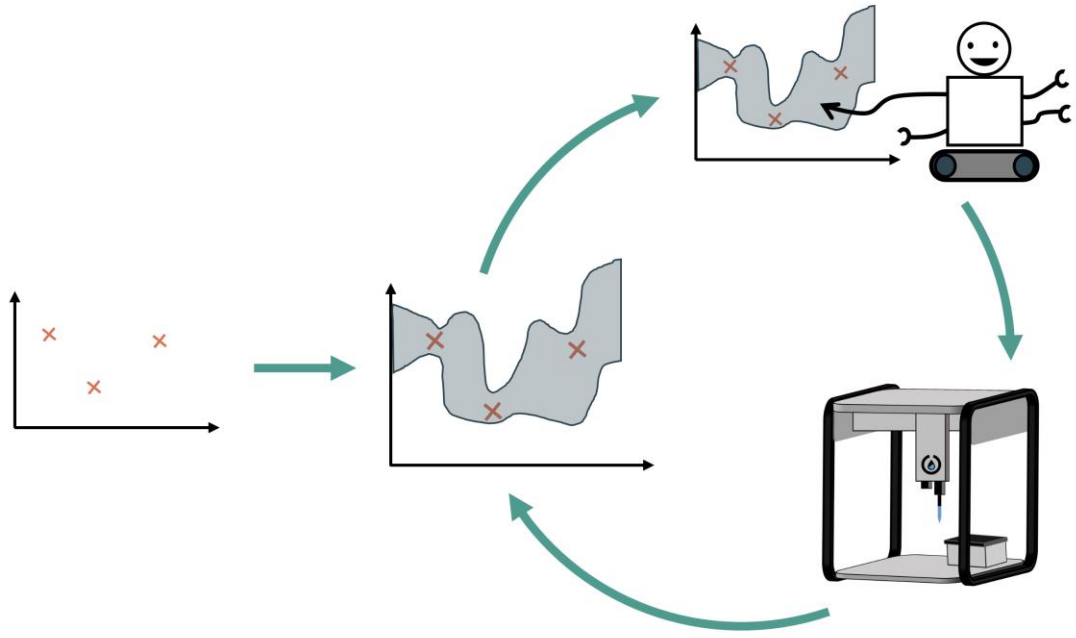


Figure 2.2: Autonomous experimentation platforms combine the algorithmic decision making of Bayesian optimization with automated experiment execution. An optimization campaign starts with some initial data which allows an initial model of the experimental system to be fit. A decision-making algorithm picks the most valuable point to test next and dispatches the experiment to an automated experimentation workflow. The model is updated with results from the experiment and the cycle repeats.

for effort expended if some human execution is still used for difficult to automate experiment steps[9].

While they can be powerful tools, autonomous experimentation platforms are not appropriate for all research objectives. Many forms of exploratory research that involve creative development of new synthesis processes, application of new methods, fundamental studies of novel materials, or other novel cutting-edge science doesn't fit well into the autonomous experimentation framework, at least as it is developed and discussed in its current form. Perhaps there will come a time in the future when flexible automation coupled with automated reasoning systems that understand physics and chemistry will play an important role in exploratory research. However, for the time being, autonomous experimentation is most effective when applied to reasonably well understood materials systems that need extensive tweaking and optimization to get from lab bench concept to real world solution.

Autonomous experimentation platforms have been successfully applied to many materials systems. In perhaps the first demonstration of an autonomous experimentation platform, a platform was developed to optimize the growth rates of carbon nanotubes[10]. In this experiment, an autonomous experimental planner learned to control the growth rate of carbon nanotubes in a chemical vapor deposition process by varying temperature, pressure, and partial pressures of reactants. Autonomous experimentation platforms have also been developed to optimize organic electronic materials synthesized from Suzuki-Miyaura cross coupling reactions[11], photostability of organic photovoltaic materials[12], and bandgap engineering of lead-halide perovskite quantum dots[5]. With regards to nanoparticle synthesis, autonomous experimentation platforms have been developed by multiple groups to optimize spectral properties of gold and silver nanoparticles[13–15]. An automated experimentation platform has been developed for synthesizing nanomaterials,

including the solid spherical and mesoporous silica nanoparticles synthesized in this work[16]. The high level of discussion surrounding autonomous experimentation in terms of perspectives, infrastructure development work, and funding[17–20] suggests that these systems will see wide adoption across materials research in the coming years.

2.4 Small-angle X-ray Scattering

Small-angle X-ray scattering (SAXS) is a material characterization method that provides information on length scales ranging from 1 to 20,000 nm. SAXS provides a spatially and orientationally averaged description of structures present in a sample. small-angle scattering uses the interaction of radiation with matter to describe structure and dynamics. In addition to X-rays, neutrons and visible light are used as radiation sources for scattering. A standard pinhole SAXS experimental setup is illustrated in Figure 2.3. A monochromated, collimated beam of X-ray radiation is placed on a prepared sample. Most of the incident radiation is transmitted directly

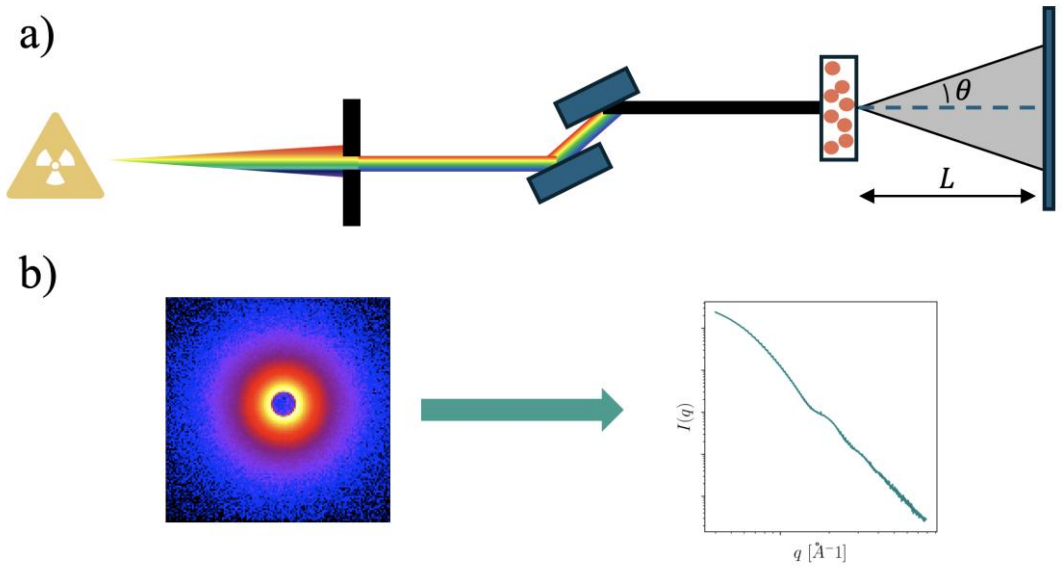


Figure 2.3: a) Typical setup for a SAXS experiment. An X-ray beam is generated from a source with some bandwidth and radial dispersity. This beam is collimated and monochromated before arriving at the sample. The scattered X-ray intensity is detected at a distance L from the sample by a detector. Additional collimation after the monochromator is omitted for clarity. B) Data collected by a 2D detector is reduced to a 1D scattering curve.

through the sample, but a small amount is scattered. A 2-dimensional detector is placed a known distance away from the sample. Intensity at the detector is measured as a function of position, giving a 2D scattering profile. For an isotropic sample, this profile has angular symmetry. This profile is azimuthally averaged to produce a 1D scattering curve. Positional variation in the scattering profile is converted to momentum transfer (between the radiation and the scatterer) by the relation $q = |\vec{\boldsymbol{q}}| = \frac{4\pi}{\lambda}\sin\frac{\theta}{2}$. Alternative instrument geometries include Bonse-Hart instruments such as the USAXS instrument discussed in chapter 6, which replace the 2-dimensional detector with an analyzer crystal and a pin diode intensity detector[21].

The reduced 1D scattering curve is generally plotted on log-log axes as measured intensity vs momentum (q). q roughly correlates to an inverse length scale. Features with large length scales appear in scattering curves at low q values, and features with small length scales appear at high q. The range of q values measured depends on the distance from the sample to the detector and can be adjusted by moving the position of the detector. SAXS measures the amplitude of the Fourier transform of the spatial autocorrelation function of scattering density in a sample. The general scattering equation is shown in Equation 2.1.

$$I(\boldsymbol{q}) = \int_V \gamma(\boldsymbol{r}) e^{-i\,\boldsymbol{q}\cdot\boldsymbol{r}} d\boldsymbol{r} \tag{2.1}$$

Here, $\gamma(\boldsymbol{r})$ is the spatial correlation function of scattering length density in the sample. This is an inherently lossy measurement as the phase component of the Fourier transform cannot be measured. This component contains most of the information[22]. Given this, scattering is an incomplete description of the structure of a sample, and SAXS scattering profiles are non-unique, meaning multiple physical structures can have the same measured profile. Interpretation of SAXS data requires a-priori information or assumptions about this system. With this information,

appropriate physical models can be selected and fit to the scattering profiles, providing quantitative results about the sample. Given this, SAXS measurements are commonly combined with complementary local imaging techniques like scanning or transmission electron microscopy.

Many models and simplifying assumptions are used in analyzing SAXS data. Because the Fourier transform is a linear operation, it is possible in some cases to separate the scattering contributions from the particles in a system from the scattering contributions of the arrangement of particles in a system. In these cases, the particle scattering is described by a quantity known as a form factor and denoted P(q), while the arrangement scattering is described by a structure factor denoted S(q). For a dilute system of identical particles, it is possible to rewrite Equation 2.1 with these factors into Equation 2.2.

$$I_m(\boldsymbol{q}) = \Phi V_{part} P(\boldsymbol{q}) S(\boldsymbol{q}) \tag{2.2}$$

Here, Φ is the volume fraction of the particle phase of interest within the sample volume and V_{part} is the volume of an individual particle. This linear decoupling of measured scattering from individual particle and overall structural contributions makes it possible to measure both for a sample. When working with an unknown sample, it is typical to first measure scattering from a very dilute sample to isolate the form factor P(q). Then, the scattering from the sample at the concentration of interest can be measured. The form factor can be divided from the overall scattering of this sample to elucidate the inter-particle structure in the sample via the structure-factor S(q). However, before any conclusions can be drawn from SAXS data, several correction and processing steps need to be completed. After a 2D SAXS scattering pattern is collected by the detector, a number of detector and experimental artifact corrections are applied[23]. These corrections are generally instrument specific and are beyond the scope of this document. Corrected

2D scattering data is then 'reduced' to the 1D scattering curve described above. Measured 1D scattering data is background subtracted before further analysis is made by subtracting scattering from a 'blank' sample that contains everything except the sample of interest. It is also often desired to measure sample features with length scales that extend beyond the q range that can be measured in a single measurement. This is handled by taking measurements at multiple sample-detector distance configurations, then merging the resulting subtracted 1D scattering profiles to result in a single scattering curve with a larger q range. This process is typically handled by manually selecting cut points to merge the contributing curves at.

With a corrected scattering curve, quantitative conclusions are drawn by fitting an appropriate model to the data. In a traditional modeling process, a parametric model that describes the scattering of the expected structure in the sample is selected. This model is fit to the data with least squares approach. The structural parameters of the fit model can be used to describe the sample, if the model reasonably fits the scattering data. This process illustrates the importance of a priori or external understanding of the structure of the sample when analyzing SAXS data. For a dilute suspension of monodisperse spherical particles, the scattering equation simplifies to $I(q) = (\Delta\rho)^2 V^2 P_0(q)$, where $P_0(q) = \left[\frac{3(\sin qR - qRcosqR)}{(qR)^3}\right]^2$. This scattering equation is plotted in Figure 2.4. The location of the minima in q is determined by the radius of the particles, so using this model it is possible to determine spherical nanoparticle size by fitting this model to data. Real particles, of course, are likely not perfectly monodisperse. Polydispersity in particle size can be accounted for by assuming a particle size distribution and integrating the scattering contributions from particles in that distribution, i.e. $I(q) = c\int_0^\infty D(R) * P_0(q,R)dR$, where $D(R)$ is the size distribution. Distribution parameters can be fit by minimizing the least squares error between the calculated scattering and the data. An alternative to fitting analytical structural models is Monte

Carlo modeling[24,25]. This approach uses Monte Carlo rejection sampling to fit a particle size distribution to measured scattering data. A parameterized model for individual scatterers is assumed. For example, a spherical scattering particle with radius as a parameter could be assumed for a nanoparticle experiment. A set of these particles is initialized. At each iteration, the overall scattering pattern is calculated by summing over the individual scatterers and a figure of merit for model fit to measured data is calculated. At each iteration, one scatterer is replaced with another with randomly selected parameter values, then the scattering and figure of merit is recalculated. If the fit improves the change is kept, if not it is rejected. This process proceeds until a particle size distribution that agrees with the measured scattering data is found or a stopping criterion is met. The major advantage of this method compared to the analytical modeling approach described above is that a form for the particle size distribution is not assumed, meaning the fit distribution can potentially more accurately fit the data. A downside is the increased computational cost and degeneracy in the possible solutions.

Scattering can be used to understand the structure of mesoporous materials, such as the mesoporous colloidal silicas discussed in chapter 6. For porous materials with highly ordered packing structures, peak indexing of the Bragg reflection peaks can be used to make phase identifications[26,27]. Other techniques can be used to extract quantitative pore volume fractions, specific surface areas, and porous structure arrangements from ordered mesoporous materials. A Porod's law analysis can be applied to extract pore volume fraction and specific surface

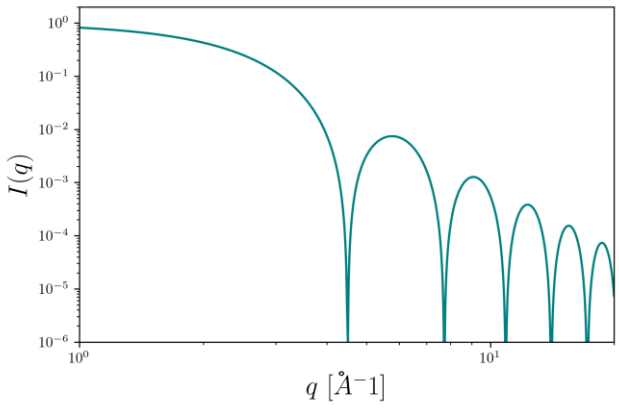


Figure 2.4: Form factor for a spherical particle.

area[28,29]. Porod's law essentially states that scattering due to an interface follows a q^{-4} relation. This relation makes it possible to identify scattering due to sharp interfaces between the medium filling a pore and the surrounding material. This makes it possible to extract specific surface area and volume fraction of pores from relevant scattering regions.

2.5 Dynamic light scattering

Dynamic light scattering (DLS) was used for particle size analysis for the silica nanoparticle synthesis described in this work. DLS calculates particle size by fitting Stokes-Einstein diffusion to a light scattering autocorrelation function. The experimental configuration is illustrated in figure 2.5. Monochromatic, coherent light is used to illuminate a sample of colloidal particles suspended in a dispersant. The particles scatter the incident light, and interference between light scattered from the particles produces a 'speckle pattern' on a detector some distance from the sample. As the particles move due to Brownian motion, the speckle pattern changes. These dynamic changes in speckle pattern are processed through a correlator to produce a correlation function for the measurement that is related to the displacement of particles. The decay rate of this correlation function is related to the size of the particles. Large particles move slower than small particles, and hence a correlation function measured from a sample with large particles will decay slower than one with smaller particles. Polydisperse samples are typically fit with the Cumulants method to

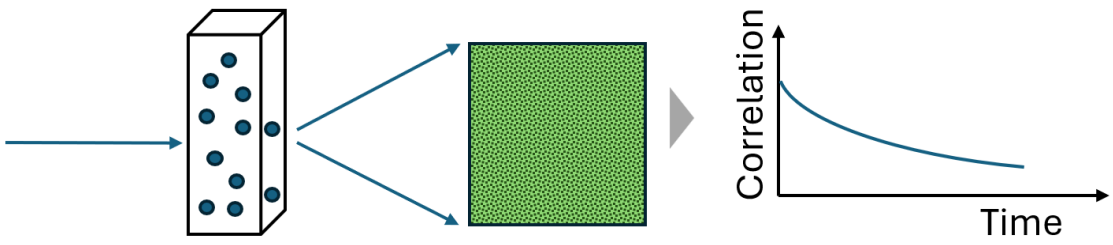


Figure 2.5: Dynamic light scattering experimental configuration.

determine size and polydispersity. In this method, the measured correlation function is fit with an expanded single exponential function shown in equation 2.3.

$$G_2(\tau) = A[1 + Bexp(-2\Gamma\tau + \mu_2\tau^2)] \quad 2.3$$

Where A is the correlation function amplitude, B is the correlation function baseline, and $\Gamma = Dq^2$, where D is the particle diffusion coefficient and q is the same as defined for SAXS. The diffusion coefficient D is related to particle size using the Stokes-Einstein equation: $D = \frac{6k_BT}{6\pi\eta R}$, which depends on the temperature and viscosity of the dispersant. Equation 2.3 is linearized and fit to the correlation function to determine a mean particle size and a measure of polydispersity known as the polydispersity index (PDI)[30]. It is important to note that DLS measures the size of a particle as if it were a sphere undergoing Brownian motion. This may not be an informative value for non-spherical particles, and the technique does not provide any information on particle shape.

2.6 References

(1) Williams, C. K.; Rasmussen, C. E. *Gaussian Processes for Machine Learning*; MIT press Cambridge, MA, 2006; Vol. 2.

(2) Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine Learning Methods for Small Data Challenges in Molecular Science. *Chem. Rev.* **2023**, *123* (13), 8736–8780. https://doi.org/10.1021/acs.chemrev.3c00189.

(3) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving Laboratories. *Trends Chem.* **2019**, *1* (3), 282–291. https://doi.org/10.1016/j.trechm.2019.02.007.

(4) *Illustration of prior and posterior Gaussian process for different kernels*. scikit-learn. https://scikit-learn/stable/auto_examples/gaussian_process/plot_gpr_prior_posterior.html (accessed 2024-01-30).

(5) Abdel-Latif, K.; Epps, R. W.; Bateni, F.; Han, S.; Reyes, K. G.; Abolhasani, M. Self-Driven Multistep Quantum Dot Synthesis Enabled by Autonomous Robotic Experimentation in Flow. *Adv. Intell. Syst.* **2021**, *3* (2), 2000245. https://doi.org/10.1002/aisy.202000245.

(6) Abolhasani, M.; Kumacheva, E. The Rise of Self-Driving Labs in Chemical and Materials Sciences. *Nat. Synth.* **2023**, 1–10. https://doi.org/10.1038/s44160-022-00231-0.

(7) Garnett, R. Bayesian Optimization. https://www.cse.wustl.edu/~garnett/cse515t/spring_2019/files/lecture_notes/12.pdf (accessed 2024-02-05).

(8) Tom, G.; Schmid, S. P.; Baird, S. G.; Cao, Y.; Darvish, K.; Hao, H.; Lo, S.; Pablo-García, S.; Rajaonson, E. M.; Skreta, M.; Yoshikawa, N.; Corapi, S.; Akkoc, G. D.; Strieth-Kalthoff, F.; Seifrid, M.; Aspuru-Guzik, A. Self-Driving Laboratories for Chemistry and Materials Science. *Chem. Rev.* **2024**, *124* (16), 9633–9732. https://doi.org/10.1021/acs.chemrev.4c00055.

(9) Scheurer, C.; Reuter, K. Role of the Human-in-the-Loop in Emerging Self-Driving Laboratories for Heterogeneous Catalysis. *Nat. Catal.* **2025**, *8* (1), 13–19. https://doi.org/10.1038/s41929-024-01275-5.

(10) Nikolaev, P.; Hooper, D.; Webber, F.; Rao, R.; Decker, K.; Krein, M.; Poleski, J.; Barto, R.; Maruyama, B. Autonomy in Materials Research: A Case Study in Carbon Nanotube Growth. *Npj Comput. Mater.* **2016**, *2* (1), 1–6. https://doi.org/10.1038/npjcompumats.2016.31.

(11) Seifrid, M.; Pollice, R.; Aguilar-Granda, A.; Morgan Chan, Z.; Hotta, K.; Ser, C. T.; Vestfrid, J.; Wu, T. C.; Aspuru-Guzik, A. Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab. *Acc. Chem. Res.* **2022**. https://doi.org/10.1021/acs.accounts.2c00220.

(12) Langner, S.; Häse, F.; Perea, J. D.; Stubhan, T.; Hauch, J.; Roch, L. M.; Heumueller, T.; Aspuru-Guzik, A.; Brabec, C. J. Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems. *Adv. Mater.* **2020**, *32* (14), 1907801. https://doi.org/10.1002/adma.201907801.

(13) Salley, D.; Keenan, G.; Grizou, J.; Sharma, A.; Martín, S.; Cronin, L. A Nanomaterials Discovery Robot for the Darwinian Evolution of Shape Programmable Gold Nanoparticles. *Nat. Commun.* **2020**, *11* (1), 2771. https://doi.org/10.1038/s41467-020-16501-4.

(14) Mekki-Berrada, F.; Ren, Z.; Huang, T.; Wong, W. K.; Zheng, F.; Xie, J.; Tian, I. P. S.; Jayavelu, S.; Mahfoud, Z.; Bash, D.; Hippalgaonkar, K.; Khan, S.; Buonassisi, T.; Li, Q.;

Wang, X. Two-Step Machine Learning Enables Optimized Nanoparticle Synthesis. *Npj Comput. Mater.* **2021**, *7* (1), 1–10. https://doi.org/10.1038/s41524-021-00520-w.
(15) Tao, H.; Wu, T.; Kheiri, S.; Aldeghi, M.; Aspuru-Guzik, A.; Kumacheva, E. Self-Driving Platform for Metal Nanoparticle Synthesis: Combining Microfluidics and Machine Learning. *Adv. Funct. Mater.* **2021**, *31* (51), 2106725. https://doi.org/10.1002/adfm.202106725.
(16) Zaki, M.; Prinz, C.; Ruehle, B. A Self-Driving Lab for Nano- and Advanced Materials Synthesis. *ACS Nano* **2025**, *19* (9), 9029–9041. https://doi.org/10.1021/acsnano.4c17504.
(17) MacLeod, B. P.; Parlane, F. G. L.; Brown, A. K.; Hein, J. E.; Berlinguette, C. P. Flexible Automation Accelerates Materials Discovery. *Nat. Mater.* **2022**, *21* (7), 722–726. https://doi.org/10.1038/s41563-021-01156-3.
(18) *Acceleration Consortium*. https://acceleration.utoronto.ca/ (accessed 2024-01-29).
(19) Kusne, A. G.; Gao, T.; Mehta, A.; Ke, L.; Nguyen, M. C.; Ho, K.-M.; Antropov, V.; Wang, C.-Z.; Kramer, M. J.; Long, C.; Takeuchi, I. On-the-Fly Machine-Learning for High-Throughput Experiments: Search for Rare-Earth-Free Permanent Magnets. *Sci. Rep.* **2014**, *4* (1), 6367. https://doi.org/10.1038/srep06367.
(20) Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Aspuru-Guzik, A. Gryffin: An Algorithm for Bayesian Optimization of Categorical Variables Informed by Expert Knowledge. *Appl. Phys. Rev.* **2021**, *8* (3), 031406. https://doi.org/10.1063/5.0048164.
(21) Ilavsky, J.; Zhang, F.; Andrews, R. N.; Kuzmenko, I.; Jemian, P. R.; Levine, L. E.; Allen, A. J. Development of Combined Microstructure and Structure Characterization Facility for in Situ and Operando Studies at the Advanced Photon Source. *J. Appl. Crystallogr.* **2018**, *51* (3), 867–882. https://doi.org/10.1107/S160057671800643X.
(22) Sivia, D. S. *Elementary Scattering Theory: For X-Ray and Neutron Users*; Oxford University Press, 2011. https://doi.org/10.1093/acprof:oso/9780199228676.001.0001.
(23) Pauw, B. R. Everything SAXS: Small-Angle Scattering Pattern Collection and Correction. *J. Phys. Condens. Matter* **2013**, *25* (38), 383201. https://doi.org/10.1088/0953-8984/25/38/383201.
(24) Bressler, I.; Pauw, B. R.; Thünemann, A. F. McSAS: Software for the Retrieval of Model Parameter Distributions from Scattering Patterns. *J. Appl. Crystallogr.* **2015**, *48* (3), 962–969. https://doi.org/10.1107/S1600576715007347.
(25) Pauw, B. R.; Pedersen, J. S.; Tardif, S.; Takata, M.; Iversen, B. B. Improvements and Considerations for Size Distribution Retrieval from Small-Angle Scattering Data by Monte Carlo Methods. *J. Appl. Crystallogr.* **2013**, *46* (2), 365–371. https://doi.org/10.1107/S0021889813001295.
(26) Hamley, I. W.; Castelletto, V. Small-Angle Scattering of Block Copolymers: In the Melt, Solution and Crystal States. *Prog. Polym. Sci.* **2004**, *29* (9), 909–948. https://doi.org/10.1016/j.progpolymsci.2004.06.001.
(27) Kim, T.-W.; Chung, P.-W.; Lin, V. S.-Y. Facile Synthesis of Monodisperse Spherical MCM-48 Mesoporous Silica Nanoparticles with Controlled Particle Size. *Chem. Mater.* **2010**, *22* (17), 5093–5104. https://doi.org/10.1021/cm1017344.
(28) M. Weigandt, K.; C. Pozzo, D.; Porcar, L. Structure of High Density Fibrin Networks Probed with Neutron Scattering and Rheology. *Soft Matter* **2009**, *5* (21), 4321–4330. https://doi.org/10.1039/B906256D.
(29) Lindner, P. (Peter). Neutrons, X-Rays and Light : Scattering Methods Applied to Soft Condensed Matter. *No Title*.

(30) *What is the Z-average?* https://www.malvernpanalytical.com/en/learn/knowledge-center/faqs/faq0015averagediameter (accessed 2025-05-14).

3 Prediction of Exact Quantum Kinetic Rate Constants for Surface-Catalyzed Reactions

The work presented in this section is reported in the publication:

Pelkie, Brenden G., and Stéphanie Valleau. "Machine learning the quantum flux–flux correlation function for catalytic surface reactions." Digital Discovery 1.6 (2022): 851-858.

Dr. Stéphanie Valleau provided support and guidance in all aspects of this work, including project supervision, quantum kinetic methods, and machine learning approaches.

3.1 Introduction

Reaction rate constants relate reactant concentrations to reaction rates as a component of a rate law[1]. Rate constants are primarily determined by temperature and the potential energy surface of a reaction. Rate laws are a required component for kinetic modeling, for example in microkinetic models of surface catalyzed reactions or for reactor design. Obtaining rate constants computationally is an expensive process. Many theoretical methods exist for calculating rate constants[2]. All require a minimum energy path to be known. A minimum energy path is the path through the (3N-6)-dimensional potential energy surface that connects the reactant minima to the product minima with the lowest maximum energy, where N is the number of atoms in the reactive system. Finding this minimum energy path requires a saddle searching algorithm such as the nudged elastic band (NEB) method[1]. These methods use the gradient of the potential energy with respect to atomic position to locate the minimum energy path. Gradients are calculated using a quantum mechanical modelling method, generally density functional theory (DFT). The NEB method starts with known reactant and product configurations, calculated from a DFT (or other method) relaxation calculation. A series of atomic configurations (images) are interpolated

between the reactant and product. Their positions are then updated based on the gradient of the potential energy surface and an imposed 'elastic' force keeping them equally spaced[3]. This process requires tens to hundreds of DFT evaluations of intermediate structures, leading to the high cost of calculating reaction rate constants. Once a minimum energy path is found, it can be used with a kinetic theory to calculate a rate constant. Transition state theory (TST) is the most used kinetic theory owing to its reasonable performance and simplicity. TST requires minimal information about the potential energy surface near the reactants and products. It makes a few key simplifying assumptions: that all dynamics are classical, that there is a dividing surface that separates the products and reactants, that the 'transition states' on the dividing surface are at equilibrium with the reactant state, and that all reaction trajectories that cross the dividing surface with no recrossing effects. In addition to potential energy surface information, the partition functions of the reactants and transition state is also needed to calculate and absolute rate constant with TST. TST provides good approximations to reaction rates for many situations. However, the assumptions made in TST, particularly that all dynamics are classical and that no recrossing occurs, can lead to deviation from true rate constants in some situations. For example, quantum tunneling can provide a significant contribution to reaction rates for light reactants at low temperatures. As TST assumes classical dynamics, it does not account for these impacts. Figure 3.1 shows the impact of this deviation. To account for these deviations, one can turn to an exact quantum rate constant theory such as the flux-flux correlation method. The flux-flux correlation method calculates an exact

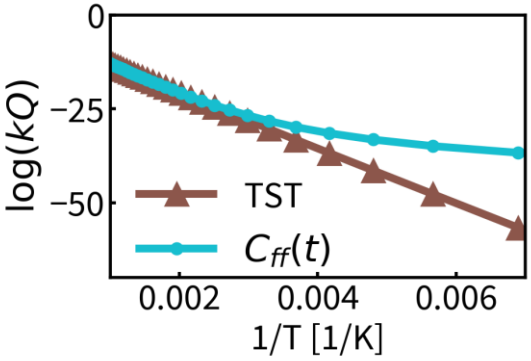


Figure 3.1: Arrhenius plot of the diffusion of H on Ni(100) shows that TST diverges from exact quantum rate constants at low temperatures.

quantum rate constant for a reaction by projecting the wave function of the reaction forward in time[4]. In this method, the rate constant of a reaction is given by Equation 3.1.

$$k(T) = \frac{1}{Q_r(T)} \int_0^\infty C_{ff}(t)dt \tag{3.1}$$

Here, $Q_r(T)$ is the partition function of the reactants, and $C_{ff}(t)$ is the flux-flux correlation function of the reaction, given in Equation 3.2.

$$C_{ff}(t) = \frac{1}{\hbar} Tr\left[e^{-\frac{\beta}{2}\hat{H}}\left[\hat{H},\hat{\theta}\right]e^{-\frac{\beta}{2}\hat{H}}\, e^{i\frac{\hat{H}t}{\hbar}}\left[\hat{H},\hat{\theta}\right]e^{-i\frac{\hat{H}t}{\hbar}}\right] \tag{3.2}$$

Here, $\hbar$ is the reduced Planck's constant, Tr is the matrix trace operator, $\beta = \frac{1}{k_B T}$, $\hat{H}$ is the Hamiltonian operator, and $\hat{\theta}$ is the Heaviside operator, which places a dividing surface between the products and reactants at the reaction coordinate that corresponds to the transition state. The square brackets $[\hat{H},\hat{\theta}]$ represent the commutator of $\hat{H}$ and $\hat{\theta}$, defined as $[\hat{H},\hat{\theta}] = \hat{H}\hat{\theta} - \hat{\theta}\hat{H}$. Figure 3.2 provides a visualization of what this method is calculating. With these operators defined,

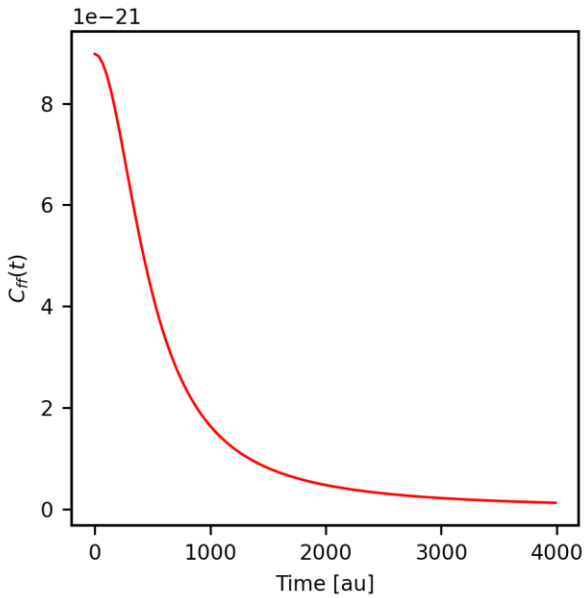


Figure 3.3: The flux-flux correlation function for the reaction CHOH* + * -> HCO* + H* on Ag(111) at 241 K.

Equation 3.2 can be evaluated at discrete time points until its value converges to zero. $C_{ff}(t)$ is then numerically integrated to find the rate constant. Figure 3.3 shows what a typical flux-flux correlation function for a reaction studied in this work looks like. While this method can provide theoretically exact rate constants, it introduces considerable computational expense compared to transition state theory. A common approximation is to represent the reaction potential energy surface as a 1D reaction barrier defined along a reaction coordinate axis. Even with this approximation, calculating a rate constant can still take weeks of computing time. Evaluation of $C_{ff}(t)$ doesn't yield an absolute reaction rate constant but rather a reaction quotient, $k(T)Q_r(T)$, where $Q_r(T)$ is the reactant partition function. This work focuses on methods for accessing $C_{ff}(t)$ only. Accessing an absolute rate constant will require calculating the partition function. Recent work has pointed to the possibility of predicting this quantity with machine learning methods as well[5].

Supervised machine learning approaches have found success in predicting many properties of relevance for kinetics or heterogeneous catalysis. ML has been used to predict reactant and transition state partition functions[5], Gibbs free activation energies[6], activation energies[7], and quantum rate constants for small systems[8–10]. ML has been used to predict many quantities in the context of catalysis[11–13], especially adsorption energy prediction[14,15]. Far less effort has been spent on machine learning approaches for predicting catalytic rate constants. We believe this is largely due to the lack of datasets available in this space.

To provide rapid access to exact quantum rate constants, in this work we sought to develop machine learning methods to predict exact rate constants. We evaluated various options for predicting exact rate constants for heterogeneously catalyzed reactions.

3.2 Establishing a dataset of exact reaction rate constants

To develop machine learning models for predicting rate constants, a dataset of rate constants calculated with the flux-flux correlation function was needed. Developing this dataset required implanting software to calculate the flux-flux correlation function and associated rate constant, selection of reactions to include, and calculation of rate constants for those reactions.

A python library for the calculation of the flux-flux correlation function and the associated rate constant quotient was developed. The sinc basis set discrete variable representation was used to represent the quantum operators involved in the flux flux correlation function. This representation uses a basis set of evenly spaced sinc functions ($sinc(x) = \frac{\sin(x)}{x}$). Its use allows all quantum operators used to be written as matrices with elements evaluated on this grid[16,17]. This allows $C_{ff}(t)$ to be evaluated with a series of matrix manipulations. These manipulations were primarily carried out in the numerical package Numpy. However, some reactions displayed divergence issues that were believed to be caused by limited numerical precision. Evaluating $C_{ff}(t)$ for these reactions using arbitrary precision numerical operations solved this divergence issue at the expense of greatly increase computational cost. High precision calculations were implemented using mpmath[18] and flint[19].

Reactions to include in the dataset were selected from the Catalysis-Hub database[20]. Catalysis hub is a web platform for sharing computational catalysis research. It hosts thousands of DFT calculated reaction and adsorption energies as well as reaction barriers. For this work, a set of 14 gas-phase heterogeneous catalytic surface reactions was selected from this database. The selected reactions had either an NEB computed minimum energy path with multiple images, or at least reactant, product, and transition state images available. The small reaction dataset size was chosen to reduce the computational cost of calculating quantum rate constants. Table 3.1 describes the reactions that were selected. For each reaction, a reaction coordinate was calculated for each available image (molecular configuration) to describe the reaction progress on a 1D axis. An Eckart barrier or skew normal function was fitted to the 1D reaction energy pathway to provide a continuous approximation to the minimum energy path. This fit barrier was used to evaluate the Hamiltonian operator for the reaction in the above-mentioned DVR representation, which was used in the calculation of the flux-flux correlation function. For each reaction, four temperatures were randomly selected from the range [150, 400] K for the calculation of the flux-flux correlation function. This resulted in 55 completed

Table 3.1: Description of selected reactions from Catalysis Hub.

N_{react}	Reaction	Catalyst	Surface	E_a [kcal / mol]
1	$CH^*+* \rightarrow C^*+H^*$	Rh	111	33.5
2	$COH^*+* \rightarrow C^*+OH^*$	Rh	111	27.9
3	$CHOH^*+* \rightarrow CHO^*+H^*$	Rh	111	19.2
4	$CH_3^*+* \rightarrow CH_2^*+H^*$	Rh	211	10.3
5	$CH_2OH^*+* \rightarrow CHOH^*+H^*$	Pt	111	27.0
6	$CH_2^*+* \rightarrow CH^*+H^*$	Ir	111	2.6
7	$CH_3^*+* \rightarrow CH_2^*+H^*$	Ir	111	14.2
8	$CH_3^*+* \rightarrow CH_2^*+H^*$	Pt	111	24.7
9	$CHOH^*+* \rightarrow HCO^*+H^*$	Ag	111	14.7
10	$CHOH^*+* \rightarrow CH^*+OH^*$	Ir	111	12.3
11	$CHO^*+* \rightarrow CO^*+H^*$	Pd	111	2.9
12	$COH^*+H^* \rightarrow CHOH^*$	Cu	100	17.0
13	$CO^*+H^* \rightarrow CHO^*$	Cu	100	22.4
14	$CH^*+H^* \rightarrow CH_2^*$	Cu	100	17.0

flux-flux correlation functions and associated rate constants. One flux-flux correlation function did not converge and was not included in the final set.

3.3 Machine learning rate constant quotients

Various approaches were evaluated for learning rate constant quotients from reaction features. A direct approach to learn rate constant quotients directly from reactant geometry features and temperature was attempted. This resulted in poor performance. In a second approach, fit parameters of a distribution function were learned to provide an approximation of the flux-flux correlation function. This fit function was then integrated to provide a rate constant. This resulted in improved performance. For both approaches, Gaussian Process Regression (GPR) was used as the modeling method. This selection was made due to the small size of the available dataset. Two chemistry-informed train/test splitting methods were applied. Randomly splitting the dataset of 55 rate constants would not have been appropriate as this would have leaked information about the reactant geometry or reaction temperature to the test set. To avoid this, we split the dataset by either temperature or reaction. In the temperature split, one temperature example for each reaction was selected as a 'test' example, while the rest remained in the 'train' set. This split evaluates the ability of a machine learning approach to learn the temperature dependence of the rate constant. In the reaction split, 4 reactions were selected as 'test' reactions. All rate constants from all 4 temperatures for each reaction were placed in the 'test' set, while the remaining 10 reactions constituted the 'train' set. This approach evaluated the predictive abilities of the model for new reaction geometries. This split was made manually to ensure that every metallic catalytic surface was represented in the train set. Given that all metals are represented in the training set, no conclusions about the ability to learn the dependence of reaction rates on catalyst can be drawn.

In molecular machine learning, representations are generally required to convert the 3D atomic coordinates used to represent molecules for structural calculations into 1D feature vectors that can be used as inputs to machine learning models. In this work, we evaluated two molecular representations: Coulomb matrices and encoded bonds. Coulomb matrices are essentially matrices of a coulomb-interaction-inspired pair interaction term for every atom in a molecule that are flattened into a vector[21]. Encoded bonds use a 1D vector to represent a smoothed histogram of pair inter-atomic distances[22]. These representations encode the geometry of a single molecule. However, when predicting reaction quantities, representations of both reactant and product geometries are generally needed. One strategy to represent both geometries is to use difference features, in which both product and reactant features are calculated and the difference between product and reactant vectors are used as features. This approach has proven successful in past work[8,23] and is the approach we used here. Input features as well as GPR kernel functions were selected by comparing the mean absolute error of train set predictions of rate constants for both temperature-wise and reaction-wise split. Based on this approach, we selected coulomb matrix geometry features for both splitting strategies. In addition to coulomb matrix features, inverse temperature (1/T) and reaction energy were also included as features. Reaction energy is a 'free' feature as it will be calculated when the reactant and product geometries are generated by DFT relaxation. For temperature-wise splits, a combination of Matérn and pairwise linear kernels was selected. For reaction-wise splits, two Matérn kernels were used. To avoid overfitting the training data, a noise term was added to the diagonal of the covariance matrix during fitting of the GPR model[24]. This term was set at $\alpha = 0.5$. This value was selected by comparing the predictive error on the train and test set for multiple noise values, and selecting the value that minimized the difference between train and test. Due to the small size of our dataset, a validation set was not

available. Use of a validation set would have been preferable for these hyperparameter selection tasks.

These model parameters were used to train models to directly predict reaction rate constants. For the temperature-wise test split, the model had a high prediction accuracy with a mean absolute percent error (MAPE) of 0.998%. Figure 3.4 shows a parity plot of predicted rate constant quotients, which shows good agreement between predicted and ground truth values. However, the performance on the reaction-wise split was lower. For this set, the MAPE was 33.8%. While this is a large error value, the uncertainty associated with the model predictions was also large. Figure 3.5 shows that the ground truth rate constants fall within a standard deviation of the predicted values for an example reaction. This large error motivated an alternative prediction strategy.

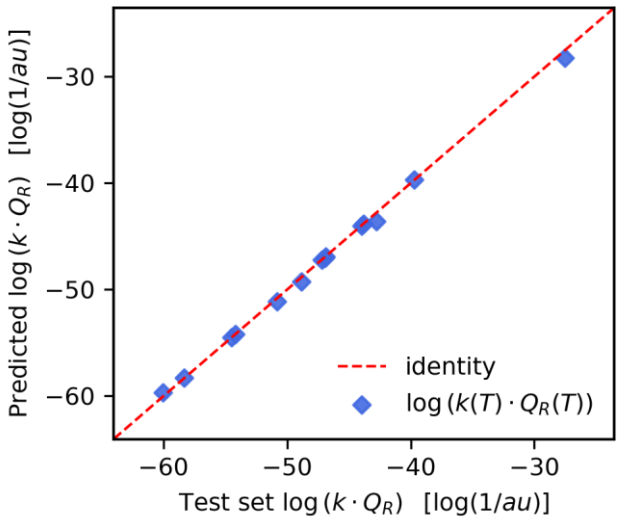


Figure 3.4: Parity plot of the predicted test set values of log((k(T) · Q_R (T)); the product of the quantum reaction rate constant k(T) with the reactant canonical partition function $Q_R(T)$ for temperature-wise split train and test sets. The predicted values are in strong agreement with the test set with a test set MAPE of 0.998%.

Instead of directly predicting rate constants, we turned to predicting fitting parameters to predict approximations of the flux-flux correlation function for reactions. We selected the Cauchy distribution as a fitting function. The equation for a Cauchy distribution is shown in Equation 3.3.

$$C_{ff}(t;T) \approx C_{ff}(0;T)\frac{1}{\pi s\left(1+\left(\frac{t-\lambda}{s}\right)^2\right)} \tag{3.3}$$

This function was selected as it provides a reasonable approximation to the shape of the flux-flux correlation function for the reactions in our dataset and is defined by a single parameter, s. (λ sets the position of the peak of this distribution and is set to zero for all reactions). In this approach, the Cauchy distribution is scaled by the value of the flux-flux correlation function at time zero. Compared to direct prediction of a rate constant, this requires a minimum energy path to be found, reaction barrier to be fit, and the first time point of the flux-flux correlation function to be calculated. This is a major limitation of this approach. However, it does avoid the need to calculate the flux-flux correlation function at additional time points and could provide a method to achieve a lower-cost approximate quantum rate constant. The same input features as used for the direct prediction approach (Coulomb matrix with inverse temperature and reaction energy) were used to predict the distribution scale factor. The same method as for the direct fit was used to find an optimal GPR kernel function. In this case, the

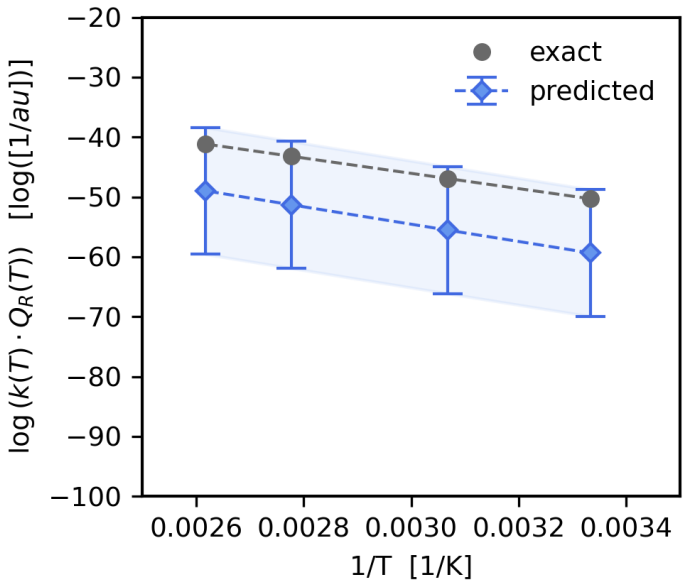


Figure 3.5: Predicted value of reaction rate constant product as a function of 1/T for the reaction of $CH_3^*+{}^*\rightarrow CH2^* + H^*$ on Pt(111), taken from the test set (Table 1 – reaction 8).

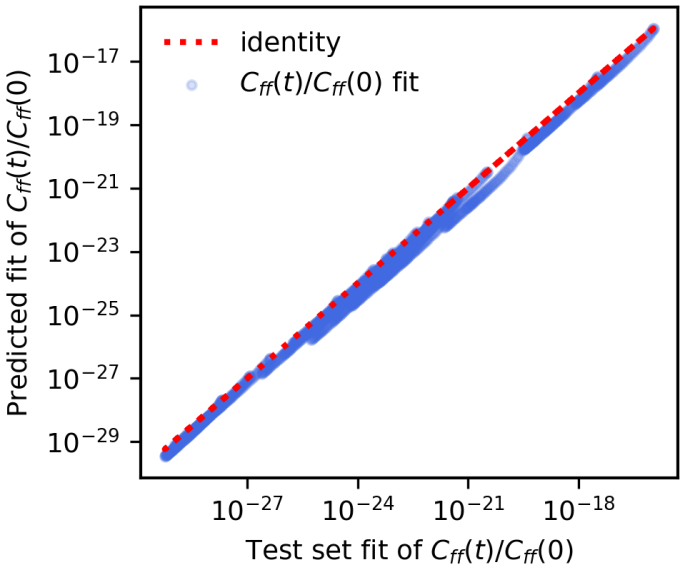


Figure 3.6: Plot of test set time series values from Cauchy fits of the flux– flux correlation function, normalized by its value at time zero (x-axis) respect to the time series values from predicted fits using GPRs (y- axis). Both axes are in log scale to emphasize data points. The predicted fits are in strong agreement with the exact Cauchy fits.

sum of a Matérn kernel with a rational quadratic kernel was selected. A regularization parameter $\alpha = 0.5$ was also used in this case.

We found that the Cauchy function approximation of the flux-flux correlation function with predicted scale parameters was able to closely approximate the ground-truth calculated flux-flux correlation function. Figure 3.6 is a parity plot of the Cauchy-

Table 3.2: The first two rows report the average mean absolute percent error on the computation of $C_{ff}(t;T)/C_{ff}(0;T)$ using a Cauchy distribution (Eq 3) with the GPR predicted scale parameter. The 'rx' subscript indicates that MAE errors are averaged over all reactions. The last two rows report the percent mean absolute error on the logarithm of the product of the reaction rate constant with the reactant partition function. Here the product is obtained by using trapezium integration of the Cauchy fit of the scaled flux-flux correlation function.

Computed quantity	Method	Error metric	% Error
$\frac{C_{ff}(t;T)}{C_{ff}(0;T)}$	Cauchy curve fit with GPR predicted scale parameter for temperature-wise split data	$\langle MAPE\left(C_{ff}^{exact}(t;T), C_{ff}^{GPR\ Cauchy}(t;T)\right)\rangle_{rx}$	9.77x10^{-1}
$\frac{C_{ff}(t;T)}{C_{ff}(0;T)}$	Cauchy curve fit with GPR predicted scale parameter for reaction-wise split data	$\langle MAPE\left(C_{ff}^{exact}(t;T), C_{ff}^{GPR\ Cauchy}(t;T)\right)\rangle_{rx}$	9.76x10^{-2}
$\log(k(T) \cdot Q_R(T))$	Trapezium integral of predicted $C_{ff}(t)$ Cauchy fit for temperature-wise split data	$MAPE(\log(k \cdot Q_R)^{exact}, \log(k \cdot Q_R)^{GPR\ Cauchy})$	6.78x10^{-1}
$\log(k(T) \cdot Q_R(T))$	Trapezium integral of predicted $C_{ff}(t)$ Cauchy fit for reaction-wise split data	$MAPE(\log(k \cdot Q_R)^{exact}, \log(k \cdot Q_R)^{GPR\ Cauchy})$	8.21x10^{-1}

function defined $C_{ff}(t)$ time values vs. the ground truth time values for the reaction wise split. Each set of points in this plot corresponds to one predicted $C_{ff}(t)$ curve. This plot shows that the 'predicted' $C_{ff}(t)$ values generally agree with the ground truth values. These predicted $C_{ff}(t)$ functions can then be integrated to yield a rate constant. The resulting MAPE is 0.82%, two orders of magnitude lower than it was for the direct prediction of rate constants for the reaction split. Figure 3.7 shows the result of predicting a Cauchy function fit to an exact $C_{ff}(t)$ function and the resulting agreement of the integrated rate constant with exact values for a reaction. Table 3.2 summarizes the metrics calculated for the $C_{ff}(t)$ curve fit parameter approach.

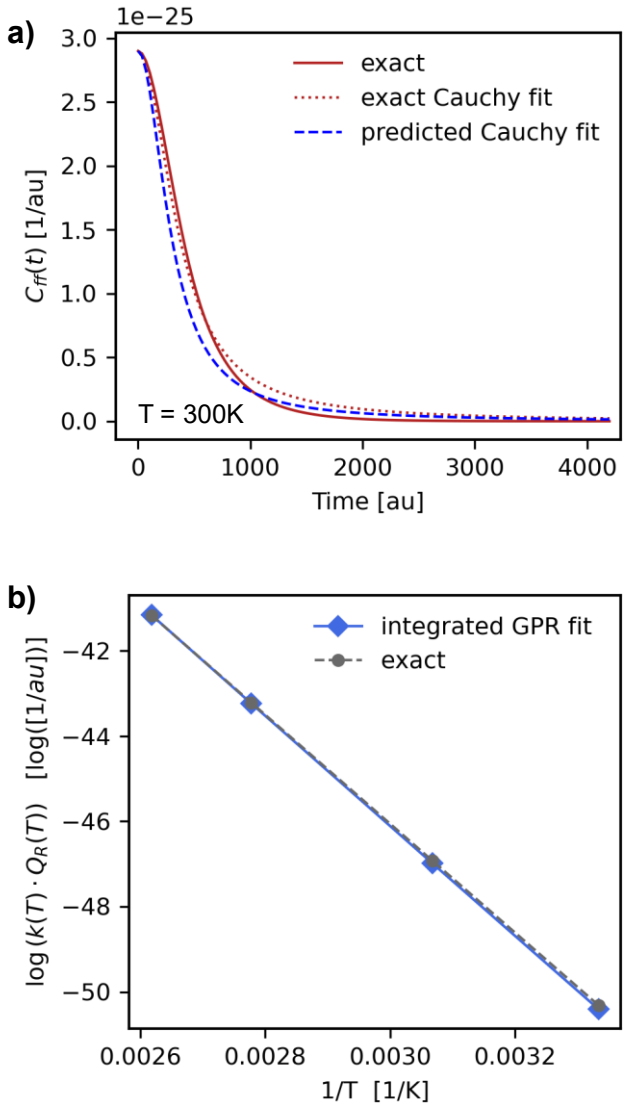


Figure 3.7: Panel a) Predicted Cauchy fit of the flux-flux correlation function (dashed blue line) for the reaction of $CH_3^* \rightarrow CH_2^* + H$ on Pt(111). (reaction 8 in table 3.1at 300K. Panel b) Comparison of the exact reaction rate constant products (grey circles and dashed line) with numerically integrated values obtained from the GPR Cauchy predicted scale parameter fit (blue diamonds and solid line). A large improvement on predicted test set reaction rate constant values respect to predicting the rate constant product directly is observed.

3.4 Discussion and Conclusions

This work evaluated the possibility of predicting quantum rate constants for catalyzed reactions. A tool to directly and accurately predict exact quantum rate constants that requires only reactant and product geometries to be calculated as input would be a potentially powerful aid in accelerating the screening and design of new reaction systems. This is a complex problem. The current work suggests that more effort would be required to address this task with a machine learning approach. While this work was able to predict reaction rate constants reasonably well for a given reaction at new temperatures, the utility of this approach over using some form of Arrhenius equation derived scaling term to account for temperature dependence is not immediately evident. The large error when predicting rate constants for new reactions suggests that this approach is not yet useful for this purpose. However, including more information about the reactive system in the form of a single time point of the flux-flux correlation function greatly improved the predictive accuracy of the model. This shows the importance of domain-specific insights when applying machine learning techniques to chemistry problems.

A valid critique of this work is that quantum rate constants are not particularly relevant for industrially meaningful catalytic reactions, where transition state theory probably does well enough. If this line of research were to be continued, I split the focus of this work into two research thrusts: one to develop methods for the prediction of quantum rate constants, focusing on simpler gas-phase reactions, and another on developing screening criteria for catalysts.

3.5 References

(1) Peters, B. *Reaction Rate Theory and Rare Events*; Elsevier, 2017.

(2) Hänggi, P.; Talkner, P.; Borkovec, M. Reaction-Rate Theory: Fifty Years after Kramers. *Rev. Mod. Phys.* **1990**, *62* (2), 251–341. https://doi.org/10.1103/RevModPhys.62.251.

(3) Mills, G.; Jacobsen, K. W. CHAPTER 16 Nudged Elastic Band Method for Nding Minimum En- Ergy Paths of Transitions. *Class. Quantum Dyn. Condens. Phase Simul.* **1997**.

(4) Miller, W. H.; Schwartz, S. D.; Tromp, J. W. Quantum Mechanical Rate Constants for Bimolecular Reactions. *J. Chem. Phys.* **1983**, *79* (10), 4889–4898. https://doi.org/10.1063/1.445581.

(5) Komp, E.; Valleau, S. Low-Cost Prediction of Molecular and Transition State Partition Functions via Machine Learning. *Chem. Sci.* **2022**, *13* (26), 7900–7906. https://doi.org/10.1039/D2SC01334G.

(6) Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P. O. Organic Reactivity from Mechanism to Machine Learning. *Nat. Rev. Chem.* **2021**, *5* (4), 240–255. https://doi.org/10.1038/s41570-021-00260-x.

(7) Lewis-Atwell, T.; Townsend, P. A.; Grayson, M. N. Machine Learning Activation Energies of Chemical Reactions. *WIREs Comput. Mol. Sci.* **2022**, *12* (4), e1593. https://doi.org/10.1002/wcms.1593.

(8) Komp, E.; Valleau, S. Machine Learning Quantum Reaction Rate Constants. *J. Phys. Chem. A* **2020**, *124* (41), 8607–8613.

(9) Houston, P. L.; Nandi, A.; Bowman, J. M. A Machine Learning Approach for Prediction of Rate Constants. *J. Phys. Chem. Lett.* **2019**, *10* (17), 5250–5258. https://doi.org/10.1021/acs.jpclett.9b01810.

(10) Nandi, A.; Bowman, J. M.; Houston, P. A Machine Learning Approach for Rate Constants. II. Clustering, Training, and Predictions for the O(3P) + HCl → OH + Cl Reaction. *J. Phys. Chem. A* **2020**, *124* (28), 5746–5755. https://doi.org/10.1021/acs.jpca.0c04348.

(11) Meyer, B.; Sawatlon, B.; Heinen, S.; Lilienfeld, O. A. von; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9* (35), 7069–7077. https://doi.org/10.1039/C8SC01949E.

(12) Schlexer Lamoureux, P.; Winther, K. T.; Garrido Torres, J. A.; Streibel, V.; Zhao, M.; Bajdich, M.; Abild-Pedersen, F.; Bligaard, T. Machine Learning for Computational Heterogeneous Catalysis. *ChemCatChem* **2019**, *11* (16), 3581–3601. https://doi.org/10.1002/cctc.201900595.

(13) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. I. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10* (3), 2260–2297. https://doi.org/10.1021/acscatal.9b04186.

(14) Chowdhury, A. J.; Yang, W.; Walker, E.; Mamun, O.; Heyden, A.; Terejanu, G. A. Prediction of Adsorption Energies for Chemical Species on Metal Catalyst Surfaces Using Machine Learning. *J. Phys. Chem. C* **2018**, *122* (49), 28142–28150. https://doi.org/10.1021/acs.jpcc.8b09284.

(15) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; Palizhati, A.; Sriram, A.; Wood, B.; Yoon, J.; Parikh, D.; Zitnick, C. L.; Ulissi, Z. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catal.* **2021**, *2020*, 6059–6072. https://doi.org/10.1021/acscatal.0c04525.

(16) Light, J. C.; Hamilton, I. P.; Lill, J. V. Generalized Discrete Variable Approximation in Quantum Mechanics. *J. Chem. Phys.* **1985**, *82* (3), 1400–1409. https://doi.org/10.1063/1.448462.
(17) Colbert, D. T.; Miller, W. H. A Novel Discrete Variable Representation for Quantum Mechanical Reactive Scattering via the Sâ□• matrix Kohn Method. *J. Chem. Phys.* **1992**, *96* (3), 1982–1991. https://doi.org/10.1063/1.462100.
(18) Johansson, F. Mpmath: A Python Library for Arbitrary-Precision Floating-Point Arithmetic (Version 0.18), December 2013. *URL Httpmpmath Org* **2013**.
(19) Johansson, F. Arb: Efficient Arbitrary-Precision Midpoint-Radius Interval Arithmetic. *IEEE Trans. Comput.* **2017**, *66* (8), 1281–1292. https://doi.org/10.1109/TC.2017.2690633.
(20) Winther, K. T.; Hoffmann, M. J.; Boes, J. R.; Mamun, O.; Bajdich, M.; Bligaard, T. Catalysis-Hub.Org, an Open Electronic Structure Database for Surface Reactions. *Sci. Data* **2019**, *6* (1), 1–10. https://doi.org/10.1038/s41597-019-0081-y.
(21) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108* (5), 058301. https://doi.org/10.1103/PhysRevLett.108.058301.
(22) Collins, C. R.; Gordon, G. J.; Von Lilienfeld, O. A.; Yaron, D. J. Constant Size Descriptors for Accurate Machine Learning Models of Molecular Properties. *J. Chem. Phys.* **2018**, *148* (24). https://doi.org/10.1063/1.5020441.
(23) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-Based Machine Learning Representations for Predicting the Enantioselectivity of Organocatalysts. *Chem. Sci.* **2021**, *12* (20), 6879–6889. https://doi.org/10.1039/d1sc00482d.
(24) Williams, C. K.; Rasmussen, C. E. *Gaussian Processes for Machine Learning*; MIT press Cambridge, MA, 2006; Vol. 2.

4 Highlighting Community Needs for Integrated Materials Data Management

This chapter also appears as a published perspective article,

Pelkie, Brenden G., and Lilo D. Pozzo. "The laboratory of Babel: highlighting community needs for integrated materials data management." Digital Discovery (2023).

There is a vast amount of noise around the topic of data management, especially as it pertains to automated and autonomous experimentation. This perspective was an attempt at sensemaking on this topic, with the potential goal of identifying a potential entry point for a more concrete contribution in this space. At this point, I have not made any meaningful changes to how I do science as a result of developing this perspective. However, the automated experimentation proposed in Chapters 6 and 7 presents an opportunity to implement many of the digital sample tracking methods described here.

4.1 Introduction

Automated experimentation methods are rapidly transitioning from being research subjects themselves to serving as indispensable tools in materials research. The availability of relatively affordable off the shelf hardware, the spread of data-hungry machine learning methods to materials science, and the ever-pressing need to accelerate the pace of materials innovation to meet a changing climate have all contributed to the adoption of automated experimental methods in our laboratories[1–3]. A dizzying array of recent research has contributed tools that enable this paradigm shift, including new open hardware platforms[4], optimization and experiment planning methods[5], and methods for sharing procedures across different laboratories[6–8]. This newfound ability to generate vast troves of experimental data comes as new machine learning and data science methods

build off that data[9,10], turning it into a first-class research product in itself[11]. However, comparably little effort has been expended on systems to collect, organize, store, and share this data effectively. As a research community, we've largely applied the existing data management methods and culture that developed around manual experimentation to automated workflows. This worked fine for initial demonstration projects and forays into the field, but as the field matures and continues producing valuable data with automated platforms, we need to adopt better data management practices. The data management path we are currently following reminds us of the 'Library of Babel' imagined by Jorge Luis Borges in the namesake short story[12]. This vast library contains an enumeration of all possible past and future human knowledge, with the catch that all of the valuable information is hidden amongst a sea of utter gibberish. In this library, generations of librarians are driven mad trying to find meaning in the expanse of text. If we continue advancing the state of automated experimentation without overhauling how we collect, organize, and share our data, we will find ourselves lost and isolated standing in an analogous 'Laboratory of Babel'. We'll know that the data we need to support our next materials innovation is out there somewhere, held in an unknown dataset in a distant repository, mixed into an expanse of useless untracked and unexplainable data. We believe that now is the right time to make technical and cultural shifts in how we handle this problem, so that we help future generations of researchers to develop an 'index' to effectively extract value from abundant 'gibberish' in materials data held in this global 'Laboratory'.

We envision a data management future where collecting and organizing experimental data and metadata is automated and effortless, data provenance is fully tracked, access to up to the minute data is enabled, and new community data sharing platforms make all experimental data findable. Having such a data management practice in place would streamline the development of

machine learning models for accelerating materials discovery, allow researchers to check the reproducibility of their results, and improve the quality and trustworthiness of the data we generate. An agreed upon system for data management would allow data related issues to fade into the background, allowing scientists to focus on the science that ultimately motivates us.

In this perspective we propose a holistic vision for how data might be collected, organized, and shared, focusing on laboratories that have adopted automated equipment. This vision is based on our research group's experience and challenges in setting up automated experimentation workflows from scratch, and from ideas proposed in the literature. We compare the current state of the field against this vision to elucidate opportunities for improvement, but also to highlight current successes. We intentionally discuss these topics at a high level and ground them in examples, rather than getting into details of technical implementations. While many recent works have explored various aspects of the data management problem[13,14], and several projects implement isolated items that are needed to make data management work[8,15], we believe that discussion about how disparate pieces of data management tooling fit together to form an integrated system is missing. Our hope is to start an accessible conversation around *what* our data management systems should do, not *how* they go about doing this. We believe this will provide

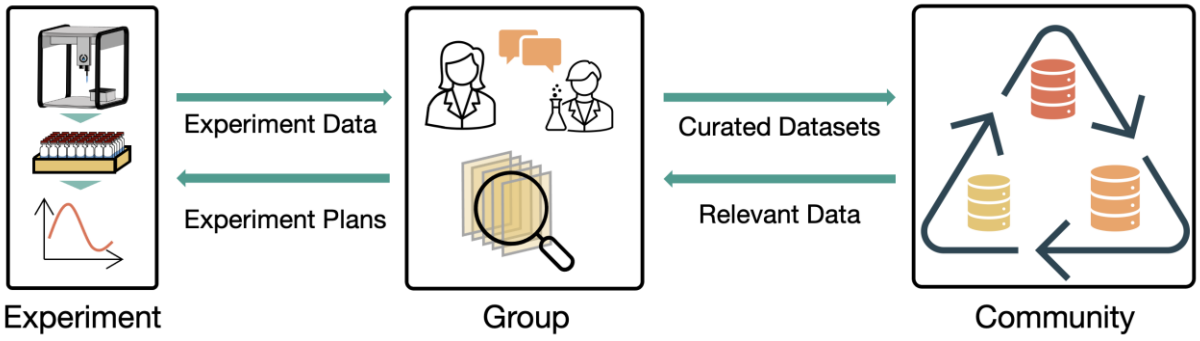


Figure 4.1: Managing materials research data is an inherently multi-scale endeavor. Data collected at the experiment-scale is organized and discussed at the group scale before being shared at the community scale. In turn, relevant data from community data shares can complement internal data at the group scale, serving to better inform experiment decision making.

useful guidance for future work on technical solutions to this problem and provide motivation for a cultural shift around integrated and holistic data management. While we hope our perspective helps guide future work on research data infrastructure, it should not replace formal customer development or user requirement scoping processes, such as those used in technology and entrepreneurship (e.g. NSF Innovation Corps)[16,17]. Developers of new data management tools should thoroughly evaluate the needs of the scientists who will be using them, so that these tools are a simple and valuable addition to research workflows.

Throughout our discussion, we talk about data management 'systems' or 'platforms' in somewhat abstract terms. Because we aim to discuss our vision in terms of capabilities rather than specific implementations, we avoid discussing how these aspects of our vision could or should be implemented. If specific implementation strategies are of interest to the reader, we recommend the tutorial perspective on databases for chemistry by Duke et al[18]. Any implementation of the ideas we discuss here would be intimately related to laboratory automation initiatives such as laboratory scheduling tools, remote equipment control capabilities, or full self-driving laboratories. The 'ideal' data management software implementation would likely include these capabilities, but these topics are out of scope for this perspective. To frame the field of experimental data management into a structured discussion, we break the task into three scales: experimental data collection, group data management, and community data sharing. Experimental data collection concerns the collection and management of data from individual experiments. Group data management concerns management of data within a laboratory, research group, collaboration, or organization. Community data sharing concerns the sharing of data among the broader community in a manner that makes it broadly accessible and reusable. Each scale has unique requirements and

challenges but relies on integration with the other two scales to realize its full potential, as illustrated in Figure 4.1.

4.2 Experiment-scale data management

In our three-part organization framework, the task of accurately and completely gathering experiment data and metadata is handled at the experiment level. Here we consider an 'experiment' to be the collection of preparation, processing, and characterization steps performed on a sample or group of samples prepared in the same campaign, and 'data' to be any recorded information associated with an experiment, including characterization and preparation information. This is the minimum granularity of data that provides context to enable downstream use of the data. For example, in a synthesis experiment with characterization by nuclear magnetic resonance (NMR), the NMR results on their own are meaningless without the context of how the sample was prepared. In our framework, experiment-scale data management tools are primarily concerned with correctly recording data, and may have limited support for enforcing quality of data, tracking the motivation behind an experiment, or otherwise providing context beyond the boundaries of an individual experiment. These tasks are mainly addressed at higher levels of our framework. An experiment-scale data management tool should ensure that any point of data recorded in a laboratory is surrounded by the context needed to interpret it. To maintain this standard, such a system must be capable of maintaining a complete record of provenance, processing, and characterization steps applied to any sample. For example, in a battery electrolyte screening study, a sample record should contain the source of the stock solutions and components a sample is made from, a detailed log of the liquid or solid handling steps used in the preparation of the sample, and the results of any characterization processes. To the greatest extent possible, the collection and organization of this data should be automated. Manual data transfer slows research and introduces opportunities

for error[19]. However, many laboratories may never be fully automated, and some experimental steps may always require human interaction. The manual entry of data and notes by researchers needs to be well supported. A graphical user interface could provide this support. Recent advances in natural language processing technologies such as GPT-4[20] may also enable new ways of recording data, such as a voice-assistant based lab notebook. Additionally, recording exploratory experiments without a preplanned structure should be straightforward. The intended steps in an experiment (e.g. target weights) should be captured for comparison to the actual executed experiment. This would allow for any deviations from the plan to be automatically flagged, aiding in the identification of systemic issues with an experiment or in the hardware that is used in its execution. To enable data-driven workflows like closed-loop optimization experiments, data should be made available in real time as it is collected. Once data from an experiment is collected, it should be stored in a flat structure to enable direct access to data attributes without parsing individual files. While data quality control is not a primary focus of an experimental data collection system in our vision, identifying 'bad' data as early as possible can avoid wasted time and effort. Thus, quality control should happen whenever possible. Users should be able to flag experiments and data points with known issues. Automatic real-time data validation could help catch mistakes as they happen to prevent wasted time and effort. However, these data collection systems should still log and store 'bad' data so that a complete record of experiments is obtained, and any data validation checks should pose a minimal interruption to the user. Implementation of a system that manages all these tasks needs to be simple to use so that adoption in the laboratory does not pose an undue burden on researchers. Such a system would streamline the collection and organization of data in the laboratory. We believe this would reduce errors associated with incorrect data recording and save researcher's time. Simpler data recording could facilitate the collection of data

from experiments that 'fail' in the eyes of researchers, contributing to a more balanced record of experimental data that would better support machine learning use cases[21]. This system would provide a single complete record of entire experiments, ensure data that might be relevant for future data science initiatives or repurposing of results is collected, and lay the foundation for the organization of data at the group and community scales. Figure 4.2 illustrates how this system could interact with laboratory processes and equipment.

Current data management workflows typically tend to rely on manual record organization and usually fail to fully integrate automation and digitalization. 15 years ago, Shankar found that record keeping in research tends to be left up to individual researchers[22]. In our experience, not much has changed since. Individuals are left to find a system they feel comfortable with. Experimental procedures and some results are usually stored in a researcher's paper laboratory

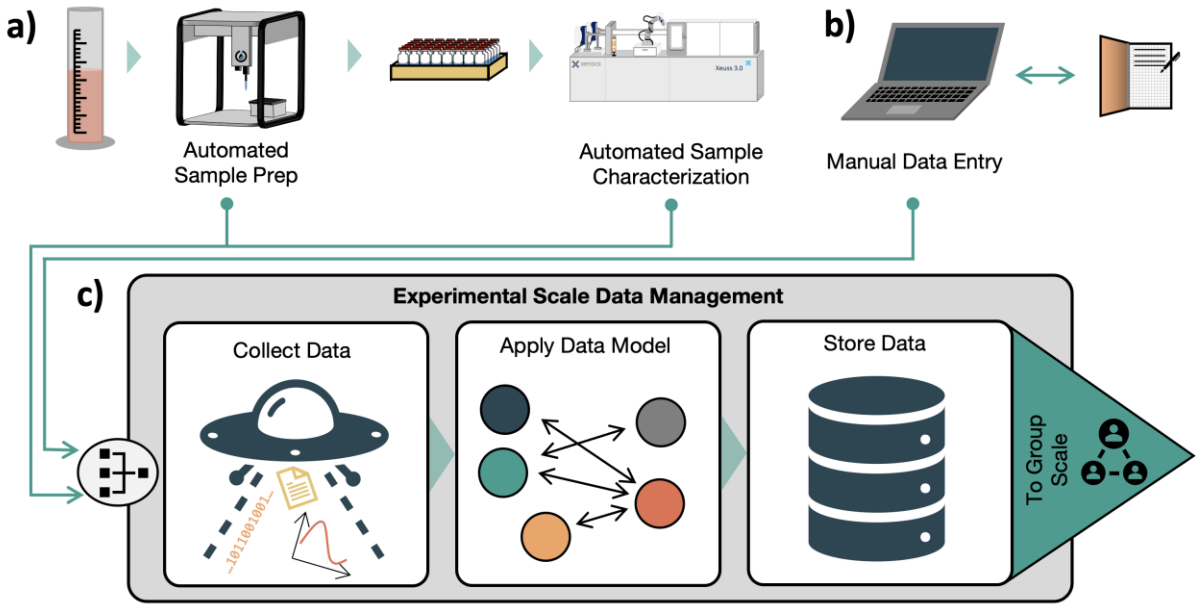


Figure 4.2: Data flow in an idealized experiment-scale data system. A) Samples are prepared and characterized in a physical workflow that involves a mix of automated and manual steps. Data is collected automatically when possible. B) Researchers can directly interact with the data pipeline and manually enter data through a user interface. C) The experimental data management system collects data, validates and organizes it into an appropriate data model, stores it, and makes it available to group-scale data management tools.

notebook, in varying levels of detail. Data files from characterization instruments are generally organized in a directory and file naming structure set up by the researcher, and are manually copied from instruments to a central location like the researcher's PC or a cloud storage provider. Log files recording processing steps on automated equipment are stored in a similar fashion, if they are retained. These files are linked together manually using a laboratory notebook as an index[23]. Records of provenance for results reside across an array of files, notes, and the researcher's memory. Extracting data from these results for future work tends to require bespoke file and data processing and can be an onerous task. There are many examples of projects and tools that address these limitations and implement aspects of our vision. A core component of an automated data management system is the ability to retrieve data from experimental equipment. A common approach for this task is to write a series of custom scripts that collect and aggregate data[24,25]. This approach is effective but can require extensive effort to implement and may be impacted by small changes in the laboratory environment, such as updated equipment configurations. Laboratory orchestration software packages and standards that enable automated experimentation already interact directly with equipment, which provides an opportunity to leverage existing capabilities to automate data capture. The Bluesky family of python packages allows users to specify and execute experiments and collect data by directly interfacing with hardware that uses the EPICS protocol as well as a few other hardware interface protocols[26]. This project was developed to standardize experiment specification and data collection from synchrotron light sources. It is widely used at the National Synchrotron Light Source ll (NSLS-II) as well as other US and international light sources[27]. The Standardization in Lab Automation (SiLA) standard and the Laboratory and Analytical Device Standard (LADS) are two competing standards that seek to provide a unified application programming interface (API) for interacting with lab equipment.

Both are primarily targeted at life science laboratories, and both build from existing network communication protocols to provide lab-specific features. SiLA has seen adoption among equipment manufacturers[28], and LADS is scheduled to be released in late 2023[29,30]. Collaborative development of competing lab equipment standards could lead to a set of widely adopted interfaces to equipment that each have specialized support for a particular use case. This would allow experimenters to pick the best tools for particular experimental tasks. For example, an automated flow-through nanoparticle synthesis experiment could communicate with a bank of syringe pumps over SiLA to control experimental conditions and a beamline with BlueSky to manage sample characterization. Each of these standards fulfills the needs of the application it is used for and alleviates the need for a single monolithic standard to handle every research task imaginable. However, development of many overlapping standards also has the potential to fracture the ecosystem for managing hardware and software, and preclude straightforward digital data management and communication. Care should be taken in standards development and adoption to avoid this.

Data collected from an experiment needs to be validated, organized, and stored. Data validation checks that collected data is in an expected format and an expected range. For example, a simple validation on the recorded mass of a sample could check first that the entry is numeric and not a text string, then that the value is within the measurable range of the balance used. This approach does not verify that the recorded number is correct but can catch major issues with data. As discussed above, invalid data should still be recorded but also flagged for review. Several data models for organizing experimental data have been proposed. A common theme among many of them is to represent each sample in an experiment as a graph of sample states connected by procedures. This is an intuitive way to represent an experiment: In the lab, a sample starts from

some feedstock materials (an initial state) before a number of processing steps are applied, each of which generates a new sample state. At states of interest, characterizations on the sample are performed. Figure 3.3 illustrates the application of a graph data model to an experimental procedure for one sample[31]. Projects or works that use this form of data model include the Event-Sourced Architecture for Materials Provenance (ESAMP)[32], Cript[15], and Citrine Informatics' GEMD data model[33]. Explicitly organizing experimental data in a format that follows this structure would make tracking the provenance of any piece of data straightforward: given a sample or measurement, the chain of processes and samples can be traced backwards to determine where the sample came from, or forwards to see how a future step turned out. This data model can be implemented in any database or storage format, and each of the mentioned projects has built its own version. A prerequisite to using such a data model is a schema to describe what data is stored and how it is related. Tools to parse data from its source and transform it into the data model are also needed to implement the data models we describe. Developing these items can be a significant challenge. A potential opportunity exists to establish a standardized representation for experiments that can be shared and reused between different software systems. Once infrastructure is in place to collect data from equipment and a conceptual model for organizing that data is agreed on, these must be implemented into a piece of usable software. The Experiment Specification, Capture, and Laboratory Automation Technology (ESCALATE) ontology and software package implements tools to specify experiment plans, record experimental execution, and link resulting files to samples in a database[34]. ESCALATE provides both a data model for organizing experimental data as well as a software implementation. It provides capabilities to specify experiment plans, record experimental executions, and manage files. Users interact with the software either through a graphical web interface or an API. Bespoke solutions in this space are also common. Several

organizations have discussed the design and implementation of custom in-house data management systems. When implementing their internal data system, The Joint Center for Artificial Photosynthesis developed a lightweight system of file types and scripts to track sample data and automate the recording of data when possible[35]. Data from this system eventually made it into the MEAD database (discussed below). The National Renewable Energy Lab (NREL) has implemented a similar file and script based workflow that enables tracking sample preparation and characterization data by having users load data files into a centralized warehouse[25].

Electronic lab notebooks (ELNs) also fit into this section of our framework. Traditional ELNs sought to entirely replace paper lab notebooks with a direct translation to a digital document. While this provides major improvements for data searchability, shareability, and security, it does

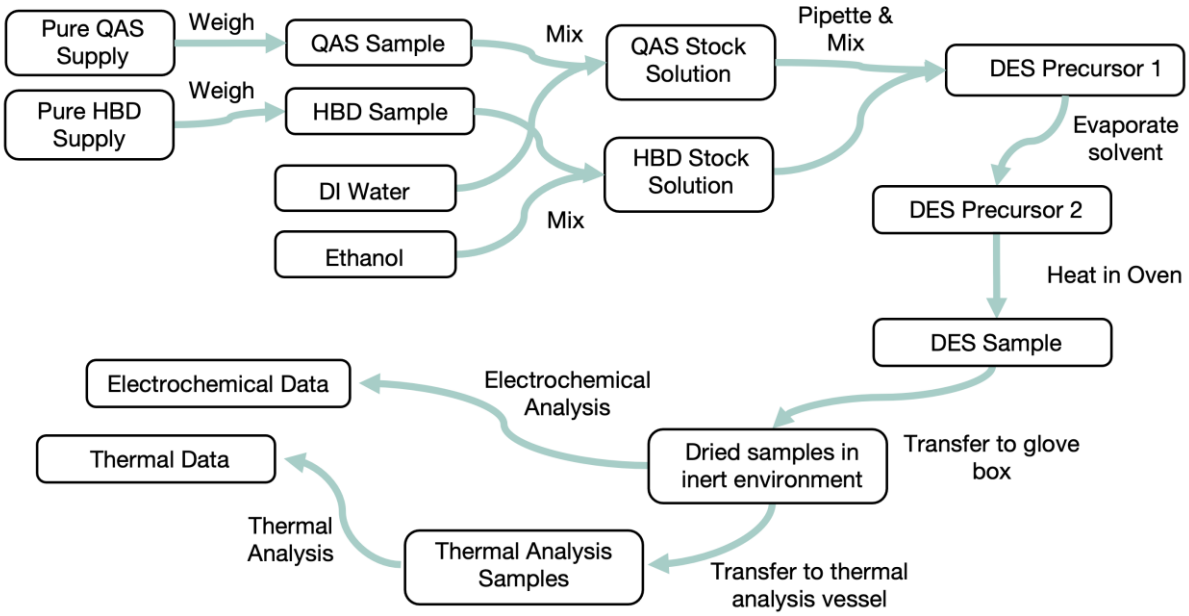


Figure 4.3: One of many possible examples of a graph data model applied to experimental data. Here nodes (gray blocks) represent sample states or outcomes of measurements, and edges (blue arrows) represent the processes connecting those states and their data. In a deep eutectic solvent (DES) screening study, samples start from solid supplies of quaternary ammonium salts (QAS) or hydrogen bond donors (HBD), are synthesized in a multistep process involving several sample states, and are thermally and electrochemically characterized. Tracking every process of the sample synthesis along with the data they generate provides a complete record of sample provenance. Applying a graph data model makes working with this complex web of data tractable.

not enable the data collection infrastructure we envision. More modern ELNs incorporate more extensive data management features, like recording data directly from instruments or supporting inline data analysis[36,37]. Modern ELNs can also interact with a vendor's laboratory information management system (LIMS) product to enable organization of data across experiments and laboratories. LIMS are discussed below. Given the wide selection of ELN systems tailored for diverse types of lab work, the lack of adoption in academic laboratories[38] raises questions. While a full exploration of issues associated with the limited adoption of ELNs is beyond the scope of this work, high cost, significant effort, low community expectations, and traditionalist attitudes are contributing factors[23].

4.3 Group Scale Data Management

In our vision, the goal of a group scale data management platform is to organize individual experiments across research projects and other group objectives. In this discussion a group is a collection of researchers actively collaborating on a project, such as an academic research group, a department/unit, or multiple collaborators spanning different institutions. Elements of editorial discretion are also introduced at this level. Group members will know and trust each other, understand the context around the collection of

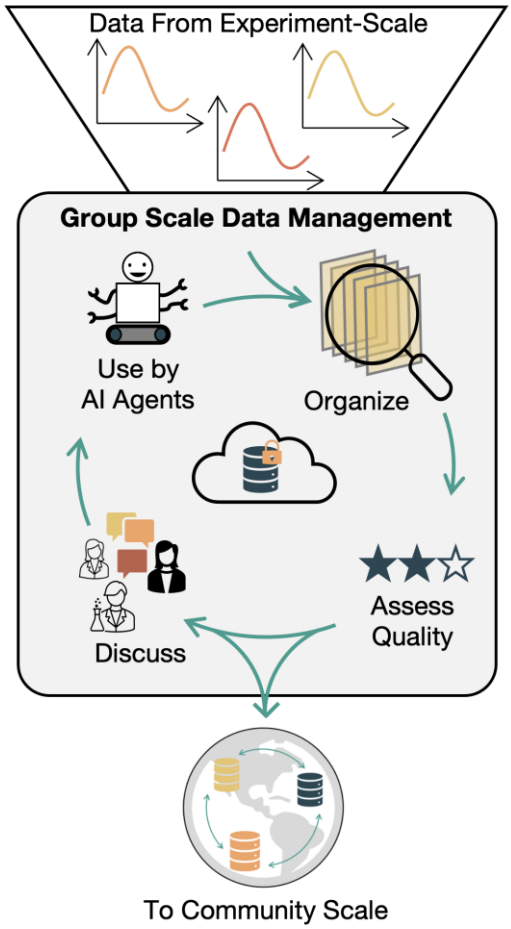


Figure 4.4: A group scale data management system supports the organization of data collected from an experiment-scale system into research projects, searching across all a group's data, scientific discussion around data quality and results, and inclusion in AI decision making. These capabilities are enabled by a secure centralized data location. Sharing data to a community scale is straightforward with this infrastructure in place.

data, and agree on how it will be used. They will also be involved in discussions to assess the quality of collected data prior to broadly disseminating it or reporting major outcomes in the scientific literature. Effective group scale data management builds from and complements strong experimental-scale data collection and management. In practice, the distinction between experimental and group scales is 'thin' and may not be apparent in real world data management systems. However, the disparate goals of these two data management scales merit separate treatments.

To support the goals of group data management, tools should enable linking data from related experiments and samples. In turn, groups of experiments should be linkable into project campaigns. For example, in an automated sample synthesis process, multiple experiment campaigns might be run, with multiple replicates of a material in each campaign. Samples that are replicates of one material should be linked to that material, as well as to the campaign where they were generated. In our examples, multiple high-throughput synthesis campaigns that are all part of the same project should be linked and accessible as one combined project. Data from different but loosely related experiments should also be grouped together. Our imaginary 'user' should be able to view all the work done with a particular sample precursor, grouped across different sample preparations and experiments. Computational results should be included in this grouping to broaden the scope of available information. Having all this data in one place will enable anyone involved in a project, be they researchers, supervisors, or artificial intelligence agents, to have access to up-to date versions of data which will enable faster and smarter decision making around future experimental plans. It is important that editorial tasks and human data interpretation be supported. Capturing the motivation and intent behind running an experiment can give context to a group of experiments in a project. Low quality or compromised data could still be relevant to a

project at this scale, but quality issues should be flagged. Tracking the quality of data should be supported. This could involve automated or human data review. Backups of data should be automated so data isn't lost by accident, and modifications to data should be tracked and version controlled so data isn't manipulated by malice or by mistake[39]. Preparation of data for downstream uses, like machine learning initiatives or for 'export' to community level databases, should be straightforward and automatable to prevent errors and to remove data processing bottlenecks.

As with experimental data collection, a common approach to managing data at the group scale is to design a custom system. This can either be a software system, or a manual workflow. Building a custom software system requires the expertise and effort to set up physical and digital data management infrastructure but can yield a system that better complements a laboratory's experimental workflows. The internal systems developed at NREL and JCAP, both discussed above, also provide group data management capabilities. They allow for linking individual experiments into campaigns, sorting experiments by criteria like experimental method, and sharing up to date data among a group[25,35]. These one-off systems can work well for groups with the resources to fully implement them but are out of reach for most researchers. Many laboratories have pieced together a manually updated data management system centered around a commercial cloud storage provider such as Google Drive or Dropbox. These platforms are attractive to use as they facilitate simple data sharing amongst laboratory members, provide a degree of data versioning and backup, and are often provided via an institutional license making them free to use. Before relying on third party cloud storage solutions, researchers need to consider the appropriateness of a particular offering for the sensitivity of the data they work with. As an example, storing protected health information on a consumer Google Drive account would violate HIPAA[40]. The vendors of these products may also choose to make changes to either the product

itself or the terms of use of the product that can be disruptive to how they are used in a laboratory or group, forcing researchers to make disruptive changes to their workflows.

Commercially available systems for managing data at the group scale are commonly referred to as Laboratory Information Management Systems (LIMS). Traditional LIMS systems provide sample tracking and provenance management capabilities, a centralized store of experimental data and other information, some level of integration with instruments for data collection, capabilities to manage experimental workflows, and a user interface. Some systems integrate with automated equipment, providing capabilities we classify as experiment management. These systems usually work in concert with a vendor's ELN solution. Like ELNs, LIMS are available from a robust array of vendors[37,41–43], and at least one open source option is available[44]. Many available LIMS systems are designed for life science or commercial laboratories, but there are options targeted at materials science research. Dotmatics offers a platform with LIMS capabilities that is designed for materials science and chemistry laboratories[45]. Citrine informatics' data management system is based on their GEMD data model (discussed above) and targeted at materials laboratories. This system has the benefit that data is extracted from files and stored directly in their data model, which makes the data more searchable and usable[46]. As with ELNs, adoption in academic laboratories is limited, likely for the same reasons (e.g. cost and/or complexity).

We believe that there is a notable lack of open-source software options that provide the capabilities we envision for laboratory data management, especially in materials science fields. An open-source software tool, built off a robust experiment capture infrastructure as described above, would make the group data management we envision accessible and customizable for a wide array of groups. An important criterion for this software will be its useability and ease of adoption, in

addition to how it handles technical data management tasks. Adoption of a group data infrastructure with the attributes we envision that is integrated with experimental data collection tools would revolutionize data management in most academic laboratories, and in our opinion would be a worthwhile investment on its own. However, even greater benefits can be realized by using this infrastructure to share research data with the broader community.

4.4 Community scale data management

Community data sharing has always been at the core of scientific communication. Traditionally, data has been shared through plots and tables in manuscripts, with important context embedded in manuscript text. While recent efforts make it possible to extract information from these documents using natural language processing methods[47–50], traditional publications are not an efficient way of transmitting data. Fortunately, a slow shift towards more open data sharing is underway. The importance of accessible data sharing has gained broad acceptance. References to making data FAIR (originally defined as findable, accessible, interoperable, and repurpose-able)[51] are common in the literature[50,52–54]. Well-designed community data sharing practices enable the aggregation and dissemination of data from multiple research groups and heterogeneous projects in a unified fashion, allowing existing data to drive unforeseen future works. Ultimately the goal of openly and accessibly sharing research data is to make it reusable in future research. Effective data sharing can enable new machine learning initiatives, make comparing new results to existing values simple, or prevent the unnecessary reproduction of existing work. What exactly this future reuse looks like is difficult to define, which is part of what makes establishing robust and useful data sharing infrastructure difficult. Thus, community data sharing initiatives should be built to be as generally useful as possible, rather than optimizing for a particular downstream use case. To support data findability, data should be stored in curated, focused databases. The domain scope for

these databases should be tuned so that relevant materials or experiments are stored together, without fractioning the ecosystem into hyper-specific datasets. Choosing a level of specificity for a database is an important consideration that impacts how data is likely to be re-used in the future. Specialized databases may make re-use simple for new applications that are similar to the original use of the data. However, being too specific can limit community contributions and engagement that is needed to sustain a database after initial support runs out, and can make it difficult to find databases that contain the desired information. Conversely, databases that are too broad in scope might not support the level of detail needed for some downstream use cases. In an example from computational materials science, the Catalysis Hub surface reactions database provides a home for relaxation and chemisorption energies of reactants on catalyst surfaces obtained from electronic structure calculations[55]. This focus strikes a balance that makes it specific enough to be useful, but broad enough to have over 100,000 entries. A critical aspect of making data accessible is ensuring the long-term existence of databases and their accessibility over the internet. Shared data is an important part of the scientific record, so community databases should be administered in a way that guarantees long term availability. To make data interoperable, data needs to be accessible as database records rather than as groups of files. This makes searching and filtering data using standard query tools like SQL or SPARQL possible. However, existing community and domain specific standards should be respected where applicable. For example, the small-angle scattering (SAXS, SANS) community has standardized beamline data around the CANSAS standard[56]. Any database dealing in a particular type of data needs to support data retrieval in the agreed standard to facilitate use with domain-specific tools. Data format exchange tools like Tiled[57] can facilitate data access in preferred formats. Human readability should be enabled through well designed user interfaces. This would make exploratory data analysis easier and enable access by users without a

coding background. To make data reusable, it is necessary to maintain a minimum standard of quality and completeness for data. Low quality data can't be tolerated at this scale because users lack the context to critically evaluate data quality and experimental nuances that is present at smaller scales. Peer-reviewing data submissions to these databases as part of journal manuscript submissions could help curate this data quality. To enable reproducibility of data (another interpretation of the 'R' in FAIR), information about how a sample was prepared and characterized should be shared in a standardized format. Intermediate data points should be shared alongside final results when possible, to support use cases that rely on them. Figure 4.5 illustrates the process for sharing data in this envisioned system. Once one establishes a new community data repository that satisfies all the points of our vision, attention needs to be turned to the social aspects of managing a community data resource. Use of these databases needs to be incentivized for researchers. Mechanisms for encouraging use might involve mandates from publishers or funding agencies, the generation of a citable digital object identifier (DOI) and other mechanisms for claiming credit for data submissions, and a seamless interface with lab data management systems to take the pain out of publishing data. When broadly sharing data as we advocate for, export controls may need to be considered. Current US export controls don't generally restrict the public sharing of basic research outputs[58]. However the US government has recently announced efforts to limit foreign access to US technology[59] and influence over research[60], so future controls on research sharing may become more restrictive. Researchers outside of the US will need to consider how their government's export control policies might impact how they share research data. The current state of experimental data sharing in materials science has been described as 'critical'[61]. A report on the materials genome initiative by the (US) National Science Foundation points to data sharing as a bottleneck in the materials innovation process, and suggests that national agencies

should establish data sharing infrastructure like what we envision above[62]. We think this is an appropriately dire assessment of the current situation, but there is a lot of work in this space that gives cause for hope. Perhaps the most widespread form of data sharing currently is through journal article supplementary information and general data repositories like Zenodo[63] and Dryad[64]. Repositories allow users to upload their data files along with a description of what the data contains, and then generate a unique identifier for that data. Materials focused repositories include the Materials Data Facility[65] and Citrination[66]. Sharing data in this form at least makes it available but doesn't address the spirit of the FAIR philosophy. Assembling datasets from files shared as supplementary information, or via a repository, requires searching the literature for relevant publications and datasets, checking that they contain the desired data, then manually collecting and parsing files that likely use unique (i.e. non-standardized) formatting. Poor metadata annotation often means one needs to read the original journal publication to understand the data. Curated databases provide solutions to these issues. These databases aggregate data from multiple experiments and projects into one database with a specific domain focus. Examples in the materials science field are limited. The NREL High Throughput Experimental Materials (HTEM) database contains records for over 82,000 unique samples of inorganic thin film materials . This database is populated using the NREL internal research data infrastructure described in the experimental data section and includes characterization data as well as some sample synthesis metadata. The Materials Experiment and Analysis Database (MEAD) was hosted by the Joint Center for Artificial Photosynthesis and populated by their internal experimental data collection system. Unfortunately,

this database was not accessible at the time of writing, which highlights the need to plan for the long-term stability and availability of community data resources. The inorganic crystal structure database (ICSD) provides a database of inorganic crystal structures compiled from literature[68]. This database is neither open to community contributions, nor open access. Outside of experimental materials science, many more examples show the promise of shared community data sharing. The Materials Project is a widely known community database for computational materials data. It contains properties for over 154,000 materials[69] and has over 200,000 registered users[70]. This database makes computed properties of materials available via both an easy to navigate web page and an API. Arguably the best success story of community data sharing is the Protein Data Bank (PDB), a database of protein structures. This database has lasted for over 50 years and has grown to over 30,000 data contributors and over a million site views per year. This database provides curated, validated data in a consistent format. It has become a core part of research in its

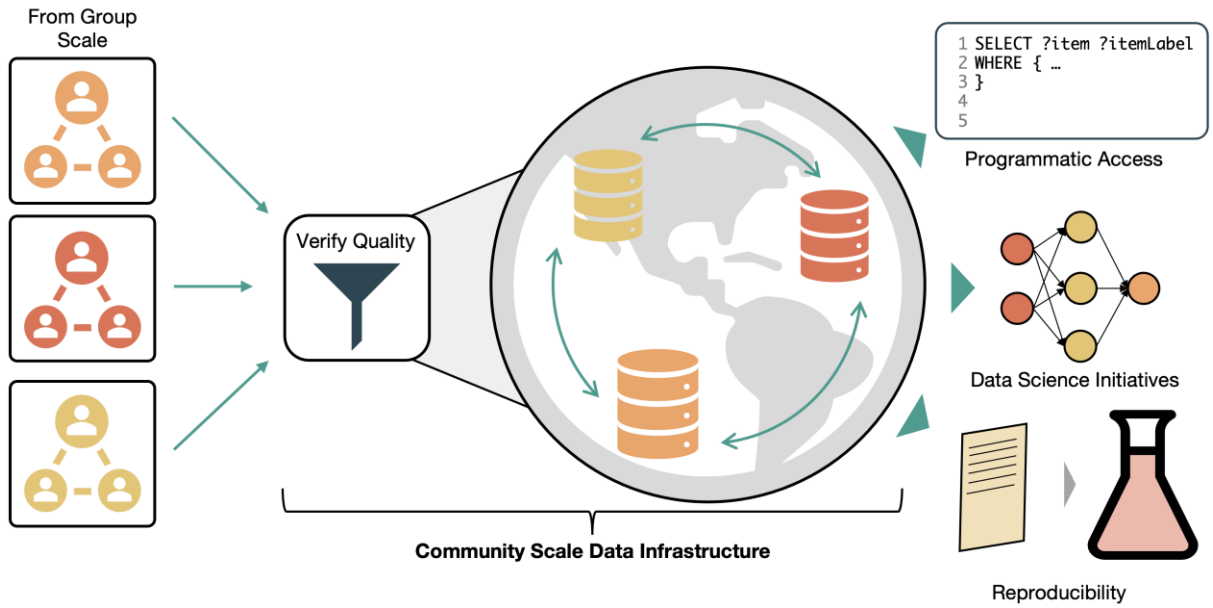


Figure 4.5: In a community data sharing ecosystem, data from individual research groups is curated and validated before being added to a network of domain databases. This makes data accessible with programmatic tools like SPARQL, makes reuse (e.g. in data science initiatives) feasible, and enables straightforward reproducibility, among other benefits.

field, as submitting a new protein structure for publication practically requires submission of that structure to the PDB[71]. Access to the PDB has also enabled groundbreaking advancements such as the accurate prediction of protein folding and *de-novo* structures based on sequence, with machine learning models[9,72].

Several projects and organizations are working toward new data sharing platforms that provide many of the capabilities we described above. The FAIRmat consortium is a German initiative to realize many of the goals for community data sharing that we describe here. This project aims to build a series of domain specific data repositories, following similar criteria as we propose to establish 'as few as possible but as many are needed' to support diverse needs of different fields. Their proposal, which is to create a federated network of databases with centrally searchable metadata, has the potential to enable domain specific databases that are still findable and reusable for applications in other contexts. This project has also considered the needs of experimental and laboratory data management infrastructure to feed these community repositories[73]. The development of knowledge graphs to store and structure experimental data is a promising approach to implanting data sharing infrastructure. Knowledge graphs structure information as a connected graph, with data points and entities represented as graph nodes and properties or relationships represented as edges[74]. This is similar to the graph data models discussed in the experimental data collection section. When coupled with a standardized definition of properties and relationships, known as an ontology, knowledge graphs promise enhanced interoperability between databases and flexibility in defining data schemas. Knowledge graphs can be queried using standard query languages such as SPARQL. Bai et. al. envision a knowledge-graph based approach to managing data, and more broadly, laboratory automation. In their vision, computational agents maintain the state of the knowledge graph and use it to drive closed-loop

experimentation[14]. The Mat-o-lab initiative seeks to develop domain-specific ontologies for materials research, and corresponding knowledge graph representations of data[54]. This initiative envisions a network of knowledge graphs using compatible ontologies to represent data and enable re-use across projects. The OPTIMADE consortium aims to solve the data interoperability challenge by providing an API specification for materials databases[75]. This specification has been adopted by several databases including the Materials Project[76]. The wide range of solutions under development to address the community data sharing problem shows that this issue is well understood by the community and makes us hopeful that truly FAIR data sharing is near.

4.5 Common obstacles and recommendations

As shown above, data management is an active area of research, and the need for the capabilities we describe herein are recognized by the community. So, what motivated this perspective? Most of these initiatives are carried out as individual efforts to solve a small subset of the problems facing the field. While this bottom-up approach is leading to innovative and exciting tools, these individual efforts generally don't integrate with other tools and do little to reduce the fractured nature of the data management field. The NREL internal research data infrastructure team recognized these limitations, and called for a top-down approach to design data management infrastructure with a holistic vision in mind[25]. While this is a noble goal, building an entire research data infrastructure from scratch as one project is a major undertaking. Further, one organization is likely unable to anticipate and build for the diverse use cases and requirements of such a tool. Rather than leave the development of future tools to one entity, we believe that future development should continue to be undertaken by diverse community projects, but with a stronger eye towards how projects will inter-operate to enable the seamless data management system we envision. To enable this interoperability between different tools and software, we should define and adopt

standards for how data is represented as a community. As we discussed in the experimental data collection section, interfacing lab equipment with any data collection software is a challenge due to vendor-specific protocols and data formats. Developing a standardized API for communicating with laboratory equipment would resolve this challenge. A standardized data model for representing collected experimental data would enable greater interoperability between competing solutions. Agreeing on a common implementation and specification of a graph data model for sharing between data management tools would enable easier data sharing and make data more reusable. Alongside data, information needed to reproduce experiments needs to be shared. Leroy Cronin's χ-DL project seeks to develop a 'compilable' language for specifying synthesis steps[6,7]. A standard means to specifying experimental procedures promises to make experiments reproducible on heterogeneous automated experimentation platforms.

We noted that there is an acute lack of software tools to organize and store data at the laboratory or group scale. This gap in the data management infrastructure compounds shortfalls in experimental data collection and community data sharing. Without an effective means to organize and use large amounts of experimental data and metadata, little motivation exists to expend effort to collect data beyond what is immediately needed. And if collected data is scattered across a wide range of files and locations, preparing data for submission to a community database can be a herculean task. A robust, user-friendly, and generalizable implementation of a laboratory data management system would bridge the gap between the two other levels of data management, encouraging wider experimental data collection and facilitating rapid dissemination of data to community databases. An open-source software tool for managing laboratory data that implements the data management standards we describe should be created, either by establishing a new project or extending an existing one. This tool should be generalizable to different laboratory

environments, but customizable to provide the specificity and workflow efficiencies needed for any given laboratory. This tool should provide both a graphical user interface for easy data management, as well as programmatic access via (at least) a python library so that use of the tool can be integrated into existing data-generating processes controlled by python scripts. In practice, such a tool could also fulfill our vision for experimental data collection and management as the two are closely related. We believe the availability of an open-source tool (as opposed to a proprietary one) in this space is critical. While proprietary software can solve many data management problems, it raises issues with vendor lock-in, laboratory equipment support, and custom use cases. Automation in research laboratories involves prototyping new hardware and workflows, so it is virtually impossible for a commercial vendor to envision and support all the possible use cases. A community-driven open-source tool could be more responsive to new applications, and individual laboratories would have a fair shot at adapting an open-source tool to creative new workflows or use cases. An open-source tool would also remove financial hurdles to using a data management infrastructure.

At present, implementing data management infrastructure requires a high level of proficiency in software engineering and systems administration. Most academic laboratories do not have access to personnel with these skills. Tasking grad students with learning them is problematic as it adds to their already full plates. Expecting laboratories to establish their own research data infrastructure[13,18] will lead to low adoption rates, poor quality systems, and frazzled graduate students. We do not think it is fair to ask this of junior researchers. Community organizations should establish a leadership stance on this front and guide the development of both software tools for use at the experimental and group scale, and community databases for data sharing. These organizations could be federal agencies like the NSF or NIST, existing community organizations

like the MRS, or newly formed consortia with data management as their express goal. For their part, the NSF has recognized the need for a leader in this space, and the role they could play[62]. The FAIRmat initiative has been referenced as a model for what that role could look like[62].

A common theme across all areas of data management is the need for broad community buy-in. Building the perfect data management tool won't make a difference if nobody chooses to use it. Getting researchers to move beyond paying lip service to data as a first-class research product is a major barrier to our envisioned data management future. Funding agencies should make effective data sharing a core project requirement by seriously considering data management plans in grant application evaluations and following up to ensure they are followed. The extra overhead this creates for researchers should also be recognized and accommodated. Publishers could require data sharing to community databases as a condition of publication[77]. Mechanisms for giving credit to researchers for their data contributions would provide further incentives. Success stories shared by early adopters of time saved, mistakes avoided, and discoveries enabled by new data management tools and practices would also help show that taking data management seriously is a worthwhile endeavor.

4.6 Conclusion

Materials science and chemistry are entering a new phase where automated and autonomous experimentation methods will multiply the capabilities of researchers and make new data driven research paradigms commonplace. To make the most out of these new research paradigms, the community needs to overhaul how data is handled at all scales of the research process. However, we are concerned that the moment to make these changes is being missed. Rather than enabling new research approaches based on easy access to data, new automated methods will lead to an incomprehensible data landscape and siloed research projects. We believe that effective data

management practices for this new era must consider the entire data lifecycle across scales from individual experiments to broad community dissemination. We envision one possible set of capabilities and norms that could contribute to such a multiscale data management system. However, this is certainly not the only vision that should be considered. Our hope is that this perspective encourages more researchers to participate in the discussion around data management by making it accessible and by presenting a tantalizing potential future where data management 'just works' and can fade into the background. We are excited to hear alternative visions from other researchers and to collaborate towards a future that embraces digital data management methods and prevents us from getting lost in the 'Laboratory of Babel'.

4.7 References

(1) Green, M. L.; Choi, C. L.; Hattrick-Simpers, J. R.; Joshi, A. M.; Takeuchi, I.; Barron, S. C.; Campo, E.; Chiang, T.; Empedocles, S.; Gregoire, J. M.; Kusne, A. G.; Martin, J.; Mehta, A.; Persson, K.; Trautt, Z.; Van Duren, J.; Zakutayev, A. Fulfilling the Promise of the Materials Genome Initiative with High-Throughput Experimental Methodologies. *Appl. Phys. Rev.* **2017**, *4* (1), 011105. https://doi.org/10.1063/1.4977487.

(2) Ashraf, C.; Joshi, N.; Beck, D. A. C.; Pfaendtner, J. Data Science in Chemical Engineering: Applications to Molecular Science. *Annu. Rev. Chem. Biomol. Eng.* **2021**, *12* (1), 15–37. https://doi.org/10.1146/annurev-chembioeng-101220-102232.

(3) Seifrid, M.; Hattrick-Simpers, J.; Aspuru-Guzik, A.; Kalil, T.; Cranford, S. Reaching Critical MASS: Crowdsourcing Designs for the next Generation of Materials Acceleration Platforms. *Matter* **2022**, *5* (7), 1972–1976. https://doi.org/10.1016/j.matt.2022.05.035.

(4) *PhasIR: An Instrumentation and Analysis Software for High-throughput Phase Transition Temperature Measurements*. https://openhardware.metajnl.com/articles/10.5334/joh.39/ (accessed 2022-11-17).

(5) Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Aspuru-Guzik, A. Gryffin: An Algorithm for Bayesian Optimization of Categorical Variables Informed by Expert Knowledge. *Appl. Phys. Rev.* **2021**, *8* (3), 031406. https://doi.org/10.1063/5.0048164.

(6) Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; Cronin, L. Organic Synthesis in a Modular Robotic System Driven by a Chemical Programming Language. *Science* **2019**, *363* (6423), eaav2211. https://doi.org/10.1126/science.aav2211.

(7) Mehr, S. H. M.; Craven, M.; Leonov, A. I.; Keenan, G.; Cronin, L. A Universal System for Digitization and Automatic Execution of the Chemical Synthesis Literature. *Science* **2020**, *370* (6512), 101–108. https://doi.org/10.1126/science.abc2986.

(8) Leong, C. J.; Low, K. Y. A.; Recatala-Gomez, J.; Quijano Velasco, P.; Vissol-Gaudin, E.; Tan, J. D.; Ramalingam, B.; I Made, R.; Pethe, S. D.; Sebastian, S.; Lim, Y.-F.; Khoo, Z. H. J.; Bai, Y.; Cheng, J. J. W.; Hippalgaonkar, K. An Object-Oriented Framework to Enable Workflow Evolution across Materials Acceleration Platforms. *Matter* **2022**, *5* (10), 3124–3134. https://doi.org/10.1016/j.matt.2022.08.017.

(9) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373* (6557), 871–876. https://doi.org/10.1126/science.abj8754.

(10) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; Palizhati, A.; Sriram, A.; Wood, B.; Yoon, J.; Parikh, D.; Zitnick, C. L.; Ulissi, Z. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catal.* **2021**, *2020*, 6059–6072. https://doi.org/10.1021/acscatal.0c04525.

(11) Beck, D. A.; Carothers, J. M.; Subramanian, V. R.; Pfaendtner, J. *Data Science: Accelerating Innovation and Discovery in Chemical Engineering*; 2016; Vol. 62, pp 1402–1416.

(12) Borges, J. L. *The Library of Babel*; 1941.
(13) Medina, J.; Ziaullah, A. W.; Park, H.; Castelli, I. E.; Shaon, A.; Bensmail, H.; El-Mellouhi, F. Accelerating the Adoption of Research Data Management Strategies. *Matter* **2022**, *5* (11), 3614–3642. https://doi.org/10.1016/j.matt.2022.10.007.
(14) Bai, J.; Cao, L.; Mosbach, S.; Akroyd, J.; Lapkin, A. A.; Kraft, M. From Platform to Knowledge Graph: Evolution of Laboratory Automation. *JACS Au* **2022**, *2* (2), 292–309. https://doi.org/10.1021/jacsau.1c00438.
(15) Walsh, D.; Zou, W.; Schneider, L.; Mello, R.; Deagen, M.; Mysona, J.; Lin, T.-S.; Pablo, J. de; Jensen, K.; Audus, D.; Olsen, B. CRIPT: A Scalable Polymer Material Data Structure. **2022**. https://doi.org/10.26434/chemrxiv-2022-xpz37.
(16) Bosman, L.; Garcia-Bravo, J. Lessons Learned: Research Benefits and Beyond Associated with Participating in the NSF I-Corps™ Customer Discovery Program. *Technol. Innov.* **2021**, *22* (1), 41–54. https://doi.org/10.21300/21.4.2021.5.
(17) Nnakwe, C. C.; Cooch, N.; Huang-Saad, A. Investing in Academic Technology Innovation and Entrepreneurship: Moving Beyond Research Funding through the NSF I-CORPS™ Program. *Technol. Innov.* **2018**, *19* (4), 773–786. https://doi.org/10.21300/19.4.2018.773.
(18) Duke, R.; Bhat, V.; Risko, C. Data Storage Architectures to Accelerate Chemical Discovery: Data Accessibility for Individual Laboratories and the Community. *Chem. Sci.* **2022**. https://doi.org/10.1039/D2SC05142G.
(19) Seidel, S.; Cruz-Bournazou, M. N.; Groß, S.; Schollmeyer, J. K.; Kurreck, A.; Krauss, S.; Neubauer, P. A Comprehensive IT Infrastructure for an Enzymatic Product Development in a Digitalized Biotechnological Laboratory. In *Smart Biolabs of the Future*; Beutel, S., Lenk, F., Eds.; Advances in Biochemical Engineering/Biotechnology; Springer International Publishing: Cham, 2022; pp 61–82. https://doi.org/10.1007/10_2022_207.
(20) OpenAI. GPT-4 Technical Report. arXiv March 16, 2023. https://doi.org/10.48550/arXiv.2303.08774.
(21) Cole, J. M. The Chemistry of Errors. *Nat. Chem.* **2022**, *14* (9), 973–975. https://doi.org/10.1038/s41557-022-01028-6.
(22) Shankar, K. Order from Chaos: The Poetics and Pragmatics of Scientific Recordkeeping. *J. Am. Soc. Inf. Sci. Technol.* **2007**, *58* (10), 1457–1466. https://doi.org/10.1002/asi.20625.
(23) Kanza, S.; Willoughby, C.; Gibbins, N.; Whitby, R.; Frey, J. G.; Erjavec, J.; Zupančič, K.; Hren, M.; Kovač, K. Electronic Lab Notebooks: Can They Replace Paper? *J. Cheminformatics* **2017**, *9* (1), 31. https://doi.org/10.1186/s13321-017-0221-3.
(24) Schwarz, N.; Veseli, S.; Jarosz, D. Data Management at the Advanced Photon Source. *Synchrotron Radiat. News* **2019**, *32* (3), 13–18. https://doi.org/10.1080/08940886.2019.1608120.
(25) Talley, K. R.; White, R.; Wunder, N.; Eash, M.; Schwarting, M.; Evenson, D.; Perkins, J. D.; Tumas, W.; Munch, K.; Phillips, C.; Zakutayev, A. Research Data Infrastructure for High-Throughput Experimental Materials Science. *Patterns* **2021**, *2* (12), 100373. https://doi.org/10.1016/j.patter.2021.100373.
(26) *Hardware Interface Packages — bluesky 1.10.0.post14+gfc4204d4 documentation*. https://blueskyproject.io/bluesky/hardware-interfaces.html (accessed 2022-11-17).
(27) Allan, D.; Caswell, T.; Campbell, S.; Rakitin, M. Bluesky's Ahead: A Multi-Facility Collaboration for an a La Carte Software Project for Data Acquisition and Management.

Synchrotron Radiat. News **2019**, *32* (3), 19–22. https://doi.org/10.1080/08940886.2019.1608121.
(28) Juchli, D. SiLA 2: The Next Generation Lab Automation Standard. In *Smart Biolabs of the Future*; Beutel, S., Lenk, F., Eds.; Advances in Biochemical Engineering/Biotechnology; Springer International Publishing: Cham, 2022; pp 147–174. https://doi.org/10.1007/10_2022_204.
(29) Brendel, A.; Dorfmüller, F.; Liebscher, A.; Kraus, P.; Kress, K.; Oehme, H.; Arnold, M.; Koschitzki, R. Laboratory and Analytical Device Standard (LADS): A Communication Standard Based on OPC UA for Networked Laboratories. In *Smart Biolabs of the Future*; Beutel, S., Lenk, F., Eds.; Advances in Biochemical Engineering/Biotechnology; Springer International Publishing: Cham, 2022; pp 175–194. https://doi.org/10.1007/10_2022_209.
(30) *Networked laboratory equipment*. SPECTARIS - Deutscher Industrieverband für Optik, Photonik, Analysen- und Medizintechnik. https://www.spectaris.de/en/association/thespectarisindustries/networked-laboratory-equipment/ (accessed 2023-02-01).
(31) *High-throughput and data driven strategies for the design of deep-eutectic solvent electrolytes - Molecular Systems Design & Engineering (RSC Publishing) DOI:10.1039/D2ME00050D*. https://pubs.rsc.org/en/content/articlehtml/2022/me/d2me00050d (accessed 2023-02-15).
(32) *ESAMP: Event-Sourced Architecture for Materials Provenance Management and Application to Accelerated Materials Discovery | Materials Chemistry | ChemRxiv | Cambridge Open Engage*. https://chemrxiv.org/engage/chemrxiv/article-details/60c73cbf842e650956db1678 (accessed 2022-10-04).
(33) *GEMD Documentation*. https://citrineinformatics.github.io/gemd-docs/ (accessed 2022-11-18).
(34) Pendleton, I. M.; Cattabriga, G.; Li, Z.; Najeeb, M. A.; Friedler, S. A.; Norquist, A. J.; Chan, E. M.; Schrier, J. Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): A Software Pipeline for Automated Chemical Experimentation and Data Management. *MRS Commun.* **2019**, *9* (3), 846–859. https://doi.org/10.1557/mrc.2019.72.
(35) Soedarmadji, E.; Stein, H. S.; Suram, S. K.; Guevarra, D.; Gregoire, J. M. Tracking Materials Science Data Lineage to Manage Millions of Materials Experiments and Analyses. *Npj Comput. Mater.* **2019**, *5* (1), 1–9. https://doi.org/10.1038/s41524-019-0216-x.
(36) *Electronic Lab Notebook (ELN)*. Labfolder. https://labfolder.com/ (accessed 2022-12-23).
(37) Inc, B. *Laboratory Information Management System | LIMS | Labguru*. https://www.labguru.com/lims (accessed 2022-11-21).
(38) Argento, N. Institutional ELN/LIMS Deployment. *EMBO Rep.* **2020**, *21* (3), e49862. https://doi.org/10.15252/embr.201949862.
(39) Bik, E. M.; Casadevall, A.; Fang, F. C. The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications. *mBio* **2016**, *7* (3), e00809-16. https://doi.org/10.1128/mBio.00809-16.
(40) *HIPAA Compliance with Google Workspace and Cloud Identity - Google Workspace Admin Help*. https://support.google.com/a/answer/3407054?hl=en (accessed 2023-02-15).
(41) Inc, L. *Automate Your Laboratory with the Global Leader for LIMS and ELN*. https://www.labware.com (accessed 2022-11-21).

(42) *Cloud-based platform for biotech R&D | Benchling*. https://www.benchling.com/ (accessed 2022-11-21).
(43) *LIMS- Laboratory Information Management Systems - US*. https://www.thermofisher.com/us/en/home/digital-solutions/lab-informatics/lab-information-management-systems-lims.html (accessed 2022-11-21).
(44) *SENAITE · Enterprise Open Source Laboratory System*. https://github.com/senaite/senaite.github.io/ (accessed 2022-11-21).
(45) *Entity Registration | Dotmatics*. Entity Registration | Dotmatics. https://www.dotmatics.com/capabilities/entity-registration (accessed 2022-11-21).
(46) *Data Management*. Citrine Informatics. https://citrine.io/product/what-is-the-citrine-platform/data-management/ (accessed 2022-11-21).
(47) Yan, R.; Jiang, X.; Wang, W.; Dang, D.; Su, Y. Materials Information Extraction via Automatically Generated Corpus. *Sci. Data* **2022**, *9* (1), 401. https://doi.org/10.1038/s41597-022-01492-2.
(48) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56* (10), 1894–1904. https://doi.org/10.1021/acs.jcim.6b00207.
(49) Venugopal, V.; Sahoo, S.; Zaki, M.; Agarwal, M.; Gosvami, N. N.; Krishnan, N. M. A. Looking through Glass: Knowledge Discovery from Materials Science Literature Using Natural Language Processing. *Patterns* **2021**, *2* (7), 100290. https://doi.org/10.1016/j.patter.2021.100290.
(50) Jacobsson, T. J.; Hultqvist, A.; García-Fernández, A.; Anand, A.; Al-Ashouri, A.; Hagfeldt, A.; Crovetto, A.; Abate, A.; Ricciardulli, A. G.; Vijayan, A.; Kulkarni, A.; Anderson, A. Y.; Darwich, B. P.; Yang, B.; Coles, B. L.; Perini, C. A. R.; Rehermann, C.; Ramirez, D.; Fairen-Jimenez, D.; Di Girolamo, D.; Jia, D.; Avila, E.; Juarez-Perez, E. J.; Baumann, F.; Mathies, F.; González, G. S. A.; Boschloo, G.; Nasti, G.; Paramasivam, G.; Martínez-Denegri, G.; Näsström, H.; Michaels, H.; Köbler, H.; Wu, H.; Benesperi, I.; Dar, M. I.; Bayrak Pehlivan, I.; Gould, I. E.; Vagott, J. N.; Dagar, J.; Kettle, J.; Yang, J.; Li, J.; Smith, J. A.; Pascual, J.; Jerónimo-Rendón, J. J.; Montoya, J. F.; Correa-Baena, J.-P.; Qiu, J.; Wang, J.; Sveinbjörnsson, K.; Hirselandt, K.; Dey, K.; Frohna, K.; Mathies, L.; Castriotta, L. A.; Aldamasy, M. H.; Vasquez-Montoya, M.; Ruiz-Preciado, M. A.; Flatken, M. A.; Khenkin, M. V.; Grischek, M.; Kedia, M.; Saliba, M.; Anaya, M.; Veldhoen, M.; Arora, N.; Shargaieva, O.; Maus, O.; Game, O. S.; Yudilevich, O.; Fassl, P.; Zhou, Q.; Betancur, R.; Munir, R.; Patidar, R.; Stranks, S. D.; Alam, S.; Kar, S.; Unold, T.; Abzieher, T.; Edvinsson, T.; David, T. W.; Paetzold, U. W.; Zia, W.; Fu, W.; Zuo, W.; Schröder, V. R. F.; Tress, W.; Zhang, X.; Chiang, Y.-H.; Iqbal, Z.; Xie, Z.; Unger, E. An Open-Access Database and Analysis Tool for Perovskite Solar Cells Based on the FAIR Data Principles. *Nat. Energy* **2022**, *7* (1), 107–115. https://doi.org/10.1038/s41560-021-00941-3.
(51) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.;

Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3* (1), 160018. https://doi.org/10.1038/sdata.2016.18.

(52) Draxl, C.; Scheffler, M. The NOMAD Laboratory: From Data Sharing to Artificial Intelligence. *J. Phys. Mater.* **2019**, *2* (3), 036001. https://doi.org/10.1088/2515-7639/ab13bb.

(53) Brinson, L. C.; Bartolo, L. M.; Blaiszik, B.; Elbert, D.; Foster, I.; Strachan, A.; Voorhees, P. W. FAIR Data Will Fuel a Revolution in Materials Research. arXiv April 6, 2022. https://doi.org/10.48550/arXiv.2204.02881.

(54) Bayerlein, B.; Hanke, T.; Muth, T.; Riedel, J.; Schilling, M.; Schweizer, C.; Skrotzki, B.; Todor, A.; Moreno Torres, B.; Unger, J. F.; Völker, C.; Olbricht, J. A Perspective on Digital Knowledge Representation in Materials Science and Engineering. *Adv. Eng. Mater.* **2022**, *24* (6), 2101176. https://doi.org/10.1002/adem.202101176.

(55) Winther, K. T.; Hoffmann, M. J.; Boes, J. R.; Mamun, O.; Bajdich, M.; Bligaard, T. Catalysis-Hub.Org, an Open Electronic Structure Database for Surface Reactions. *Sci. Data* **2019**, *6* (1), 1–10. https://doi.org/10.1038/s41597-019-0081-y.

(56) *canSAS.org*. https://www.cansas.org/ (accessed 2022-11-22).

(57) *Tiled — tiled 0.1.0a87 documentation*. https://blueskyproject.io/tiled/ (accessed 2023-03-20).

(58) *Export Administration Regulations*; Vol. 15.734.8.

(59) *Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification*. Federal Register. https://www.federalregister.gov/documents/2022/10/13/2022-21658/implementation-of-additional-export-controls-certain-advanced-computing-and-semiconductor (accessed 2023-02-09).

(60) *An Update on Research Security: Streamlining Disclosure Standards to Enhance Clarity, Transparency, and Equity | OSTP*. The White House. https://www.whitehouse.gov/ostp/news-updates/2022/08/31/an-update-on-research-securitystreamlining-disclosure-standards-to-enhance-clarity-transparency-and-equity/ (accessed 2023-02-09).

(61) Horton, M. K.; Woods-Robinson, R. Addressing the Critical Need for Open Experimental Databases in Materials Science. *Patterns* **2021**, *2* (12), 100411. https://doi.org/10.1016/j.patter.2021.100411.

(62) National Academies of Sciences, E. *NSF Efforts to Achieve the Nation's Vision for the Materials Genome Initiative: Designing Materials to Revolutionize and Engineer Our Future (DMREF)*; 2022. https://doi.org/10.17226/26723.

(63) *Zenodo - Research. Shared.* https://zenodo.org/ (accessed 2023-02-02).

(64) *Dryad | Our mission*. https://datadryad.org/stash/our_mission (accessed 2023-02-02).

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

(66) *Search Citrination*. https://citrination.com/search/simple?searchMatchOption=fuzzyMatch (accessed 2022-11-22).

(67) Zakutayev, A.; Perkins, J.; Schwarting, M.; White, R.; Munch, K.; Tumas, W.; Wunder, N.; Phillips, C. High Throughput Experimental Materials Database, 2017, 2 files. https://doi.org/10.7799/1407128.
(68) *Home | ICSD*. https://icsd.products.fiz-karlsruhe.de/ (accessed 2022-11-22).
(69) *Materials Project - Materials Explorer*. Materials Project. https://materialsproject.org/materials (accessed 2023-02-14).
(70) *Materials Project - Community*. Materials Project. https://materialsproject.org/community (accessed 2023-02-14).
(71) Burley, S. K.; Berman, H. M.; Christie, C.; Duarte, J. M.; Feng, Z.; Westbrook, J.; Young, J.; Zardecki, C. RCSB Protein Data Bank: Sustaining a Living Digital Data Resource That Enables Breakthroughs in Scientific Research and Biomedical Education. *Protein Sci. Publ. Protein Soc.* **2018**, *27* (1), 316–330. https://doi.org/10.1002/pro.3331.
(72) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.
(73) Scheffler, M.; Aeschlimann, M.; Albrecht, M.; Bereau, T.; Bungartz, H.-J.; Felser, C.; Greiner, M.; Groß, A.; Koch, C. T.; Kremer, K.; Nagel, W. E.; Scheidgen, M.; Wöll, C.; Draxl, C. FAIR Data Enabling New Horizons for Materials Research. *Nature* **2022**, *604* (7907), 635–642. https://doi.org/10.1038/s41586-022-04501-x.
(74) Hitzler, P. A Review of the Semantic Web Field. *Commun. ACM* **2021**, *64* (2), 76–83. https://doi.org/10.1145/3397512.
(75) Andersen, C. W.; Armiento, R.; Blokhin, E.; Conduit, G. J.; Dwaraknath, S.; Evans, M. L.; Fekete, Á.; Gopakumar, A.; Gražulis, S.; Merkys, A.; Mohamed, F.; Oses, C.; Pizzi, G.; Rignanese, G.-M.; Scheidgen, M.; Talirz, L.; Toher, C.; Winston, D.; Aversa, R.; Choudhary, K.; Colinet, P.; Curtarolo, S.; Di Stefano, D.; Draxl, C.; Er, S.; Esters, M.; Fornari, M.; Giantomassi, M.; Govoni, M.; Hautier, G.; Hegde, V.; Horton, M. K.; Huck, P.; Huhs, G.; Hummelshøj, J.; Kariryaa, A.; Kozinsky, B.; Kumbhar, S.; Liu, M.; Marzari, N.; Morris, A. J.; Mostofi, A. A.; Persson, K. A.; Petretto, G.; Purcell, T.; Ricci, F.; Rose, F.; Scheffler, M.; Speckhard, D.; Uhrin, M.; Vaitkus, A.; Villars, P.; Waroquiers, D.; Wolverton, C.; Wu, M.; Yang, X. OPTIMADE, an API for Exchanging Materials Data. *Sci. Data* **2021**, *8* (1), 217. https://doi.org/10.1038/s41597-021-00974-z.
(76) *OPTIMADE*. materials-consortia.github.io. https://optimade.org/ (accessed 2023-02-22).
(77) Empty Rhetoric over Data Sharing Slows Science. *Nature* **2017**, *546* (7658), 327–327. https://doi.org/10.1038/546327a.

5 Democratizing accelerated experimentation with open-source automation infrastructure

5.1 Introduction

Laboratory automation and autonomous experimentation systems are quickly gaining widespread adoption throughout materials and chemicals research communities[1–3]. As research groups, companies, universities, and other organizations build and unveil new self-driving labs, alternative visions for how researchers will access these tools and perform their science as they become commonplace are demonstrated. Given the varied and at times conflicting approaches to experiment automation, orchestration, and collaboration researchers are pursuing in their automation development projects, it is clear the future of how science is done is at a crossroads. Many lab automation developers are pursuing a path of automated experimentation as 'big science'. Several examples of automated workflows use industrial automation equipment and professionally engineered custom fixtures and integrations[4,5]. These systems can provide advanced automation capabilities and reliable operation. However, they require significant resources to build and maintain. Several of these examples involve six or seven figures worth of hardware and require teams of dedicated staff engineers to manage integration and operation. For some projects, this approach to automation is appropriate. However, this level of complexity is likely excessive for many experiment automation needs. This level of resource investment is also out of reach for most researchers. As an alternative to individual investment in commercial automation solutions, some researchers foresee a future of 'cloud laboratories' and automation user facilities. In the cloud lab model, researchers upload experiments as tasks to be executed by a commercial provider using automated equipment, in a manner analogous to running computing jobs on a remote high

performance computing cluster[6]. In a user facility model, scientists would be awarded time on shared equipment and likely receive support from facility staff to execute their experiment. While these approaches appear to provide broad accessibility to automated experimentation by enabling researchers to use these tools without an upfront infrastructure investment, they risk ultimately divorcing the researcher from the science they are doing. In a cloud lab context, scientists are completely removed from their experiment. They submit a protocol and receive results a few days later. This may be acceptable for performing well-defined screening campaigns using established protocols, but scientists lose a full understanding of how their experiment is run that is important for understanding new or unexpected findings. Cloud labs also tend to cater to well-established and widely marketable experimental protocols and techniques[6], which is not conducive to supporting novel experimental methods that are critical to moving materials research forward. User facilities, if executed well, could provide a brighter path forward. This approach could pair researchers with experienced automation experts and equipment owners who can help scientists translate their experimental goals into effective automated experiments. However, this approach may still limit the flexibility and availability of automation as automated experiment setup requires non-trivial hardware and sample integration. This barrier will likely result in user facilities that either cater to specific automated execution workflows or are limited in availability due to extensive setup overhead.

Between these three automation implementation styles, important science can be accomplished. Well-resourced labs can execute large, well-coordinated screening and optimization campaigns. Cloud labs enable access to automation for researchers needing well-established protocols. And user facilities can serve an intermediate role by providing infrastructure

access to promising research. What is left out here? Flexible, creative, exploratory research done by researchers without the resources to purchase or access to utilize these high-complexity commercial lab automation systems[7]. If this is the approach to lab automation implementation that the research community pursues and normalizes, we risk deprioritizing early-stage experimental research that is critical for developing new materials and limiting the pool of researchers able to leverage automation in their work to the exceptionally well resourced.

5.2 Democratized automation and autonomous experimentation through open-source infrastructure

An alternative path forward for the automated experimentation community is one characterized by democratized access to these tools and resources. 'Democratized' automation infrastructure and self-driving labs are loosely defined. They are generally characterized by reasonably affordable and extensible systems. This makes automation accessible to more researchers and allows them to modify and extend systems to meet the needs of new research projects. Well-supported open-source hardware systems are exemplars of democratized automation.

Commercial proprietary systems can also play a role in democratized automation. For example, the Opentrons OT2 liquid handling robot is a moderate-cost automation platform. While it is not fully open source and is not intended for modification, it has served as an entry point to lab automation for many researchers and has been extended, for example to enable automated electrochemical testing[8]. However, many challenges arise in adapting proprietary equipment to perform tasks it was not designed for. Democratized access to automated experimentation may best be achieved through fully open-source automation infrastructure that is designed and intended to be modified and extended by end users.

Open-source automation infrastructure is hardware and related software for which designs are freely released and licensed[9]. Effective open-source hardware (open hardware) projects provide a complete bill of materials with guidance on purchasing or acquiring components, along with full documentation that covers the entire build, setup, and operation procedures for the equipment. Many open hardware projects are designed from the ground up to utilize standard off the shelf components where possible, which lowers the cost and increases the availability of components. When custom fabricated components are required, detailed designs along with instructions for procurement or options to purchase through a manufacturing partner facilitate access. Open-source automation infrastructure can provide several benefits to lab automation users. Open-hardware equipment inherently supports modification and extension. When automation equipment is assembled from scratch, users have the option of excluding unneeded components or selecting more appropriate alternatives for their use case. This also makes it, in principle, feasible to extend equipment's capabilities with new components and apply it to new use cases beyond those originally intended. Open-source automation equipment is especially powerful when it serves as baseline automation infrastructure that can be shared between many different experimental workflows and even domains of science. For example, Science-Jubilee has been used for both nanomaterial synthesis and plant biology research[10]. In both cases, the existence of an open-source platform to perform basic XYZ motion allowed researchers to take the basics of automation for granted and focus on the automation development required for their specific project. The open nature of Science-Jubilee allowed users to modify and extend the platform as needed to support their experiments, for example by replacing vulnerable PLA 3D printed components with more robust SLA resin versions. Open-source equipment that is assembled by

the end user requires the user to understand the low-level assembly and operation of the device. This understanding can be important to understand the capabilities of a device when used in new research areas and helps facilitate effective troubleshooting. Open-source equipment also typically has a lower price compared to commercial alternatives.

The advantages of open-source automation equipment do not come without tradeoffs, however. Because these systems require end users to assemble, set up, and troubleshoot the hardware, significant time investment is required to implement open-source solutions. This investment may more than fully offset any cost savings in a simple time value comparison with commercial solutions. This time investment in system assembly does allow users to build new skills that they would not have by setting up an off-the-shelf solution, including mechanical assembly and soldering. While the time and skill investment to get started with a piece of open-hardware equipment may seem severe at first glance when compared to using off-the-shelf solutions, the difference is not as stark when considering a highly integrated automated experimental workflow. Automated workflows comprised of multiple off-the-shelf components require a significant investment in system integration and orchestration work. Often these systems require custom hardware development and modifications to get commercial off the shelf systems working together in ways they were not designed to. These efforts are often complicated by poor or missing documentation and issues related to modifying systems designed without extension in mind, such as a lack of an available software API to automate equipment operation. When the whole end-state system is considered, the time investment required to use an open-source, built from scratch component may be a small part of the overall system development effort. The effort invested in building open-source equipment also builds detailed knowledge of the system that

facilitates troubleshooting and maintenance. This is required, because there is no support contract for open-source hardware like one might expect for commercial equipment. Leveraging open-source equipment also requires an acceptably mature open-source option to exist for the type of equipment needed. Successfully designing a piece of open-source equipment requires a substantial amount of effort and fully documenting it and supporting it so that new, non-expert users can find success with it requires an order of magnitude more effort. This work does not fit into traditional roles in academic research, which makes it difficult to dedicate the time to these projects that they require.

When defining and categorizing open-source hardware solutions, it is important to acknowledge that not all custom automation equipment projects are open source. In fact, perhaps most examples of custom, end-user-developed automation infrastructure used in materials and chemicals research does not meet the licensing and documentation requirements to be considered 'open source'. Many project developers do not fully document their hardware due to the extensive effort required and the lack of incentive to undertake this effort in academic contexts. Some may also avoid sharing their designs because they view them as a proprietary competitive advantage that should be protected. Regardless of the motivation, many contributions to the advancement of automated experimentation are made through such systems. These systems can provide the same benefits related to novel capabilities and affordability that fully open-source systems do. However, they lack the community enablement advantages of openly shared projects. They also raise unique reproducibility challenges. Using fully custom, undocumented equipment that is not generally available results in data that can only be produced in that lab. The catch-all term 'User-developed automation infrastructure' has been coined to describe all systems that fall under the umbrella of

custom, end-user-built, non-commercial laboratory automation equipment, including both fully open-source and non-open projects. It is important to acknowledge that a wide gradient of community sharing and reproducibility exists within the user-developed automation space. While all user-developed automation can be critical for enabling new science, the community should strive to build truly open-source projects that fill the needs of custom lab automation and avoid settling for closed-source, one-off systems.

Given their unique advantages and capabilities, open-source lab automation equipment will play an important role in automated labs of the future. Open automation will enable automated experiments in contexts where commercial solutions are simply unavailable, as well as where they are out of reach given the resources available for an experiment. However, open automation adoption won't simply run on autopilot. To build healthy projects that contribute to the advancement of lab automation, a significant portion of the community needs to agree that open automation is critical. We need to accept open automation as 'real' solutions to automating experiments, not write them off as low-cost alternatives that are 'built out of Legos' for less-resourced researchers and undergraduate teaching labs. We need to value and prioritize open-source automation efforts by recognizing them as important contributions, providing incentives to researchers to develop and maintain documentation for their custom automation hardware projects, and provide funding to support these efforts alongside more traditional materials research. Throughout my PhD, democratizing laboratory automation through open-source infrastructure development has been a main objective. My primary contribution to this effort has been contributions to the development and maintenance of the Science-Jubilee open-source automation platform project. I have also supported and led several initiatives to foster a community and build

collaborations around open-source automation infrastructure, including co-organizing a Science-Jubilee focused workshop at UW, co-organizing a workshop on democratizing self-driving labs at the 2024 Accelerate conference (the premier conference in this field), and showcasing the capabilities of the Science-Jubilee platform internationally.

5.3 Science-Jubilee: Developing open, flexible automation infrastructure

A key advantage of open-source automation infrastructure is that it allows researchers to implement custom automated experiments without needing to start from scratch. Rather than spending time designing, building, and validating a new piece of hardware to handle the automation task they need, researchers can use existing designs that have been tested and refined, spending more time on the novel automation integration aspects of their experiment and the science they are ultimately interested in doing. A core component of lab automation required by many experiments is basic 3-dimensional motion. Many experiments involve performing operations on samples with a tool and require moving the tool between samples to automate the task. Given how fundamental this basic motion capability is to successfully implementing lab automation, making a simple, robust, well-tested, and extensible system to perform this motion available to the community in an open-source format would greatly lower

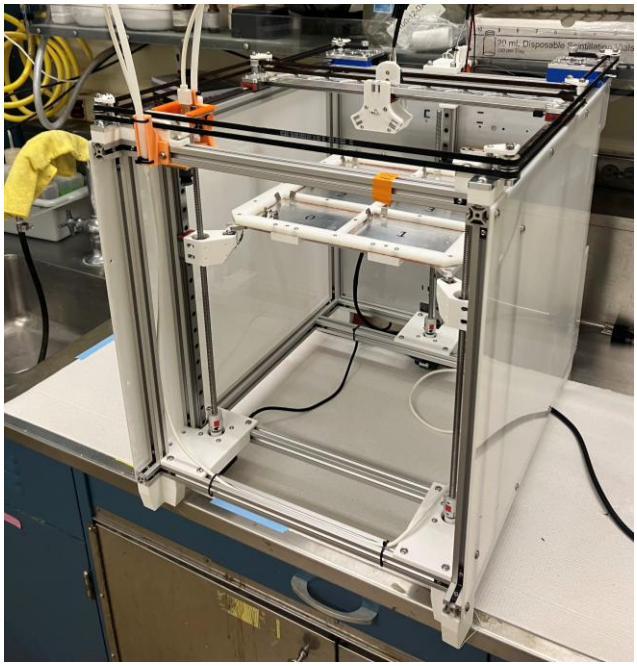


Figure 5.1: The base Jubilee platform with no tools attached.

barriers to entry in automation. The Science-Jubilee project seeks to do just this. Science-Jubilee is an active collaboration between multiple researchers at the University of Washington and beyond. This project aims to provide an open-source platform for automating diverse experiments through a 3D motion platform, compatible tools to perform many common laboratory tasks, and a python software interface to program experiments.

The Science-Jubilee project builds off the success of the Jubilee tool-changing 3D printer platform developed by Sonya Vasquez and Machine Agency[11]. Jubilee was originally developed as an extensible tool for digital fabrication. The core Jubilee platform is shown in figure 5.1. It consists of a 3D motion platform with a tool changing head that enables it to swap tools on the fly during a workflow. It is open-hardware, meaning that it is freely licensed and all documentation required to build it is freely available online[12]. It is assembled from commercially available components, including several standard off-the-shelf hardware items and a few custom-fabricated components that are available from partner vendors. The electronic components are controlled by Duet control boards[13] using RepRap firmware[14]. In the context of digital fabrication, tool changing enables techniques like 3D printing with multiple filament colors or materials, advanced multi-tool methods like additive subtractive manufacturing by layer, and 2D methods including drawing and laser engraving. The use of an open source, easily adaptable control board and a documented mechanical tool changing interface (figure 5.2) makes it feasible for users to develop new tool capabilities for Jubilee. This open hardware extensibility has attracted a large community of makers. The Jubilee discord server has close to 400 members and hosts regular discussions on platform enhancements and is a source of support for first time Jubilee users.

The base Jubilee system provides a promising platform for laboratory automation as it addresses many issues with traditional lab automation. The tool changing capability provides a path to solving the single-task specificity of most lab automation equipment. A Jubilee can perform multiple lab tasks in a single workflow by integrating different tools into an experiment. For example, a pipette can be used to prepare samples in the same manner as a commercial liquid handling robot, then a spectroscopic tool can be used to characterize samples. The open hardware and software of Jubilee along with the tool changing hardware interface makes it feasible to develop new modular tools to automate lab tasks for which automated equipment is not available. Contributors to the Science-Jubilee project have adapted a sonication horn into a Jubilee compatible tool format which enabled the automated synthesis of cadmium selenide nanocrystals on a Jubilee instance nicknamed 'sonication station'[15]. Other work has developed a platform for automating duckweed experimentation with Jubilee[16]. The open nature of jubilee lowers the monetary cost of the platform. The base motion platform kit can be purchased for approximately $1,700 and existing lab equipment can potentially be adapted to work with the platform. However, researchers considering investing in Jubilee should consider the large time commitment required to build and commission the platform as well as the lack of paid support options should issues arise.

Figure 5.2: Close up of the tool changing head that enables Jubilee to swap tools on the fly.

Our work with the Science-Jubilee project has focused on developing hardware and software infrastructure that enables the core Jubilee motion platform to be used in experimental workflows, and on enabling community engagement and

contributions to the development of Jubilee. This project is a collaboration between current and past Pozzo group members and Machine Agency, Nadya Peek's research group in UW Human Centered Design and Engineering. The 'core Jubilee motion platform' describes the basic motion infrastructure of Jubilee that handles XYZ movements and does not include capabilities provided by tools. This encompasses the frame, XYZ motion motors and components, tool changing mechanism, and control boards. This system is well established and generally reliable. We mostly avoid modifying this system in our work, instead focusing on making the system more accessible to scientists. We have done this by developing Jubilee's scientific hardware capabilities, developing a software interface to enable science on Jubilee, and working to enable community usage and contributions toward Jubilee for science. My contributions entail contributing to the initial development and ongoing refinement of the Science-Jubilee python control software, designing and integrating new tools, and co-organizing events to build community around Science-Jubilee.

5.3.1 Hardware development

Effective automation platforms for autonomous experimentation platforms need to integrate capabilities to perform the sample preparation and characterization tasks required for a particular experiment. As described above, the Jubilee platform makes this theoretically possible by changing tools. Hence, our work has included tool development. Currently, Jubilee has a handful of experimentation tools available, with new tools being developed on an as-needed basis.

Liquid handling tools: Science-Jubilee has three main liquid handling tool options: a peristaltic pump tool, a pipette interface for an OT2 pipette, and a syringe pump tool adapted from an external open-source hardware initiative.

I developed the peristaltic pump tool for use in a color-matching experiment, which is described below. The peristaltic pump tool is well suited to low precision, high volume liquid transfers. This tool consists of an array of 3 peristaltic pumps attached with tubing to 3 dispensing nozzles on a tool head, pictured in Figure 5.3. The pump inlet is connected by tubing to a reservoir of liquid. Liquid is dispensed by placing the appropriate dispensing nozzle over the desired location then activating the desired pump. This dispensing method is well suited to applications that require large volumes (O[mL]) or where time is important, as this method does not require the aspiration steps that a pipette does. However, the precision of delivered volumes is relatively poor. Initial measurements estimate the precision at +/- 0.1mL, which is unacceptable for many experiments.

The Opentrons OT2 pipette tool and Digital Pipette syringe tools both offer precision liquid handling and are suitable for different tasks. The Opentrons OT2 tool adapts an Opentrons pipette to mount on Jubilee. The Opentrons liquid handling systems use modular, standalone pipettes. These pipettes use a standard stepper motor that can be controlled by the Duet board used for the Jubilee platform. A mounting bracket and wiring harness has been designed that makes it possible to use these pipettes on Jubilee. Once mounted, users can perform a simple gravimetric calibration to enable precise liquid

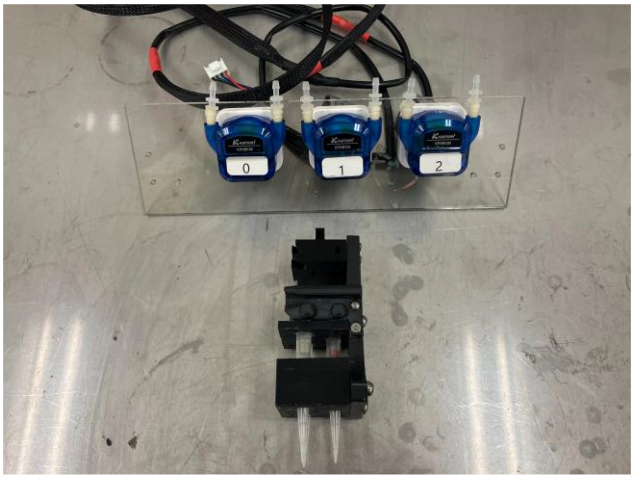


Figure 5.3: Peristaltic pump liquid dispensing components. The pumps (blue objects in back) are independently controlled and connected to dispenser nozzles (black array in front) by tubing (not pictured).

handling. This tool provides a true pipetting solution that enables transfer of microliter quantities with disposable pipette tips. While the OT2 pipette integration has several advantages, the relatively high cost ($2,750 USD per pipette) motivates lower-cost, open-source alternatives.

As an open-source alternative for precise liquid handling, the Digital Pipette liquid handling tool was integrated onto Jubilee. The Digital Pipette tool is an open-hardware liquid handling tool developed at the University of Toronto[17]. The original design consists of a linear servo motor, a 10cc disposable syringe, and a 3D printed frame to mount the syringe to the servo. The tool was originally designed to be a low-cost liquid handling solution that could work with grippers on robotic arms. The cost per tool is around $80 USD, in addition to a required microcontroller module such as an Arduino or Raspberry Pi (additional $40-80, can be shared between multiple Digital Pipette tools). This tool has several advantages over the OT2 Pipette tool. Selecting syringes with standard Luer-lock tip connectors enables compatibility with a wide range of needles and tips, allowing the tool to work with septa. The full syringe barrel volume can be filled with liquid, avoiding pipette head space evaporation leading to dripping when handling volatile liquids. Glass syringes can be mounted, allowing compatibility with most liquids. The low cost per tool makes it feasible to mount several tools on a Jubilee and dedicate each one to a specific liquid, which can greatly speed up the dispense process for larger amounts of liquid. The fully enclosed fluid handling volume also minimizes evaporation of volatile compounds which enables more accurate transfers and safe operation outside of a fume hood. To work with Science-Jubilee, the original Digital Pipette system required extensive modification. The 3D printed frame was modified to accept a Jubilee tool mount plate and parking post wings (Figure 5.4). Three versions of the frame were designed to accommodate different syringe types: a 1cc disposable

syringe, a 1cc Hamilton glass syringe, and a 10cc disposable syringe. Glass syringes are used when a liquid is incompatible with the polypropylene or Buna rubber components of the disposable syringes. Due to the design of the syringe mount, the mount needs to be adjusted for every new syringe design, including syringes of equivalent volume from different manufacturers. The control software was also modified to integrate with the Science-Jubilee python control software. The original Digital Pipette control software was designed to run a single pipette tool from an Arduino. To integrate with Jubilee, the capability to control multiple syringes from a separate control computer was needed. This was accomplished by re-implementing the control software as an HTTP web server running on a Raspberry Pi. The Raspberry Pi GPIO pins are used to control the servo motors, and liquid handling commands are sent as HTTP requests. This web server integrates with a client that is structured as a Science-Jubilee Tool class, making it seamless to control the syringe tools from a python process running the Science-Jubilee control software. Gravimetric calibrations were performed to calibrate the syringes and assess their accuracy and precision. Results of these calibrations are shown in Appendix 1.

Figure 5.4: Five adapted Digital Pipette tools mounted on a Jubilee.

Sample characterization tools: Jubilee currently has multiple camera tools for sample characterization. These tools are designed around Raspberry Pi cameras. There are many variants

of this tool design available, for both the basic Raspberry Pi camera as well as the 'high-quality' camera. Camera tools are useful anytime images of samples are required as characterization, for example in duckweed growth analysis[16] or Bayesian optimization color matching experiments (see below). Camera tools also provide the potential of automating experimental setup tasks.

An automated sample loading system has also been implemented that allows integration of the Jubilee platform with measurement techniques that support flow-cell measurements. This tool was implemented with the intention of integrating small-angle X-ray scattering measurements but is compatible with other systems that have flow-cells for measurement, such as dynamic light scattering. This tool is discussed further in chapter 6.

Jubilee frame alterations for science: An exploratory project to adapt the Jubilee motion platform frame for better compatibility with lab automation use cases was pursued. This work was primarily done by a summer REU student. Two issues with the original Jubilee frame were identified. First, the assembly and alignment process is delicate and precise, requiring careful manual alignment of the aluminum extrusion frame components. Misaligned frames lead to inaccurate positioning and require extensive rework to correct. This challenge is amplified in a lab automation context, where the frame assembly may be the first mechatronic assembly work a researcher attempts, so issues can cause frustration and disappointment. Second, the stock Jubilee

components are printed out of polylactic acid filament (PLA), which has poor compatibility with many organic solvents. To address the frame alignment issue, the top and bottom frame components were replaced with single-piece laser cut aluminum plates (Figure 5.5). The plates were designed to integrate with the rest of the existing frame components, without modifying the shape or kinematics of the motion platform. These plates greatly simplify the assembly process, helping new builders get to an assembled frame around 8 hours sooner. They can be ordered from online laser cutting vendors like Send-Cut-Send, making them as accessible as the original Jubilee frame is. To address the chemical compatibility issue, critical PLA components were replaced with SLA resin-printed components. This work is still ongoing. However, the modified Jubilee has been successfully used for 3D printing and lab automation tasks without any issues associated with the new frame components.

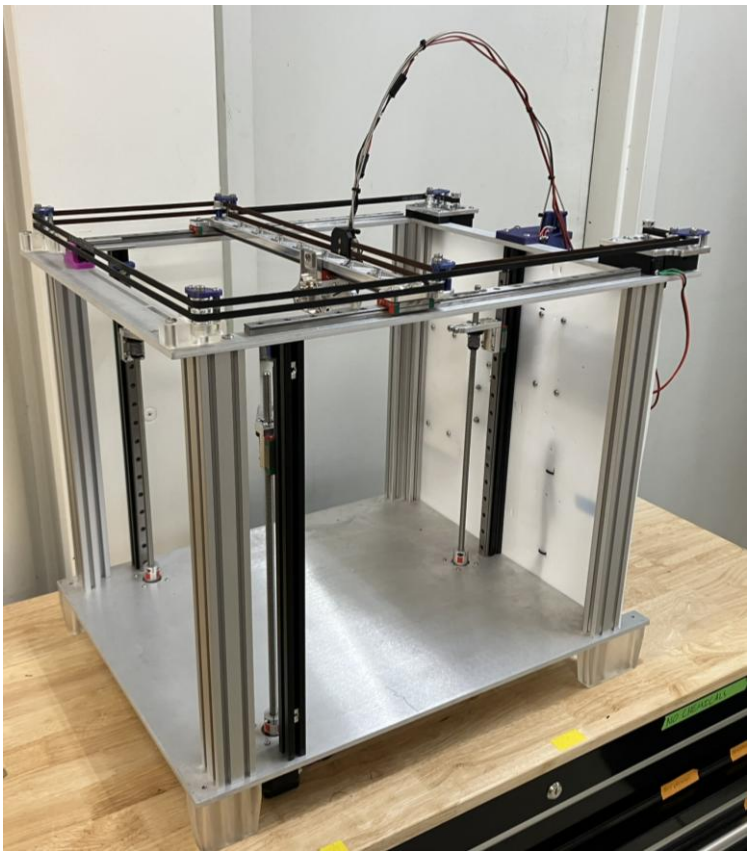


Figure 5.5: Modified Jubilee with aluminum top and bottom frame components and resin-printed components.

5.3.2 Software Development

Successful execution of experiments on Jubilee requires effective software control. The Science-Jubilee python package was developed to facilitate the programming of experiments. The goal of this interface is to enable researchers using Jubilee to define experiments using an intuitive pythonic programming interface and to handle the required state management, sub-tasks, and movement planning that is required to execute the experiment but not directly of interest to the experimenter. There are plenty of examples of similar programming interfaces to draw inspiration from. Any commercial lab automation hardware will have its own programming interface that will most likely manage the points mentioned above. PyLabRobot is an open source software project that attempts to be provide an abstracted interface to control liquid handling robots, essentially solving the same problem that we are solving for Jubilee[18]. We have chosen to implement our own solution from scratch to enable greater flexibility in managing the tool-changing capabilities of Jubilee. This is a unique capability among lab automation hardware, and we anticipated challenges in adapting existing interfaces to Jubilee. Our interface builds from an existing Jubilee software interface written for sonication station[19]. In developing our interface, we chose to follow the Opentrons API as closely as possible by using the same deck configuration and pipetting function definitions. The current version of the Science-Jubilee control software has proven to reliably manage core experiment workflows[20]. This implementation includes capabilities to manage deck layout, interact with labware such as well plates, and control the tools that have been developed. Documentation has been developed alongside the code. This software package significantly lowers the barrier to entry for new users of the Science-Jubilee platform. However, improvements are still possible. While the core functionality has been refined, control software for some tools still

contains bugs and unexpected behaviors. Currently, there are no unit tests and no defined protocol for hardware integration tests when changes are made. The largest issue, however, is with the overall architecture and intended use of the package. This python package was developed with the intent of being run interactively in a Jupyter notebook. This interactivity provides several advantages. Users can run subcomponents of experiment steps and carefully manage the software state of the system to ensure it matches the physical state of the Jubilee. However, it has become apparent over a few years of using this approach that notebooks are terrible structures for building and maintaining mirrored physical state. As a user sets up an experiment, there are many pieces of state that are changed on the Jubilee that also need to be updated in the Python object model. For example, when loading a new piece of labware, the user needs to both place the labware on the deck and set the location of the labware in the software. If the physical and software states do not match, the Jubilee might crash into an unknown piece of physical labware or may try to interact with a software-defined piece of labware that does not exist. This state management is an inherent part of writing control software for physical systems and is not inherently problematic. However, managing it in a notebook is. Notebooks allow users to skip cells, run cells out of order, and arbitrarily reset values. This is great for experimental debugging but quickly leads to problems when the notebook and physical state diverge. Running all Jubilee control from a notebook also poses problems when researchers need to integrate a Science-Jubilee platform with other equipment in an orchestrated fashion, for example when synthesizing samples determined from a separate optimization system and transferring them to an X-ray scattering instrument for measurement. This can become complicated to implement from a notebook, for example requiring that all steps be orchestrated from the Jubilee main run loop or mediated through message queue

to manage the jubilee component of the workflow. As an alternative to the notebook-based approach, the main Science-Jubilee control software should be repackaged to run as a web-accessible service that runs on a control computer that is physically attached to the Jubilee, e.g., a Raspberry Pi bolted to the back panel. Interactive access can be maintained through a transparent client library that allows users to control their jubilee through a notebook over a network connection. This client would provide a well-scoped subset of the full Science-Jubilee functionality so that users minimize their chances of unintentionally breaking their state. This would also facilitate easier automated control for machine-machine communication.

5.3.3 Jubilee Community Development Initiatives

Having a reliable hardware platform with unparalleled scientific capability and flawless control software doesn't matter if nobody uses it. Developing a strong community of users and contributors for Jubilee is a critical step towards expanding usage of the platform. Existing community driven hardware programs can provide resources to guide or bootstrap development around a community for Science Jubilee. In particular, the existing Jubilee printing community is a group of dedicated individuals contributing to the

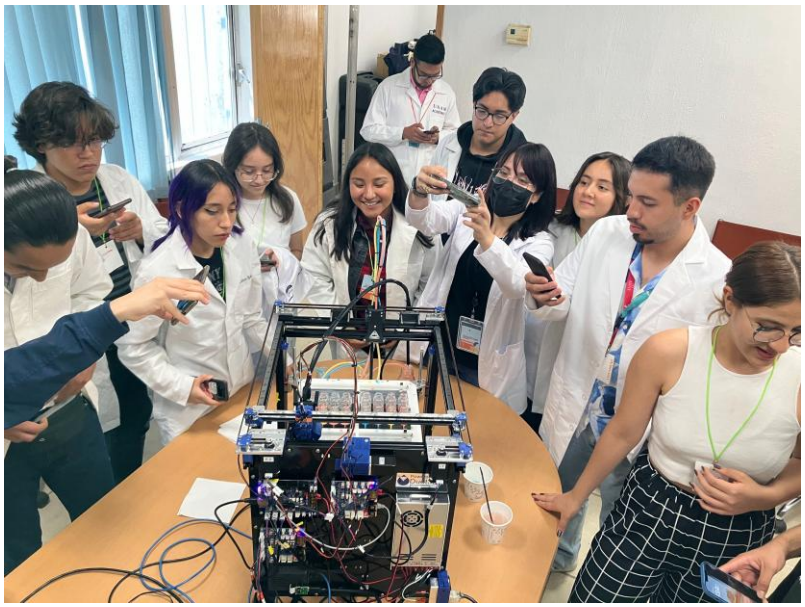


Figure 5.6: Students observe a Jubilee color mixing experiment

development of the Jubilee platform and guiding new members in getting started. An active Discord server is a hub for resources and feedback on Jubilee related topics. This community has been immensely helpful in our efforts to get started with Jubilee from a hardware perspective. The Gathering for Open Science Hardware (GOSH) is an organization that aims to build and support a global community for the development of open science hardware[21]. This organization works to coalesce a community around open hardware by hosting meetings, advocating for institutional support, and providing resources to newcomers. We aim to build a similarly strong community around Jubilee for Science without duplicating the efforts or resources of these existing spaces. Our efforts have focused on providing resources for interested new users to make getting started as straightforward as possible, and to raise awareness of the Jubilee platform through live interactions with relevant researchers. To facilitate new user onboarding, the project contributors have developed thorough documentation to walk new users through the process of building a Jubilee and using it for science applications. We also have a discord server for support and a GitHub repository to facilitate community contributions. To raise awareness of the Science-Jubilee project, we have provided live demonstrations of the system at conferences and events. I have presented the system at conferences in Canada and Slovenia, and incorporated the system into a

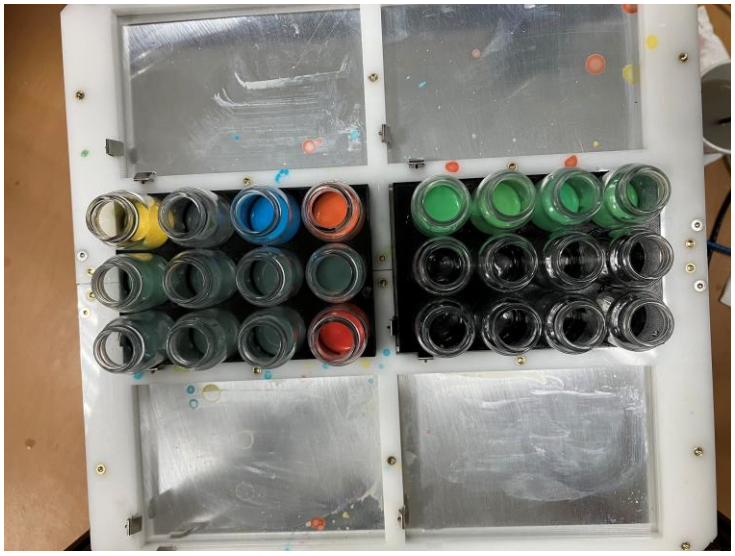


Figure 5.7: Experiment samples after a color mixing optimization campaign in which the objective was to make the greenish color on the right.

weeklong outreach event in Mexico. These demonstrations are immensely valuable for sparking excitement about the platform. At every event, I have watched researchers faces light up as they realize the possibilities that a flexible, approachable automation system would unlock for their work. These live interactions seem to be critical for demonstrating the potential of the project. During these demonstrations, a color matching experiment is typically used. Color mixing is a standard demonstration for autonomous experimentation platforms[22,23]. In a color mixing experiment, a Bayesian optimization campaign learns how to make a target color by mixing appropriate amounts of component colors, typically red, yellow, and blue for liquid examples. This demo is popular because it works easily without a lot of optimization algorithm tuning, and it is safe, intuitive, and cost effective. For this demo, liquid handling was done with the peristaltic pump tool, and sample characterization was done with a camera. Experimental control was managed with an early version of the Science-Jubilee control software. Outside of 'adult' research demonstrations, this demo is great for introducing children to automated research at outreach events like Discovery Days.

5.4 Advocating for broad adoption of democratized automation

To realize the vision laid above for a democratized future of autonomous experimentation, a robust ecosystem of community-driven infrastructure is needed. I co-organized a workshop dedicated to democratizing self-driving labs at the 2024 Accelerate conference with the goal of catalyzing community development in this space. This workshop was a joint effort between many contributors to the autonomous experimentation field who are similarly motivated by the benefits of democratized access to these tools. The workshop consisted of presentations and discussions focused on exploring the role of open-source automation infrastructure in autonomous

experimentation research and reducing barriers to the deployment and adoption of these systems. The workshop also featured live hardware demonstrations of over 14 user-developed automation infrastructure projects. This demonstration provided an important opportunity for interested members of the community to learn more about available open-source solutions and to talk with project developers about adapting them to their own uses. It also facilitated idea exchanges between builders of these tools. One of the conclusions of the workshop's discussion sections was that missing incentives hinder the development of complete open-source infrastructure projects. As a first step toward addressing this issue, I led the compilation of a hybrid perspective-showcase manuscript that compiled outcomes of the workshop along with highlights of the contributed projects for the live demonstration[24]. Inclusion in this manuscript provides the developers of these projects with a citable reference for their work. As authors, we hope that this contribution will help to motivate more researchers to adopt democratized automation practices. A larger conversation about democratizing this field will both inspire researchers to consider adopting democratized approaches where feasible and normalize contributing infrastructure advancements back to the wider community as a core ethos in the field.

5.5 References

(1) Rupnow, C. C.; MacLeod, B. P.; Mokhtari, M.; Ocean, K.; Dettelbach, K. E.; Lin, D.; Parlane, F. G. L.; Chiu, H. N.; Rooney, M. B.; Waizenegger, C. E. B.; Hoog, E. I. de; Soni, A.; Berlinguette, C. P. A Self-Driving Laboratory Optimizes a Scalable Process for Making Functional Coatings. *Cell Rep. Phys. Sci.* **2023**, *4* (5). https://doi.org/10.1016/j.xcrp.2023.101411.

(2) Bennett, J. A.; Orouji, N.; Khan, M.; Sadeghi, S.; Rodgers, J.; Abolhasani, M. Autonomous Reaction Pareto-Front Mapping with a Self-Driving Catalysis Laboratory. *Nat. Chem. Eng.* **2024**, *1* (3), 240–250. https://doi.org/10.1038/s44286-024-00033-5.

(3) Strieth-Kalthoff, F.; Hao, H.; Rathore, V.; Derasp, J.; Gaudin, T.; Angello, N. H.; Seifrid, M.; Trushina, E.; Guy, M.; Liu, J.; Tang, X.; Mamada, M.; Wang, W.; Tsagaantsooj, T.; Lavigne, C.; Pollice, R.; Wu, T. C.; Hotta, K.; Bodo, L.; Li, S.; Haddadnia, M.; Wołos, A.; Roszak, R.; Ser, C. T.; Bozal-Ginesta, C.; Hickman, R. J.; Vestfrid, J.; Aguilar-Granda, A.; Klimareva, E. L.; Sigerson, R. C.; Hou, W.; Gahler, D.; Lach, S.; Warzybok, A.; Borodin, O.; Rohrbach, S.; Sanchez-Lengeling, B.; Adachi, C.; Grzybowski, B. A.; Cronin, L.; Hein, J. E.; Burke, M. D.; Aspuru-Guzik, A. Delocalized, Asynchronous, Closed-Loop Discovery of Organic Laser Emitters. *Science* **2024**, *384* (6697), eadk9227. https://doi.org/10.1126/science.adk9227.

(4) Szymanski, N. J.; Rendy, B.; Fei, Y.; Kumar, R. E.; He, T.; Milsted, D.; McDermott, M. J.; Gallant, M.; Cubuk, E. D.; Merchant, A.; Kim, H.; Jain, A.; Bartel, C. J.; Persson, K.; Zeng, Y.; Ceder, G. An Autonomous Laboratory for the Accelerated Synthesis of Novel Materials. *Nature* **2023**, *624* (7990), 86–91. https://doi.org/10.1038/s41586-023-06734-w.

(5) Vescovi, R.; Ginsburg, T.; Hippe, K.; Ozgulbas, D.; Stone, C.; Stroka, A.; Butler, R.; Blaiszik, B.; Brettin, T.; Chard, K.; Hereld, M.; Ramanathan, A.; Stevens, R.; Vriza, A.; Xu, J.; Zhang, Q.; Foster, I. Towards a Modular Architecture for Science Factories. *Digit. Discov.* **2023**, *2* (6), 1980–1998. https://doi.org/10.1039/D3DD00142C.

(6) *Emerald Cloud Lab: Remote Controlled Life Sciences Lab*. https://www.emeraldcloudlab.com/ (accessed 2025-05-14).

(7) Lo, S.; G. Baird, S.; Schrier, J.; Blaiszik, B.; Carson, N.; Foster, I.; Aguilar-Granda, A.; V. Kalinin, S.; Maruyama, B.; Politi, M.; Tran, H.; D. Sparks, T.; Aspuru-Guzik, A. Review of Low-Cost Self-Driving Laboratories in Chemistry and Materials Science: The "Frugal Twin" Concept. *Digit. Discov.* **2024**, *3* (5), 842–868. https://doi.org/10.1039/D3DD00223C.

(8) Abed, J.; Bai, Y.; Persaud, D.; Kim, J.; Witt, J.; Hattrick-Simpers, J.; H. Sargent, E. AMPERE: Automated Modular Platform for Expedited and Reproducible Electrochemical Testing. *Digit. Discov.* **2024**, *3* (11), 2265–2274. https://doi.org/10.1039/D4DD00203B.

(9) *Open Source Hardware Definition*. https://oshwa.org/resources/open-source-hardware-definition/ (accessed 2025-05-14).

(10) Subbaraman, B.; Lange, O. de; Ferguson, S.; Peek, N. The Duckbot: A System for Automated Imaging and Manipulation of Duckweed. *PLOS ONE* **2024**, *19* (1), e0296717. https://doi.org/10.1371/journal.pone.0296717.

(11) Vasquez, S.; Twigg-Smith, H.; Tran O'Leary, J.; Peek, N. Jubilee: An Extensible Machine for Multi-Tool Fabrication. In *Proceedings of the 2020 CHI Conference on Human Factors*

in Computing Systems; CHI '20; Association for Computing Machinery: New York, NY, USA, 2020; pp 1–13. https://doi.org/10.1145/3313831.3376425.
(12) Oellermann, M.; Jolles, J. W.; Ortiz, D.; Seabra, R.; Wenzel, T.; Wilson, H.; Tanner, R. L. Open Hardware in Science: The Benefits of Open Electronics. *Integr. Comp. Biol.* **2022**, *62* (4), 1061–1075. https://doi.org/10.1093/icb/icac043.
(13) *Duet3D · Duet3D*. Duet3D. https://www.duet3d.com/ (accessed 2024-02-05).
(14) *RepRapFirmware.org*. https://www.reprapfirmware.org/ (accessed 2024-02-05).
(15) Politi, M.; Baum, F.; Vaddi, K.; Antonio, E.; Vasquez, S.; Bishop, B. P.; Peek, N.; Holmberg, V. C.; Pozzo, L. D. A High-Throughput Workflow for the Synthesis of CdSe Nanocrystals Using a Sonochemical Materials Acceleration Platform. *Digit. Discov.* **2023**.
(16) Subbaraman, B.; de Lange, O.; Ferguson, S.; Peek, N. The Duckbot: A System for Automated Imaging and Manipulation of Duckweed. *PloS One* **2024**, *19* (1), e0296717. https://doi.org/10.1371/journal.pone.0296717.
(17) Yoshikawa, N.; Darvish, K.; Vakili, M. G.; Garg, A.; Aspuru-Guzik, A. Digital Pipette: Open Hardware for Liquid Transfer in Self-Driving Laboratories. *Digit. Discov.* **2023**, *2* (6), 1745–1751. https://doi.org/10.1039/D3DD00115F.
(18) Wierenga, R. P.; Golas, S. M.; Ho, W.; Coley, C. W.; Esvelt, K. M. PyLabRobot: An Open-Source, Hardware-Agnostic Interface for Liquid-Handling Robots and Accessories. *Device* **2023**, *1* (4). https://doi.org/10.1016/j.device.2023.100111.
(19) Machineagency/Sonication_station, 2023. https://github.com/machineagency/sonication_station (accessed 2024-02-02).
(20) Machineagency/Science_jubilee, 2023. https://github.com/machineagency/science_jubilee (accessed 2024-02-05).
(21) *Home - Gathering for Open Science Hardware*. https://openhardware.science/ (accessed 2024-02-05).
(22) *SDL for KIDS*. https://sites.google.com/matterhorn.studio/sdl4kids/home (accessed 2024-02-05).
(23) Baird, S. G. Sparks-Baird/Self-Driving-Lab-Demo: V0.8.4, 2023. https://doi.org/10.5281/ZENODO.7855492.
(24) Pelkie, B.; Baird, S.; Aissi, E.; Aspuru-Takata, K.; Cao, Y.; Chang, J. H.; Gambhir, K.; Hale, W. S.; Hao, L.; Hattrick, C.; Hein, J.; Luo, D.; Melville, O.; Ngan, M.; Nyeland, L. L. B.; Peek, N.; Politi, M.; Rajkumar, E. E.; Siemenn, A.; Subbaraman, B.; Vasquez, S.; Watchorn, J.; Zhang, W.; Ziskason, R.; Pozzo, L.; Buonassisi, T.; Vegge, T. Democratizing Self-Driving Labs through User-Developed Automation Infrastructure. ChemRxiv February 12, 2025. https://doi.org/10.26434/chemrxiv-2025-zhkrf.

6 Open-hardware automation platform for accelerated sol-gel nanomaterial synthesis

This work is also in preparation as a manuscript for publication.

6.1 Introduction

Mesoporous and nanoporous colloidal silicas are synthesized via sol-gel methods that incorporate micelle-forming surfactants to produce an ordered (nano)porous internal structure. Mesoporous silicas are critical in a wide range of applications. In chemical separations, mesoporous silica chromatography column packings enable efficient separations due to their narrow particle and pore size distributions[1,2] and improve the performance of pervaporation separations when included in membranes[3]. Mesoporous silica nanoparticles (MSNs) are also promising drug delivery platforms. Pores can be loaded with pharmaceutical ingredients to avoid solubility and other delivery issues[4], and particle surfaces can be functionalized to promote targeted delivery[5,6] and coated with stimuli-responsive coatings to control drug release[7]. Pore and particle size control is important to optimize delivery and release characteristics[4,8]. While drug delivery applications are still in development, mesoporous silica has been used for decades in tablet formulations as an excipient[9]. In catalysis, mesoporous silica can serve as a support substrate for metallic heterogeneous catalysts[10], where the material's high specific surface area and controllable pore sizes contributes to high mass transfer and can prevent metal sintering[11]. MSN supported metallic catalysts have shown promising results for various CO_2 utilization reactions including CO_2 methanation[12], CO_2 photoreduction[13] and electrocatalytic reduction of CO_2 with Ag or Au nanoparticles[14].

Colloidal silica particles are typically synthesized through a surfactant templated sol-gel process illustrated in figure 6.1. In the traditional sol-gel synthesis, commonly referred to as the Stöber process[15], a silicon alkoxide such as tetraethyl orthosilicate (TEOS) undergoes a base-catalyzed hydrolysis reaction (often using ammonia) followed by a condensation step to form nanoparticles. Mesoporous silica can be produced by including a micelle-forming surfactant, such as cetyltrimethylammonium bromide (CTAB), in the synthesis. The surfactant micelles provide a template for silicic acid monomers to condense around, before the resulting oligomer-micelle complexes aggregate and condense to form highly ordered porous structures[16,17]. The packing structure of the porous phase is thus controlled by the micelle structure and reaction conditions. Common pore phases for these materials include the hexagonal rods of the canonical Mobil Composition of <atter (MCM)-41[18] and Santa Barbara Amorphous (SBA)-15[19] mesoporous silicas, and the gyroid cubic packing of MCM-48[18]. The synthesis of a mesoporous silica material was first reported in 1992 with the development of the MCM mesoporous silicas[18]. These materials were subsequently deployed commercially in 2002[20], a notably fast turnaround for a novel material.

Several approaches to synthesizing mesoporous silicas have been investigated with the goal of gaining control over particle morphology and aggregation, pore size, and pore arrangement, include evaporation induced self assembly[21], simple monophasic solution synthesis[22], heterogeneous oil-water biphasic systems[23], semi-batch continuous TEOS addition[24], and other schemes[25]. This complex space of synthesis routes multiplies when the potential for including new surfactants as templates to control pore phase properties. Many of the reported mesoporous silica synthesis routes have been well characterized and the general factors that impact particle

morphologies are understood in a 'directional' sense. For example, it is generally accepted that ammonia, TEOS concentration, and temperature control particle size[26], and that surfactant chain length controls internal pore size[16]. However, these understandings are often stated as general trends, rather than prescriptive guidelines for achieving target mesoporous silica morphologies. Many studies do not quantify the reproducibility of their results and draw conclusions from single replicates. As Noureddine and coworkers hypothesize[26], the complexity of synthesis methods present in the literature may hinder the adoption of MSNs for downstream applications, as selecting appropriate synthesis conditions to produce MSNs with specific target morphologies requires navigating a vast design space.

An automated synthesis and characterization procedure for mesoporous silica could help overcome these challenges by facilitating repeatable, high-throughput explorations of synthesis parameter space and enabling closed-loop optimization to discover appropriate synthesis conditions for target morphologies. Mesoporous silicas are well suited to study with a high-

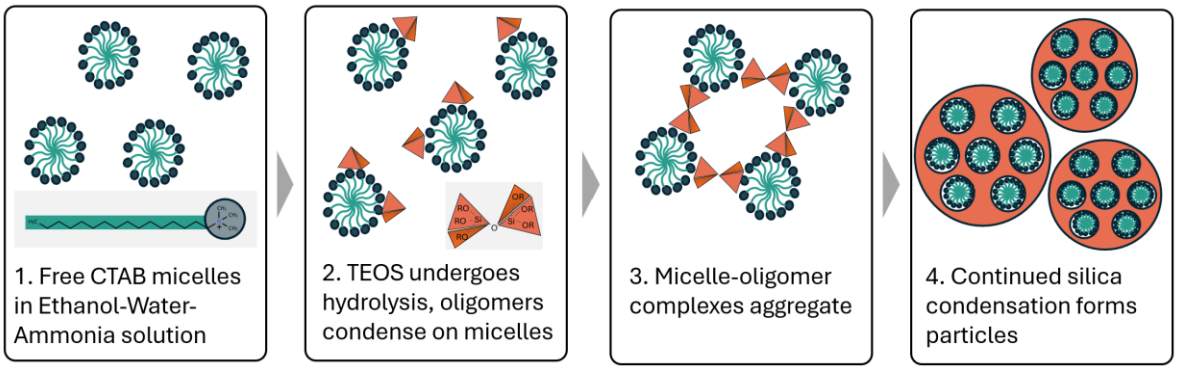


Figure 6.1: Illustration of the cooperative assembly mechanism of particle formation during the synthesis process. 1) CTAB solution is added to a mixture of ethanol, water, and ammonia, resulting in free micelles. 2) TEOS is added to the micelle-containing solution. TEOS undergoes hydrolysis to form silicic acid monomers, which interact with the micelles. 3) Micelle-oligomer complexes aggregate to form assemblies. 4) Continued TEOS condensation produces spherical particles.

throughput or accelerated experimentation approach due to liquid-based, ambient-condition synthesis procedure, complex dependency of material morphologies on synthesis conditions, and large parameter space of compositions and component identities. In this work, we describe such a system for the automated synthesis and characterization of mesoporous materials. This system performs sol-gel synthesis with an open-source liquid handling platform and integrates with small-angle X-ray scattering instruments for sample characterization, as illustrated in Figure 2. The Science-Jubilee open-hardware laboratory automation platform is used to automate the synthesis process[27,28] utilizing an adapted version of the Digital Pipette syringe tool[29] to provide liquid handling capabilities. Small-angle X-ray scattering instruments, including a lab scale instrument and a synchrotron beamline, are used for sample characterization. A modified version of the NIST Autonomous Formulations Laboratory (AFL) sample loading module is used to automate sample loading from the Jubilee synthesis platform into the X-ray scattering instruments[30]. This platform enables the fully automated and reproducible synthesis of mesoporous colloidal silicas, allowing for high-throughput investigations and providing an important component of a self-driving lab for this material. Additionally, the use of open-source automation components showcases the promise of democratized approaches to accelerated experimentation[31]. This platform integrates three open-hardware automation components, including the internally co-developed Science-Jubilee and ones developed by the broader community (Digital Pipette and NIST-AFL). Using open hardware provides flexibility to meet the unique requirements of this sol-gel synthesis and facilitates extension of the platform by other researchers to study alternative sol-gel synthesized systems such as solid silica nanoparticles or other metal-oxide nanomaterials.

Small-angle X-ray scattering is a powerful technique for the automated characterization of porous nanomaterials. X-ray scattering provides volume-averaged structural information over a broad range of length scales with a single instrument, from atomic crystalline structure (wide-angle X-ray scattering or X-ray diffraction) through mesopores (small-angle X-ray scattering or SAXS) to particle size, shape, and aggregation (ultra- small-angle X-ray scattering or USAXS). Samples can be quickly (seconds to minutes) measured as liquid dispersions, making the method amenable to automation.

The automated synthesis and characterization system described in this work was used to perform an initial composition space screening for a mesoporous colloidal silica synthesis. This synthesis campaign validates that the system is capable of reproducibly synthesizing monodisperse colloidal silicas with highly ordered mesoporous structures, supports insights into synthesis – morphology relationships for this system, and provides an initial starting point for future autonomous optimization campaigns. The binary surfactant mesoporous silica synthesis procedure

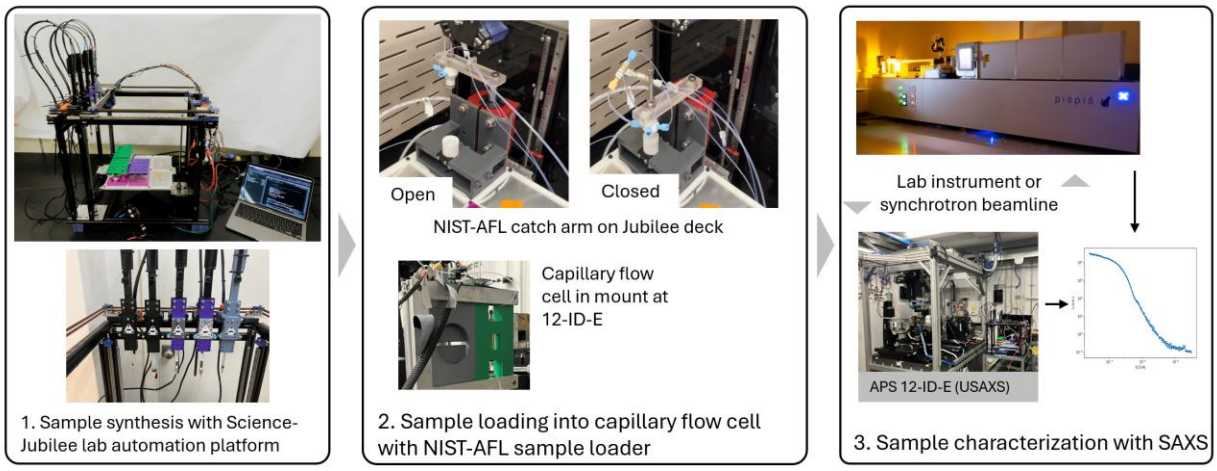


Figure 6.2: Overview of automated synthesis platform. Silica nanoparticle samples are synthesized on Jubilee, transferred for characterization using the NIST-AFL sample loader, and characterized with small-angle X-ray scattering using either a laboratory instrument or synchrotron beamline.

first reported by Kim et al.[25] was used for this exploration. This procedure modifies the traditional Stöber synthesis by adding two surfactants: A cationic surfactant such as CTAB to serve as a pore template, and Pluronic F127 to aid in particle dispersity. It is performed at ambient conditions, can produce monodisperse, colloidally stable particles, and can be performed in around 20 minutes, all of which make it amenable to automated and autonomous experimentation.

6.2 Materials and Methods

6.2.1 Mesoporous colloidal silica synthesis

TEOS (Sigma Aldrich), Anhydrous Ethanol (Decon labs), CTAB (TCI), Pluronic F127 (Sigma), and Ammonium Hydroxide aqueous solution (Sigma) were all used as received. Ultra-pure water (resistivity > 18 $M\Omega$*cm) was used throughout. The synthesis was performed on a Science-Jubilee laboratory automation platform. Science-Jubilee is an extension of the Jubilee open-hardware 3D printer project. The science-jubilee project extends the Jubilee open-hardware 3D printer project[28,32] to include tools, software, and documentation that facilitates laboratory automation[27]. Liquid handling for the sol-gel synthesis was performed with a modified implementation of the Digital Pipette liquid handling tool[29]. The 3D printed components of the tool were modified to integrate with the Jubilee tool changing mechanism, and the control software was adapted into the science-jubilee python library. Five Digital Pipette tools were used in this synthesis procedure: a dedicated 1 cc disposable plastic syringe dispensed water, dedicated 1 cc glass syringes (Hamilton) dispensed TEOS and ammonium hydroxide solution, a shared 1 cc disposable plastic syringe dispensed both aqueous CTAB and Pluronic F127 solutions, and a shared 10 cc disposable plastic syringe dispensed ethanol, mixed samples, and transferred samples to the NIST-AFL sample loader (described below) for characterization. 1 cc syringes provide more precision for low-volume

transfers compared to 10 cc syringes. Glass syringes were used with TEOS and ammonia due to incompatibility of these reactants with the plastic syringes. Accuracy for the used Digital Pipette tools is shown in tables S3-S5. Documentation for the Science-Jubilee/Digital Pipette integration is available on the Science-Jubilee project website[33]. The Science-Jubilee was used with the lab automation deck component, which retains 6 standard SLAS labware plates in a grid.

Sample synthesis was performed in 20 mL glass scintillation vials. Custom 3D printed labware was used to hold the vials. A synthesis volume of 10 mL per sample was used. This synthesis involves volatile reactants – ammonium hydroxide solution and ethanol. Evaporation of these components needs to be prevented to avoid changes to sample compositions during synthesis. TEOS vapors are also hazardous, and mitigation is needed to limit exposure. Pre-slit silicone septa were used on all reaction and stock solution vials to minimize evaporation. This effectively prevented reactant evaporation and allowed for the synthesis procedure to be run outside of a fume hood, facilitating beamside sample preparation. The automated synthesis procedure was controlled from a Jupyter notebook using the science-jubilee python package to control the Jubilee platform[27].

6.2.1.1 Workflow for mesoporous silica nanoparticle synthesis

6.2.1.1.1 Experimental design

For the batch synthesis campaign, a space filling experimental design strategy was used to select sample compositions. An experimental parameter space was selected by defining composition bounds on each component. Composition bounds (Table 6.1) were selected by expanding the compositions explored by Kim et al[25]. and identifying ranges likely to produce dispersed colloids through initial experiments. A set of randomly selected compositions from within these bounds was generated using a constrained Sobol rejection sampling strategy which selected samples that met the constraints imposed by the selected composition bounds and the necessity that all volume fractions sum to 1. First, a set of 128 composition values for the first five components (excluding water) was generated with a Sobol sequence[34,35]. For each sampled composition, the volume fraction of water required to meet the additive constraint was then calculated, accounting for water added during the addition of aqueous surfactant solutions. Compositions meeting all composition bound and additive constraints were retained, and the rest were discarded. This process was repeated until 80 valid sample compositions were selected. 63 of these selected samples were successfully synthesized within available time at a synchrotron beamline. Additionally, further experiments were run to validate the reproducibility of the synthesis protocol.

Table 6.1: Composition bounds. Bounds for liquid components are reported as volume fraction of final sample mixture, while surfactant components are reported as concentrations in [mg/mL] in final sample mixture.

Component	**Lower bound**	**Upper bound**
TEOS	0.0035	0.04
Ammonium hydroxide solution	0.01	0.08
Ethanol	0.2	0.4
CTAB [mg/mL]	1	10
Pluronic F127 [mg/mL]	0	30
Water	0.3	1

6.2.1.1.1.1 Synthesis execution

Stock solutions of TEOS, ammonium hydroxide solution, ethanol, water, CTAB solution (aqueous) and Pluronic F127 solution (aqueous) were prepared. TEOS was dissolved in ethanol at a ratio of 1 part TEOS:1.85 parts ethanol by volume. This dilution was necessary to increase the syringe delivery volume for TEOS dispensing, which increases the relative precision of low-volume dispenses. This dilution amount was selected such that the minimum dispense volume for the lower bound of the TEOS composition space would be 100 uL, which was identified as the minimum reasonable volume for precise transfers with the syringe tool (Table A1.4). CTAB was dissolved in water at a concentration of 15 g CTAB per liter of water. The mixture was lightly heated on a hot plate to dissolve. Recrystallization during the synthesis campaign was prevented by periodically re-heating the CTAB solution on the hot plate during synthesis campaigns. Pluronic F127 solution was prepared at a concentration of 50 g/L by dissolving Pluronic F127 powder in water. Ethanol, ammonium hydroxide solution, and water were used without preparation. Stock solutions were transferred to 20 mL septa-covered vials and placed on the Jubilee deck. The actual delivery volume for each reactant was calculated from the target sample composition and the stock solution compositions. For example, the water delivered with the aqueous CTAB solution is accounted for in the total water delivered to the sample. A typical nanoparticle synthesis involves the following steps and is illustrated in figure 6.3. 1) Ethanol is transferred from the ethanol stock vial to the sample vial using the 10 cc syringe tool. 2) Water is dispensed into the sample vial from the dedicated 1 cc water syringe tool. The syringe is refilled from the water stock vial as needed. 3) Ammonia solution is dispensed with the dedicated syringe tool. 4) CTAB solution is dispensed with the shared surfactant syringe, then the syringe is rinsed

by aspirating and dispensing ethanol from a series of three rinse vials dedicated for use before TEOS is added to a sample. 5) Pluronic F127 solution is dispensed with the shared surfactant syringe, then the syringe is rinsed as it was after CTAB. 6) The precursor solution is mixed via aspirate/dispense cycles with the 10 cc syringe used for ethanol. 7) TEOS is added with the dedicated syringe tool. The Jubilee immediately swaps for the 10 cc syringe tool, then mixes the sample with aspirate/dispense cycles. 8) The 10 cc syringe is rinsed in a series of 4 post-TEOS addition ethanol vials. 9) Particle growth is complete and the sample is ready to be loaded for measurement two hours after TEOS addition. The entire synthesis process takes approximately 20 minutes of active time per sample.

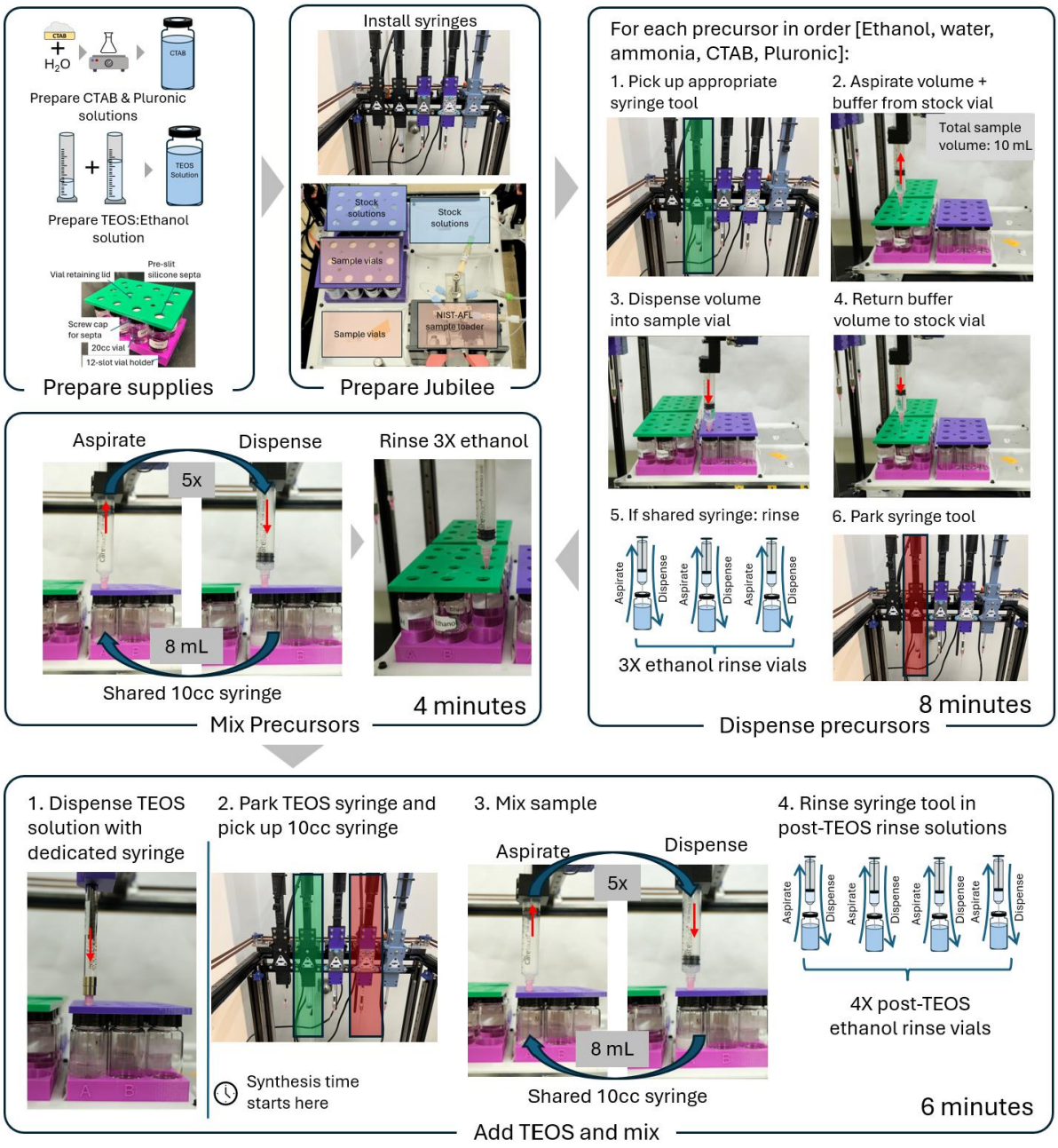


Figure 6.3: Mesoporous colloidal silica synthesis protocol.

6.2.2 Sample transfer to characterization instruments

The NIST-AFL sample loading system was used to transfer samples from the Jubilee synthesis platform to X-ray scattering instruments for sample characterization. The sample loader is a component of the NIST-AFL platform, an open-source initiative to develop autonomous experimentation capabilities for formulation development tasks[30]. The sample loader is a pneumatic system that uses air pressure to push a sample through flexible tubing into a quartz capillary flow cell. To load a sample, the Jubilee robot transfers a sample into the NIST-AFL 'catch' cell using the 10 cc mixing syringe tool. The catch cell is closed with a rotating clamp arm. Air pressure is used to move the sample through tubing until it reaches a capillary flow cell positioned in the X-ray instrument. After measurement, the sample contact path is rinsed and dried. More information on our implementation of this system for Jubilee can be found on the project GitHub repository, and additional information on the NIST-AFL system can be found in the hardware and software repositories for this project[36,37]. The sample loader design used in this work is an update to the version described in the original AFL publication. The sample loader is well suited to run in a sequential one-sample-at-a-time mode. Alternatively, custom liquid cartridge cells were also used to run samples in a 'batch' mode. Documentation to reproduce these cells is available online[38].

6.2.3 Sample Characterization

6.2.3.1 Small-angle X-ray scattering

Small-angle X-ray scattering measurements were made at the 12-ID-E USAXS beamline at the Advanced Photon Source[39], and with a laboratory instrument. The USAXS instrument enables observation of larger structural features than the laboratory instrument, while the laboratory

instrument provides higher resolution data at the high-q regions corresponding to scattering from the internal porosity of mesoporous samples. USAXS measurements were made with an X-ray energy of 21 keV (λ = 0.5895 Å). USAXS flyscan data were reduced and corrected using Igor. Data were corrected for instrumental and empty container background scattering and de-smeared to correct for instrumental slit smearing. Laboratory SAXS measurements were made using a Xenocs Xeuss 3.0 instrument with a Cu-Kα source (λ = 1.5406 Å). Data were reduced and corrected for empty container background scattering using XSACT 2.10 from Xenocs.

Operating this system in a fully automated manner requires integrating with X-ray scattering instruments to trigger measurements and collect completed data. This integration is instrument-specific due to varied sample mounting stages and control interfaces. Mechanically integrating the synthesis system with an instrument requires mounting the capillary flow cell and NIST-AFL electronic components on the instrument sample stage. This can be achieved with appropriate sample stages or custom 3D printed holders[40], and requires the Science-Jubilee and NIST-AFL control hardware to be located next to the instrument. Software integration with the instrument control system is needed to trigger a measurement once a sample is loaded. The details of this implementation are instrument-specific, but common instrument control softwares including Bluesky and SPEC provide provisions for initiating measurements through a network request. The integrations developed in this work for the 12-ID-E and the Xenocs instrument are documented in the project GitHub repository[41].

Raw X-ray scattering data needs to be corrected and reduced then analyzed before it can be used to understand the structure of the sample it was generated from. For closed-loop autonomous experimentation, this data processing process needs to be automated to allow for fully

autonomous optimization. Instrument-specific data reduction and correction processes perform the necessary transformations to convert raw detector data into 1-dimensional, intensity vs. scattering vector (q) data. Fully automated reduction and correction pipelines exist for many instruments, making it straightforward to attain reduced 1D data from a measurement. The larger challenge in integrating SAXS characterization into autonomous experiments is automation of the data analysis and interpretation process. This process traditionally requires extensive researcher involvement because it relies on external or a-priori information about the sample and expert model selection. Nevertheless, the scattering community is making significant progress in developing methods to automate this process. Possible approaches include machine-learning classifier guided model fitting[42–44], domain-specific structural solvers such as CREASE[45,46], and shape metric approaches to compare measured and target scattering[47]. While this problem is not addressed in this work, continued progress by the community is likely to enable SAXS integration into autonomous experimentation workflows in the near future.

6.2.3.2 Dynamic light scattering

Dynamic light scattering (DLS) was used to measure particle sizes in initial exploratory experiments and reproducibility verification campaigns. DLS measurements were made with a Malvern Instruments Zetasizer Nano ZS fitted with a 633nm laser. As-prepared samples were diluted in water before measurements. Analysis was performed with Malvern Zetasizer Software version 8.0. Z-average diameters and PDI are reported here.

6.2.3.3 Electron Microscopy

Scanning electron and transmission electron micrographs were collected from samples supported on a carbon grid. SEM images were collected on a Thermo Fisher Scientific Phenom Pharos

desktop instrument in secondary electron detector mode operated at 20 kV. Samples were prepared for SEM by dispersing as-prepared samples in ethanol then drop-casting them on the grid. Transmission electron microscopy (TEM) was performed using a Tecnai G2 F20 SuperTwin TEM at an operating voltage of 200kV using a Gatan Ultrascan CCD digital camera. TEM samples were first washed to remove surfactant using an acid wash procedure[26]. Washed samples were dispersed in ethanol and drop-cast onto grids.

6.2.3.4 Integrating automated experimentation with a synchrotron beamline

The integration of the Jubilee and NIST-AFL systems with the 12-ID-E USAXS beamline was a significant undertaking, and important details are worth discussing here to document decisions made and provide a reference point for similar future experiments. This experiment was scheduled consecutively with another Pozzo group USAXS experiment, providing 96 hours of continuous beamtime. The other experiment involved pre-prepared static samples that were relatively simple to set up and could be run uninterrupted. The concurrent scheduling allowed timing flexibility with the execution of automated sol-gel synthesis. When problems were encountered with the automated equipment, alternative samples could be swapped in quickly to make the best use of available beamtime. This allowed the 48 beamtime hours of this experiment to be spread over the 96 hours allocated to our group, enabling significant flexibility for troubleshooting. Pursuing a similar backup scheduling strategy is highly recommended for future work. This trip would not have resulted in useful sol-gel scattering data without this flexibility.

Transporting equipment to a national laboratory facility and operating it there presents logistical challenges that need to be addressed. Options for transporting equipment are to ship it via ground transportation or to fly with it as checked baggage. An effective system for flying with

the Jubilee platform has been developed. The NIST-AFL platform and auxiliary experimental equipment (e.g., vials, tools) were shipped. It is necessary to allow sufficient time for shipping and is recommended to ship equipment to a FedEx or UPS store in the vicinity of the national lab and to pick it up on the drive to the laboratory. This avoids issues with the facilities shipping receiving department. In addition, any hazardous chemicals needed for the experiment must be delivered well in advance. From experience on this trip, it is recommended that hazardous materials be scheduled for delivery to Argonne National Lab 3 business days before they are needed for the experiment to allow for intra-lab delivery. Plan well in advance to manage the hazmat shipping process for these materials. Arrangements also need to be made with the beamline staff to provide air and network access inside the instrument hutch. While the experiment safety form does ask about this, it is also suggested to directly communicate with the staff well in advance. The equipment complexity involved in this system is beyond the scope of most beamline experiments, so it is imperative to clearly communicate the experiment setup plans. Equipment used at APS also needs to pass an electrical safety inspection. Electrical equipment generally needs to be NRTL listed. The 24V power supply used on Jubilee and the NIST-AFL implementation used here is not listed and is also used in a non-approved manner. While the equipment used here was approved for use with some additional wiring protection, future APS trips should replace all 120V components with an NRTL listed 24V power supply. Future trips involving the Jubilee system will also need to develop shielding that covers the exposed belts and prevents users from putting appendages in the Jubilee deck space while the machine is active.

Integrating the automated synthesis system with the instrument to trigger measurements and retrieve data is an administrative and integration challenge. In principle, triggering

measurements should be trivial. At the time of this experiment, the USAXS instrument control was done with Bluesky. Through the NIST-AFL project, various drivers have been developed to allow for triggering over the network. The barrier to this is security restrictions on access to the national lab network. Due to these restrictions, it was impossible to place the control computer used for the sol-gel synthesis on the network or otherwise gain direct access to the USAXS control system. The circumvention developed on the fly for this experiment was to develop a cloud-brokered task messaging architecture. The Jubilee experiment control computer updated a file on an AWS EC2 instance when a sample was loaded and ready for measurement. A script running on the USAXS control machine monitored the file for changes and triggered the measurement when changes were detected. While more elegant solutions exist, this was a reliable solution implemented on a time crunch. Unfortunately, this solution will not work for future experiments due to more security restrictions that will prevent instrument control computers from communicating with the general internet. Future experiments will need to work with APS IT/security to clear the experiment control computer for access to the internal network. This will raise another issue: GPU access for closed-loop experiments. In this work, had an autonomous optimization been attempted, Amazon Web Services GPU resources would have been used to run the BoTorch based optimization workflow. Fully segregating the experiment control system from the general internet will preclude this resource. One possibility is to gain access to GPU resources at Argonne's computing user facilities to run the experiment. Data reduction is another challenge for fully automated experiments. At the time of this experiment, an automated data reduction pipeline for USAXS flyscan data was prototyped and available for testing. Again, it is imperative

to communicate with the beamline staff well in advance to make sure this capability is in place and available to integrate into the experimental workflow.

6.3 Results and discussion

6.3.1 Workflow validation

6.3.1.1 Assessing synthesis reproducibility

Two assessments were used to understand the reproducibility of the automated synthesis workflow. In the first, a sample composition observed to produce highly monodisperse, colloidally stable (non-aggregating) particles was replicated. Sample 1 (composition listed in table 6.2) was selected. This sample was replicated five times in a single synthesis run. Particle size was measured with DLS, and the mesoporous structure was characterized with SAXS. Across the five samples, the particle diameters were similar (average particle diameter 778 nm, with a standard deviation of 12nm), and the scattering data from the mesoporous region is nearly identical for all particles (figure 6.4). These results suggest that the synthesis workflow is capable of reproducibly synthesizing samples.

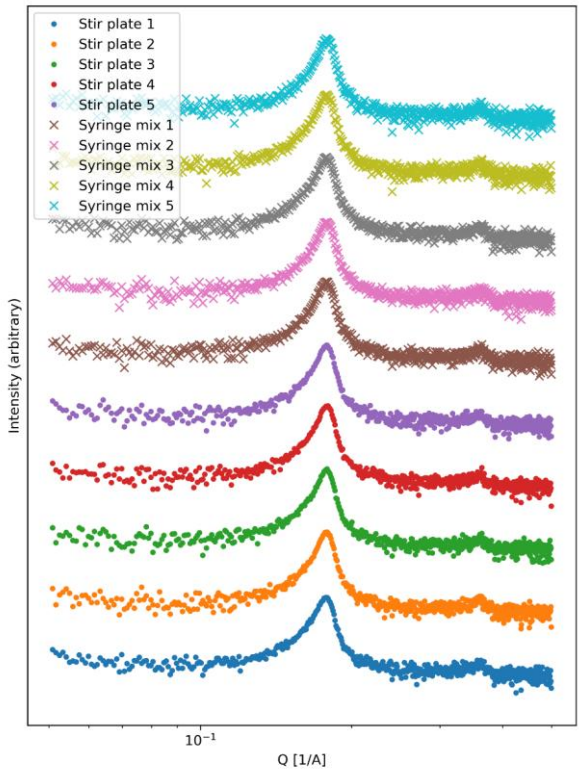


Figure 6.4: Mesophase SAXS scattering from reproducibility control experiment

In the second reproducibility check, control samples were synthesized throughout the main batch synthesis experiment (described below). Samples with identical compositions were made in triplicate at the start of the experiment, then at regular intervals over the course of the 24-hour synthesis campaign. The control composition was selected as the most monodisperse sample from early exploratory experiments and is shown in Table S1. The scattering data suggests all samples were polydisperse in diameter. While samples 1-4 were well dispersed, the scattering for sample 5 suggests particle aggregation (figure 6.5). This control experiment also indicates that the synthesis procedure is generally reproducible both between individual synthesis executions and over the course of a 24-hour campaign.

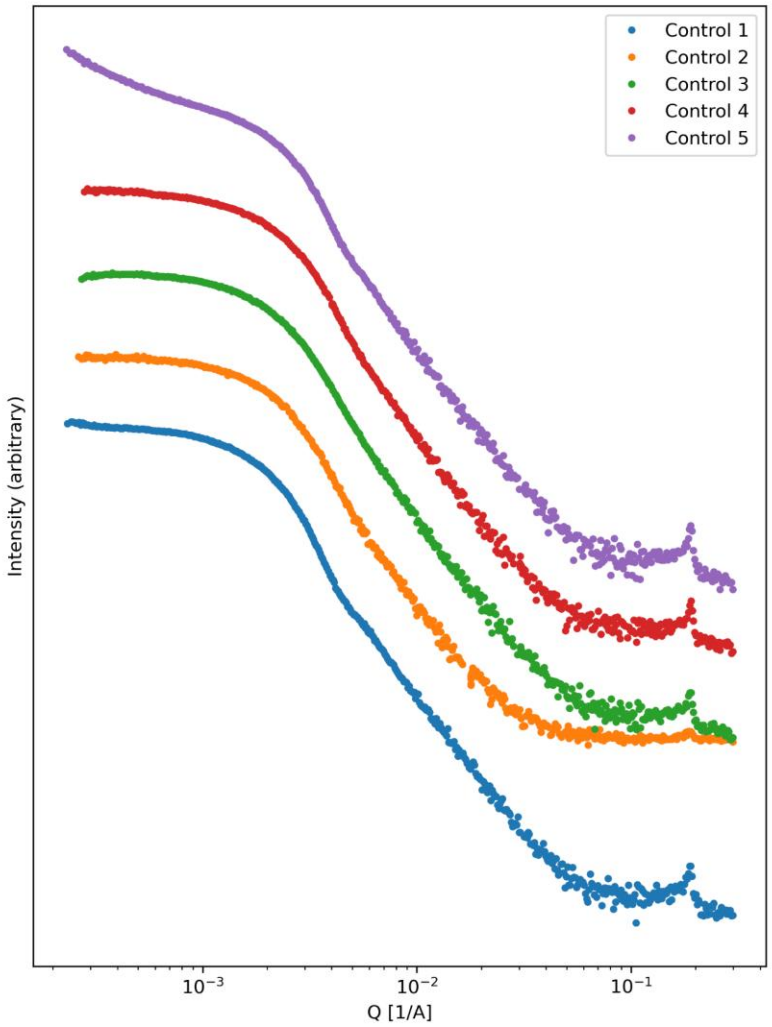


Figure 6.5: Controls from USAXS batch experiment.

Initial development experiments indicated that effective control of ammonia evaporation is important for achieving reproducibility in this synthesis as ammonia concentration has a large influence over particle size

and ammonia is highly volatile. Researchers working on similar sol-gel syntheses, whether using automated systems or traditional manual methods, should take precautions to prevent ammonia concentration changes due to evaporation from impacting results. In this work, silicone septa were used to minimize evaporation.

6.3.1.2 Comparison of this workflow to traditional stirred workflow

Many reports suggest that the stirring of the reaction mixture during and after TEOS addition is an important factor in determining the packing structure of the porous mesophase and potentially particle size[25,48]. The synthesis procedure presented in this work, however, replaces continuous stirring with repeated syringe aspirate/dispense cycles. To evaluate the impact of this procedural adaptation on sample morphologies, samples with identical compositions were synthesized using both traditional stirring and the syringe mixing used here. The composition of sample 1 (table 6.2) was used. For the traditionally stirred samples, the precursor mixture was prepared on the Jubilee platform, then TEOS was added with a manual pipette while stirring at 1000 RPM. Stirring was continued for 1 minute after TEOS addition then stopped. The Jubilee syringe mixed samples were prepared following the general procedure presented in this work. Samples were characterized with dynamic light scattering to measure particle size and SAXS to compare mesopore phase structure. The resulting sizes were similar between methods (726 +/- 45 nm for stirred, 778 +/- 12 nm for jubilee, n=5), and scattering from the porous structures was essentially identical (disordered porous structure, figure 6.5). This suggests that particles synthesized using this automated synthesis procedure are similar to those made with traditional manual procedures.

6.3.2 Exploring the design space for mesoporous silica nanoparticles

The automated synthesis platform was used to perform a batch parameter space exploration experiment. This experiment was performed at the APS 12-ID-E beamline. 63 samples were synthesized in the available experiment time from the 80 compositions generated with the constrained Sobol sampling method discussed above. Samples were continuously synthesized over a contiguous 24-hour period. Samples were characterized with USAXS using the custom liquid cartridge plates described in the methods section. Fully automated synthesis-characterization integration was achieved at 12-ID-E, and several samples were synthesized in this manner. However, the samples described here were prepared offline and measured in batch mode to more efficiently utilize beamtime for the initial screening experiment. A minimum of two hours elapsed between sample synthesis and USAXS measurement. To supplement the USAXS scattering data, additional SAXS data was collected from the same samples 11 days after synthesis on the Xenocs Xeuss 3.0 instrument. SEM and TEM images of selected samples were also collected 26-60 days after sample synthesis. Particle dispersion and size monodispersity are both important sample characteristics. This experiment campaign demonstrates that the automated synthesis presented is capable of producing stable, monodisperse particles, and that USAXS is an effective method for understanding these properties for colloidal silica samples in high throughput. Figure 5 presents USAXS scattering data from four selected samples. SEM images of these samples are also shown. The scattering data presented in Figure 6.6a suggests a sample consisting of monodisperse particles with minimal aggregation. Size monodispersity is indicated by the oscillations present in the intensity (y axis), and minimal aggregation is indicated by the turnover to a flat slope at low q

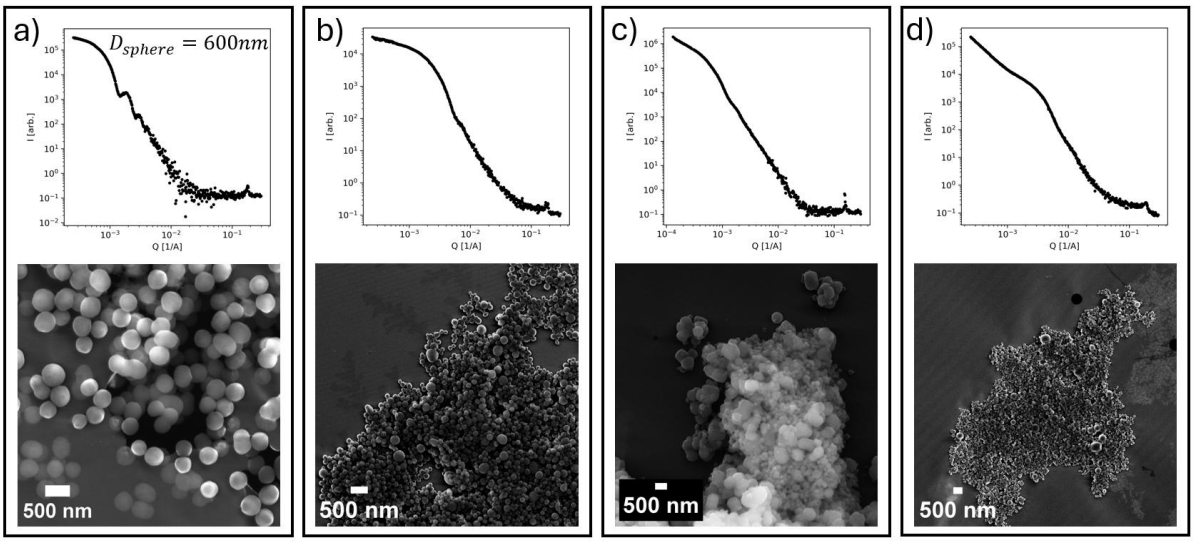


Figure 6.6: SEM images and USAXS scattering for selected samples. A) highly monodisperse sample (sample composition 1), b) Moderately monodisperse sample (sample composition 2), c) aggregated polydisperse sample (sample composition 3), and d) aggregated polydisperse sample (sample composition 4).

values (x axis) (left side of plot). The associated SEM image confirms that this sample contains uniformly sized, spherical particles with minimal aggregation. A polydisperse sphere model fit of the USAXS data suggests particles with a mean diameter of 602 nm. Figure 6.6b-d show samples whose scattering data suggests varying degrees of aggregation and polydispersity. The SEM images of these samples indeed confirm polydisperse, aggregated particles. Compositions of select samples discussed are shown in table 6.2.

Table 6.2: Compositions of selected samples

Sample	**TEOS vol. frac.**	**Ammonia vol. frac.**	**Ethanol Vol. Frac**	**[CTAB] [mg/mL]**	**[F127] [mg/mL]**	**Water vol. frac.**
1	0.008	0.018	0.37	5.0	6.9	0.60
2	0.019	0.064	0.21	2.8	6.3	0.70
3	0.034	0.055	0.21	7.5	0.5	0.70
4	0.029	0.022	0.23	4.2	5.3	0.72
5	0.011	0.059	0.29	5.4	1.8	0.64
6	0.007	0.021	0.40	7.0	1.0	0.57
7	0.020	0.026	0.28	8.5	5.4	0.68
8	0.012	0.020	0.27	3.6	11.9	0.70

6 out of the 63 synthesized samples have USAXS scattering data that is consistent with stable, size monodisperse particles. Polydisperse sphere model fits made with the Sasview software[49] indicate that the mean diameters of these samples range from 220 nm to 630 nm, and that their associated polydispersity indices range from 0.08 to 0.19. Model fits are shown in figures A1.3-A1.6.

The influence of sample composition on particle size monodispersity can be evaluated qualitatively. Figure 6.7 presents the relationship between sample composition and USAXS scattering by plotting the measured scattering data directly in the composition space. Sample compositions low in TEOS, high in ethanol, and low in F127 favor monodisperse particles. Monodisperse scattering data is primarily present in the lower left corner of the composition-

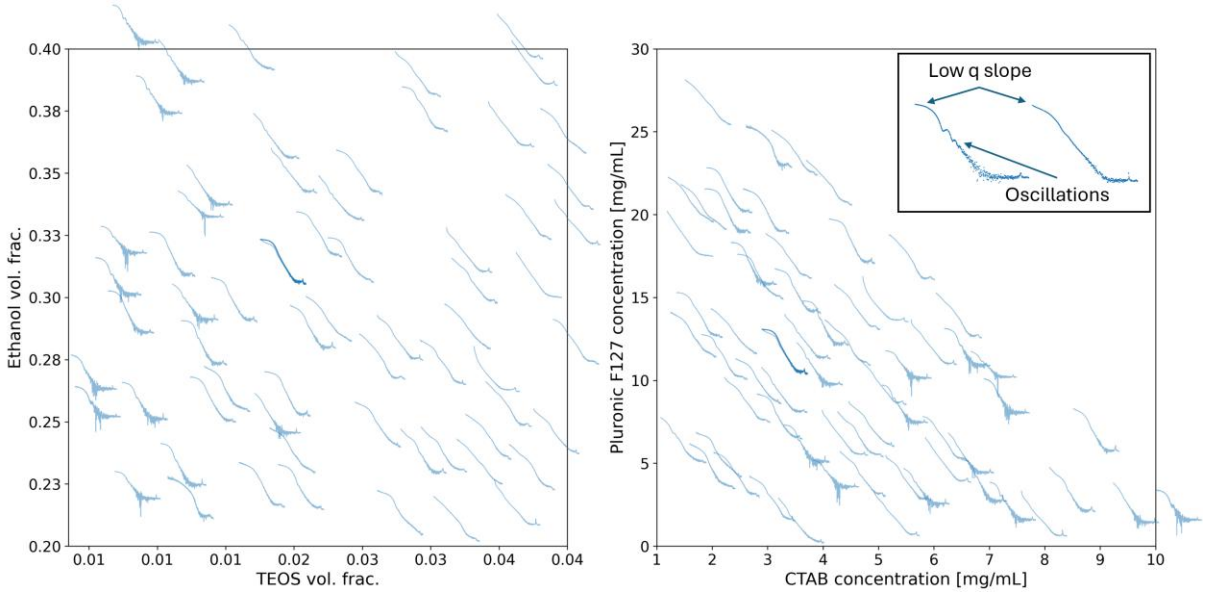


Figure 6.7: Composition-scattering plots for USAXS scattering data demonstrates relationships between composition and particle morphology. USAXS scattering data for each sample is plotted at a position corresponding to that sample composition with respect to the components in each axis. The inset highlights distinct features in the scattering data. Scattering from highly size monodisperse samples has clear oscillations while polydisperse samples lack this feature. Scattering from aggregated samples has a negative slope in the low-q region (the left portion of the scattering data) while stable samples have a slope approaching zero.

scattering plots, indicating low compositions. This campaign did not generate enough monodisperse samples to facilitate quantitative conclusions about composition – diameter relationships. Composition-aggregation relations can also be inferred from these plots by comparing the low-q slope of the scattering data. Stable particles are produced primarily from samples with low TEOS content. Other components do not significantly impact aggregation.

Samples with a variety of ordered internal pore mesophases were formed, including those with hexagonal *p6mm* and gyroid $Ia\bar{3}d$ cubic structures. These phases were identified by multiple

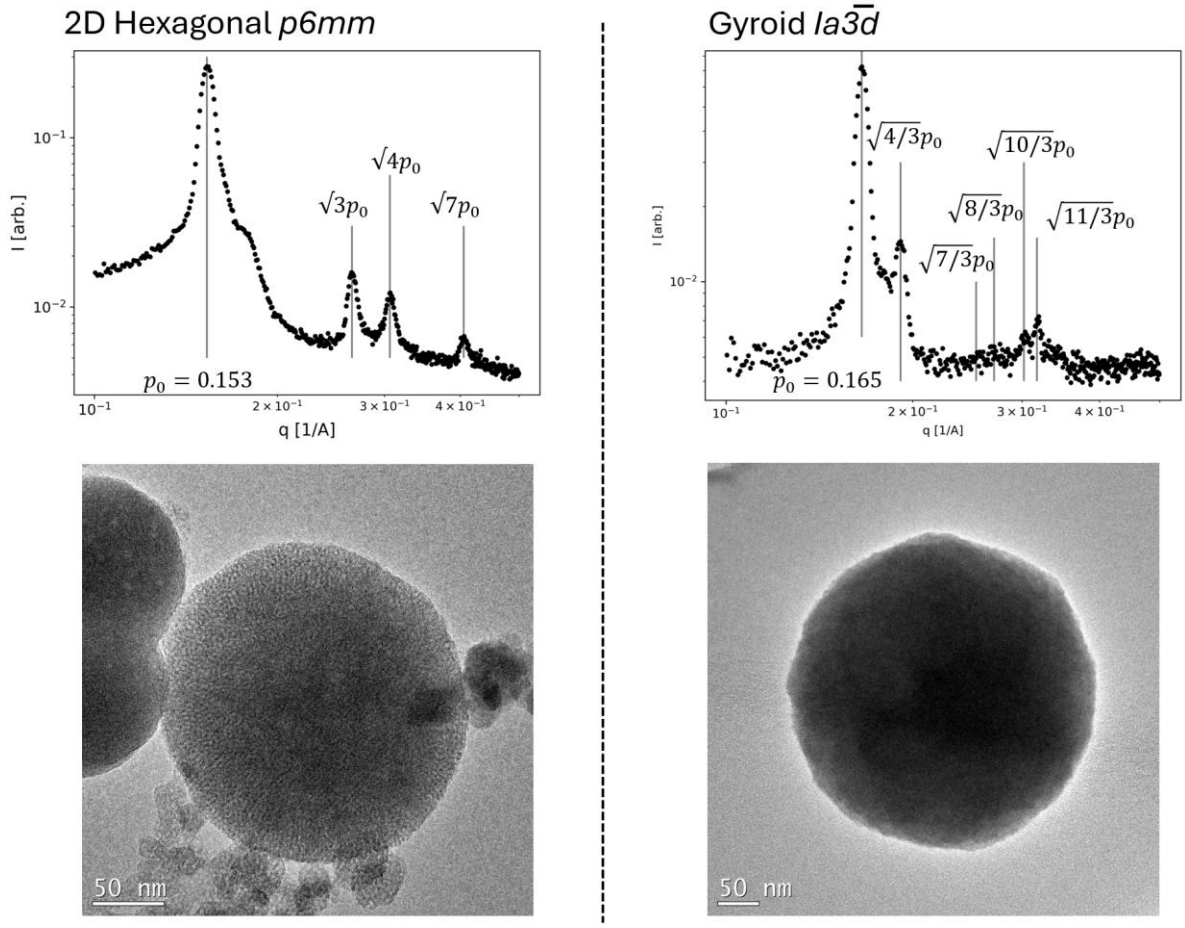


Figure 6.8: Examples of samples exhibiting ordered 2D hexagonal (Sample 3) and gyroid (Sample 5) internal pore orderings. Internal porosity is visible in a TEM image of the hexagonal sample, but cannot be seen in the gyroid sample.

sharp peaks that index to known peak spacings[50]. Additionally, many samples exhibit an unidentified two-peak mixed phase, and others show a single, often broad peak suggesting poorly ordered pores. Figure 6.8 shows examples of scattering from samples with ordered internal porous phases, along with TEM images showing their internal porous structure. Several samples do not have any high-q scattering peaks, suggesting a lack of ordered pore formation.

Qualitative composition – structure relationships for pore ordering can also be investigated with composition-scattering plots (figure 6.9). These plots show that high CTAB content is necessary to form ordered mesoporous phases, which is expected given CTAB's role as the micelle-forming template surfactant in this system. Low Pluronic F127 content also favors ordered pore phase formation. We hypothesize that high Pluronic F127 content may interfere with CTAB micelle formation to inhibit mesoporous structure growth. Pluronic F127 was originally included in this synthesis to minimize particle aggregation[25] However, stable samples were synthesized with

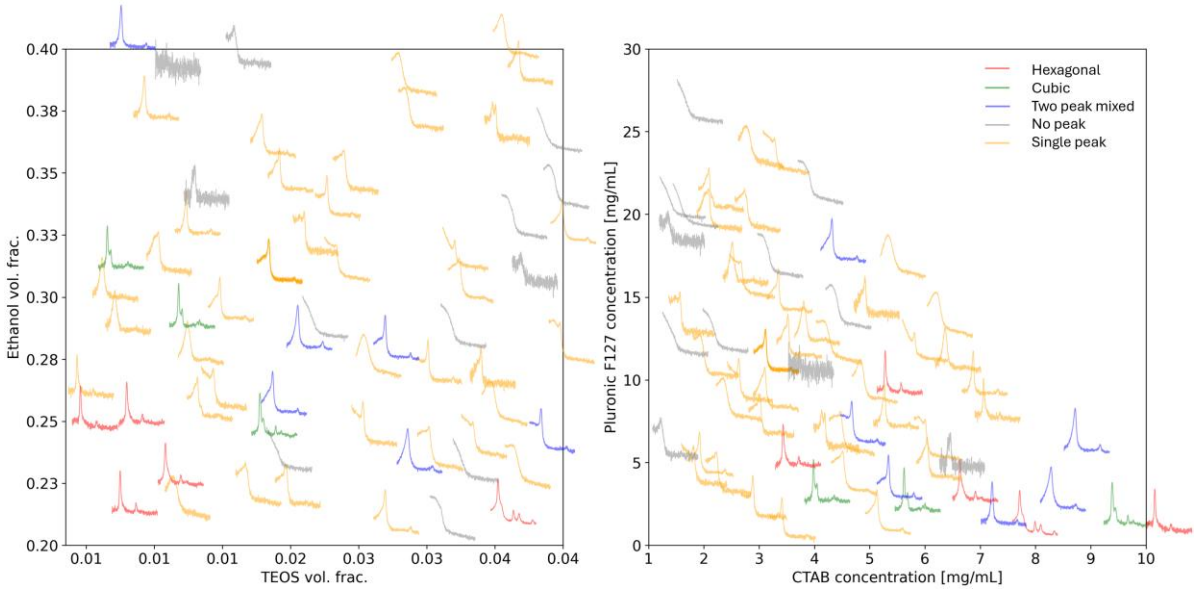


Figure 6.9: Composition-scattering plots for SAXS measurements of samples. This plot is analogous to figure 6 but shows high-q scattering associated with porous phase order from the same samples.

low Pluronic F127 content, suggesting that this may not be a necessary component of the system, or that the selected upper composition bound for this experiment needs to be lowered. Low ethanol and TEOS content with moderate to high ammonia content and high water content also favor highly ordered pore phases. These qualitative conclusions are supported by a SHAP feature importance analysis applied to a peak sharpness score. A peak sharpness score was defined as $Sharpness = \frac{Peak\ prominence}{Peak\ width}$ for the first peak in the SAXS data region (q>0.1 1/A). Peaks were identified using the scipy find_peaks function with parameters [prominence = 0.1, width = 10]. Prominence was calculated with the scipy peak_prominences function and width was calculated with the scipy peak_widths function. A SHAP feature importance analysis was used to help understand the impact of composition on peak sharpness[52]. An XGboost regressor was fit to predict peak sharpness from sample composition[53]. SHAP values were calculated from this model and are shown in figure 6.10. Figure 6.9 also suggests a compositional dependence for phase selection. In

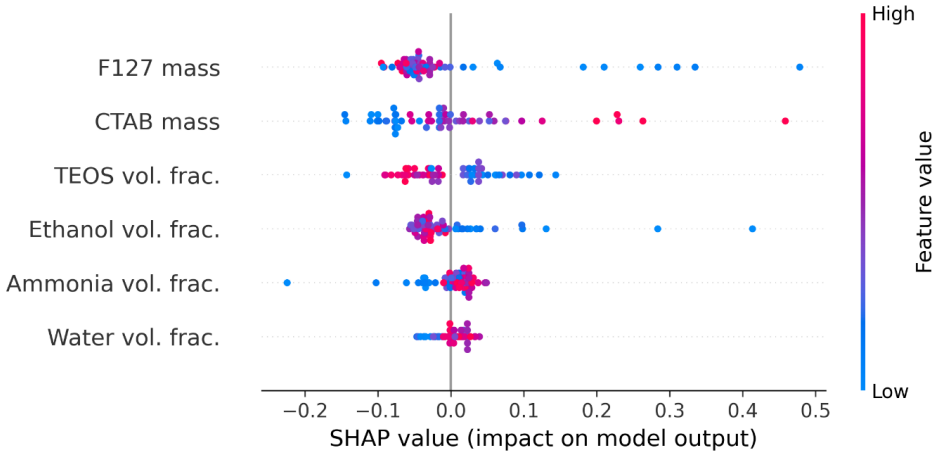


Figure 6.10: SHAP plot for peak sharpness analysis.

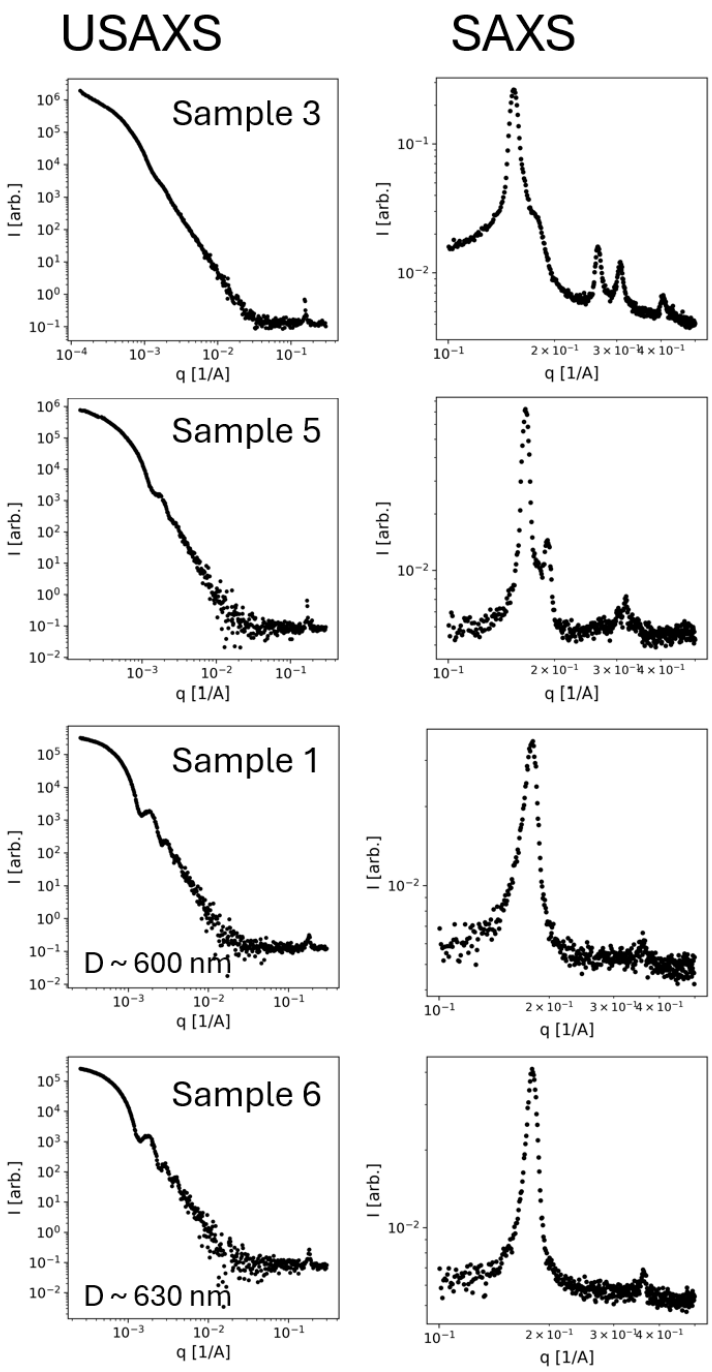


Figure 6.11: The most monodisperse samples observed have poorly ordered porous phases, and the most ordered porous phase samples observed occur in polydisperse to moderately monodisperse particles. However, moderate monodispersity in sample 28 suggests further optimization may yield monodisperse, highly ordered particles

particular, the ethanol-TEOS panel suggests that as ethanol and TEOS compositions are jointly increased, the porous phase arrangement undergoes transitions from hexagonal to gyroid to 2-peak mixed to disordered phases.

Comparing the composition-morphology relationships for overall particle morphology and pore phase order suggests the existence of a trade-off between compositions that generate monodisperse, dispersed particles and those that generate highly ordered pore phases. For example, a high ethanol content supports monodisperse particles, but a low content is associated with ordered pore phases. This effect can be observed in figure

6.11, which plots USAXS and SAXS data for examples of highly monodisperse samples as highly pore-ordered samples. This shows that the observed monodisperse samples have moderate pore ordering, while the highly ordered samples have at best moderate polydispersity. This suggests that further investigation, potentially with an autonomous optimization approach, would be necessary to identify compositions supporting both high diameter monodispersity as well as highly ordered pore phases.

6.4 Conclusions

An automated synthesis and characterization system was developed for sol-gel nanomaterial synthesis and was used to study the composition space of a mesoporous colloidal silica synthesis procedure. This study demonstrated that the platform is capable of synthesizing size monodisperse mesoporous silica particles with a wide range of particle sizes and ordered pore phases. This work also shows that small-angle X-ray scattering is an effective characterization method for automated sol-gel synthesis experiments.

This system and the initial exploratory dataset described in this work will enable future research to understand and optimize synthesis conditions for mesoporous colloidal silica and other sol-gel synthesized nanomaterials. The initial exploration conducted here will allow further investigations to study 'interesting' parameter spaces with active learning approaches. Possible objectives include phase boundary mapping to develop a better understanding of the conditions that lead to specific pore phase ordering, and morphology optimization to discover synthesis conditions for size monodisperse, colloidally stable particles with targeted particle sizes and pore structures. Many opportunities exist to expand on the synthesis conditions investigated here, for example by incorporating additional surfactants or adding a reaction quenching step to the

synthesis. Our hope is also that other researchers adapt and extend this platform to accelerate research across a diverse range of sol-gel chemistries beyond the colloidal silicas studied here. The open-hardware nature of this system makes it possible for researchers to explore new synthesis steps and conditions or alternative characterization techniques by building off the existing capabilities discussed here. This integration of new accelerated experimentation methods with traditional sol-gel synthesis techniques will yield exciting advancements in nanomaterials development.

6.5 Data, code, and hardware availability

Complete experiment information is available in an associated GitHub repository at https://github.com/pozzo-research-group/automated-mesoporous-silica, including the code and detailed procedures used to execute the experiments described here and resulting X-ray scattering data. Extensive hardware documentation for the Science-Jubilee project is available at https://science-jubilee.readthedocs.io/en/latest/, and the NIST-AFL implementation described here is documented at https://github.com/pozzo-research-group/AFL-sample-loader.

6.6 References

(1) Wang, D.; Chen, X.; Feng, J.; Sun, M. Recent Advances of Ordered Mesoporous Silica Materials for Solid-Phase Extraction. *J. Chromatogr. A* **2022**, *1675*, 463157. https://doi.org/10.1016/j.chroma.2022.463157.

(2) Gallis, K. W.; Araujo, J. T.; Duff, K. J.; Moore, J. G.; Landry, C. C. The Use of Mesoporous Silica in Liquid Chromatography. *Adv. Mater.* **1999**, *11* (17), 1452–1455. https://doi.org/10.1002/(SICI)1521-4095(199912)11:17<1452::AID-ADMA1452>3.0.CO;2-R.

(3) Flynn, E. J.; Keane, D. A.; Tabari, P. M.; Morris, M. A. Pervaporation Performance Enhancement through the Incorporation of Mesoporous Silica Spheres into PVA Membranes. *Sep. Purif. Technol.* **2013**, *118*, 73–80. https://doi.org/10.1016/j.seppur.2013.06.034.

(4) Vallet-Regí, M.; Schüth, F.; Lozano, D.; Colilla, M.; Manzano, M. Engineering Mesoporous Silica Nanoparticles for Drug Delivery: Where Are We after Two Decades? *Chem. Soc. Rev.* **2022**, *51* (13), 5365–5451. https://doi.org/10.1039/D1CS00659B.

(5) López, V.; Villegas, M. R.; Rodríguez, V.; Villaverde, G.; Lozano, D.; Baeza, A.; Vallet-Regí, M. Janus Mesoporous Silica Nanoparticles for Dual Targeting of Tumor Cells and Mitochondria. *ACS Appl. Mater. Interfaces* **2017**, *9* (32), 26697–26706. https://doi.org/10.1021/acsami.7b06906.

(6) Xiong, L.; Du, X.; Kleitz, F.; Qiao, S. Z. Cancer-Cell-Specific Nuclear-Targeted Drug Delivery by Dual-Ligand-Modified Mesoporous Silica Nanoparticles. *Small Weinh. Bergstr. Ger.* **2015**, *11* (44), 5919–5926. https://doi.org/10.1002/smll.201501056.

(7) Lai, C.-Y.; Trewyn, B. G.; Jeftinija, D. M.; Jeftinija, K.; Xu, S.; Jeftinija, S.; Lin, V. S.-Y. A Mesoporous Silica Nanosphere-Based Carrier System with Chemically Removable CdS Nanoparticle Caps for Stimuli-Responsive Controlled Release of Neurotransmitters and Drug Molecules. J. Am. Chem. Soc. 2003, 125 (15), 4451–4459. https://doi.org/10.1021/ja028650l.

(8) Bouchoucha, M.; Côté, M.-F.; C.-Gaudreault, R.; Fortin, M.-A.; Kleitz, F. Size-Controlled Functionalized Mesoporous Silica Nanoparticles for Tunable Drug Release and Enhanced Anti-Tumoral Activity. *Chem. Mater.* **2016**, *28* (12), 4243–4258. https://doi.org/10.1021/acs.chemmater.6b00877.

(9) *Excipients*. https://grace.com/industries/pharmaceutical-solutions/purification-formulation-delivery/excipients/, https://grace.com/industries/pharmaceutical-solutions/purification-formulation-delivery/excipients/ (accessed 2025-05-06).

(10) Mohan, A.; Jaison, A.; Lee, Y.-C. Emerging Trends in Mesoporous Silica Nanoparticle-Based Catalysts for CO2 Utilization Reactions. *Inorg. Chem. Front.* **2023**, *10* (11), 3171–3194. https://doi.org/10.1039/D3QI00378G.

(11) Yu, X.; T. Williams, C. Recent Advances in the Applications of Mesoporous Silica in Heterogeneous Catalysis. *Catal. Sci. Technol.* **2022**, *12* (19), 5765–5794. https://doi.org/10.1039/D2CY00001F.

(12) Wang, X.; Zhu, L.; Zhuo, Y.; Zhu, Y.; Wang, S. Enhancement of CO2 Methanation over La-Modified Ni/SBA-15 Catalysts Prepared by Different Doping Methods. *ACS Sustain. Chem. Eng.* **2019**, *7* (17), 14647–14660. https://doi.org/10.1021/acssuschemeng.9b02563.

(13) Wang, X.; Xuan, X.; Wang, Y.; Li, X.; Huang, H.; Zhang, X.; Du, X. Nano-Au-Modified TiO2 Grown on Dendritic Porous Silica Particles for Enhanced CO2 Photoreduction. *Microporous Mesoporous Mater.* **2021**, *310*, 110635. https://doi.org/10.1016/j.micromeso.2020.110635.
(14) Pal, M.; Ganesan, V. Zinc Phthalocyanine and Silver/Gold Nanoparticles Incorporated MCM-41 Type Materials as Electrode Modifiers. *Langmuir* **2009**, *25* (22), 13264–13272. https://doi.org/10.1021/la901792b.
(15) Stöber, W.; Fink, A.; Bohn, E. Controlled Growth of Monodisperse Silica Spheres in the Micron Size Range. *J. Colloid Interface Sci.* **1968**, *26* (1), 62–69. https://doi.org/10.1016/0021-9797(68)90272-5.
(16) Narayan, R.; Nayak, U. Y.; Raichur, A. M.; Garg, S. Mesoporous Silica Nanoparticles: A Comprehensive Review on Synthesis and Recent Advances. *Pharmaceutics* **2018**, *10* (3), 118. https://doi.org/10.3390/pharmaceutics10030118.
(17) Hollamby, M. J.; Borisova, D.; Brown, P.; Eastoe, J.; Grillo, I.; Shchukin, D. Growth of Mesoporous Silica Nanoparticles Monitored by Time-Resolved Small-Angle Neutron Scattering. *Langmuir* **2012**, *28* (9), 4425–4433. https://doi.org/10.1021/la203097x.
(18) Kresge, C. T.; Leonowicz, M. E.; Roth, W. J.; Vartuli, J. C.; Beck, J. S. Ordered Mesoporous Molecular Sieves Synthesized by a Liquid-Crystal Template Mechanism. *Nature* **1992**, *359* (6397), 710–712. https://doi.org/10.1038/359710a0.
(19) Zhao, D.; Feng, J.; Huo, Q.; Melosh, N.; Fredrickson, G. H.; Chmelka, B. F.; Stucky, G. D. Triblock Copolymer Syntheses of Mesoporous Silica with Periodic 50 to 300 Angstrom Pores. *Science* **1998**, *279* (5350), 548–552. https://doi.org/10.1126/science.279.5350.548.
(20) Schüth, F. The Evolution of Ordered Mesoporous Materials. In *Studies in Surface Science and Catalysis*; Terasaki, O., Ed.; Mesoporous Crystals and Related Nano-Structured Materials; Elsevier, 2004; Vol. 148, pp 1–13. https://doi.org/10.1016/S0167-2991(04)80190-3.
(21) Lu, Y.; Fan, H.; Stump, A.; Ward, T. L.; Rieker, T.; Brinker, C. J. Aerosol-Assisted Self-Assembly of Mesostructured Spherical Nanoparticles. *Nature* **1999**, *398* (6724), 223–226. https://doi.org/10.1038/18410.
(22) Vallet-Regi, M.; Rámila, A.; del Real, R. P.; Pérez-Pariente, J. A New Property of MCM-41: Drug Delivery System. *Chem. Mater.* **2001**, *13* (2), 308–311. https://doi.org/10.1021/cm0011559.
(23) Shen, D.; Yang, J.; Li, X.; Zhou, L.; Zhang, R.; Li, W.; Chen, L.; Wang, R.; Zhang, F.; Zhao, D. Biphase Stratification Approach to Three-Dimensional Dendritic Biodegradable Mesoporous Silica Nanospheres. *Nano Lett.* **2014**, *14* (2), 923–932. https://doi.org/10.1021/nl404316v.
(24) Kim, C.; Yoon, S.; Lee, J. H. Facile Large-Scale Synthesis of Mesoporous Silica Nanoparticles at Room Temperature in a Monophasic System with Fine Size Control. *Microporous Mesoporous Mater.* **2019**, *288*, 109595. https://doi.org/10.1016/j.micromeso.2019.109595.
(25) Kim, T.-W.; Chung, P.-W.; Lin, V. S.-Y. Facile Synthesis of Monodisperse Spherical MCM-48 Mesoporous Silica Nanoparticles with Controlled Particle Size. *Chem. Mater.* **2010**, *22* (17), 5093–5104. https://doi.org/10.1021/cm1017344.

(26) Noureddine, A.; Maestas-Olguin, A.; Tang, L.; Corman-Hijar, J. I.; Olewine, M.; Krawchuck, J. A.; Tsala Ebode, J.; Edeh, C.; Dang, C.; Negrete, O. A.; Watt, J.; Howard, T.; Coker, E. N.; Guo, J.; Brinker, C. J. Future of Mesoporous Silica Nanoparticles in Nanomedicine: Protocol for Reproducible Synthesis, Characterization, Lipid Coating, and Loading of Therapeutics (Chemotherapeutic, Proteins, siRNA and mRNA). *ACS Nano* **2023**, *17* (17), 16308–16325. https://doi.org/10.1021/acsnano.3c07621.
(27) Machineagency/Science_jubilee, 2023. https://github.com/machineagency/science_jubilee (accessed 2024-02-05).
(28) Vasquez, S.; Twigg-Smith, H.; Tran O'Leary, J.; Peek, N. Jubilee: An Extensible Machine for Multi-Tool Fabrication. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*; CHI '20; Association for Computing Machinery: New York, NY, USA, 2020; pp 1–13. https://doi.org/10.1145/3313831.3376425.
(29) Yoshikawa, N.; Darvish, K.; Vakili, M. G.; Garg, A.; Aspuru-Guzik, A. Digital Pipette: Open Hardware for Liquid Transfer in Self-Driving Laboratories. *Digit. Discov.* **2023**, *2* (6), 1745–1751. https://doi.org/10.1039/D3DD00115F.
(30) Beaucage, P. A.; Martin, T. B. The Autonomous Formulation Laboratory: An Open Liquid Handling Platform for Formulation Discovery Using X-ray and Neutron Scattering. *Chem. Mater.* **2023**, *35* (3), 846–852. https://doi.org/10.1021/acs.chemmater.2c03118.
(31) Pelkie, B.; Baird, S.; Aissi, E.; Aspuru-Takata, K.; Cao, Y.; Chang, J. H.; Gambhir, K.; Hale, W. S.; Hao, L.; Hattrick, C.; Hein, J.; Luo, D.; Melville, O.; Ngan, M.; Nyeland, L. L. B.; Peek, N.; Politi, M.; Rajkumar, E. E.; Siemenn, A.; Subbaraman, B.; Vasquez, S.; Watchorn, J.; Zhang, W.; Ziskason, R.; Pozzo, L.; Buonassisi, T.; Vegge, T. Democratizing Self-Driving Labs through User-Developed Automation Infrastructure. ChemRxiv February 12, 2025. https://doi.org/10.26434/chemrxiv-2025-zhkrf.
(32) *Jubilee*. https://jubilee3d.com/index.php?title=Main_Page (accessed 2025-05-13).
(33) *Science Jubilee — Science Jubilee 0.3.2.post1.dev170+g0ee6bd1 documentation*. https://science-jubilee.readthedocs.io/en/latest/index.html (accessed 2025-06-02).
(34) M, S. I. The Distribution of Points in a Cube and the Approximate Evaluation of Integrals. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 86–112.
(35) *Sobol — SciPy v1.15.3 Manual*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.qmc.Sobol.html#re15be05a07a0-1 (accessed 2025-06-02).
(36) Usnistgov/AFL-Automation, 2025. https://github.com/usnistgov/AFL-automation (accessed 2025-05-13).
(37) Usnistgov/AFL-Hardware, 2024. https://github.com/usnistgov/AFL-hardware (accessed 2025-05-13).
(38) *Automation-Hardware/Cartridge Sample Holder for SAS Experiments/SAXS-USAXS Liquids 48 well plate holder at master · pozzo-research-group/Automation-Hardware*. https://github.com/pozzo-research-group/Automation-Hardware/tree/master/Cartridge%20Sample%20Holder%20for%20SAS%20Experiments/SAXS-USAXS%20Liquids%2048%20well%20plate%20holder (accessed 2025-05-05).
(39) Ilavsky, J.; Zhang, F.; Andrews, R. N.; Kuzmenko, I.; Jemian, P. R.; Levine, L. E.; Allen, A. J. Development of Combined Microstructure and Structure Characterization Facility for in

Situ and Operando Studies at the Advanced Photon Source. *J. Appl. Crystallogr.* **2018**, *51* (3), 867–882. https://doi.org/10.1107/S160057671800643X.
(40) *Automation-Hardware/Cartridge Sample Holder for SAS Experiments/SAXS-USAXS_AntonPaar_FlowCellHolder/USAXS_flow cell holder v5.stl at master · pozzo-research-group/Automation-Hardware*. GitHub. https://github.com/pozzo-research-group/Automation-Hardware/blob/master/Cartridge%20Sample%20Holder%20for%20SAS%20Experiments/SAXS-USAXS_AntonPaar_FlowCellHolder/USAXS_flow%20cell%20holder%20v5.stl (accessed 2025-05-05).
(41) *pozzo-research-group/automated-MSN-synthesis: Documentation and example code for forthcoming automated mesoporous silica nanoparticle synthesis paper*. https://github.com/pozzo-research-group/automated-MSN-synthesis (accessed 2025-06-02).
(42) Roberts, G.; Nieh, M.-P.; W.K. Ma, A.; Yang, Q. Automated Structural Analysis of Small Angle Scattering Data from Common Nanoparticles via Machine Learning. *Digit. Discov.* **2025**. https://doi.org/10.1039/D5DD00059A.
(43) Tomaszewski, P.; Yu, S.; Borg, M.; Rönnols, J. Machine Learning-Assisted Analysis of Small Angle X-ray Scattering. In *2021 Swedish Workshop on Data Science (SweDS)*; 2021; pp 1–6. https://doi.org/10.1109/SweDS53855.2021.9638297.
(44) Archibald, R. K.; Doucet, M.; Johnston, T.; Young, S. R.; Yang, E.; Heller, W. T. Classifying and Analyzing Small-Angle Scattering Data Using Weighted k Nearest Neighbors Machine Learning Techniques. *J. Appl. Crystallogr.* **2020**, *53* (2), 326–334. https://doi.org/10.1107/S1600576720000552.
(45) Akepati, S. V. R.; Gupta, N.; Jayaraman, A. Computational Reverse Engineering Analysis of the Scattering Experiment Method for Interpretation of 2D Small-Angle Scattering Profiles (CREASE-2D). *JACS Au* **2024**, *4* (4), 1570–1582. https://doi.org/10.1021/jacsau.4c00068.
(46) Ye, Z.; Wu, Z.; Jayaraman, A. Computational Reverse Engineering Analysis for Scattering Experiments (CREASE) on Vesicles Assembled from Amphiphilic Macromolecular Solutions. *JACS Au* **2021**, *1* (11), 1925–1936. https://doi.org/10.1021/jacsau.1c00305.
(47) Vaddi, K.; Chiang, H. T.; Pozzo, L. D. Autonomous Retrosynthesis of Gold Nanoparticles via Spectral Shape Matching. *Digit. Discov.* **2022**, *1* (4), 502–510.
(48) Candela-Noguera, V.; Alfonso, M.; Amorós, P.; Aznar, E.; Marcos, M. D.; Martínez-Máñez, R. In-Depth Study of Factors Affecting the Formation of MCM-41-Type Mesoporous Silica Nanoparticles. *Microporous Mesoporous Mater.* **2024**, *363*, 112840. https://doi.org/10.1016/j.micromeso.2023.112840.
(49) SasView. *SasView*. https://sasview.github.io/ (accessed 2025-05-12).
(50) Giacomelli, C.; Borsali, R. Disordered Phase and Self-Organization of Block Copolymer Systems. In *Soft Matter Characterization*; Borsali, R., Pecora, R., Eds.; Springer Netherlands: Dordrecht, 2008; pp 133–189. https://doi.org/10.1007/978-1-4020-4465-6_3.
(51) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.;

van Mulbregt, P. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17* (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.
(52) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD '16; Association for Computing Machinery: New York, NY, USA, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.
(53) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017; Vol. 30.

7 Toward autonomous experimentation for morphological optimization of silica nanoparticles

7.1 Introduction

Fully automated synthesis and characterization of mesoporous nanomaterials was demonstrated in the previous chapter. This automated approach to studying nanomaterials enables sample synthesis in high throughput, which on its own enables high-resolution grid studies to understand the phase behavior of porous phase ordering, accelerated screening of previously unstudied templating surfactants, and more replicates to better understand the reproducibility and sensitivity of synthesis processes. However, many important questions pertaining to porous nanomaterial development could be more effectively addressed with an *autonomous* experimentation approach that integrates an active experimental design strategy with the automated experiment execution platform to selectively guide experiments in pursuit of a particular goal. In the context of mesoporous colloidal particles, properties of interest include pore ordering and connectivity, pore size, particle size and size polydispersity, and dispersibility, among others[1]. Precise retrosynthetic control over these properties is needed when developing a mesoporous material for specific applications. For example, a particle for drug delivery may need to be below 100 nm in diameter to exhibit enhanced permeability and retention[2], while containing sufficiently large pores to retain an active pharmaceutical ingredient at a controlled overall porosity to determine loading. A porous nanomaterial designed for use in catalysis, on the other hand, may need to exhibit specific pore ordering and associated connectivity to control mass transfer properties with particle size being of little importance[3]. A self-driving lab for sol-gel synthesis would provide a tool to optimize synthesis conditions for property outcomes.

To extend the automated experimentation process described in chapter 6 into an autonomously operating system, additional components are needed. The automated experimentation system is capable of synthesizing and characterizing particles. The input to this process is a sample composition, and the output is reduced one-dimensional scattering data. To integrate these capabilities into an autonomous system, scattering data processing and experimental design components are needed at a minimum. These components will enable the closed-loop selection of experiments without requiring researcher decision making. To perform truly autonomous experiments without human involvement, software orchestration to integrate and manage all interrelated components is also needed. However, human-orchestrated 'autonomous' experimentation, in which individual components are managed manually by a researcher, is also a viable approach to implementing such systems[4].

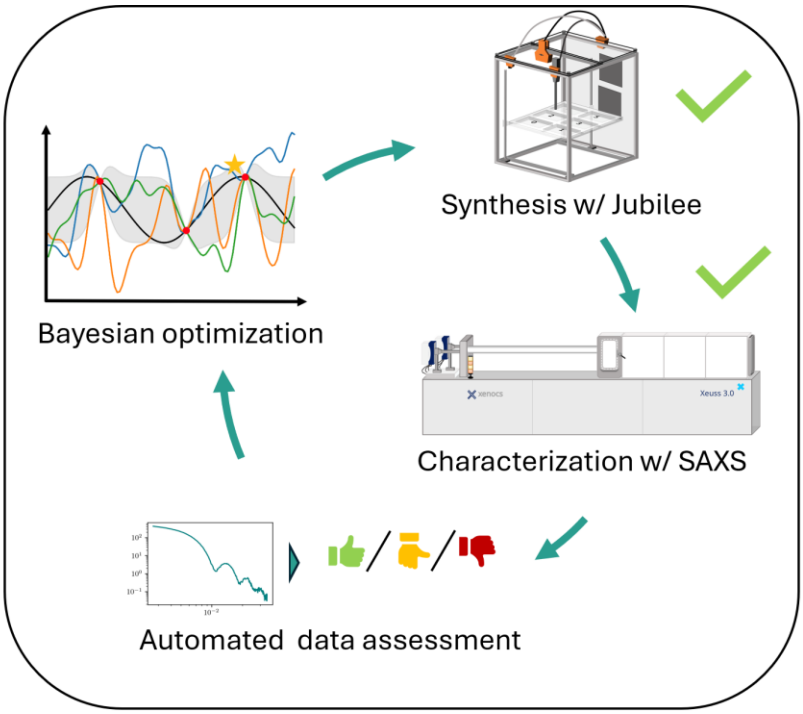


Figure 7.1: Necessary components to enable closed-loop optimization of

Implementing automated small-angle scattering data analysis is by far the most difficult and important task to enable autonomous experimentation for experiments using scattering characterization methods. Fully automating small-angle scattering data analysis to provide machine-learning ready information is a significant barrier to the successful development of small-angle scattering based autonomous experimentation systems[5]. Once reliable data analysis methodologies are developed, they can be coupled with an appropriate active experimental design strategy such as Bayesian optimization to enable autonomous experiment selection. Integrating these components with the experiment execution workflow using an orchestration software system will be challenging but is relatively approachable compared to the needs for data processing capability development.

In this chapter, progress toward the execution of autonomous experiments for the optimization of silica nanoparticle morphologies is discussed. This work focuses on the optimization of particle diameter for solid spherical silica nanoparticles, in contrast to the mesoporous colloidal particles of the previous chapter. These particles are synthesized in a similar manner to mesoporous particles, except that the surfactant micelle templates of the mesoporous synthesis are excluded. This synthesis is known colloquially as the Stöber synthesis, after its original developer[6]. Solid spherical particles provide a simpler starting point for developing the additional capabilities required for autonomous optimization. They involve optimizing over a composition space with 3 independent parameters instead of 5 for mesoporous. Optimizing solid particle morphology also avoids the additional complexity of analyzing porous contributions to scattering and balancing multiple objectives during optimization.

An appropriate X-ray scattering data processing method has been identified, performant experimental and Bayesian optimization configurations have been selected through a virtual instrument optimization simulator, experimental optimization campaigns have been executed, and experimental reproducibility challenges have been addressed. While a successful morphology optimization campaign has not been demonstrated, the components to enable one have been developed and validated. The order of work presented here has been selected to best describe the findings and results of this work and is not a historical recounting. Chronologically, the work presented in this chapter predates the mesoporous work in chapter 6. The workflow for solid spherical nanoparticle synthesis was developed initially. After some initial validation experiments, several experimental optimization campaigns were run. Concerns over the performance of these campaigns prompted an investigation into campaign structure with an in-silico simulator. This simulation suggested that experimental reproducibility is an important factor for convergence, which prompted an investigation into reproducibility. Continued development of the autonomous experimentation work was deprioritized in favor of advancing the automated mesoporous synthesis, as mesoporous particles were of greater interest for practical applications and for a scheduled synchrotron beamline trip.

7.2 Synthesis and characterization workflow

7.2.1 Experimental synthesis methods

The solid spherical nanoparticle synthesis process is analogous to the synthesis of mesoporous nanoparticles described in chapter 6, except that the surfactants (Pluronic F127 and cetyltrimethylammonium bromide (CTAB)) are omitted. Also, ethanol serves as the primary dispersant, and water is added in small quantities to participate in the tetraethyl orthosilicate

(TEOS) hydrolysis reaction, in contrast to the mesoporous synthesis where water and ethanol are both present in large quantities. The synthesis proceeds in a similar fashion: Ethanol, water, and ammonia are added to a reaction vial and mixed. TEOS is added to the precursor mixture and mixed. For appropriate synthesis compositions, colloidally stable particles are formed over a few hours. The Jubilee platform was used to automate this synthesis. As with the mesoporous synthesis, Digital Pipette syringe tools were used to perform liquid handling. The synthesis was initially performed using disposable plastic syringes for all components. After identifying compatibility issues between the rubber seal on the syringe plunger and TEOS, the TEOS and ammonia syringes were replaced with glass syringes for later experiments. Early on, it was identified that reaction vials needed to be sealed to prevent evaporation of the high ethanol content reactant solution over the extended (5 hours) reaction time. Later experiments also suggest that evaporation control is critically important to prevent ammonia evaporation from impacting particle size. Various options were tested for evaporation prevention including well plates with sealing mats or sealing stickers and vials with septa caps. Ultimately, 2 mL glass vials with snap-top pre-slit septa caps were selected as reaction vials, and 20 mL scintillation vials with pre-slit silicone septa caps were used to store stock solutions on the Jubilee deck. A 1,700 μL total sample volume was used for all samples. This value was selected as the largest volume that did not lead to liquid ejection from the vial during the syringe mixing process. Implications of reaction volume for experimental reproducibility are discussed later in the chapter. The use of septa-covered vials does preclude the use of disposable tip pipettes like the Opentrons OT2 pipette adaptation for Jubilee. Friction between the pipette tip and the silicone septa causes labware to be pulled out of its deck slot or the pipette tip to be removed from the pipette when the pipette is withdrawn from the vial.

This friction issue is still present, albeit to a lesser extent, with the syringe tips used with the digital pipette tools. The labware lifting issue can be mitigated by using a retainer lid on the vial labware and mounting the labware to the deck, and the syringe tip is not unintentionally removed from the syringe because it is threaded on with a Luer-lock fitting. Custom 3D printed labware holders were used to hold the stock and reaction vials, which are available on the Pozzo research group automation hardware repository[7]. These holders incorporated a lid that held the vials in place and could be screwed to the Jubilee deck to prevent movement. Samples were synthesized in a 'reactant-wise' batch mode, meaning that each reactant was dispensed to each vial in order of reactants in contrast to the 'sample-wise' synthesis order used for the mesoporous synthesis. A typical synthesis was performed by 1) preparing the Jubilee deck with fresh stock solutions and reaction vials, and loading the syringes with stock solution, 2) dispensing ethanol to reaction vials using a 10 cc syringe, 3) dispensing water to all reaction vials using a 1 cc syringe, 4) dispensing ammonium hydroxide solution to all reaction vials using a 1 cc syringe, 5) mixing each reaction vial with consecutive aspirate/dispense cycles using the same 10 cc syringe used for ethanol delivery, followed by rinsing the syringe by aspirating/dispensing from a series of 3 ethanol rinse vials, and 6) Dispensing TEOS from a dedicated syringe, then immediately mixing the sample with the 10 cc syringe before rinsing the syringe in ethanol. After TEOS addition, 5 hours were allowed for the nanoparticle synthesis to proceed. After 5 hours, the samples were diluted in preparation for SAXS measurements. A dilution ratio for each sample was calculated based on the TEOS composition of the sample and a target silicon loading. Each sample was diluted in ethanol by transferring the appropriate volume of sample to a new dedicated dilution vial and adding a balance volume of ethanol. The dilution concentration used was 0.1 mol silicon per liter of diluted

sample. This value was determined by comparing multiple dilution amounts for early samples and selecting the largest dilution ratio that still resulted in strong scattering. Note that dilution was not done in the mesoporous case.

SAXS measurements were made on a Xenocs Xeuss 3.0 instrument using a Cu-$k\alpha$ source. Measurements were made with the instruments 'Biocube' autosampler system. This system uses a robotic arm mounted air displacement pipette system to load samples into a capillary flow cell. Samples were manually loaded into a well plate after synthesis for use with the autosampler. Reduced, unsubtracted 1D q-intensity data was used as prepared by the instruments onboard data reduction process. Subtraction was performed using the data processing software described below.

Dynamic light scattering (DLS) was used to measure average particle size. A Malvern Zetasizer Nano ZS instrument was used with polystyrene disposable cuvette cells. Samples were diluted in ethanol before DLS measurements. Scanning electron microscopy (Thermo Phisher Phenom Pharos operated at 20kV in secondary electron detector mode) was used to image select samples.

7.2.2 Automated assessment of small-angle scattering data for optimization campaigns

Automating the assessment of small-angle scattering data is a major challenge, and a key requirement for integrating this characterization technique into a closed loop autonomous workflow. Integrating small-angle scattering data with commonly used optimization algorithms requires converting the functional form of reduced scattering data into a scalar-valued score that assigns a figure of merit to the sample. This requires an appropriate processing method that is capable of integrating the nuanced information contained in the scattering curve into a meaningful score.

Before scattering data can be analyzed, it must first be 'reduced' to the familiar 1D Intensity vs. q scattering curve form from raw instrument and detector data, and background scattering must be subtracted. The data reduction process depends on the specifics of the instrument (e.g., pinhole camera vs. Bonse-Hart geometry) and may require manual input. Many instrument control softwares, including that of the Xenocs instrument used here, provide data reduction capabilities that allow users to start with reduced 1D data. A data reduction pipeline was in development for the 12-ID-E USAXS instrument at the time of the mesoporous silica experiment but was not generally available yet. For the automated scattering analysis work described here, it is assumed that reduced, background-subtracted 1D data is available.

Quantitative analysis of scattering data is traditionally done through the fitting of structural models developed and selected by domain experts who possess significant a-priori knowledge and intuition about the sample. Parametric scattering models calculate the scattering that expected for given values of model parameters. For example, a sphere model with a polydispersity factor calculates the scattering from a sample with a given radius and radius distribution width[8]. These models are fit to the measured data to determine appropriate values for the parameters using a software implementation such as Sasview[9]. This approach provides a way to solve the 'inverse problem' inherent to scattering methods. Scattering measures the amplitude of the forward Fourier transform of the spatial arrangement of material in a sample. As phase information is not measured, assigning a structure to scattering data requires solving the inverse Fourier transform with incomplete information. This gives rise to an inherent degeneracy in model selection, as multiple models and thus structures may explain the same scattering data. Due in part to this degeneracy, selecting an appropriate model requires external information about the sample. For example,

researchers may use their knowledge about a synthesis process (e.g., an experiment is known to produce spherical particles) or complementary techniques (e.g., SEM imaging showing spherical particles) when selecting models. Basic model fitting methods can frequently converge to incorrect local optima that do not adequately describe the scattering data, requiring manual quality control and coaxing of initial conditions to achieve quality fits. These factors make reliably automating model selection and fitting challenging. Nevertheless, many approaches have been attempted. Many researchers have pursued a machine-learning classifier driven approach to select models for scattering data.[5,10–12]. It is also in principle possible to simply brute-force fit a wide range of models, for example, all models in the popular sasmodels package[13], and select the best fit. These approaches have shown promise when tested on constrained simulated data, but problems arise when applying them to real, complicated scattering data. Real material systems are not perfectly represented by idealized models and may contain mixed geometries or geometries not effectively represented by analytical models, making model selection nontrivial. The degeneracy of model fitting makes an entirely black-box model fitting approach challenging, although this may be overcome by selecting a subset of reasonable models expected to be applicable to a system under study. Further, and perhaps most critically in an optimization context, a score value that is continuous across the entire experimental parameter space is needed. Many material systems of interest may exhibit phase changes or other transitions in geometry over a large synthesis parameter space. Identifying and understanding these phase changes may be a primary goal of an experiment, such as in an autonomous phase mapping effort[14]. Even correctly selected and fit models may not be helpful in this case, as optimizing over the multiple regions of parameter space would require comparing disparate structural parameters. Early in this work, an automated fitting

approach was attempted. Polydisperse sphere models were fit to experimental scattering data, with the intention of using the resulting diameter and polydispersity values directly for optimization. Challenges included achieving correctly converged fits and incorporating results from samples that were not fit well by the selected model. Resulting particle sizes for highly polydisperse or gel-like samples had little physical significance, making it unclear how to incorporate these results into an optimizer. Rather than pursue this strategy further, an alternative approach was explored.

To develop an alternative approach to performing optimization over small-angle scattering data, consider the ultimate objective of an autonomous optimization campaign: to synthesize samples that have a desired structural property. The contribution of the scattering data analysis process to this end goal is to provide a score that describes how 'close' a particular sample is to a target structure. Rather than fitting data and comparing measured structural parameters to target structural parameters, another approach is to compare the measured scattering from a sample to the scattering that is expected from a sample with optimal structure. Target scattering can be calculated from a target geometry using the same parametric models. An appropriate comparison tool – a distance metric - can then be used to compare the scattering curves and compute a score for the sample. Defining this distance metric is important for achieving physically meaningful scores. A simple distance metric is the Euclidean distance or root mean square error (RMSE) between the scattering intensity vectors for the target and measured scattering. However, this

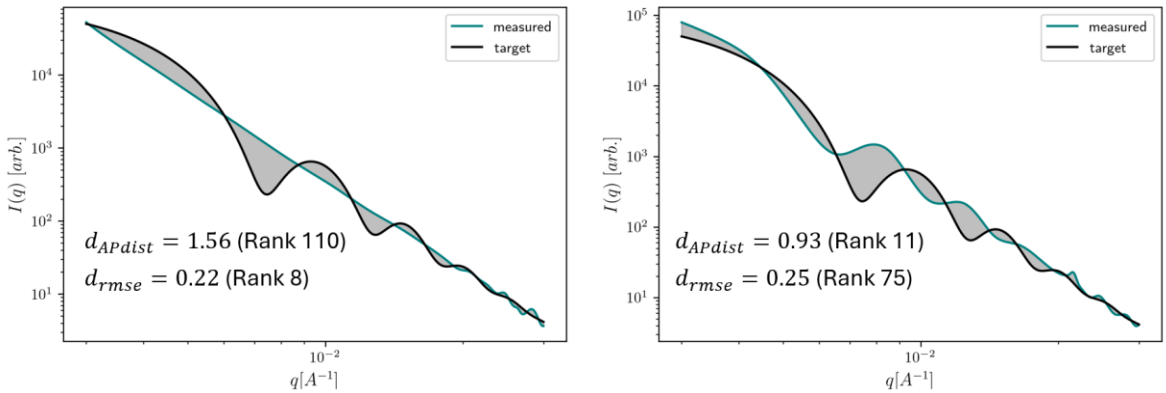


Figure 7.2: Comparison of Amplitude-Phase distance and Euclidean distance (RMSE) in evaluating distance between target and measured scattering. The oscillations in the measured scattering on the right indicate that the sample contains monodisperse particles of a larger size than the target, while the monotonic slope on the right suggests polydisperse particles. Experimental intuition would consider the sample on the right to be much closer to the target than the sample on the left. The amplitude-phase distance reflects this intuition, ranking the sample on the right as 'closer' to the target than the one on the left. The RMSE distance does not.

approach is ineffective because it only accounts for differences in intensities. Information in SAXS data is encoded in the overall shape of the scattering curve in addition to the absolute intensity. The example in Figure 7.2 compares two real measured samples: a monodisperse sample and a polydisperse sample, to a monodisperse target. The measured data has been smoothed as described below. Intuition suggests that the monodisperse sample is much 'closer' to the target in the sense that it is structurally similar. It is highly monodisperse, but with a different diameter than the target. By comparison, the polydisperse sample would be considered further away, as it demonstrates neither the target diameter nor polydispersity. However, by the RMSE distance metric the polydisperse sample is closer to the target. This example illustrates that shape-aware distance metrics are needed to make this direct scattering data comparison approach to optimization effective.

One such shape-aware metric is the Amplitude-Phase distance developed by Kiran Vaddi[15]. This approach compares the shape of two functions by considering the amount of 'warping'

required to align the functions along their amplitude (y-axis) and phase (x-axis) directions. To calculate a distance between a query and reference function, the calculation first determines an optimal warping function that re-maps the x-axis of the query function to minimize the distance to the reference function. The area under the curve of the warping function provides a 'phase' distance, or measure of distance along the x-axis. The remaining distance between the aligned functions provides an 'amplitude' distance or distance along the y-axis. This alignment is illustrated in figure 7.3. These distances can be summed to provide a single scalar distance. The warping process provides a means to account for the overall shape of the function. For more details refer to the publication[15] or GitHub repository[16]. For the SAXS scattering observed from the silica nanoparticle samples, this metric has proven to be effective at ranking sample similarity in a manner that agrees with experimentalist intuition. It consistently ranks samples exhibiting oscillations, a sign of size monodispersity, as closer to a monodisperse target than straight-line 'power law' samples indicative of polydisperse scatterers. In comparison, the RMSE distance

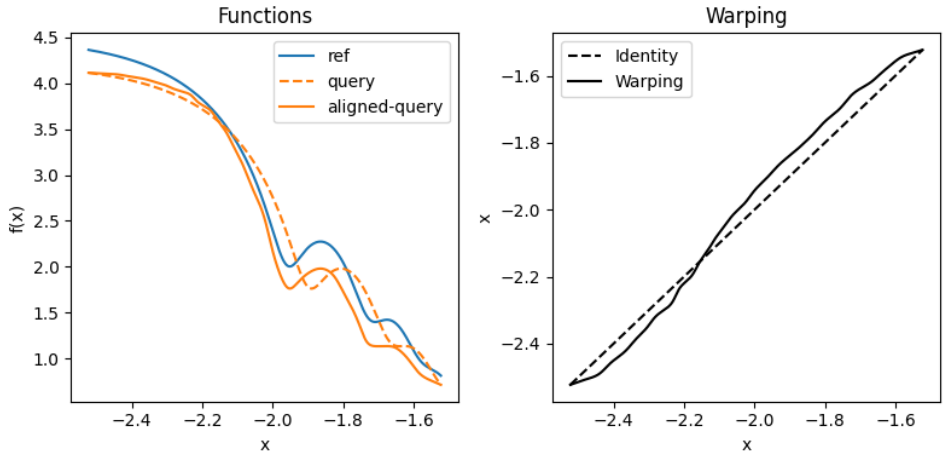


Figure 7.3: Amplitude-phase distance calculation process. Left panel: Reference, query, and aligned query functions. Right: Optimal warping function.

metric accurately identifies monodisperse samples that are very close to the target as close matches, but ranks polydisperse samples over monodisperse samples that are shifted in q. These results are discussed further in the in-silico simulation and optimization campaign sections below. A feature of the distance metric approach to SAXS-based autonomous optimization is that it merges the assessment of multiple structural optimization parameters into a single scalar-valued distance. In the case of this silica nanoparticle work, particle size as well as polydispersity are both captured in this value. This can be an advantage for optimization, as it sidesteps the need for a multi-objective optimization strategy. However, for more complex optimization problems or systems with degenerate scattering, this merging may hinder optimization. Further work would be needed to understand the impact.

Significant pre-processing of measured SAXS data is required to successfully use the amplitude phase distance metric. Experimental SAXS data is noisy at high q values, which interferes with the distance calculation. The amplitude-phase distance calculation also requires that the x-axis (q values) be equally spaced, and that the target and measured scattering intensities have the same q grid, neither of which is the case for reduced data. Because the SAXS intensity scale is arbitrary, the target and measured scattering data needs to be scaled to match. To support these requirements, a data processing pipeline was developed for the experimental scattering data. This pipeline starts by subtracting the capillary and solvent background from the data. The subtracted measured data is smoothed with a Savitsky-Golay filter[17,18] to reduce noise, then fit with a B-spline to interpolate onto an evenly spaced q grid and perform additional smoothing. The interpolated smoothed intensity is then scaled to match the intensity of the target scattering at the high-q region. The amplitude-phase distance calculation is done in the '$\log(I)\log(q)$' space. This processing

pipeline works well for the smooth continuous features observed in this work. However, the smoothing performed might flatten sharp features, such as the pore diffraction peaks seen in the mesoporous silica samples discussed in chapter 6. Alternative data preparation approaches may be needed in this case. This pipeline contains components from the saxs_data_processing repository[19] and is demonstrated in the optimization processing notebooks of the main silica nanoparticle project repository[20].

7.2.3 Active learning optimization

Composition optimization was done with Bayesian optimization. Bayesian optimization generally is described in the methods chapter. Composition optimization problems need to satisfy an equality constraint in the parameter space, as the total volume fraction of all components added to a solution needs to sum to 1. There are four components in the solid spherical nanoparticle synthesis (TEOS, water, ethanol, ammonia solution). In this work, the optimization problem is posed as an unconstrained 3-parameter optimization problem, with the composition of the fourth component determined by the volume balance. TEOS, ammonia, and water compositions are directly optimized. Ethanol composition, the 4^{th} component, was calculated from the volume balance. Thus, ethanol composition is not directly selected by the optimization algorithm but can be implicitly selected through the selection of the other 3 components. The upper composition bounds of the 3 optimized components for the solid spherical case summed to 0.41, allowing for unconstrained optimization which simplifies optimization implementation as any selected composition was physically accessible. For future extension to the mesoporous case, this will not be the case because the upper composition bounds of directly optimized components will sum to more than 1 if the same composition bounds used for the batch synthesis experiment are used. In

this case, a constrained optimization scheme will need to be used to select points that satisfy the composition constraint. Bayesian optimization was implemented using the Botorch package[21].

7.3 Building an in-silico simulator to assess data processing components of workflow

7.3.1 Methods

Executing an autonomous optimization campaign requires selecting appropriate values for many configuration parameters. In this work, experimental structure (e.g., composition bounds, experiment batch size), data processing pipeline (e.g., RMSE vs. amplitude-phase distance metric approach), and Bayesian optimization (e.g., acquisition function, GPR kernel) all require the selection of parameters that could have a large impact on optimization performance. Performing this parameter tuning and campaign design validation on real experiments would be impractical due to the high time and effort investment. Recall that the end goal of this entire endeavor is to reduce the experimental investment required to develop desired materials, not to prove that a particular means of performing optimization is the best. In-silico simulators or 'virtual instruments' for autonomous experiments are useful tools to help build and tune the computational parameters of these systems without relying on real experiments. They are a form of 'frugal twin', allowing for cheap and rapid development work[22]. An in-silico simulator replaces the entire experiment component of the optimization loop with a computed surrogate model that realistically captures important features of the experimental data and outcomes. This model is not necessarily a physical model that accurately accounts for the physics and chemistry of the synthesis but rather can be a simple mathematical function selected through intuition. Integrating this model with the real data processing and experimental planning components allows for these components of the

system to be easily tested and modified. Of course, optimizing actual experiment workflow parameters still needs to be done on the real equipment.

An in-silico simulator was developed to simulate the solid spherical nanoparticle synthesis. This model was used to validate the use of the amplitude-phase distance metric to evaluate small-angle scattering data for use with Bayesian optimization, and to understand the impact of experiment parameters on convergence. The experiment simulator module takes a sample composition as input and returns corresponding simulated scattering data. This replaces the experiment and characterization components of the workflow while retaining the data processing and Bayesian optimization modules. To simulate the experiment, arbitrary functions were defined to map a sample composition to a diameter and polydispersity. SAXS scattering was calculated using a polydisperse sphere model with these parameters. This calculated scattering data was used in the data processing, amplitude-phase distance, and Bayesian optimization pipelines described above. Separate 'experiment' functions were defined for the particle diameter and the polydispersity. The diameter experiment function was defined as

$$d = 57.36 * \log(x_{TEOS}) + 344.16 * \exp(x_{Ammonia}) + 344.16 * x_{ethanol}$$

Where x is the volume fraction of a component. The PDI experiment function defined as a negative multivariate normal distribution centered at $x_{TEOS}, x_{ammonia}, x_{ethanol} = (0.007, 0.025, 0.03)$ and standard deviations (0.08, 0.1, 5). A bias of 1.1 was added to the PDI to set the minimum achievable PDI to 0.1. The definitions of these functions are entirely arbitrary. They were selected to provide a reasonable approximation to the shape of the composition-scattering plots seen experimentally. Contour plots of the diameter function, PDI function, and resulting amplitude-phase distance landscape for an 80 nm target diameter is shown in figure 7.4. Experimental noise

was simulated by introducing uncertainty about the composition evaluated. For each composition component, the actual composition 'simulated' was sampled from a random normal distribution centered at the requested composition value. This noise level was tunable as a parameter to investigate its impact on optimization performance. In addition to simulated scattering, simulated dynamic light scattering was also evaluated. This was done by defining a DLS evaluation function that calculated a score for a sample based on the simulated diameter and PDI. The score function for the DLS simulation was

$$Distance_{DLS} = w_{dls}\left(\frac{|PDI_{measured} - PDI_{target}|}{PDI_{target}}\right) + (1 - w_{dls})\left(\frac{|Diameter_{measured} - Diameter_{target}|}{Diameter\_target}\right)$$

Where w_{DLS} is a weighting factor. DLS weight values of 0.1 and 0.5 were tested, and 0.5 showed the best convergence.

This in-silico simulation tool was used to explore several questions about the design of optimization campaigns. The primary objective was to demonstrate that 'reasonable' convergence to a target particle size could be achieved for the 'SAXS + Amplitude-Phase distance' sample characterization method with Bayesian optimization. RMSE was also evaluated as an alternative distance metric and compared to amplitude-phase distance. Further, dynamic light scattering was tested as an alternative to SAXS-based characterization methods, the impact of experimental noise or reproducibility on convergence was investigated, and the experimental parameters of batch size and parameter bounds were both tested. Batch size presents a tradeoff

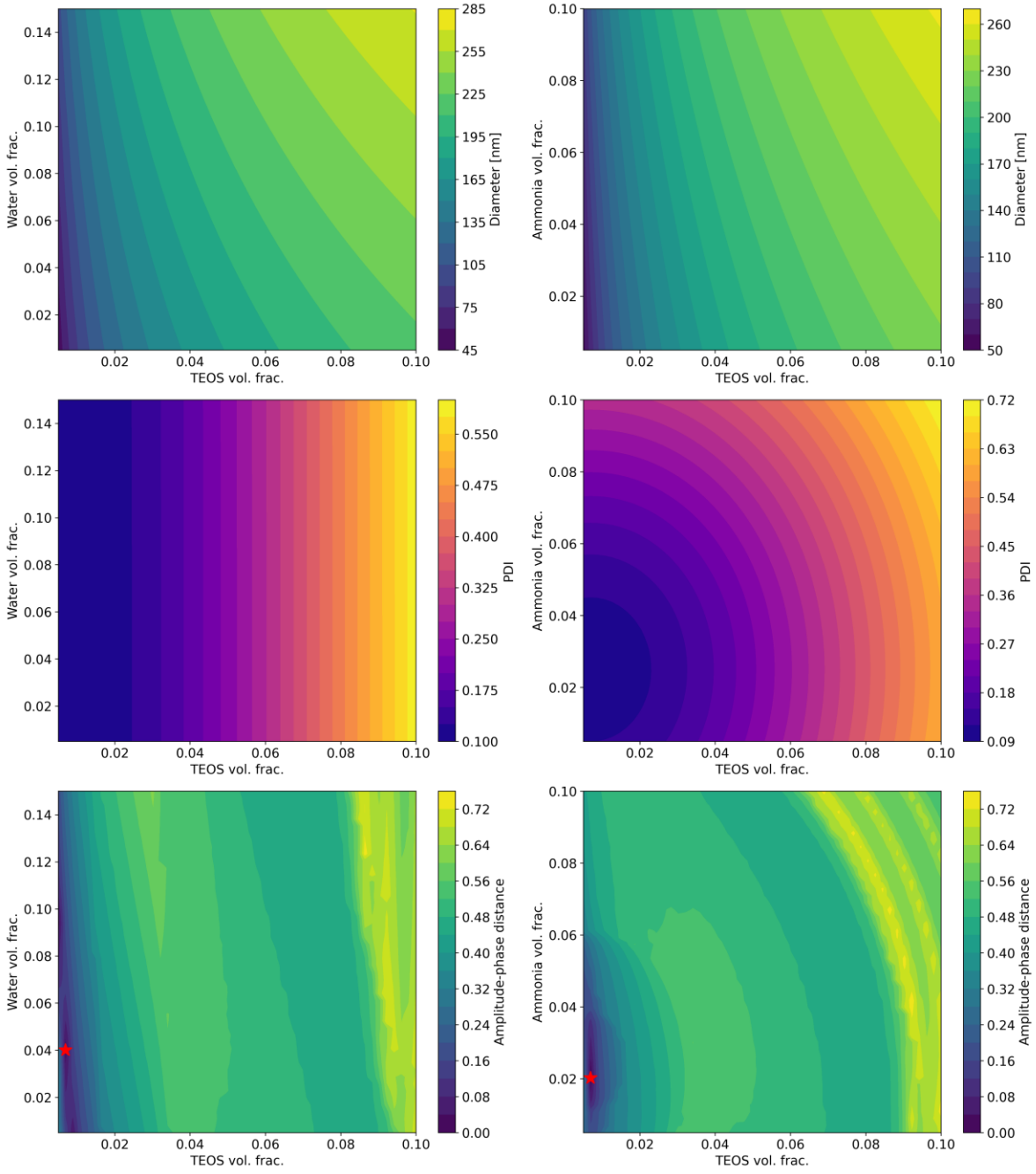


Figure 7.4: Contour plots of the in-silico optimization simulation functions. Top row: Diameter, second row: PDI, third row: Resulting amplitude-phase distance landscape for an 80 nm target, with red star indicating global minimum.

between hypothetical optimization performance and campaign efficiency. In theory, smaller batch sizes enable optimization convergence in fewer samples because the optimization algorithm has more information available when picking each sample. However, larger batches require less experimental overhead so are more efficient to run. Restricting the parameter bounds lowered the maximum concentration of each component, focusing the parameter space on the region where optimal conditions were known to be located. Each of these factors were evaluated by comparing an experiment with the relevant parameter modified to a baseline experiment. For the modified experiment, all other parameters were kept the same as the baseline. Three replicates were run for each experiment. Each trial consisted of 32 initial seed samples selected with a Sobol sampling strategy, followed by 100 samples selected with Bayesian optimization. All optimization experiments used the same Bayesian optimization configuration. The batch Log Noisy Expected Improvement acquisition function was used as implemented in Botorch[23,24] using a Gaussian process regressor with Matérn kernel ($\nu = 5/2$). For all amplitude-phase distance experiments, an amplitude weight of 0.1 was used, with corresponding phase weight of 0.9. This weight adjusts the contributions of the amplitude (y-axis) and phase (x-axis) contributions to the amplitude-phase distance. The resulting amplitude-phase distance value was given by $AP_{dist} = 0.1 * AmplitudeDist + 0.9 * PhaseDist$. These values were selected through initial experiments. A control experiment was also run using Sobol sampling to select 128 samples to compare the optimization methods against a random search. The specific parameters used in each trial are summarized in Table 7.1.

Table 7.1: Parameters evaluated in experiment conditions in-silico experiments. Bounds are volume fractions and are in the format [TEOS lower, ammonia lower, water lower] → [TEOS upper, ammonia upper, water upper].

Trial Name	**Batch size**	**Bounds**	**Noise level**	**Characterization**	**Distance metric**
Baseline	5	[0.005, 0.005, 0.005] → [0.1, 0.1, 0.15]	0.05	SAXS	AP dist.
Increased Noise	5	[0.005, 0.005, 0.005] → [0.1, 0.1, 0.15]	1	SAXS	AP dist.
DLS	5	[0.005, 0.005, 0.005] → [0.1, 0.1, 0.15]	0.05	DLS	DLS weighted score
RMSE	5	[0.005, 0.005, 0.005] → [0.1, 0.1, 0.15]	0.05	SAXS	RMSE
Increased batch size	10	[0.005, 0.005, 0.005] → [0.1, 0.1, 0.15]	0.05	SAXS	AP dist.
Restricted bounds	5	0.005 → [0.06, 0.08, 0.08]	0.05	SAXS	AP dist.
Sobol baseline	0	[0.005, 0.005, 0.005] → [0.1, 0.1, 0.15]	0.05	SAXS	AP dist.

7.3.2 In-Silico optimization results

The optimization performance was evaluated quantitatively by calculating the amplitude-phase distance between the simulated 'experiment' scattering and the calculated target scattering, using an amplitude weight of 0.1. Three results were monitored to assess performance: The number of iterations taken to identify a sample with $APdistance_{measured,target} < 0.05$, the overall convergence trajectory as monitored on a distance vs. iteration number plot, and the scattering observed for the best sample identified throughout each 132 sample campaign. In cases where optimization was performed with a metric other than amplitude-phase distance, the convergence was still monitored using amplitude-phase distance to maintain consistency. Qualitatively, the samples selected with this metric show excellent agreement with the target scattering. The average iterations to convergence for each experiment is listed in table 7.2

Table 7.2: Iterations to convergence to amplitude-phase distance of less than 0.05.

Trial	**Average iterations to convergence [3 trials]**
Baseline	52
Increased noise	102
DLS	48
RMSE	113
Batch_size_10	56
Restricted bounds	50
Sobol baseline	132+ (none converged)

Two implementations of the Amplitude-Phase distance metric were used to generate the results presented here. Initially, an implementation using numpy was used in all experiments shown here. However, during post-processing an issue was identified with anomalously large amplitude and phase distances for some 'close' samples. This issue was resolved by using a Torch implementation that performs an iterative optimization of the warping function. This torch implementation is sensitive to optimization settings, including the number of iterations allowed and the number of random restarts. Results shown here for the 'Baseline', 'Increased noise', and 'Sobol baseline' cases were run with the new torch implementation, while all other trial results shown here were generated with the numpy implementation. The resulting amplitude-phase distances and convergence performance appear to be analogous between the two methods.

7.3.2.1 Results

7.3.2.2 'Baseline' optimization compared to random search

The 'baseline' campaign configuration converges to the convergence criteria in ~ 52 total samples on average, while none of the three trials run with the Sobol baseline campaign converged to this criterion. The campaign convergence trajectories for these trials are shown in figure 7.5. This convergence plot plots the best observed amplitude-phase distance between a virtually measured sample and the target scattering in a campaign at a given sample number. The gray box on the left indicates the initial random sampling executed before the optimization is enabled. Figure 7.6 shows the best sample observed from representative campaigns for each experiment. This shows that the 'baseline' case identified a composition which mapped to a nearly exact scattering match, and nearly exactly reproduced the target 80 nm diameter. The

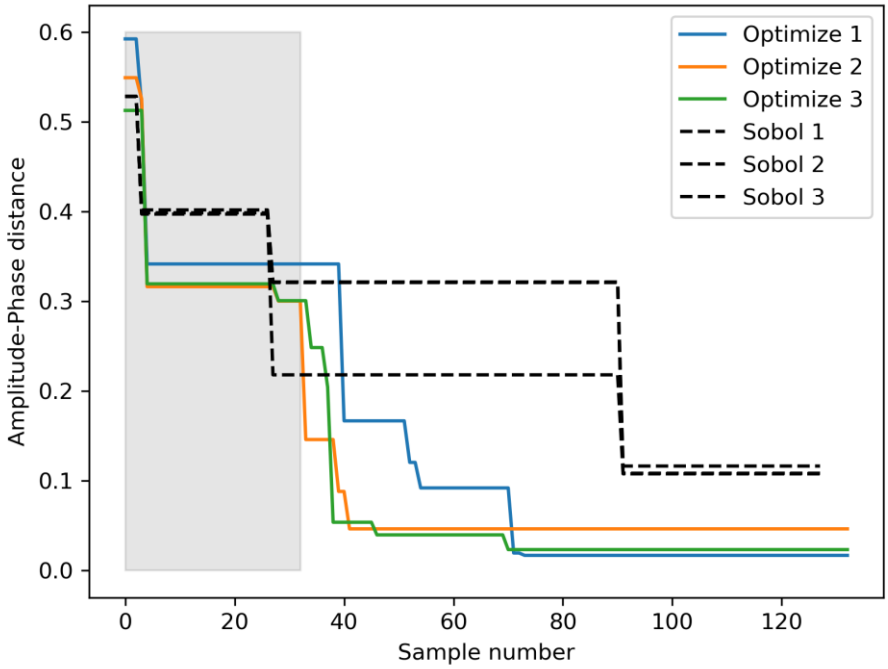


Figure 7.5: Convergence plot for baseline and Sobol comparison virtual campaigns.

'Sobol' comparison case also identified a close match to the target. However, the scattering does not match as closely as the 'baseline' case, resulting in a larger amplitude-phase distance. The Sobol comparison campaign also took longer to identify this sample than the baseline campaign.

7.3.2.3 Results for Amplitude-Phase vs. RMSE distance comparison

The convergence performance of the campaigns run using the RMSE distance metric are shown in figure 7.7. The best-observed samples from the three RMSE optimization campaigns are also plotted in figure 7.8. These results show that the amplitude-phase distance metric enables more effective optimization than RMSE. Only one of the three RMSE campaigns identified a reasonably close match to the target scattering, and it took longer to do this than the amplitude-phase distance

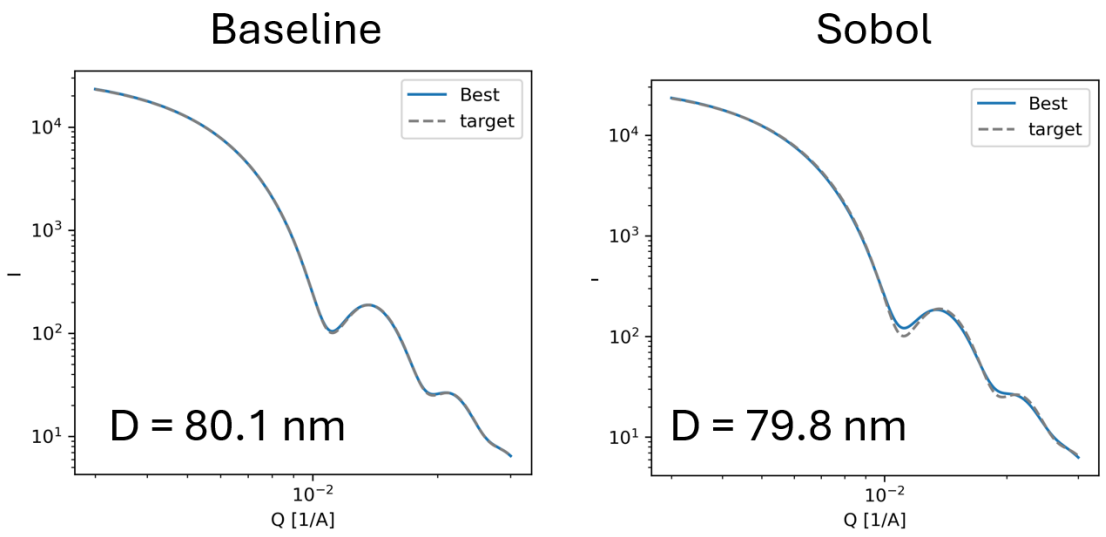


Figure 7.6: Examples of best observed simulated samples from baseline and Sobol sampling comparison campaign.

campaign. These results and observations from other initial exploratory investigations suggest that the RMSE distance metric correctly ranks scattering that is 'close' to the target scattering and identifies curves with nearly exact matches. However, as demonstrated above and below in the experimental optimization results, it ranks polydisperse scattering as 'closer' to a monodisperse target than scattering from monodisperse particles with a different average diameter than the target. In the one trial that converged in this campaign, a sample that was sufficiently close to the target for the RMSE distance to correctly select it as the closest was identified, then the optimization algorithm could fine-tune the composition of that sample using the now-performant RMSE distance. This close-match sample identification likely did not occur in the other two campaigns, resulting in non-converged optimization. The results from this optimization campaign suggest that the improved ranking performance of the amplitude-phase distance method in terms of intuition-based 'closeness' does indeed translate into effective optimization convergence.

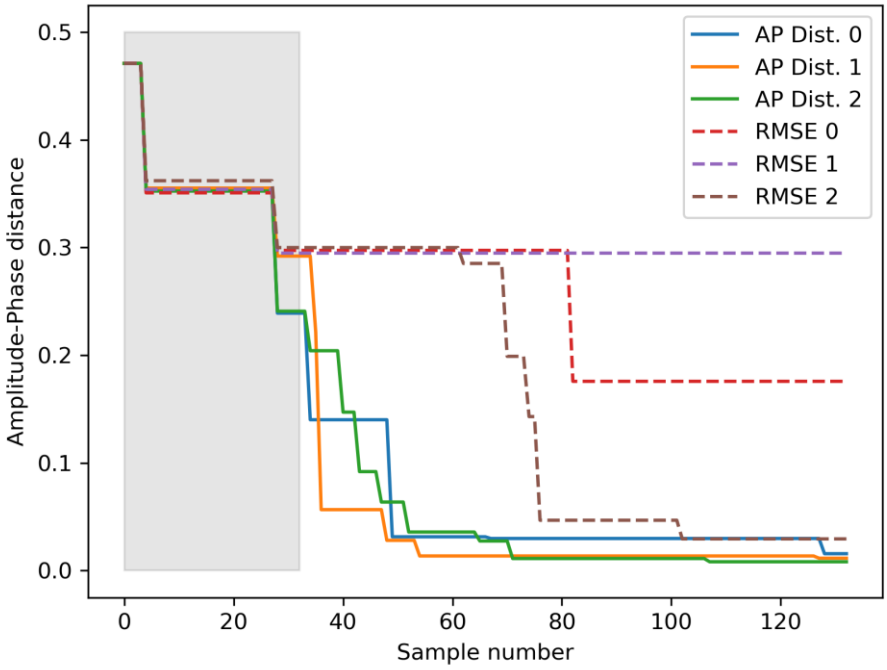


Figure 7.7: Convergence plot for RMSE campaigns, plotted against baseline amplitude-phase distance campaigns. Note that the AP-distance campaign run here is a different optimization run (using identical conditions except for the use of the Torch amplitude-phase distance implementation) than the 'baseline' campaign discussed above.

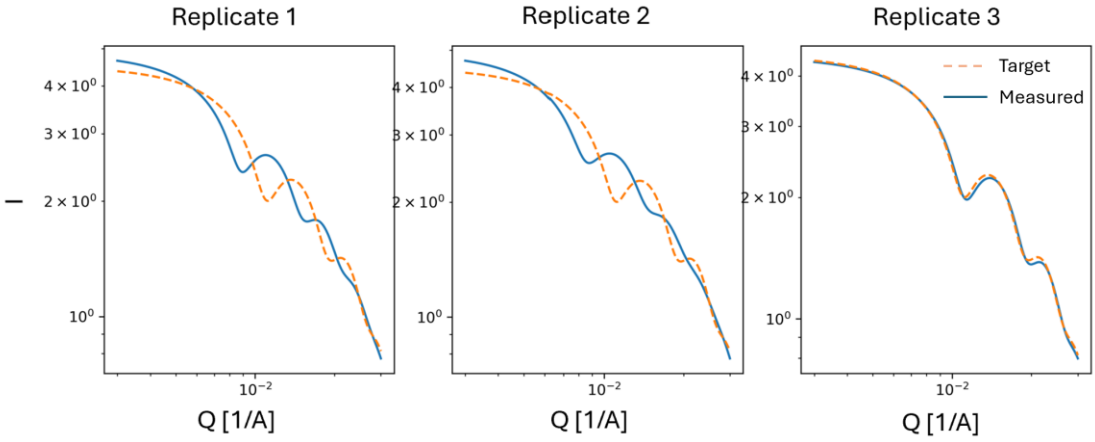


Figure 7.8: Best observed samples for RMSE campaigns.

7.3.2.4 Results for increased noise case

Increasing the simulated experimental noise has a negative impact on optimization performance. Experimental noise was simulated by introducing uncertainty to the actual sample composition that was simulated. Before an experiment for a composition was simulated, the requested composition was modified by selecting new composition values for each component from a normal distribution centered at the requested composition values. The standard deviation of this distribution was determined as a function of both the requested composition as well as a 'noise fraction' value that controlled the magnitude of the introduced noise. The function was defined as a piecewise linear function such that a noise fraction value of 1 was fit to the measured precision of the Digital Syringe tool, and a noise fraction of 0 resulted in noise-free experiment simulations. The 'baseline' noise value used throughout the virtual instrument investigation was 0.05. Early investigations indicated that larger noise fraction values adversely impacted optimization, so this value was selected to allow for a fair comparison of the impact of other parameters on convergence. The 'increased noise' case shown here was simulated with a noise fraction value of 1. Figure 7.9 shows the overall convergence of the increased noise campaigns compared to the baseline campaigns. Only one of the trials from the increased noise campaigns converged to the amplitude-phase distance < 0.05 threshold, and it took around 110 samples to do so. The resulting best sample does not match the target scattering as closely as the best samples from the baseline campaign as shown in figure 7.10. These results are in agreement with the adverse impact of experimental noise on Bayesian optimization convergence that has been documented in the literature on self-driving labs[25].

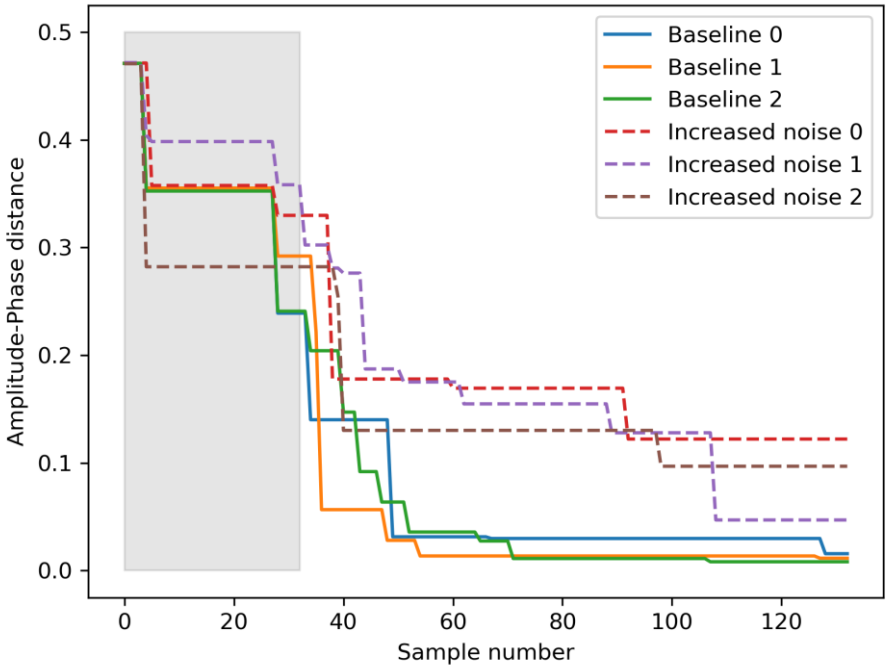


Figure 7.9: Convergence plot comparing campaign convergence of increased noise case to baseline case.

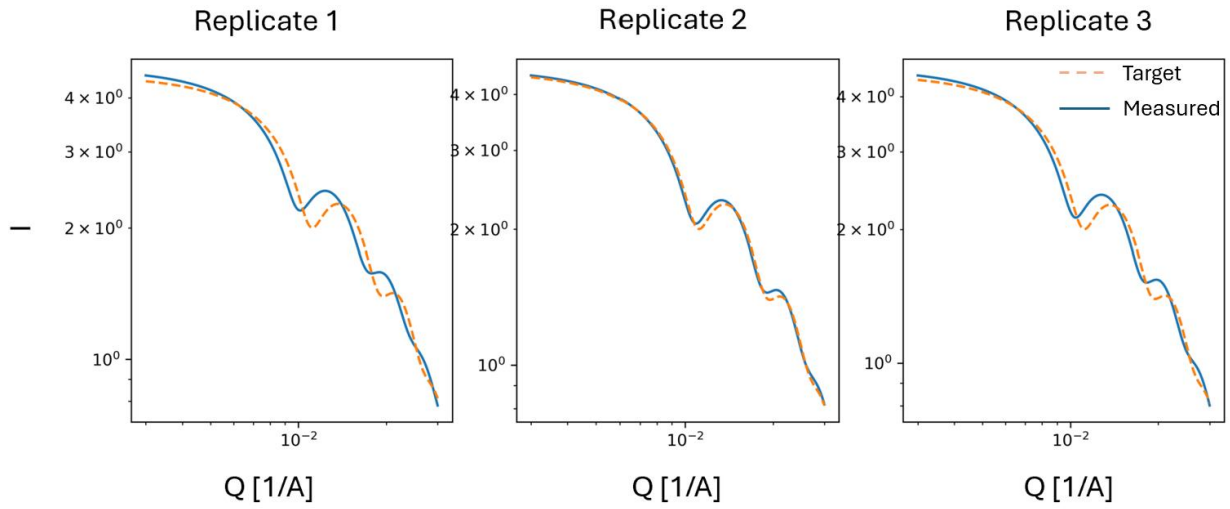


Figure 7.10: Best observed samples from increased noise campaigns.

7.3.2.5 Impact of other investigated parameters on optimization convergence

The other tested campaign configuration parameters were not observed to have a meaningful impact on optimization convergence. The performance of the overall scattering to distance metric characterization pipeline was evaluated by comparison with a direct optimization over particle diameter and polydispersity, with optimization performed through a scalarized score function defined above. This simulated characterization was referred to as the 'DLS' method due to its similarity to dynamic light scattering characterization. The convergence plot for these campaigns is shown in Figure 7.11. The convergence shows that the direct optimization over particle properties performs similarly to the indirect optimization of scattering with the amplitude-phase

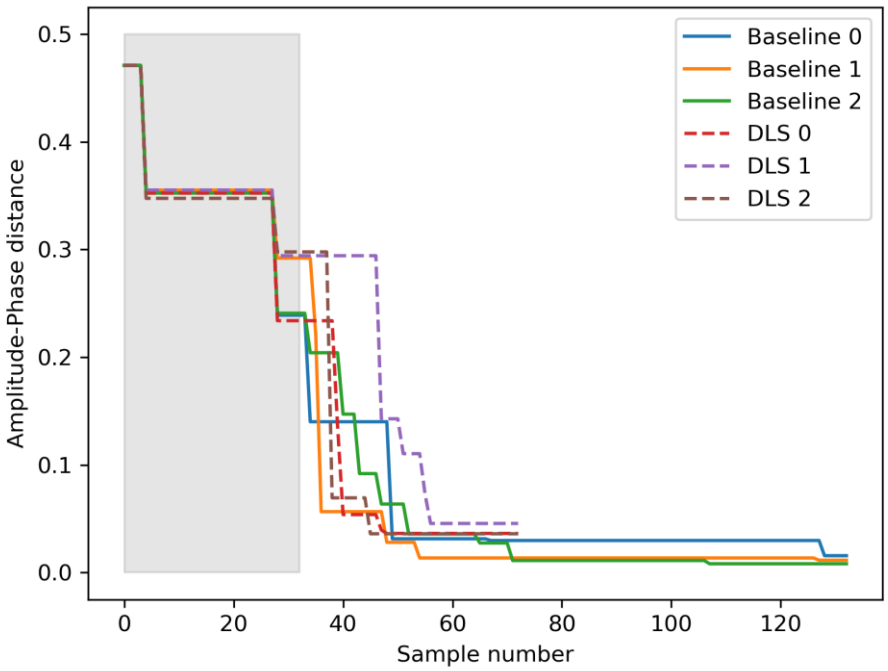


Figure 7.11: Convergence plot for simulated DLS characterization case. Note restricted campaign with 40 samples was run due to time constraints.

distance. However, the direct particle size optimization required objective function tuning to achieve this level of performance. Varying the weights or structure of the scalarizing objective function can have a large, negative impact on convergence, suggesting care needs to be taken here if DLS is used for experimental optimization.

Increasing the batch size of individual experiment batches from 5 to 10 samples, or reducing the bounds on the composition space, does not observably impact optimization performance. Figures 7.12 and 7.13 show the convergence results for these experiments. While the restricted bounds campaigns do identify closer samples during the initial random search phase, the overall convergence during the optimization phase of the campaign is similar to the full

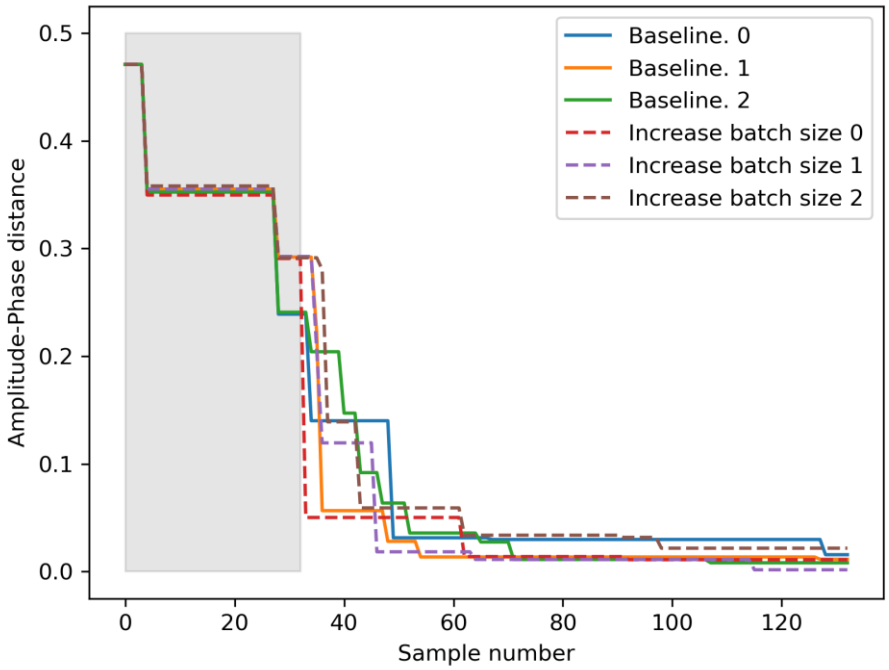


Figure 7.12: Convergence plot for batch size 10 experiments

composition bound case. These results suggest that an overall ‘reasonable’ campaign structure will enable a simple Bayesian optimization implementation to identify samples close to the target structure. The negligible impact of experimental batch size on convergence relaxes constraints on experiment scheduling. Batching sample synthesis together can improve sample throughput by reducing the number of tool changes and washing steps required and can provide more flexibility in sample synthesis sequencing compared to a one-sample-at-a-time approach.

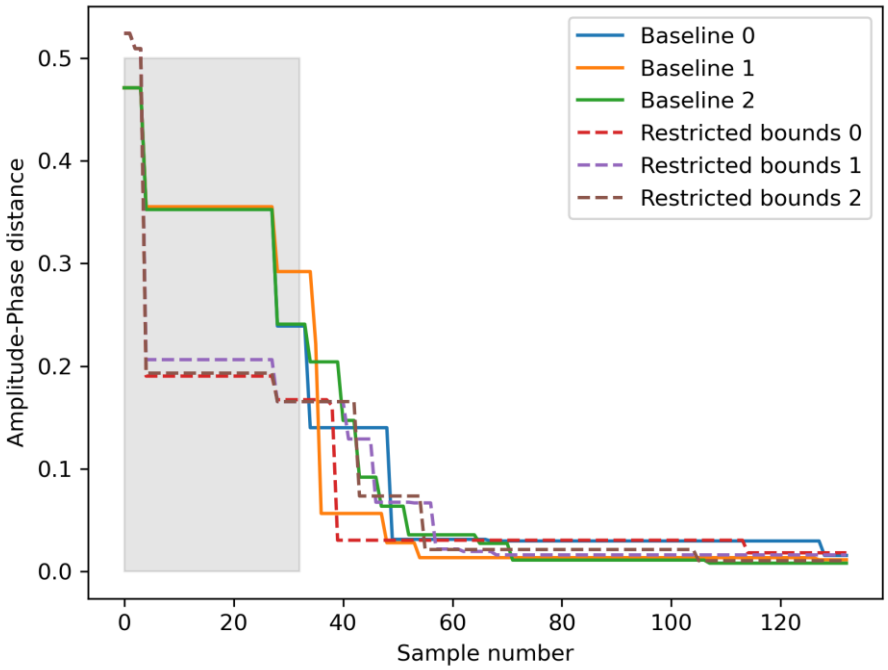


Figure 7.13: Convergence plot for restricted bounds experiments.

7.3.3 Virtual instrument conclusions

Overall, the results from the in-silico simulation validate the general approach to nanoparticle optimization and help identify important experimental conditions to monitor. Reasonable optimization was achieved with a range of experimental parameters, suggesting that the premise of optimizing nanoparticle structure using SAXS coupled with the amplitude-phase distance metric for characterization is sound. Amplitude-phase distance is generally an effective metric. These results do highlight experimental reproducibility as an important factor for optimization performance. This can be understood by considering that excessive noise or irreproducibility in an experiment will lead to an effectively uncontrollable synthesis, which would be impossible to optimize. The virtual experiments reported here were intended as an initial validation, and the in-silico optimization tool may be of further use to develop the optimization campaigns. Further work could investigate the impact of additional Bayesian optimization configuration parameters. A single model, kernel, and acquisition function was used here. Alternatives may impact performance. No assumptions about experimental noise were made in the Gaussian process regressor model. Modeling expected noise could improve optimization performance, although the interaction of this noise accommodation with the noise assumptions made in the noisy expected improvement acquisition function should be considered. The in-silico simulator itself could be improved by replacing the ad-hoc experiment surrogate functions with a model trained on real experimental data. This might provide a more realistic surrogate, but the benefits should be weighed against the effort required, and the end goal of optimizing real nanoparticles through experiments should be kept in sight. Virtual instruments should also be used moving forward to

optimize campaign scheduling. Given the long sample synthesis times (2+ hours reaction time) and Jubilee machine times (20 minutes) for each sample, synthesis sequencing and scheduling should be optimized to make effective use of instrument resources while enabling optimization performance. Options for batching synthesis to increase overall sample throughput should be investigated. Given the large difference between machine-on time and total reaction time per sample, new samples will need to be selected by the optimizer before results from the most recently selected samples are available. The virtual instrument approach should be used to evaluate the impact of this lag on optimization performance and identify strategies to reduce any negative impacts.

7.4 Experimental optimization campaigns

A series of experimental optimization campaigns were run, with an optimization objective to produce particles with a target diameter and polydispersity. The general optimization workflow described above in the experimental methods section was used. Samples were synthesized on Jubilee and characterized with SAXS using the 'Biocube' autosampler. Reduced 1D scattering data was retrieved from the SAXS instrument, processed using the above-described data workflow, and compared to the calculated scattering profile for the target particle using the amplitude-phase distance. Bayesian optimization was used to select batches of sample candidates. The data processing and Bayesian optimization was managed manually using notebooks. Two sets of campaigns were run, with experimental parameters and protocols modified between the sets. The same data processing and Bayesian optimization configurations were used in both sets of campaigns. The amplitude-phase distance was computed as a simple sum of the amplitude and phase distances, with equal weighting. A single-task Gaussian process regressor was used as the

Bayesian optimization surrogate model, with a Matérn kernel ($\nu = 5/2$). The batch Log Noisy Expected Improvement acquisition function was used. The optimization problem was posed as a maximization problem by negating the amplitude phase distance and scaling the observed values to the range [0,1].

7.4.1 Experimental campaign set 1

7.4.1.1 Methods:

The composition space parameter bounds for the first set of campaigns were selected by expanding the range of compositions of monodisperse samples synthesized during initial protocol development work, and by considering compositions of monodisperse samples reported in literature. These bounds are shown in table 7.3. The same set of samples was used to initialize optimization for all campaigns in the first set. This set consisted of 84 compositions selected randomly from a uniform distribution over the composition space. During the optimization phase of each campaign, a batch size of 20 samples was used. This batch size was selected because it is an efficient batch size to run on the Jubilee platform. Three campaigns were run in the first set. The first targeted a 120 nm particle, and the second also targeted a 120 nm particle and added a comparison between RMSE and amplitude-phase distance metrics. In this campaign, parallel optimization campaigns were run, one with sample scattering assessment through amplitude phase

Table 7.3: Composition bounds for campaign set 1.

Component	**Lower limit [volume fraction]**	**Upper limit [volume fraction]**
TEOS	0.01	0.15
Ammonia	0.005	0.11
Water	0.005	0.15

distance, and the other with sample scattering assessment through RMSE. The third campaign targeted 80 nm particles, and also compared amplitude-phase distance and RMSE.

7.4.1.2 Results:

In the first campaign (targeting 120 nm particles with only the amplitude-phase distance metric), the optimization phase of the campaign did not improve on the best sample identified during the random search phase, as shown in figure 7.14. In this campaign, samples from a reproducibility trial were included in the set of initial samples. The best sample identified in the campaign was from the random sampling phase and synthesized during the reproducibility experiment. A

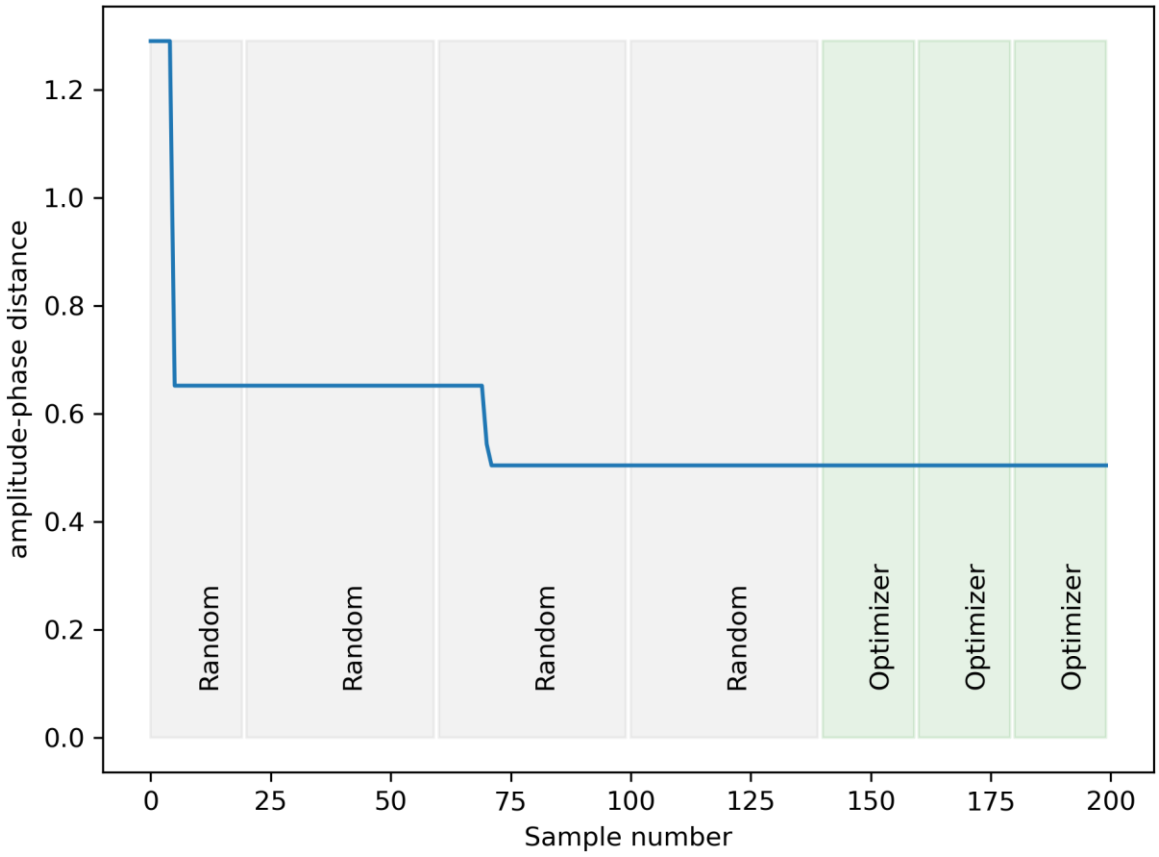


Figure 7.14: Campaign convergence plot for 120 nm target, campaign round 1. Each gray box labeled 'Random' represents an initial batch of experiments selected randomly, and each 'Optimizer'-labeled green box represents a batch of experiments selected with Bayesian optimization.

polydisperse sphere model fit to the scattering from this sample (figure 7.15) suggests a mean particle diameter of 130 nm, with a PDI of 0.11. Samples that separately met the diameter target (figure 7.16) and size monodispersity target (figure 7.17) were also identified in this campaign. These results suggest particles meeting the target are obtainable in the parameter space, but the optimization in this campaign was not effective.

The second campaign also targeted 120 nm particles and ran a comparison between the amplitude-phase distance metric and RMSE. During the execution of this campaign, an error was made in synthesis execution code while loading the sample composition files. The amplitude-phase distance and RMSE samples were swapped, resulting in the cross-contamination of the

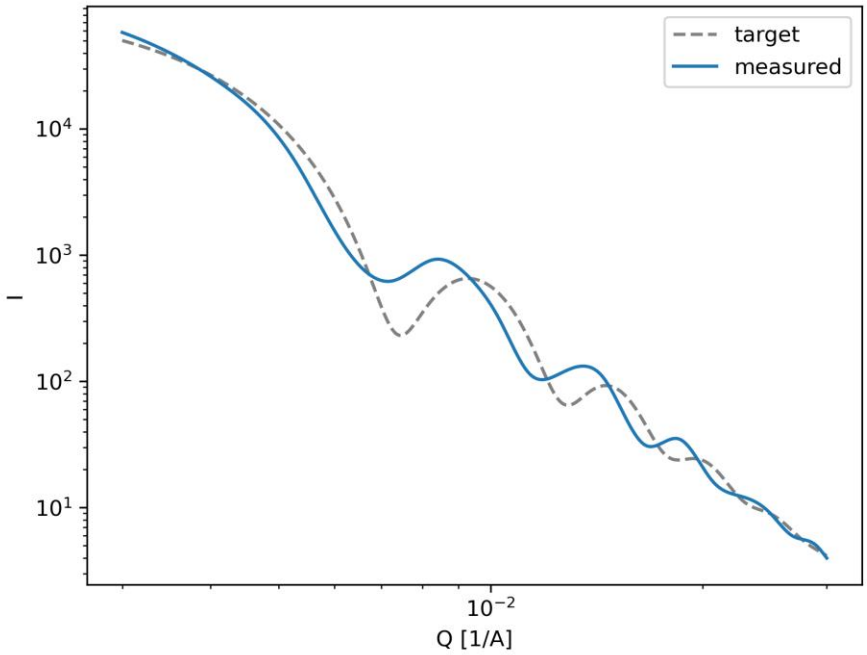


Figure 7.15: Best observed sample from campaign 1

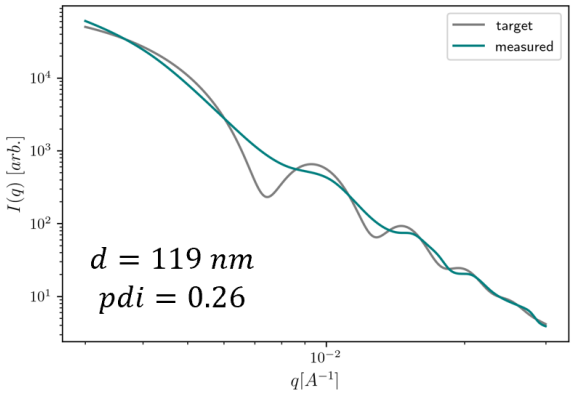


Figure 7.16: Best diameter observed in campaign 1.

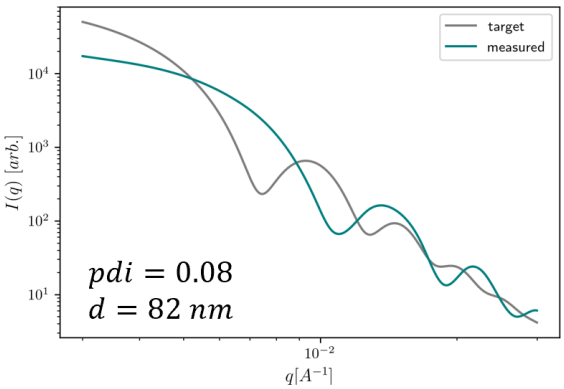


Figure 7.17: Best PDI observed in campaign 1.

optimization datasets. Due to this error, only the first round of optimization was valid. Figure 7.18 shows that minimal improvements were made to the best observed sample in this first round. The results still provide a useful comparison between the RMSE and amplitude-phase distance metrics. Figure 7.19 shows the top four samples from both the amplitude-phase distance and RMSE campaigns, as ranked by the respective metric. The top amplitude-phase distance samples all exhibit either high size monodispersity or a close match to the target particle diameter, as indicate by the location of oscillation maxima and minima that align with the target scattering. On the other hand, the RMSE top candidates include polydisperse samples in 2nd and 4th place. Both metrics do identify a top sample that does closely match the target.

The third campaign targeted 80 nm particles and also compared amplitude-phase distance and RMSE. This campaign suffered the same error of mixed-up file loading, so the optimization results are invalid. However, the same comparison between amplitude-phase distance and RMSE results ranking can be made. Figure 7.20 shows the scattering from the top-ranked samples in each campaign. As with the 120 nm target, the amplitude-phase distance metric ranks highly

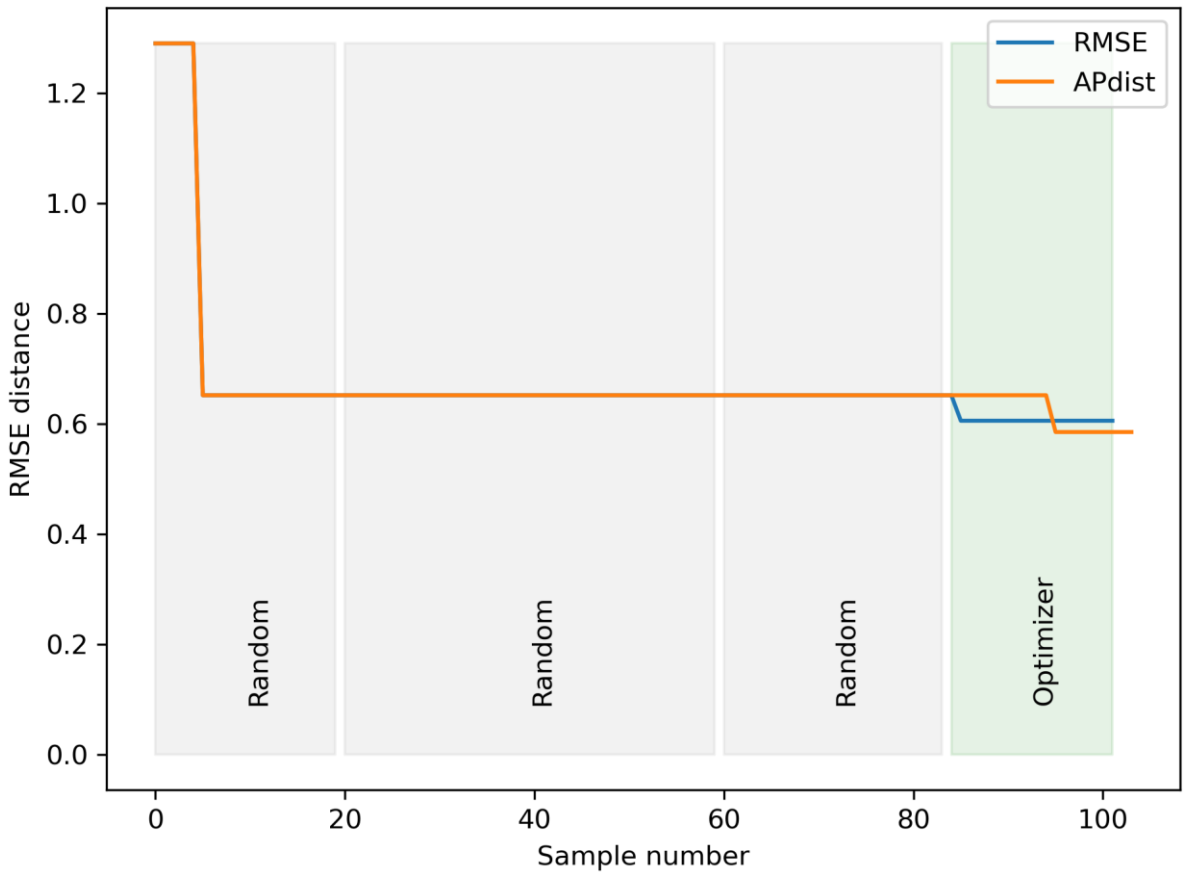


Figure 7.18: Campaign convergence plot for 120 nm target comparing RMSE and amplitude phase distance performance (evaluated with amplitude phase distance).

monodisperse samples highly, while the RMSE metric selects multiple samples that exhibit

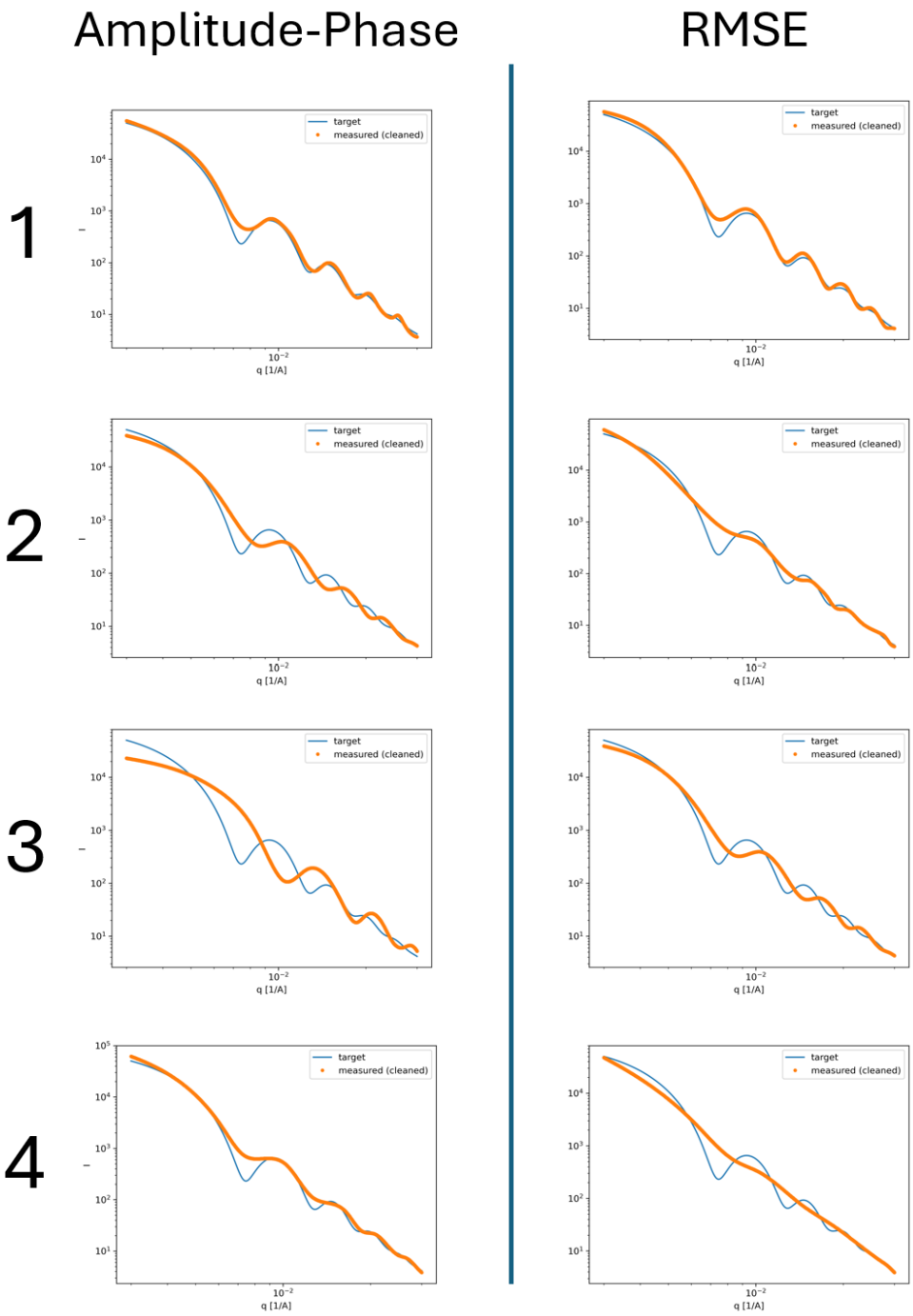


Figure 7.19: Top-ranked samples by respective metric for the amplitude-phase distance and RMSE campaigns. Results for the second 120 nm target campaign with comparison between RMSE and amplitude-phase distance.

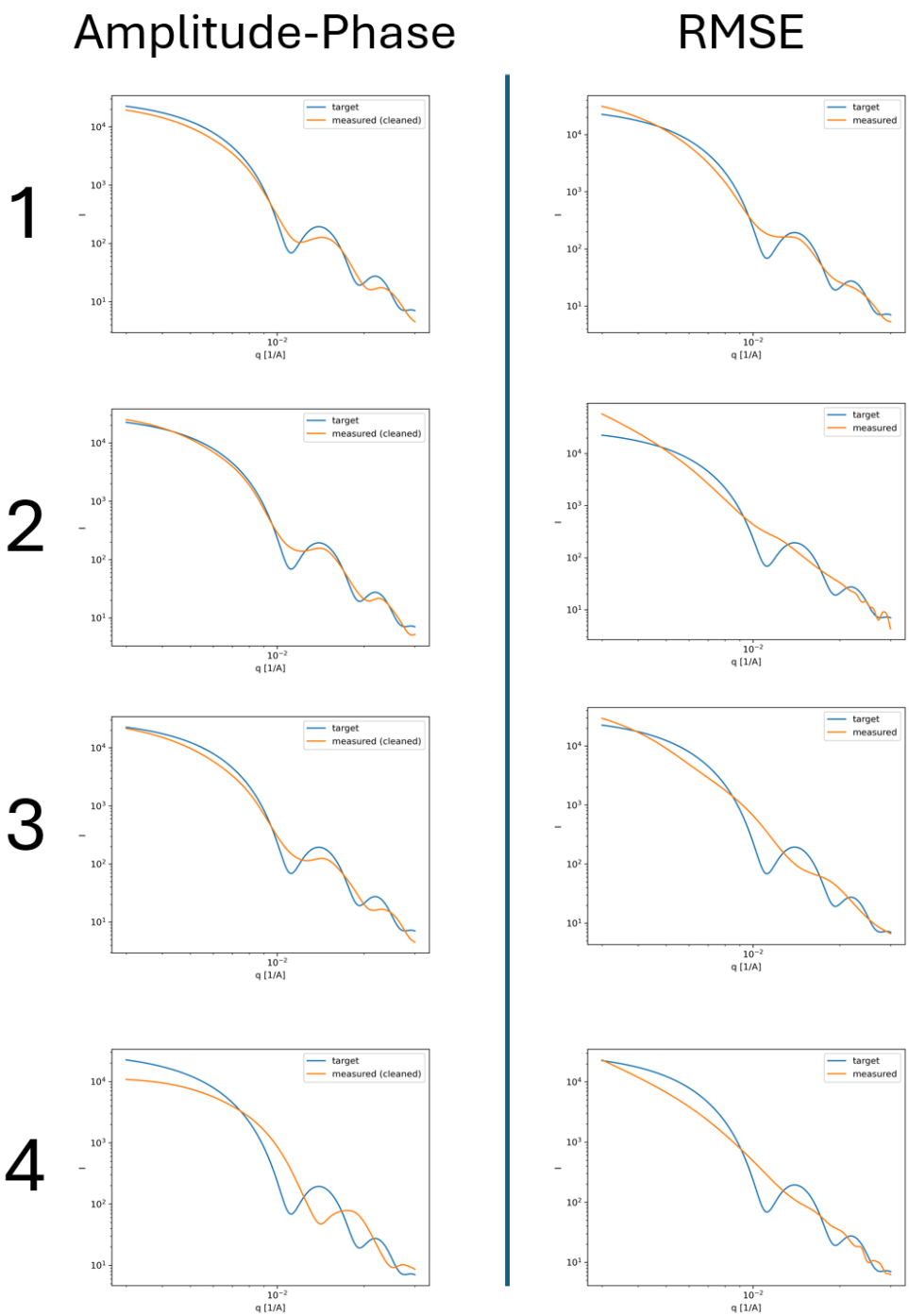


Figure 7.20: Comparison of top-ranked samples from amplitude-phase distance and RMSE, from 80 nm target comparison campaign.

polydisperse scattering. In both campaigns, samples that have reasonable agreement with the target scattering are identified. Figure 7.21 shows scattering and SEM images for samples which have scattering that is similar to the target. The SEM images show that the samples do indeed consist of spherical particles with sizes close to their respective targets.

7.4.1.3 Campaign set 1 discussion

In all 3 campaigns, no or insignificant improvement was observed in the optimization phase. However, samples that somewhat matched the target particle morphology or scattering were identified in all cases. These results could be caused by an overly extensive random search phase that oversaturates optimal regions of parameter space, or by an issue with the experiment or optimization, such as poor experimental repeatability. The results from these campaigns do suggest that the amplitude-phase distance metric is an effective tool for scoring and ranking scattering, and that it outperforms RMSE at identifying samples that are similar but not close to the target scattering.

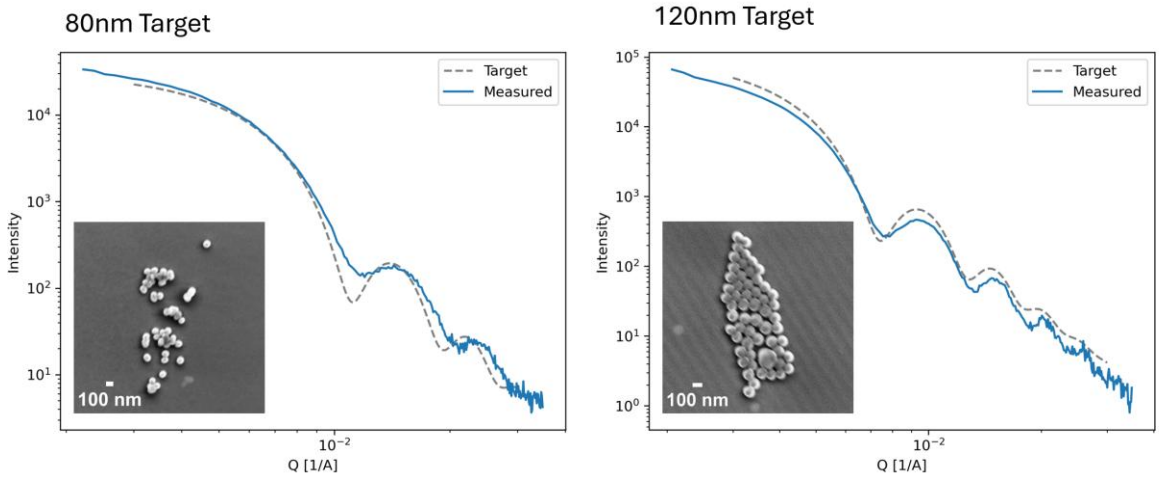


Figure 7.21: Scattering and SEM images from selected samples from 80 nm and 120 nm target campaigns.

7.4.2 Experimental optimization campaign 2

Following the first round of campaigns, updates were made to the experimental setup to improve optimization performance. Experimental protocol issues uncovered during the first round of experiments were also addressed. In the second set of campaigns, the lower bounds on TEOS composition were reduced. It was observed in the first campaign that monodisperse particles were synthesized on the lower boundary of the TEOS composition space. The updated parameter space bounds are listed in Table 7.4. It was discovered that the rubber seal on the plunger of the disposable plastic syringes in use with the Digital Pipette tool were incompatible with TEOS, so the TEOS and Ammonia syringes were replaced with glass syringes. A glass 10 cc mixing syringe was also tested, but the plunger seized, potentially due to nanoparticle contamination. Due to this, a plastic 10 cc syringe was used in this set of experiments. To mitigate any swelling, a new syringe was used for every batch of experiments. It was suspected that the Digital Pipette syringe dispense precision may have been a cause of the poor convergence seen in the first set of campaigns. To improve the relative dispense precision, the TEOS, ammonia, and water stocks were diluted in ethanol. This dilution increased the dispense volume of a given dispense step, increasing the relative precision. The stocks were diluted at a ratio of 1:1.85 stock component:ethanol by volume.

Table 7.4: Composition bounds for campaign set 2.

Component	**Lower limit [volume fraction]**	**Upper limit [volume fraction]**
TEOS	0.005	0.1
Ammonia	0.005	0.1
Water	0.005	0.15

7.4.2.1 Campaign 2 results

One optimization campaign was run with this configuration. This campaign targeted 80 nm particles. The same optimization and data processing configuration as the first set of campaigns was used. An initial set of 32 seed samples was selected using a Sobol sequence. Batch sizes of 10 samples were used in the optimization phase. Figure 7.22 shows the optimization performance of the campaign. Over the three optimization campaigns, essentially no improvements were made. None of the scattering from synthesized samples was similar to the target scattering, and a

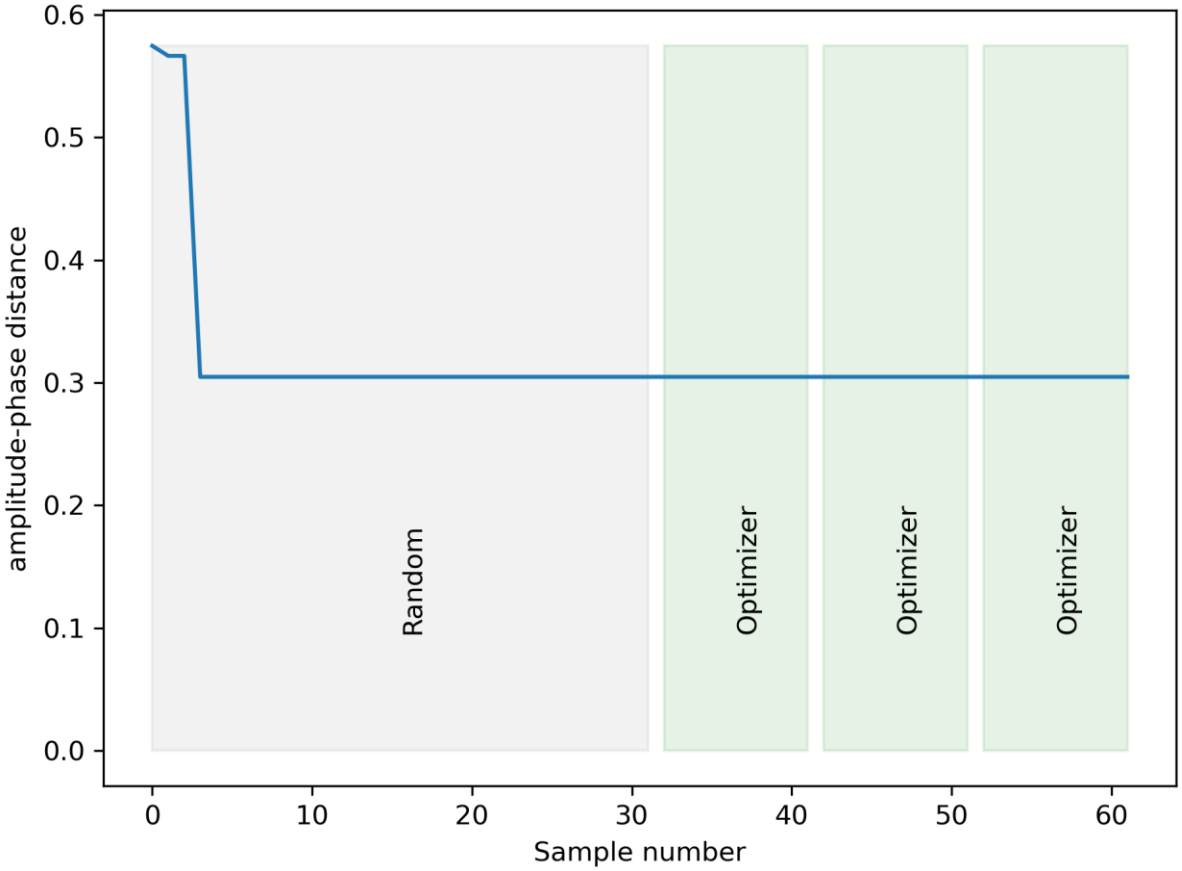


Figure 7.22: Convergence plot for second round optimization with 80 nm target.

significant amount of polydisperse, gel-like scattering was observed. Figures 7.23 and 7.24 plot the observed samples in composition-scattering plots. These plots show that monodisperse samples are mostly clustered in the lower portion of the composition space (lower left corner of plots), but that polydisperse samples are made throughout the parameter space.

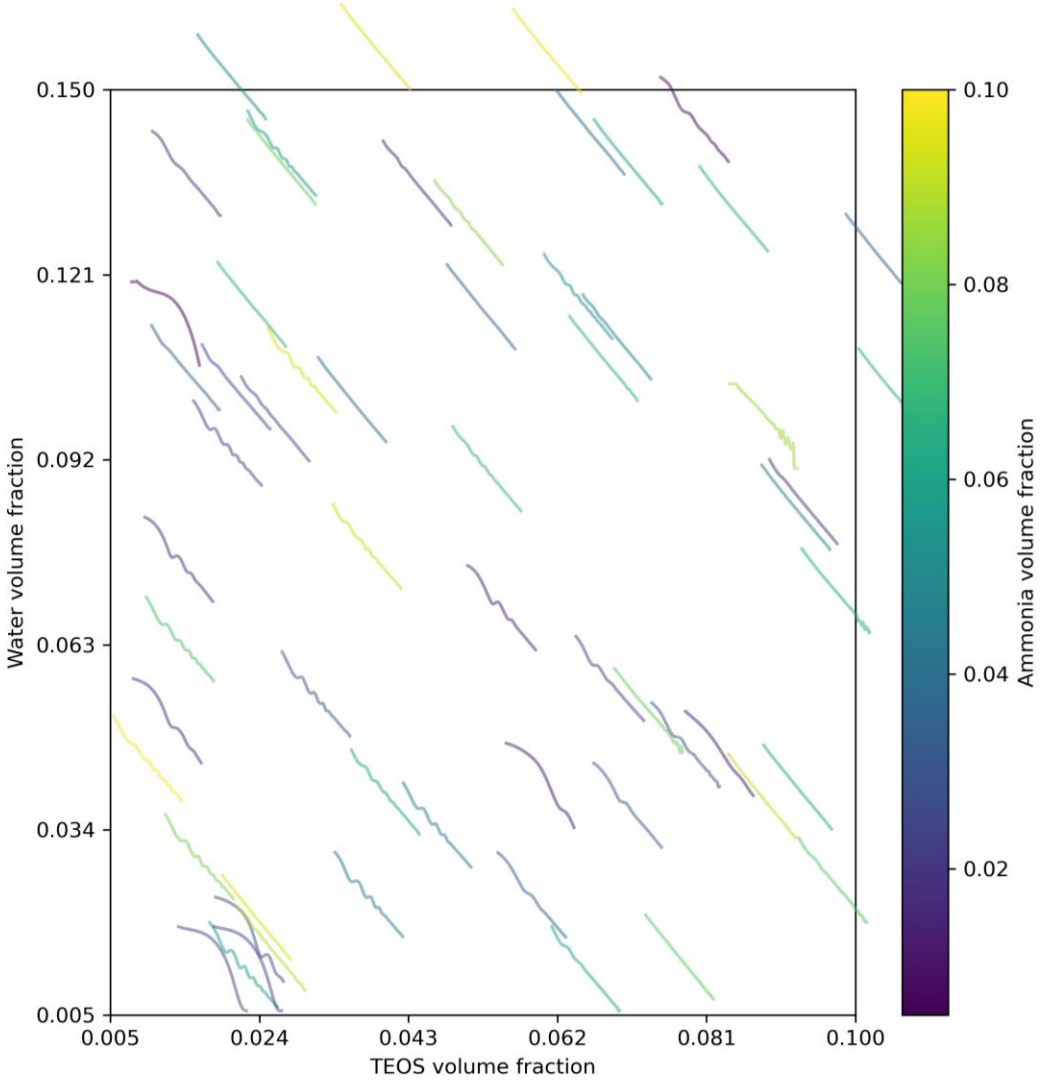


Figure 7.23: composition-scattering diagram for campaign 2, showing the regions of parameter space where monodisperse particles are made.

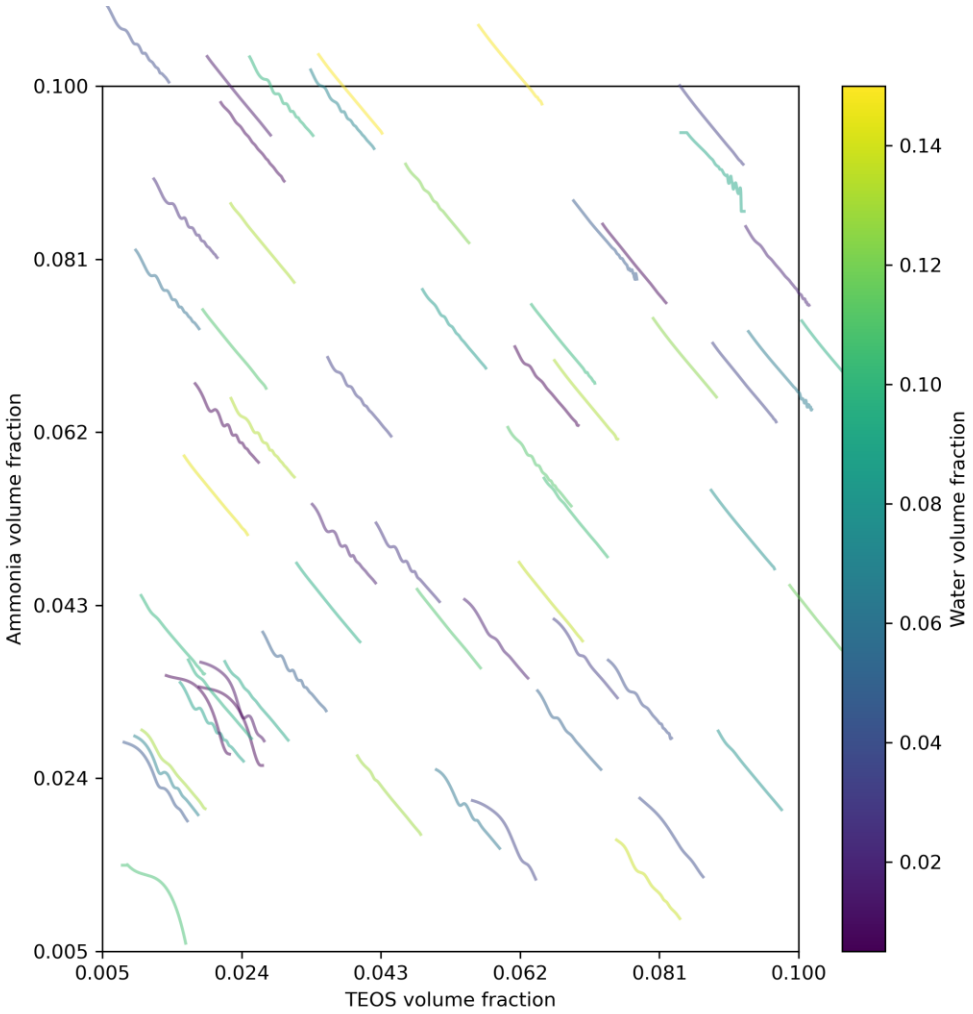


Figure 7.24: Composition-scattering diagram for campaign 2, (TEOS-Ammonia axes).

7.4.3 Experimental synthesis campaigns - conclusions

After the three rounds of optimization were run in the second set of campaigns, it became clear that larger issues needed to be addressed to enable effective morphological optimization. In the experimental campaigns that were run, none demonstrated effective convergence. Possible causes of this poor performance include inappropriate campaign configuration, failure of the amplitude-phase distance metric as an optimization scoring function, and experimental reproducibility issues.

The in-silico simulation results suggest that the campaign configuration, including amplitude-phase distance, enables the optimization of synthesis conditions in a system that behaves similarly to the experimental sol-gel synthesis system which that morphological optimization is achievable with the workflow tested here. In the following section, factors impacting the reproducibility of this synthesis will be discussed. Despite the lack of optimization convergence, this work still provided valuable insights for the development of autonomous nanoparticle optimization. The work suggests that monodisperse samples can be synthesized for a range of nanoparticle diameters using the Jubilee synthesis protocol and selected composition bounds. These results also suggest that the amplitude-phase distance metric is an effective approach to analyzing small-angle scattering data for optimization objectives.

7.5 Investigating synthesis repeatability

The experimental reproducibility of the solid spherical nanoparticle synthesis procedure was repeatedly and extensively investigated. An initial, informal reproducibility check was run during

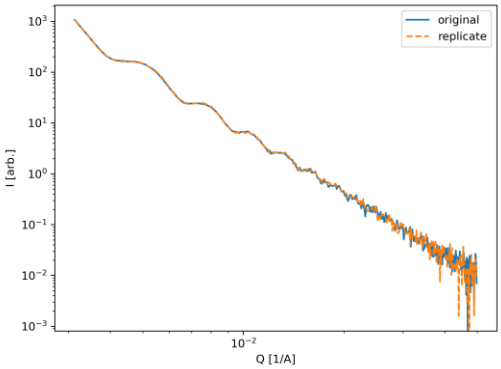


Figure 7.25: Nearly exact reproduction of initial scattering in second sample. This sample had composition TEOS:Ammonia:Water:Ethanol [uL] of 32:130:89:1448.

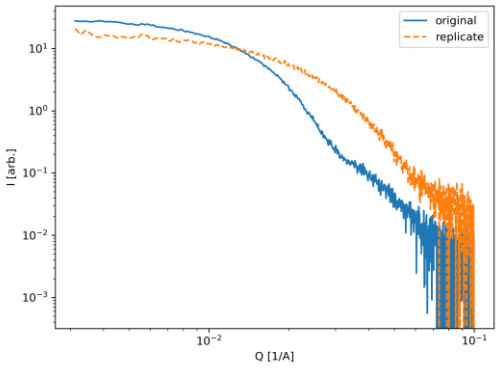


Figure 7.26: Sample with least reproducibility of scattering between initial and replicate sample. This sample had composition TEOS:Ammonia:Water:Ethanol [uL] of 23:33:18:1625.

the initial synthesis setup and workflow validation. In this check, 20 compositions were selected from the first batch of random initial screening samples to be duplicated. The replicated samples were selected because their initial synthesis had scattering that suggested monodisperse particles. Reproducibility was analyzed by qualitatively comparing the original scattering to the scattering of the replicate. Overall, reproducibility was deemed to be reasonable. Scattering from replicated samples generally matched that of the original. The closest match is shown in figure 7.25, and the worst match is shown in figure 7.26. In the best case, the scattering is nearly identical. In the worst case, the scattering is similar, but suggests different particle sizes between the two samples. These results suggested that the synthesis procedure was reasonably reproducible. Further investigation was not pursued before starting experimental optimization campaigns.

However, concerns about reproducibility were raised after observing the poor experimental optimization performance and the large impact experimental noise had on convergence in the in-silico simulation. It was noted that some samples in the lower regions of the composition space (lower left corner in figures 7.23, 7.24) seemed to vary in scattering shape from close neighbors. An additional reproducibility investigation was run, specifically focusing on compositions in the lower region of the parameter space that had been observed to create small (<100 nm) particles. A spot check to reproduce one such composition in triplicate showed that one sample (replicate 1) had notably different scattering from the other 3 samples synthesized with this composition, as shown in figure 7.27. An extensive investigation was conducted to better understand the reproducibility of the nanoparticle synthesis process as well as the factors that controlled it.

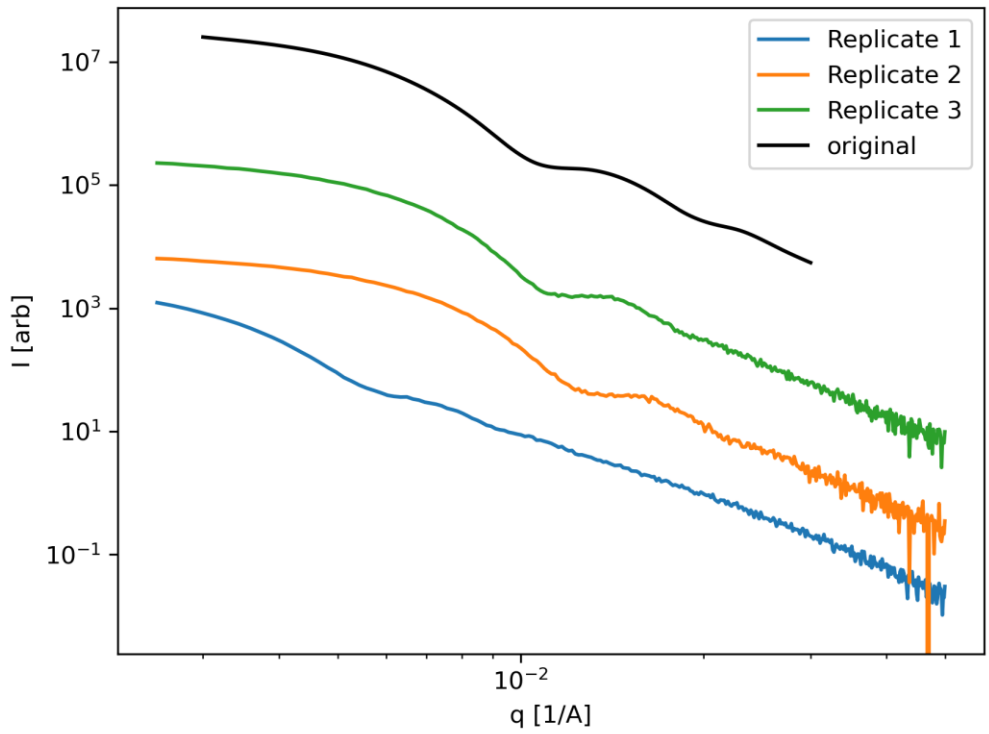


Figure 7.27: Scattering from replicates of a sample made for the second repeatability check. Sample composition [volume fractions TEOS:Ammonia:Water:Ethanol] is 0.018:0.028:0.010:0.84.

Several factors were considered that could potentially impact reproducibility. The volumetric precision of the Digital Pipette syringe tools could have a large impact on repeatability. It was known that the syringes had a low precision (10%+) at low volumes (< 100 uL) as shown in Appendix A1. Large variances in the actual delivered composition of the sample could lead to variable synthesis outcomes. The age of diluted stocks was also hypothesized to be a factor. The dilution of TEOS in ethanol could potentially 'age' the TEOS, potentially pre-hydrolyzing it or otherwise altering it in a way that would impact the synthesis. The syringe mixing step could potentially be contaminating samples with residue from previous samples. In this case, the primary concern was that the syringe would transfer seed particles to later samples, impacting the growth

kinetics of the synthesis. Evaporation of ammonia from the ammonium hydroxide solution was considered. Ammonia is volatile, and evaporation could potentially alter the concentration of delivered ammonia solution.

These factors were investigated through a series of repeatability trials. Each trial generally involved synthesizing 5 replicates of a sample composition, then characterizing the mean particle diameter and polydispersity of each sample using dynamic light scattering. The sample composition used was [volume fractions TEOS:Ammonia:Water:Ethanol] 0.0075:0.0188:0.044:0.93. This composition was selected because it exhibited the closest scattering to the 80 nm target during the final experimental optimization campaign. With the exception of the syringe precision study, all experiments were run using a 10,000 μL sample volume. Synthesis was performed with TEOS, Ammonium hydroxide, and water stocks diluted in ethanol at a 1:1.85 reactant:ethanol ratio by volume.

7.5.1 Impact of Syringe precision

Syringe precision was investigated by scaling the sample volume. This increased the volume dispensed of each reactant. The synthesis was originally performed at 1,700 μL total sample volume. This volume was selected for compatibility with the 2000 μL vials used for the experimental optimization trials. Increased volumes of 5,000 μL and 10,000 μL were tested. The larger volume syntheses were performed in the same 20,000 μL scintillation vials with pre-slit silicone septa caps used for stock solutions. The mass of each dispense was measured during the synthesis process to compare the precision of the sample composition to the reproducibility of the mean diameter. Table 7.5 contains the results of this experiment. These results show that increasing the sample volume leads to a large improvement in the precision of dispense volumes, and that

this correlates with a large reduction in the standard deviation of particle sizes. The differences in particle diameter are within the standard deviation range. These results suggest that the precision of the reactant dispensing has a large impact on sample reproducibility.

Table 7.5: Syringe precision experiment results

Sample volume [uL]	**Mean particle diameter [nm]**	**Diameter standard deviation [nm]**	**Diameter percent standard deviation**	**Volume delivered percent standard deviation (average for Ethanol, water, ammonia)**
1700	156	25	16%	5%
5000	158	14	9%	2%
10,000	148	1.9	1.3%	0.4%

7.5.2 Impact of stock age

The TEOS, ammonium hydroxide, and water stocks were diluted in ethanol to improve the volumetric precision of syringe dispenses during the second set of experimental optimization campaigns. It was hypothesized that the stock age from dilution and mixing could impact sample diameter or reproducibility. This hypothesis was evaluated in two ways. In the first, the same stock solution preparations were used to run multiple syntheses on different days. This controlled for any variability in the stock dilution. In the second, multiple batches of stock solutions were prepared at varying times in advance to run multiple synthesis experiments in one day. This controlled for any spurious day to day variability. Tables 7.6 and 7.7 show results for synthesis with the same stock solution across multiple days, and tables 7.8 and 7.9 show results for syntheses

using different stock solutions on the same day. Note that some individual sample results are included in multiple tables, such as the day 0 sample in tables 7.7 and 7.8.

Table 7.6: Synthesis across multiple days, Stock solution batch #1

Stock age at synthesis [days]	**Mean particle diameter [nm]**	**Diameter standard deviation [nm]**
0	148.7	1.9
7	153.9	11.4
8	144.9	4.9

Table 7.7: Synthesis across multiple days, Stock solution batch #2:

Stock age at synthesis [days]	**Mean particle diameter [nm]**	**Diameter standard deviation [nm]**
0	115.5	5.5
3	143.8	14.5

Table 7.8: Same-day synthesis experiment #1

Stock age at synthesis [days]	**Mean particle diameter [nm]**	**Diameter standard deviation [nm]**
0	115.5	5.5
8	144.9	4.9

Table 7.9: Same-day synthesis experiment #2

Stock age at synthesis [days]	**Mean particle diameter [nm]**	**Diameter standard deviation [nm]**
0	154.8	14.7
0	133.1	2.9
3	143.8	14.5

These experiments are limited and lead to inconclusive results. The results suggest that stock age may not play a significant role, and any impact it does have is small compared to variability in the synthesis. The two experiments that used the same stock preparations across multiple days have mixed results, with the 7-day aged sample in table 7.6 having a large diameter standard deviation, but the 8-day aged sample having a similar standard deviation to the 0-day sample. The 0-day sample in experiment 2 with a 115 nm particle average diameter is something of an outlier – this diameter is much smaller than all other samples synthesized. This makes drawing conclusions from the experiments represented in tables 7.7 and 7.8 challenging. The experiments shown in table 7.9 suggest that experiment variability is significant, as there is a larger difference in particle size and standard deviation between the two samples synthesized from fresh stock preparations as there is for the 3-day old stock preparation. While a larger set of experiments would be needed to conclusively rule out stock solution age, these results suggest it is not an important contributor.

A notable trend not visible in the descriptive statistics shown here is that the average particle size of samples tends to decrease with sample synthesis order. For many experiments, the first sample synthesized in a batch is the largest, and particle sizes steadily trend down. Table 7.10 shows an example of this for a batch of samples. This trend is not present in all experiments. It appears in around half of the reproducibility experiments and is present in all experiments with large (>10 nm) diameter standard deviations. The presence of this trend suggests that either the samples are somehow correlated, or that some change is occurring to the synthesis conditions within the timescale of the synthesis (~one hour). Contamination of subsequent samples through re-use of the shared mixing syringe was identified as a possible cause of sample correlation.

Evaporation of ammonia from the stock solutions was considered as a factor that could alter synthesis conditions.

Table 7.10: Measured particle diameters for a batch of 5 sample replicates synthesized using the standard Jubilee synthesis.

Sample order	**DLS diameter [nm]**
1	163
2	148
3	137
4	132
5	128

7.5.3 Syringe mixing with shared mix syringe

The use of a shared mixing syringe could potentially alter the synthesis process by contaminating downstream samples. The mix syringe was washed in a series of three ethanol rinse stock vials, so contamination with a meaningful amount of reactant was unlikely. However, it was possible that nanoparticle seeds transferred from initial samples could alter the growth kinetics of downstream samples. To evaluate this, parallel experiments were run. One synthesis used the standard Jubilee syringe mixing protocol used throughout this work, and the alternate mixed the sample after precursor addition and after TEOS addition with a stir plate and magnetic stir bar. Samples in the stir plate experiment were stirred for one minute on a stir plate after all precursors were added, and again after TEOS was added. All reactant dispensing was done with the Jubilee and Digital Syringe tools for the stir plate experiment. The same two-day old stock solution was used for both experiments. Table 7.11 presents the results. The size standard deviation is large and similar for both experiments. The downward diameter trend noted above is also present in both experiments. This suggests that the syringe mixing process is not leading to contamination or other issues that significantly impact the reproducibility of experiments.

Table 7.11: Results from syringe mix vs. stir plate experiment

Sample	Mean Diameter [nm]	Diameter standard deviation [nm]	Sample 1 diameter	Sample 5 diameter
Stir plate	133	12.0	154	125
Syringe mix	142	14.1	163	128

7.5.3.1 Ammonia evaporation

Ammonia evaporation from the stock solution or mixed sample precursors could be decreasing the ammonia content of the sample and impacting particle diameter. An experiment was run to confirm that ammonia content does impact particle diameter. The results are shown in figure 7.28. This experiment demonstrates that decreasing ammonia concentration does have a large impact on particle size, which is in agreement with literature consensus[26]. Just between the highest ammonia content and second-highest ammonia content, the volume fraction of ammonia changes by 0.0025, and the particle diameter drops by 27 nm. This experiment was not run with replicates so care must be taken in interpreting these results, but this does suggest that small changes in ammonia concentration could be causing the downward particle trend observed.

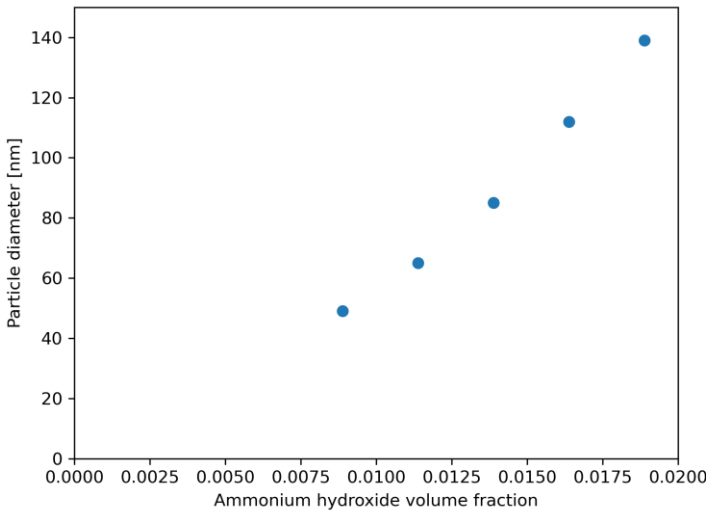


Figure 7.28: Ammonia concentration impacts particle diameter.

Two approaches to mitigating ammonia evaporation were tested. In the first, the ammonia solution was capped with a layer of mineral oil to limit

evaporation. In the second, the ammonia solution was replaced with ethanolamine. Ethanolamine has been successfully used in modified Stöber synthesis procedures[27], and is less volatile than ammonia. The vapor pressure of Ethanolamine at 20°C is 0.4 mmHg[28], while the vapor pressure of ammonia at 25°C is 7500 mmHg[29]. The mineral oil capped Ammonia experiment produced samples with average particle diameters of 147.3 +/- 10.5. A downward diameter trend was present. The first synthesized particle diameter was 159 nm, and 5th was 141 nm. The ethanolamine experiment was run by swapping the ammonia solution for an ethanolamine:ethanol solution. Previous work found that directly replacing ammonia with ethanolamine resulted in larger particles[27], so for this experiment the dilution of ethanolamine was doubled from that of ammonia to 1 part ethanolamine: 5.7 parts ethanol. All other stock preparations and experimental protocols remained the same. Two batches of ethanolamine particles were synthesized. Results from this experiment are shown in table 7.12. While the diameter standard deviations are smaller for ethanolamine than for similar experiments with ammonia, there is still a notable downward trend in particle diameters, particularly in the second experiment. This suggests that ammonia evaporation may play a role in the observed downward trend and may drive large standard deviations in inter-experiment diameters, but other factors are also likely causing this effect.

Table 7.12: Results for ethanolamine synthesis

Experiment	**Mean diameter [nm]**	**Diameter standard deviation [nm]**	**First particle diameter [nm]**	**Fifth particle diameter [nm]**
Ethanolamine 1	168	6.8	179	164
Ethanolamine 2	159	5.7	165	151

7.5.3.2 Synthesis reproducibility conclusions

This investigation provided results that develop an understanding of the reproducibility of the Stöber sol-gel nanoparticle synthesis and factors that influence this reproducibility. The results

suggest that composition precision is an important factor in enabling reproducible syntheses. On the Jubilee synthesis platform, precision can be improved by increasing the sample volume. This does come at a trade-off of decreased sample capacity. The Jubilee deck has a limited amount of space and increasing sample volume increases both the amount of deck space required per sample (12 samples per deck slot at 10,000 μL vs 40 at 1700 μL) and the amount of stock solution required for the experiment. Based on the results from this work, 10,000 μL is the recommended sample volume for future experiments. However, the possibility of realizing additional reproducibility improvements by increasing sample volume further should be investigated. The results suggest that ammonia evaporation may be influencing reproducibility. The mineral oil cap mitigation did not make a large difference in reproducibility and introduced the risk of contaminating the sample with another component. Replacing ammonia with ethanolamine as the base catalyst may improve reproducibility, and this could be further investigated. However, this is a major change to the synthesis and results in working in an entirely different chemical space. Given that most of the silica nanoparticle research in literature has used ammonia as a base, solutions that enable the continued use of ammonia are preferred. Pre-slit septa were used in this work for compatibility with the blunt-tipped syringe needles used. Further improvements may be realized by using non-slit septa and a sharp, septa piercing needle. Stock age and syringe mixing do not appear to have significant impacts on synthesis reproducibility.

An overall synthesis reproducibility for the Jubilee synthesis protocol can be estimated from the experiments executed in this investigation. Samples synthesized using a 10,000 μL sample volume with ammonia are included in this average. The average diameter of particles for all such samples made as part of this reproducibility was 142 nm +/- 11.3 nm. This is a relative standard

deviation of 8%. No known similar synthesis reproducibility investigations are available to compare this value against.

This line of investigation was not followed further due to time constraints. Further investigation is warranted to validate the findings of this work before further optimization campaigns are attempted. The continued presence of the downward diameter trend is a primary concern that should be investigated moving forward. Further steps to control ammonia evaporation, such as the solid septa described above, should be tested. Also, larger batches of samples should be run to understand how the trend continues. Synthesis timing could also be investigated to test the evaporation hypothesis. Time delays between synthesis could be added. If increasing delay time exacerbates the downward trend, this would support the evaporation hypothesis. Ultimately, a certain amount of experimental stochasticity is likely to remain. Appropriate modeling and optimization approaches should be selected to work within these constraints.

7.6 Progress towards an orchestration and data management platform for nanoparticle synthesis experiments

The campaigns described in this chapter have been run in a human orchestrated fashion. Samples were manually loaded into a well plate and transferred to the SAXS instrument for characterization, and data processing and Bayesian optimization workflows were run manually through notebooks. While the selection of experiments was autonomous and the synthesis was automated, each step of the process required human orchestration and oversight. To achieve full autonomy, a system to automatically orchestrate and manage these steps is needed. The capability to integrate the Jubilee synthesis platform with an X-ray scattering instrument with the NIST-AFL

sample loader has been developed and demonstrated. As described in chapter 6, this platform provides a reliable means of transferring samples from the Jubilee deck into a flow cell for characterization. To fully close the loop, additional infrastructure is needed to collect and manage sample data, run the data processing and Bayesian optimization jobs, and pass results back to the Jubilee system for synthesis. In this section, the architecture used to manage the integration and orchestration of these components in the existing implementation is discussed, and progress toward a data and compute management platform is described.

Orchestration describes the integration required to coordinate the disparate steps involved in an autonomous synthesis experiment. It involves coordinating multiple synthesis and characterization steps across different hardware and managing the data processing and experimental design steps that happen on the 'back end' to enable an autonomous workflow. Orchestration involves networking, communication, state, and data management administration that enables complex automation. Any autonomous experimentation endeavor involves some form of orchestration infrastructure. Many examples of these orchestration software systems have been shared with the intention of providing a shared community resource to build from[30,31]. However, adoption outside of the originating labs has been slow. This is likely due to hardware-specific design decisions made during the development of the orchestration software, and unclear long-term support and development commitments from the academic labs authoring this software. Many SDL orchestration systems use a centralized approach, in which a manager process controls all aspects of the experiment automation. Centralized orchestration can allow for sophisticated experimental concurrency and effective error handling but can require significant integration and setup work to integrate control over every hardware component into a single controller. For this

project, re-using an existing open-source lab orchestration system with centralized control was briefly considered, but due to concerns over integration effort and flexibility, it was decided to implement an alternative approach from scratch.

The orchestration approach taken for the automated synthesis – NIST-AFL – SAXS integration grew organically into a decentralized, modular system. Each hardware module that plays a discrete role in the workflow process workflow (synthesis on Jubilee, NIST-AFL sample loading, SAXS measurement, data processing and management) is independently responsible for executing its tasks. Each component has a means to receive a task request from an upstream module, and to notify relevant modules of its state and task completion status. Some modules (NIST-AFL and data management platform, discussed below) have formal task queues and status systems to manage this interface. Others, like the USAXS integration developed at APS, use simple informal methods like file uploads to communicate status. The autonomous experiment orchestration network for this experiment works as follows. The processing of a sample is started by the jubilee synthesis system (running from a notebook) querying the sample server to fetch the next sample to be made (either from a randomly selected baseline list of samples to make, or the active learning optimizer). Once the synthesis is complete, the synthesis process calls the AFL to load the sample. The AFL loads the sample, communicating with the Jubilee to manage the shared deck space of the AFL sample catch arm and the Jubilee tool. Once the sample is loaded, the AFL sends a request to the SAXS driver to start the measurement by means of an instrument-specific communication protocol. (In the APS integration, due to network security restrictions, this was done by updating a timestamp stored in a file on a jointly accessible AWS EC2 instance. A USAXS driver script watched the timestamp and triggered a measurement on a change). The SAXS driver

uploads the data to the sample server when the measurement is complete. On data upload, the sample server initiates data processing and experiment planning, adding a new proposed sample to the queue and closing the loop. The NIST-AFL process also watches for the measurement file to be uploaded, which triggers the rinse cycle for the sample cell.

This decentralized orchestration system may seem like a 'hack'. In fact, significant portions of it were developed on-site at the APS beamline to circumvent last-minute integration challenges. While some of the communication methods should be formalized (such as replacing a shared scp-updated timestamp file with a proper task queue), this approach has proven to be simple, extensible, and reliable, and has several advantages over a more formal centralized approach to task orchestration. This decentralized approach is simple to modify and extend. Adding another component to the workflow orchestration is a matter of modifying upstream task software to call the new component and configuring the new component to trigger subsequent downstream tasks. Each module maintains complete control over its own internal state, and only exposes high-level tasks that are needed for integration control. For example, the NIST-AFL sample loader only provides a 'load sample' task for external control and maintains all solenoid valve and state management internally. This facilitates modular development and makes modules easily interchangeable. The NIST-AFL sample loader system could easily be swapped onto an Opentrons OT2 robot without modifying the workflow orchestration system, for example. Perhaps most importantly, this is an effective 'minimum viable implementation' for workflow orchestration. It allows researchers to get up to speed with new integrations quickly and easily make changes that will surely be discovered through trial and error. This helps focus work on developing correct,

reliable automation that supports the end-goal science case, instead of bogging scientists down in another layer of software engineering and systems administration complexity.

The above proposed orchestration workflow for the sol-gel synthesis process requires a 'sample server' component that manages the data processing, Bayesian optimization, and data management steps. A prototype of such a service has been developed[32]. This service runs as a Flask web application. It provides endpoints for uploading sample data, running a data processing and Bayesian optimization pipeline, and managing a queue of proposed samples. Sample information is stored in a relational database table that tracks sample identifiers, synthesis status, and the location of associated data files. This service provides the missing computational backend component needed to enable fully autonomous execution of this sol-gel synthesis process. Development was not fully finished. However, it may provide a useful component for future closed-loop experiments with this platform, or a template for similar sample server components of decentralized experiment orchestration platforms.

7.7 Conclusions

The results described in this chapter lay the groundwork for future silica nanoparticle optimization campaigns. Most of the infrastructure to enable autonomous experimentation with this platform has been developed. Through initial optimization campaigns, core assumptions about integrating Bayesian optimization with X-ray scattering data were validated, and important factors influencing optimization performance were identified.

With the sample server component described above, all components exist to run the sol-gel synthesis process described in Chapter 6 and this chapter in a fully automated, closed-loop

manner. The full automated synthesis, sample transfer, and characterization have been successfully demonstrated at the APS 12-ID-E USAXS beamline, and a proof-of-concept integration has been developed for the UW Xenocs Xeuss 3.0 instrument. The sample server module has not been integrated into this process but has been tested in isolation. The experimental and in-silico campaigns described in this chapter indicate that the amplitude-phase distance metric approach for optimization with small-angle scattering data is effective. This approach provides a straightforward framework for managing the complex task of scattering data analysis in a fully automated process, allowing for scattering data to be reliably integrated into a nanoparticle optimization campaign. Extending this method to include optimization of mesoporous nanoparticles will require more development. The amplitude-phase metric has not been evaluated with the diffraction peaks seen for scattering from porous materials. It may be possible to simply use an amplitude phase distance between measured scattering and computed scattering for a mesoporous nanoparticle target. However, this may overload the metric. Alternative approaches would involve splitting the low-q 'particle' scattering and high-q 'porous' scattering into separate analysis pipelines and managing the separate evaluations with a multi-objective optimization approach. This work also demonstrated that experimental reproducibility is critically important for successful optimization. An investigation suggested that reproducibility for the experimental conditions used in the initial experimental optimization campaigns was likely poor, possibly contributing to the lack of convergence. Increasing the volume of samples greatly improves repeatability, and managing ammonia evaporation may yield further improvements.

This project is nearly ready for additional experimental optimization experiments to be run. It is recommended to pursue additional reproducibility trials to better understand the causes of the

downward diameter trend observed. However, all major infrastructure components are in place and several synthesis-specific issues have been addressed, making autonomous optimization of silica nanoparticles a feasible goal for near-term research.

7.8 References

(1) Narayan, R.; Nayak, U. Y.; Raichur, A. M.; Garg, S. Mesoporous Silica Nanoparticles: A Comprehensive Review on Synthesis and Recent Advances. *Pharmaceutics* **2018**, *10* (3), 118. https://doi.org/10.3390/pharmaceutics10030118.

(2) Vallet-Regí, M.; Schüth, F.; Lozano, D.; Colilla, M.; Manzano, M. Engineering Mesoporous Silica Nanoparticles for Drug Delivery: Where Are We after Two Decades? *Chem. Soc. Rev.* **2022**, *51* (13), 5365–5451. https://doi.org/10.1039/D1CS00659B.

(3) Yu, X.; T. Williams, C. Recent Advances in the Applications of Mesoporous Silica in Heterogeneous Catalysis. *Catal. Sci. Technol.* **2022**, *12* (19), 5765–5794. https://doi.org/10.1039/D2CY00001F.

(4) Scheurer, C.; Reuter, K. Role of the Human-in-the-Loop in Emerging Self-Driving Laboratories for Heterogeneous Catalysis. *Nat. Catal.* **2025**, *8* (1), 13–19. https://doi.org/10.1038/s41929-024-01275-5.

(5) Roberts, G.; Nieh, M.-P.; W.K. Ma, A.; Yang, Q. Automated Structural Analysis of Small Angle Scattering Data from Common Nanoparticles via Machine Learning. *Digit. Discov.* **2025**. https://doi.org/10.1039/D5DD00059A.

(6) Stöber, W.; Fink, A.; Bohn, E. Controlled Growth of Monodisperse Silica Spheres in the Micron Size Range. *J. Colloid Interface Sci.* **1968**, *26* (1), 62–69. https://doi.org/10.1016/0021-9797(68)90272-5.

(7) *Automation-Hardware/Vial Holders/2mLdram_44_wellplate_withRetainer at master · pozzo-research-group/Automation-Hardware*. https://github.com/pozzo-research-group/Automation-Hardware/tree/master/Vial%20Holders/2mLdram_44_wellplate_withRetainer (accessed 2025-06-05).

(8) Glatter, O. Chapter 3 - The Inverse Scattering Problem. In *Neutrons, X-rays, and Light (Second Edition)*; Lindner, P., Oberdisse, J., Eds.; Elsevier, 2025; pp 61–90. https://doi.org/10.1016/B978-0-443-29116-6.00004-7.

(9) SasView. *SasView*. https://sasview.github.io/ (accessed 2025-05-12).

(10) Monge, N. Development of a Machine Learning Methodology for Nanoparticle Classification from Small-Angle X-ray Scattering (SAXS) Data. phdthesis, Université Grenoble Alpes [2020-....], 2024. https://theses.hal.science/tel-05039356 (accessed 2025-05-01).

(11) Archibald, R. K.; Doucet, M.; Johnston, T.; Young, S. R.; Yang, E.; Heller, W. T. Classifying and Analyzing Small-Angle Scattering Data Using Weighted k Nearest Neighbors Machine Learning Techniques. *J. Appl. Crystallogr.* **2020**, *53* (2), 326–334. https://doi.org/10.1107/S1600576720000552.

(12) Tomaszewski, P.; Yu, S.; Borg, M.; Rönnols, J. Machine Learning-Assisted Analysis of Small Angle X-ray Scattering. In *2021 Swedish Workshop on Data Science (SweDS)*; 2021; pp 1–6. https://doi.org/10.1109/SweDS53855.2021.9638297.

(13) SasView/Sasmodels, 2025. https://github.com/SasView/sasmodels (accessed 2025-05-15).

(14) Martin, T. B.; Sutherland, D. R.; McDannald, A.; Kusne, A. G.; Beaucage, P. A. Autonomous Small-Angle Scattering for Accelerated Soft Material Formulation Optimization. arXiv March 14, 2025. https://doi.org/10.48550/arXiv.2503.11859.

(15) Vaddi, K.; Chiang, H. T.; Pozzo, L. D. Autonomous Retrosynthesis of Gold Nanoparticles via Spectral Shape Matching. *Digit. Discov.* **2022**, *1* (4), 502–510.
(16) Vaddi, K. Kiranvad/Amplitude-Phase-Distance, 2025. https://github.com/kiranvad/Amplitude-Phase-Distance (accessed 2025-06-05).
(17) Savitzky, Abraham.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36* (8), 1627–1639. https://doi.org/10.1021/ac60214a047.
(18) *savgol_filter — SciPy v1.15.3 Manual*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.savgol_filter.html (accessed 2025-06-12).
(19) Pozzo-Research-Group/Saxs_data_processing, 2024. https://github.com/pozzo-research-group/saxs_data_processing (accessed 2024-02-06).
(20) *pozzo-research-group/silica-np-synthesis*. https://github.com/pozzo-research-group/silica-np-synthesis (accessed 2025-06-05).
(21) Balandat, M.; Karrer, B.; Jiang, D.; Daulton, S.; Letham, B.; Wilson, A. G.; Bakshy, E. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2020; Vol. 33, pp 21524–21538.
(22) Lo, S.; G. Baird, S.; Schrier, J.; Blaiszik, B.; Carson, N.; Foster, I.; Aguilar-Granda, A.; V. Kalinin, S.; Maruyama, B.; Politi, M.; Tran, H.; D. Sparks, T.; Aspuru-Guzik, A. Review of Low-Cost Self-Driving Laboratories in Chemistry and Materials Science: The "Frugal Twin" Concept. *Digit. Discov.* **2024**, *3* (5), 842–868. https://doi.org/10.1039/D3DD00223C.
(23) Ament, S.; Daulton, S.; Eriksson, D.; Balandat, M.; Bakshy, E. Unexpected Improvements to Expected Improvement for Bayesian Optimization. arXiv January 7, 2025. https://doi.org/10.48550/arXiv.2310.20708.
(24) *botorch.acquisition — BoTorch documentation*. https://botorch.readthedocs.io/en/latest/acquisition.html#ament2023logei (accessed 2025-05-15).
(25) Volk, A. A.; Abolhasani, M. Performance Metrics to Unleash the Power of Self-Driving Labs in Chemistry and Materials Science. *Nat. Commun.* **2024**, *15* (1), 1378. https://doi.org/10.1038/s41467-024-45569-5.
(26) Ghimire, P. P.; Jaroniec, M. Renaissance of Stöber Method for Synthesis of Colloidal Particles: New Developments and Opportunities. *J. Colloid Interface Sci.* **2021**, *584*, 838–865. https://doi.org/10.1016/j.jcis.2020.10.014.
(27) Meier, M.; Ungerer, J.; Klinge, M.; Nirschl, H. Synthesis of Nanometric Silica Particles via a Modified Stöber Synthesis Route. *Colloids Surf. Physicochem. Eng. Asp.* **2018**, *538*, 559–564. https://doi.org/10.1016/j.colsurfa.2017.11.047.
(28) PubChem. *Monoethanolamine*. https://pubchem.ncbi.nlm.nih.gov/compound/700 (accessed 2025-05-04).
(29) PubChem. *Ammonia*. https://pubchem.ncbi.nlm.nih.gov/compound/222 (accessed 2025-05-04).
(30) Fei, Y.; Rendy, B.; Kumar, R.; Dartsi, O.; Sahasrabuddhe, H. P.; McDermott, M. J.; Wang, Z.; Szymanski, N. J.; Walters, L. N.; Milsted, D.; Zeng, Y.; Jain, A.; Ceder, G. AlabOS: A

Python-Based Reconfigurable Workflow Management Framework for Autonomous Laboratories. *Digit. Discov.* **2024**, *3* (11), 2275–2288. https://doi.org/10.1039/D4DD00129J.
(31) Sim, M.; Vakili, M. G.; Strieth-Kalthoff, F.; Hao, H.; Hickman, R. J.; Miret, S.; Pablo-García, S.; Aspuru-Guzik, A. ChemOS 2.0: An Orchestration Architecture for Chemical Self-Driving Laboratories. *Matter* **2024**, *7* (9), 2959–2977. https://doi.org/10.1016/j.matt.2024.04.022.
(32) *brendenpelkie/usaxs_processing: Run processing for USAXS data for silica nanoparticle optimization campaign*. https://github.com/brendenpelkie/usaxs_processing (accessed 2025-05-10).

8 Conclusions and Outlook

Autonomous experimentation is poised to accelerate the development of mesoporous silica materials for a variety of applications. This approach promises a means of retrosynthetic control over important material structure properties like particle size, size polydispersity, and pore ordering. These properties are determined by complicated interactions between reactant identities and composition as well as synthesis conditions. Determining appropriate conditions to generate silica materials with specific target properties is a challenging process that could be streamlined with automated experimentation and closed loop optimization. More broadly, autonomous experimentation is of great interest to researchers throughout materials science and chemistry fields and is likely to gain widespread adoption in the coming years. Flexible, accessible automation infrastructure will be critical to making these systems widely accessible. Open-hardware automation platforms are important for enabling accessible deployment and democratization of autonomous experimentation. The current moment in infrastructure development is a great time to build and advocate for these systems. Standards and best practices are still being developed, so new systems can still be adopted before alternative proprietary solutions become entrenched.

The work presented here represents advancements toward these goals. It was demonstrated that automated synthesis and characterization of mesoporous colloidal silicas is an effective way to explore synthesis parameter spaces. The workflow developed for mesoporous silicas is capable of synthesizing dozens of samples in a 24-hour period. It was used to identify interesting sample compositions, including those that synthesize a range of monodisperse spherical particles and multiple ordered pore phases. The integration of this system with an ultra- small-angle X-ray

scattering instrument demonstrates a powerful system for the synthesis and characterization of a range of sol-gel accessible materials in high throughput.

This work also demonstrates and advocates for the promise of a ‘democratized’ future for autonomous experimentation. The Science-Jubilee project presents a vision and a model for how open-hardware automation infrastructure can play a key role in autonomous laboratories. This platform provides common baseline automation capabilities, namely liquid handling, that are frequently provided by stand-alone equipment in self-driving labs. The system’s tool changing mechanism and extensible software interface presents new possibilities for researchers to integrate automation into their experiments. This makes it possible to build novel, multi-process workflows without needing to handle the complexity of robotic sample transfer. The open-source nature of the Science-Jubilee project is key to this flexibility. Modular, extensible motion platforms like Jubilee will be important to enable the application of automation to the long tail of difficult, project-specific research. The integrated sol-gel synthesis platform also reinforces the importance of openly shared automation equipment. This platform makes use of 3 open-hardware projects (Science-Jubilee, Digital Pipette, and NIST-AFL) to enable a novel automated workflow. Integrating existing, well-developed automation equipment enabled this project to progress to a functional, reliable automation platform rapidly. This approach can provide a model for the development of task-specific automation by domain scientists in the future.

Appendix 1 Additional details on automated sol-gel synthesis platform

A1.1 Polydisperse sphere model fits of selected USAXS scattering data

Model fits were performed in SasView, version 6.

Sample 1

Fit chi-squared: 311

Particle radius: 3014 A +/- 2.4

Radius PDI: 0.0332+/-0.00069

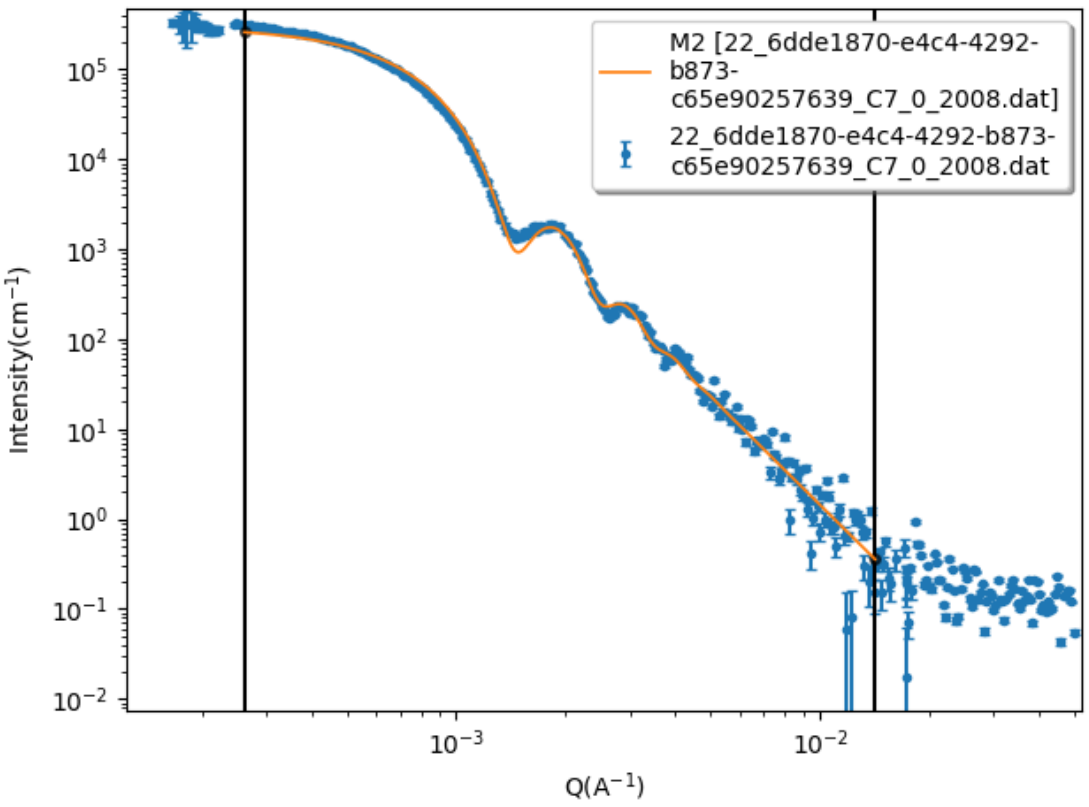


Figure A1.1: Polydisperse sphere model fit for sample 1 USAXS data.

Sample 6

Radius : 3163 +/- 2.5 A

Radius PDI: 0.083 +/- 0.0006

Chi sq: 26.8

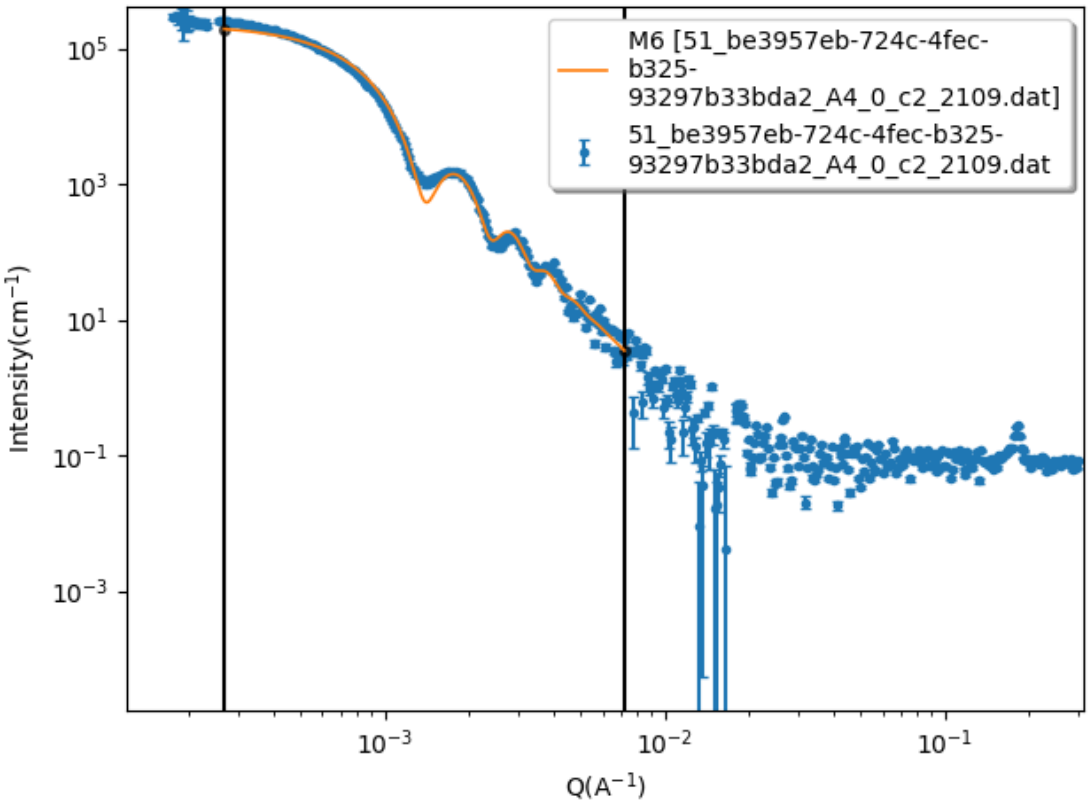


Figure A1.2: Polydisperse sphere model fit for sample 6 USAXS data

Table A1.1: Sasview polydisperse sphere model fit results for additional monodispersed disperse samples not discussed directly in the manuscript:

Sample UUID	Radius [A]	Radius error [A]	PDI	PDI error	Fit chi-squared
586e06c8-0f02-43a3-9f2f-81e46d43ef64	1579.8	1.989	0.192	0.0011	368.43
c5ce72e5-18c6-44cb-b802-1ee061e79d7c	1091.6	0.99	0.185	0.00095	244.28
1c775223-74ff-4c55-9018-cc6c61ad0bb5	2512.3	6.02	0.173	0.0017	173.07
4356786c-177a-45a2-a2c9-36ce1e03c05b	1115.2	3.8	0.188	0.003	170.76

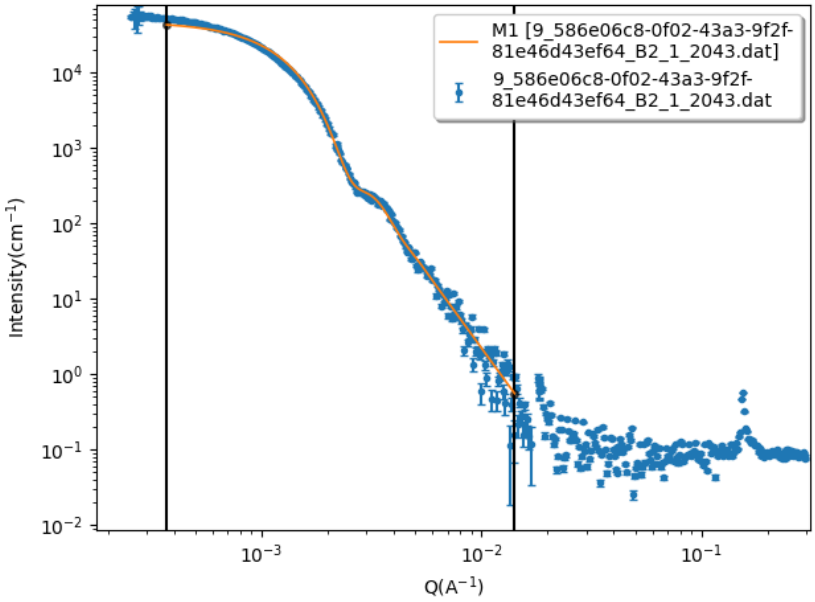


Figure A1.3: Fit for sample 586e06c8-0f02-43a3-9f2f-81e46d43ef64

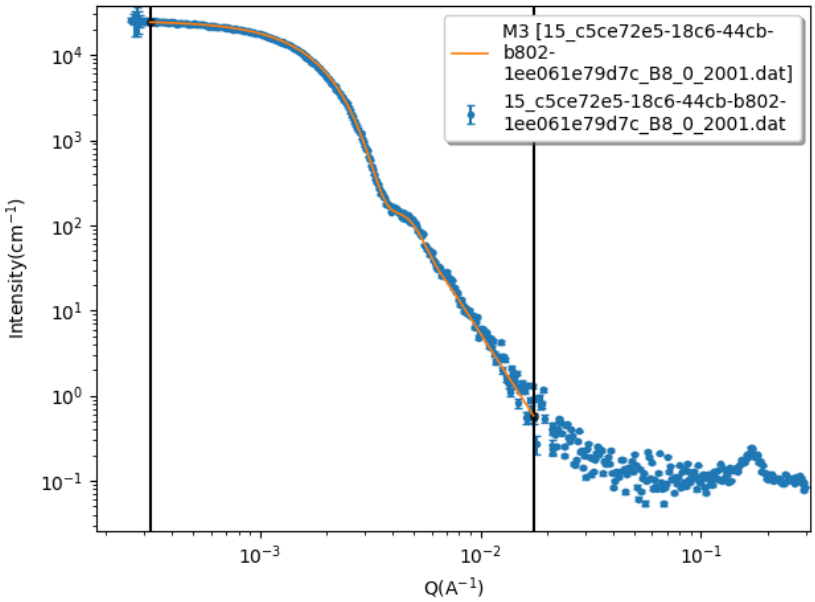


Figure A1.4: Fit for sample c5ce72e5-18c6-44cb-b802-1ee061e79d7c

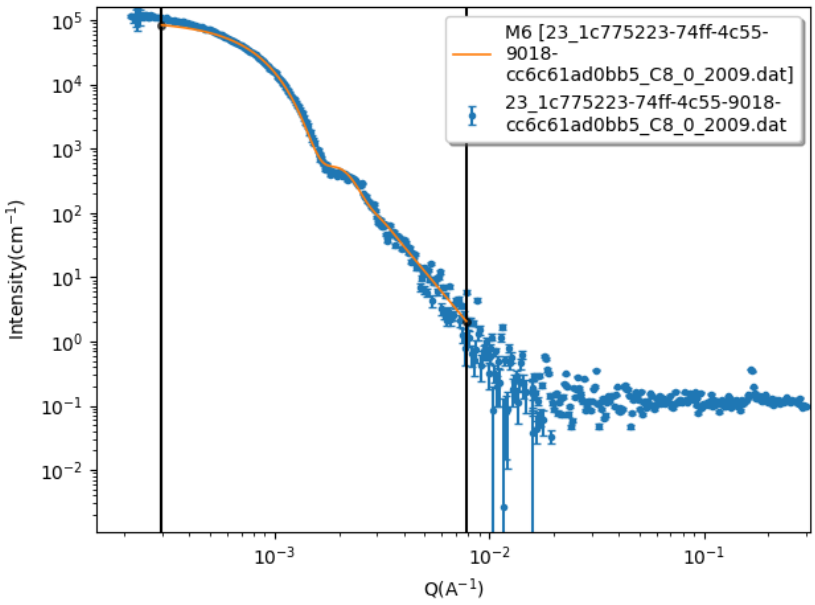


Figure A1.5: Fit for sample 1c775223-74ff-4c55-9018-cc6c61ad0bb5

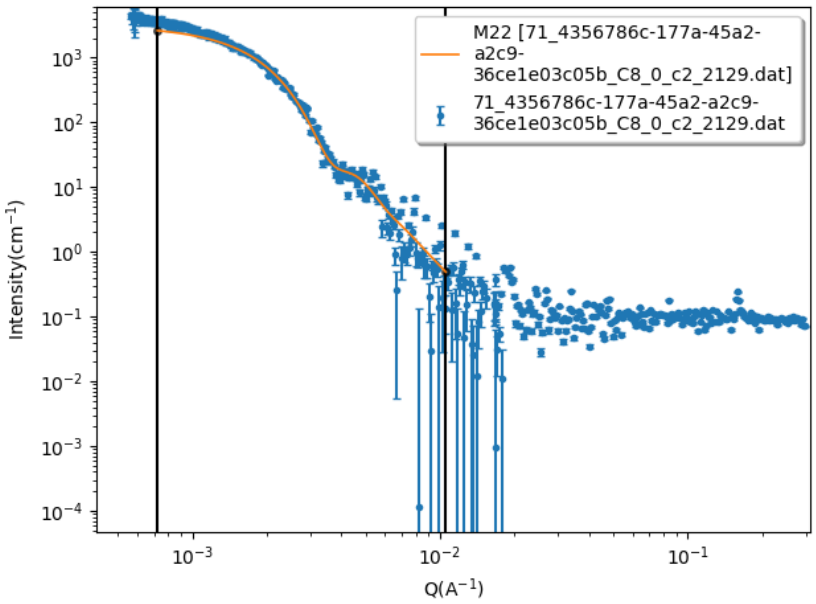


Figure A1.6: Fit for sample 4356786c-177a-45a2-a2c9-36ce1e03c05b

A1.2 Syringe tool accuracy and precision validation:

The digital pipette syringe tools were calibrated through gravimetric calibration, then validated. Validation results are shown here.

Definitions:

- $Percent\ Error\ (accuracy) = \frac{|Mean\ delivered\ volume - target\ volume|}{target\ volume} * 100$
- $Percent\ Error\ (precision) = \left|\frac{standard\ deviation\ of\ delivered\ volumes}{mean\ delivered\ volume}\right| * 100$

Table A1.2: Accuracy and precision for 1 cc disposable plastic syringe:

N = 5

Target volume [uL]	Percent error (accuracy)	Percent error (precision)
2	29.0	225.6
5	2.3	80.0
10	12.2	51.0
15	7.1	23.8
25	2.0	15.4
50	4.3	8.5
100	1.3	5.2
500	1.5	2.5

Table A1.3: Accuracy and precision for 1 cc glass syringe:

N = 5

Target volume [uL]	Percent error (accuracy)	Percent error (precision)
5	13.4	92.8
10	9.2	37.3
25	1.3	13.2
50	1.5	7.3
100	2.6	6.7

Table A1.4: Accuracy and precision for 10 cc plastic syringe

N = 5

Target volume [uL]	Percent error (accuracy)	Percent error (precision)
20	28.5	223
50	7.6	90.3
100	1.0	39.1
150	11.3	2.9
250	1.1	5.5
500	1.2	3.8

A1.3 Composition-scattering diagrams for additional components

Diagrams for TEOS vs. Ethanol and CTAB vs. Pluronic are shown in the main text. Diagrams for TEOS vs. ammonium hydroxide and TEOS vs. water are shown here.

A1.3.1 USAXS scattering

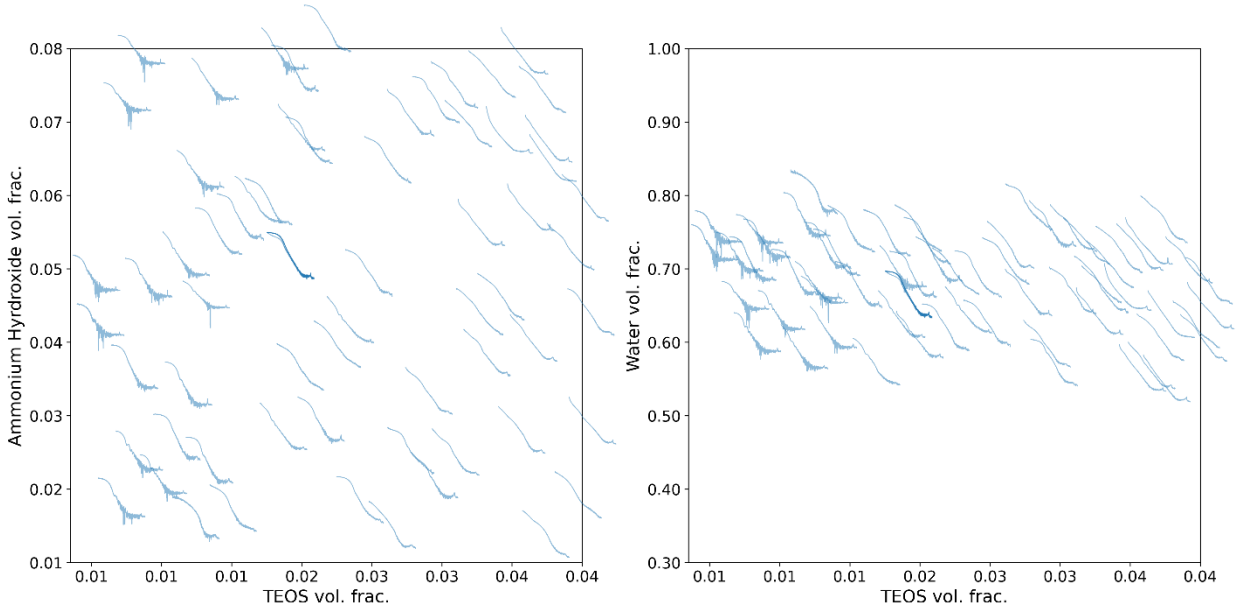


Figure A1.7: USAXS composition-scattering plot for mesoporous batch synthesis described in chapter 6.

A1.3.2 SAXS scattering

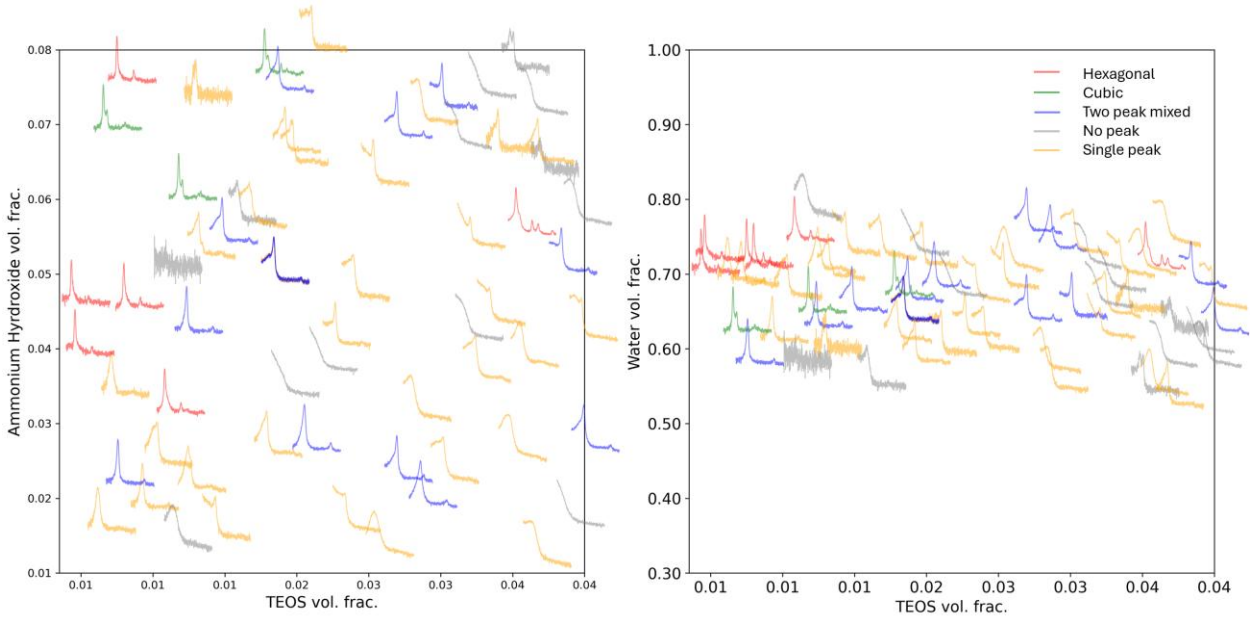


Figure A1.8: SAXS composition-scattering plot for mesoporous batch experiment described in chapter 6.

A1.4 Time growth experiment

The growth kinetics of the colloidal silica synthesis was assessed by measuring the particle size with dynamic light scattering at regular intervals after TEOS addition. The results from this experiment (figure A1.9) suggest that particle growth is complete after 2 hours. This experiment was performed using the composition of sample 1.

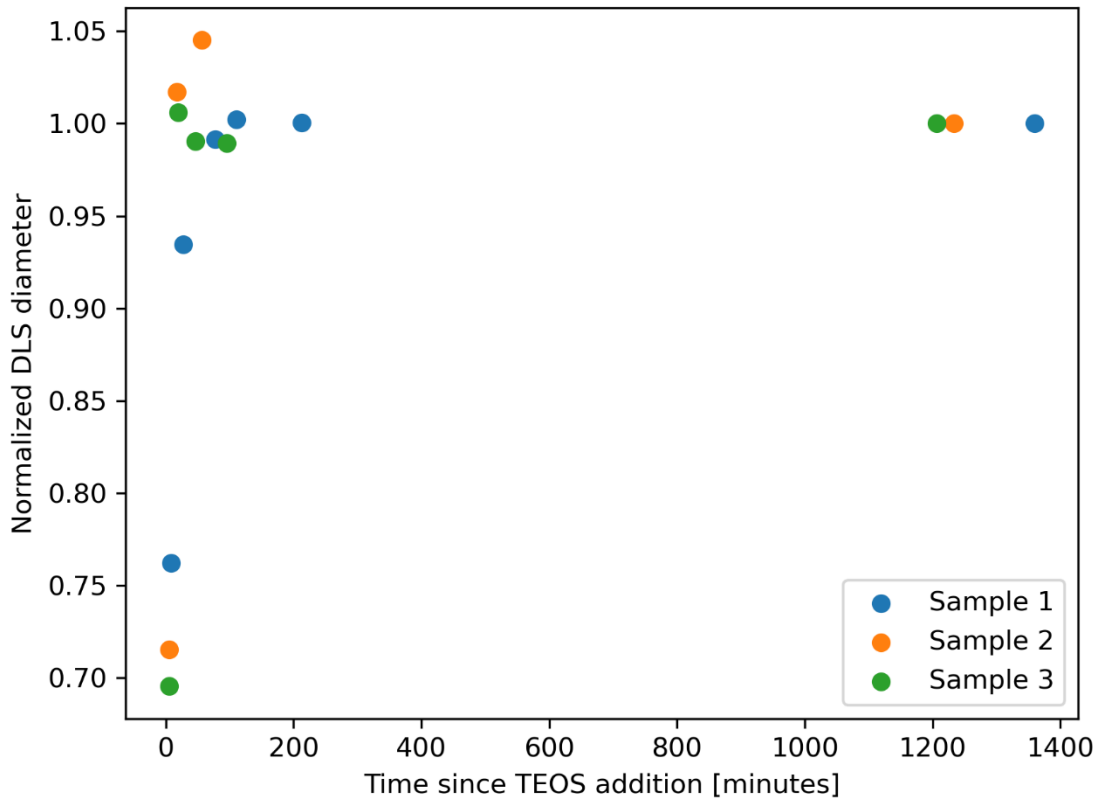


Figure A1.9: Particle size vs time.

Appendix 2 Engineering-driven chemical design space constraints for high throughput experiment planning

The methods described in this section were largely developed by Dr. Jaime Rodriguez. My contribution has been to formalize the methodology described below and to apply it to new areas of research interest. This work is included here to include new results developed with this method for future reference.

A2.1 Introduction

With the rise of high throughput experimentation in materials research and the increase experimental capability it enables, answering the question 'what should we test in this experiment?' can become challenging. Traditional manual experimentation, particularly research that seeks to identify or develop compounds for a specific purpose, is generally planned by researchers applying their knowledge and intuition about a small subset of chemical space[1]. The expanded bandwidth of high throughput experimentation makes it possible to consider a broader range of chemical species in these experiments[2]. To fully leverage the capabilities of high throughput experimentation, researchers should explore beyond well-known chemical families. However, with hundreds of millions of chemical species to choose from, a strategy to select and prioritize chemicals for experimentation is necessary. The selection of species for inclusion in an experiment can be thought of as part of the experimental design space, in an analogous way to how continuous parameter ranges can be considered to define a design space in traditional experimental design[3]. Many material discovery or development problems could potentially be addressed by

searching through the space of known chemicals in a guided manner. If successful, this approach could allow known, potentially industrially developed chemicals to address needs in new engineering applications. A structured, efficient search is needed to enable this. Careful consideration of the criteria used in outlining this design space should be done. At a minimum, successful candidate species need to exhibit the chemical and physical properties necessary to function in their desired application. Additionally, these materials are often required to meet additional constraints on safety, availability, practicality, or other 'engineering' conditions that are not directly related to the material's ability to meet the required functionality but are necessary for the implementation of a safe, reliable, and cost-effective system. For example, a battery electrolyte solvent needs to exhibit good solubility for charge carrying ions as well as ionic and electronic conductivity within a required range. However, solvent with the highest recorded solubility for species of interest and 'perfect' conductivity for the application will be inappropriate in mainstream applications if it is pyrophoric, impossible to produce at industrial quantities, or acutely toxic in small amounts. Given the importance of these engineering constraints in determining fitness of a candidate material for an application, they should be considered early on during an initial screening process before effort is spent pursuing dead-end candidate compounds with excellent chemical or physical properties but show stopping violations of other constraints. This approach has a strong analogy to the practice of screening drug leads on ADMET (Absorption, distribution, metabolism, excretion, and toxicity) properties before investing significant expense in their development as therapeutics[4]. Over the course of several material search projects of the nature described here, our group has developed a screening strategy and implementation that enables high throughput experimentation researchers to identify and prioritize

promising candidate compounds that are likely to meet all relevant criteria for a given end application[5]. This chapter describes this screening framework and showcases a few examples of how it has been applied to real-world problems.

A2.2 Selection strategy and implementation

The goal of this work is to develop a tool that provides a set of candidate compounds to experiments that screen chemical species for fitness in a particular application. This tool ideally would consider every known chemical compound and assign it a priority for inclusion in an experiment based on the likelihood of it demonstrating the desired chemical or physical properties and the degree to which it meets identified engineering constraints. To use this hypothetical tool, researchers would supply information about what chemical or physical properties they are looking for, along with a set of engineering constraints that the ideal material for their application would meet. Our approach makes aspects of this ideal tool a reality.

Our approach applies a multi-step screening process to first expand then down select a chemical design space of candidate species. This strategy incorporates as much available information about relevant criteria as is reasonably possible to select the most promising candidates. Figure A2.1 summarizes this strategy. The first step in this approach is to identify a basis set of chemical species that display the desired chemical and physical properties. This basis set can be compiled from any source. In our work so far, this set usually comes from a literature search for compounds known to exhibit desired properties or chemical intuition and initial results from intuition-guided manual experimentation. In most projects we have applied this tool to, establishing a basis set with tens to hundreds of species is possible and provides enough information to return good results from the next step. The initial set of known-working species is

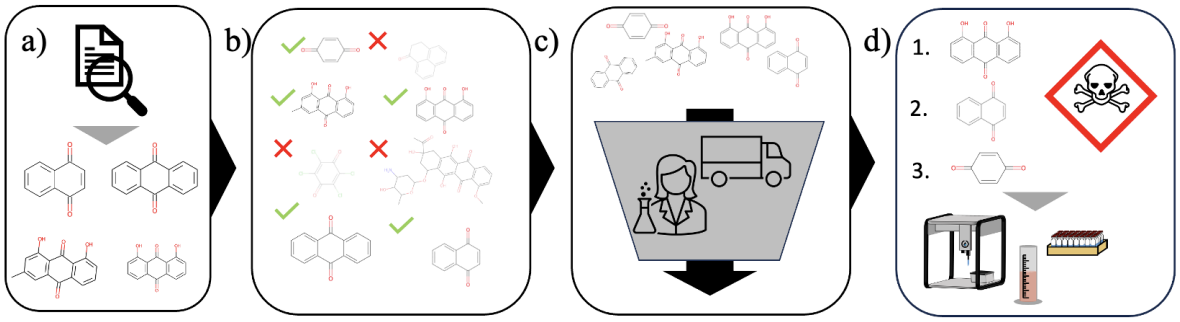


Figure A2.1: The main steps in our screening framework are to a) collect a basis set of known-working compounds, b) expand the initial basis set by similarity search then apply structural filtering criteria, c) remove candidates that do not meet practical engineering criteria like synthesizeability, and d) rank the remaining candidates and send to experimentation.

used as the basis for a similarity search of chemical space. This step is guided by the intuition that chemical structure guides chemical functionality, so 'similar' species should have similar functionality in terms of the target properties. The similarity search approach is simple and easy to conduct. It does not require sufficient data to train a predictive model for molecular screening, which is a major advantage over machine learning based screening approaches. It also does not need a complete understanding of the chemical rules governing the desired properties, and advantage over a manually defined structural rule-based screening approach. This similarity search is performed against a large chemical database. In our work, we have used PubChem for this purpose due to its free and open access, large size, and provisions for programmatic access[6,7]. The similarity search generally returns a very large set of possible candidate species, potentially with tens of thousands of 'hits' depending on the basis set and desired specificity in the search. A disadvantage of the similarity search is the non-specificity of the results: while the search is likely to return molecules with similar properties, it also likely returns species that do not display the properties of interest. Further, it provides no way to rank the returned hits based on the properties of interest. To increase the specificity of the candidates returned from the similarity search, we

next apply structural rules to screen out non-match candidates. These rules are based on the researcher's understanding of the desired chemistry or physics and can vary in complexity from simple molecular weight cutoffs to complex substructure bond conjugation criteria. These rules are implemented in the RDKit software[8] and can provide an important first pass at reducing the size of the candidate set returned from the similarity search. Next, the candidate set is screened on 'must pass' constraints. These constraints have been identified as must haves, such that any candidate that does not meet them is removed from consideration. A common example in our work is commercial availability. PubChem contains many compounds that are not available for purchase from any supplier. As we are a chemical engineering lab, not a synthetic chemistry lab, we are unable to test compounds we cannot buy. This criterion may also serve as a proxy for cost. If it is impossible to buy this compound at a laboratory scale, it will likely be prohibitively expensive or require a significant process development investment before it is usable on the commercial scale required for the types of applications this screening process is applied to. Successful application of engineering criteria at this phase is dependent on the availability of sufficient information on relevant properties. Lack of this information is a major challenge and makes effective screening on some properties impossible. For example, we are not aware of a comprehensive source for chemical cost information. While this would be an obvious screening criterion, the lack of information makes it impossible to apply. After candidates that do not meet the 'hard' engineering criteria have been removed, the remaining 'soft' criteria are used to define a ranking metric and rank the remaining candidates to prioritize them for experimental testing. These 'soft' criteria are properties that have been identified as important but exist on a continuous scale and don't have a hard cutoff of inclusion in the set of candidates. Safety is a common example of soft criteria. All

else being equal, ‘safer’ things are better than ‘unsafe’ things, so it makes sense to prioritize safer chemicals. It is possible to assign a safety score to compounds that corresponds to the overall hazards associated with them, but it is difficult to set a general hard criterion for what makes a chemical ‘too dangerous’ for consideration. A common approach we use is to gather safety information from PubChem in the form of Globally Harmonized System (GHS) hazard codes. Hazard codes can be queried from PubChem for compounds that have this information available. To form a safety figure of merit, hazard codes are assigned a numerical score based on their severity as determined by previous work[9] and the sum of hazard scores for each candidate is used as an overall score for safety. To create an overall ranking metric, we apply a modified digital logic approach to weight the relative importance of the criteria considered in the ranking[10]. Here, the researcher designing the screening campaign considers every pair-wise combination of possible ranking criteria and assigns a higher priority criterion in each comparison. The number of comparisons a criterion ‘wins’ is summed and used to weight the importance of that criteria. The values or scores of individual candidate compounds are normalized for each criterion. A final score is calculated by summing the weighted normalized scores across categories, which is used to rank the final candidates and prioritize the compounds suggested for experimental screening.

Effective visualizations of the resulting chemical design space are useful for sanity checking the results of screening criteria implementations and communicating results with researchers. An effective strategy for visualizing the final results of a screening campaign is to plot final molecules on a 2D grid representing dimensionality reduced structural descriptors. Molecular descriptors are computed feature vectors that describe structural features of molecules. Reducing these features vectors to 2 dimensions with the dimensionality reducing/clustering algorithm t-

Stochastic Nearest Neighbors (t-SNE) groups structurally similar molecules close together and makes it possible to visualize the chemical space on 2 axes[11]. The final ranking criteria of molecules can be encoded in color, and 2D molecular drawings can be made accessible via tooltips. These interactive visualizations can be trivially shared by hosting on the web service Streamlit[12]. Figure A2.2 shows an example of such a visualization with structural motifs identified.

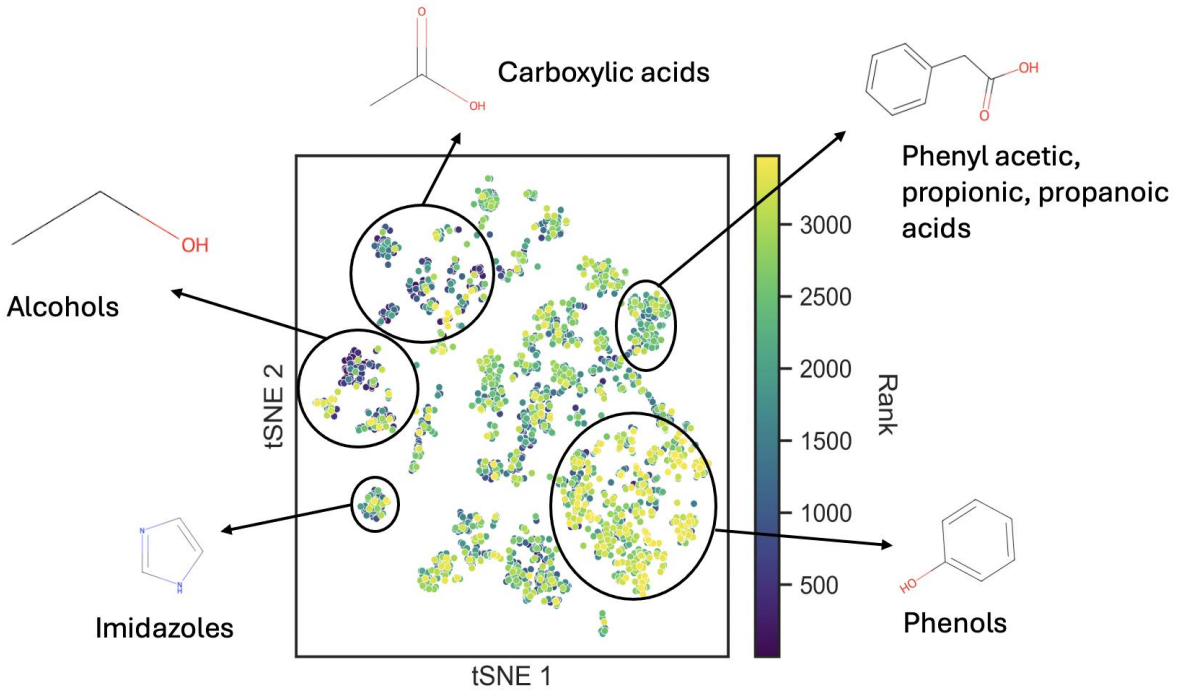


Figure A2.2: 2-dimensional t-SNE plot representing structural diversity in final design space of chemicals screened using our screening framework. Clusters of molecules with a similar structure are highlighted. Figure adapted from Rodriquez et. al.[5]

A2.3 Applications

A2.3.1 Selection of Near Infrared Upconverting Donor Molecules

I collaborated with members of Cody Schlenker's research group to develop a version of our screening method to assist in chemical design space for a project to identify upconverting molecules. Most photovoltaic materials absorb in the visible light region. However, a significant amount of solar flux is in the near infrared (NIR) region. Developing photovoltaic systems that absorb at both visible and NIR wavelengths would increase the efficiency of solar energy infrastructure[13]. One approach to solving this problem is to combine traditional photovoltaic materials with NIR upconverting materials. These materials absorb NIR light and re-emit it at visible wavelengths, enabling it to be absorbed by traditional photovoltaic materials. Existing upconverting materials either rely on expensive metals, are unstable, or suffer from parasitic absorption. The Schlenker group is developing NIR sensitizers for upconverting applications based on squarine chemistries. In this system, absorbing donor molecules are bound to a squarine core to form a sensitizer or NIR absorbing molecule. Examples of these molecules are shown in Figure A2.3. Selection of donor molecules for this system presents a large molecular selection problem that we are working towards addressing with our screening approach. For this system, an initial basis set of known working donor molecules was not available, which required a modification from the baseline screening method. Structural screening rules were used instead. Important

engineering criteria were identified in collaboration with researchers from this project. Synthetic accessibility was identified as an important factor in deciding which candidate donor molecules to investigate. The researchers were willing to invest a moderate amount of effort to synthesize promising candidates in-house, but wanted to set a cutoff on the amount of effort required. General safety was also identified as important, as the goal of this project is to develop materials that will eventually be widely deployed in solar panels. With these criteria in mind, a screening strategy was implemented.

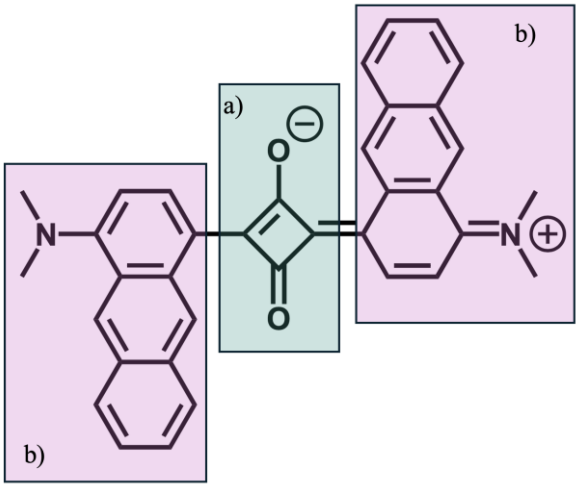
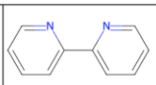
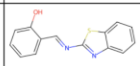
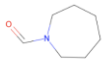


Figure A2.3: Dimethyl-aminoanthracene squaraine, an example sensitizer. A) Squarine core, b) Donor group. Selection of appropriate donor groups is the goal of this cheminformatics work.

Without a basis set of known-working examples to inform a similarity search, a robust set of structural requirements was required. A collaborator in the Schlenker group with a deep understanding of the impact of structure on absorbing behavior and described a thorough set of structural requirements developed structural selection rules. These rules were implemented in RDkit and used to screen the entirety of the PubChem database. The specific rules are described in table A2.1. These rules were used to screen the entire 115 million compound PubChem database.

Table A2.1: Structural screening criteria for selection of donor molecules.

Molecular weight < 300 g/mol	 ✓
Allowed atoms: {C,N,H,S,P,B,O}	 ✓
Group exclusion rules (ex. no carbonyl groups)	 ✗
No rings larger than 6 atoms	 ✗
Must have at least 2 aromatic rings with 5 or 6 members	
Molecules with more than 1 ring system must have ring systems connected with conjugated bonds or nitrogen or boron bridging	

No NO_2 (O=N=O) groups	 ✗
No S=O groups	 ✗
No Azides (N=N=N)	 ✗
No triple bonds to N	 ✗
No S=N groups	 ✗
No C=S groups	 ✗

Given the large size of this database, reasoned selection of the order in which rules were applied was required to maintain computational efficiency. Molecular weight was used as a first pass screening criteria as it is computationally trivial to evaluate. Atomic composition and bond membership rules were evaluated next. Complicated rules like the ring system conjugation requirement were evaluated last. This ordering combined with some embarrassing parallelization made it possible to process all 115 million compounds in a few hours on a desktop computer. Next, remaining candidates were screened on synthetic accessibility. Synthetic accessibility tools use rule based method, machine learned models, or both to assign a synthetic accessibility score to molecules that represents how difficult synthesizing the molecule is likely to be. Here, scores from four common tools were considered: RAscore[14], SAscore[15], SCScore[16], and SYBAscore[17]. Each tool uses a different approach to assign scores on a different scale. Reasonable threshold criteria, shown in Table A2.2, were selected for each score. Candidates that met all 4 thresholds were kept, those that did not were

Table A2.2: Synthetic accessibility scoring cutoff criteria

Score	Cutoff criteria
RAscore	> 0.5
SAscore	< 6
SCScore	< 3

discarded. Finally, remaining candidate molecules were ranked on safety. Figure 4.4 shows the set of candidates at each step of the screening process. Each plot shows an arbitrary 2D representation of the molecular structure. These plots show that as the screening criteria are applied, the diversity of structures remains relatively constant even as the total number of candidates drastically drops. This project is still in progress and has not generated experimental results yet. Collaborators are pursuing additional computational screening of molecular donor character to further down-select candidates for experimentation. This demonstrates the flexibility of combing this screening approach with application appropriate screening methods to select candidate molecules. The donor strength assessment pursued by collaborators is expected to be computationally intensive, so screening all 115 million compounds in PubChem would not have been feasible. Starting with our screening approach provided a structurally diverse set of candidates that have been screened for practical criteria as a starting point for downstream higher cost methods. This is the role we envision our approach playing in similar projects.

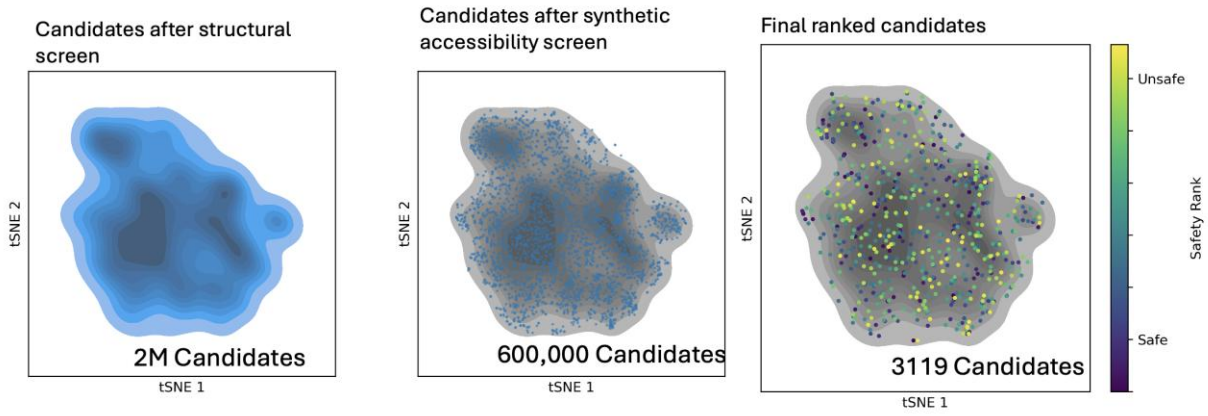


Figure A2.4: tSNE plots of donor selection space at various points throughout the screening process. Gray background in right two panels represents the space after structural screening shown in left panel, showing subsequent selection steps do not significantly reduce structural diversity of the final candidate set.

A2.3.2 Identification of Monomers for Mechano-Redox Polymerization

In another collaborative application of this screening tool, potential monomers for a novel polymerization process were identified. Collaborators in Matt Golder's research group are working with a mechano-redox polymerization process that enables new block copolymers to be synthesized[18,19]. Block copolymers are composed of two or more monomers assembled into homopolymer 'blocks' that are connected into a single polymer chain. Traditional solution synthesis methods require all monomers involved in a synthesis to be soluble in a common solvent, which limits the possible monomer combinations. Mechano-redox polymerization uses a piezo-electric catalyst and mechanical motion to drive a synthesis reaction[20]. This process does not require monomer dissolution, which relaxes constraints on monomer solubility and makes the use of new monomers and monomer combinations possible. This presents an underexplored design space of possible monomers. A screening campaign was designed to help explore possible monomers. In this screening campaign, four different monomer chemistries were explored. For use in this mechano-redox synthesis process, monomers must have one of the following exposed end groups to enable polymerization: acrylate, acrylamide, methacrylate, or methacrylamide. Known monomers with these end groups are available from commercial vendors. To form a basis set for a similarity search, available monomers of each category were web scraped from the Tokyo Chemical Industry U.S. website. A similarity search was run using the web-scraped monomers as a basis set. A structural screening rule was applied to the resulting set of similar molecules to enforce the appropriate terminal group requirement. Commercial availability of monomers was an important constraint for this project, so candidates that were not purchasable according to vendor information available on PubChem were removed. Finally, candidate monomers were ranked on

hydrophilicity. A goal of this project is to synthesize block copolymers with large differences in hydrophilicity between blocks. Selecting monomers with disparate hydrophilicities is required to do this. A hydrophilicity score was calculated based on calculated octanol/water partition coefficient, total polarizable surface area, and the difference between the maximum and minimum calculated partial charge. This score was used to rank the final candidates. The final candidates were displayed in an interactive t-SNE plot, such as the one shown in Figure A2.5 for the acrylate-based monomers. This project is ongoing and further selection of candidate monomers has been delayed while a high throughput synthesis workflow is developed.

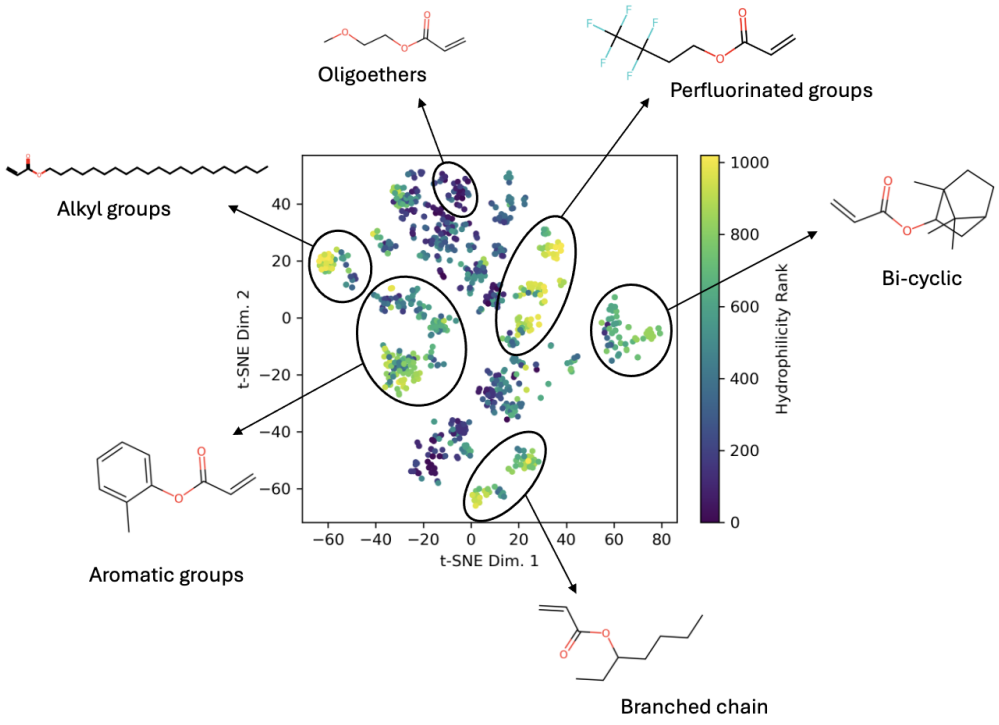


Figure A2.5: Potential acrylate monomer candidates resulting from the described screening criteria.

A2.4 Conclusions and Future Work

Future work on this project should focus on formalizing the software implementation of the screening workflow, identifying and developing sources for frequently used screening data, and validating the assumptions made in this screening strategy.

There are a few key assumptions underpinning the validity and usefulness of this strategy that should be formally validated. First, it is assumed that for any given material application, there is a latent space of existing, unidentified chemicals that will fill the role well. This would be a challenging assumption to disprove, but I am not aware of work that backs it. A major trend in the molecular search space is generative molecular design[21]. Proponents of this approach would argue that molecular design can provide a molecule tailor made for a specific task. I argue that it is worth at least searching known chemical space first, as something that may work well enough could turn up. It would be a waste to wind up in a situation where a year of molecular generative design saved a few days crunching through PubChem[22]. A more critical assumption is that chemical similarity searches are effective tools for identifying functionally similar molecules with enough specificity to be efficiently plan experimental screening. To validate this, I would find an application with a large set of known-working compounds, select a small subset of this collection to form the basis set for a similarity search, and compare the set returned from the similarity search to the known set of working molecules. This would provide some real-world insight as to how effective this campaign is. Finally, finding data sources to support the engineering constraints used in this approach is difficult. Safety constraints have used data available on PubChem, which is unavailable for the vast majority of compounds. Cost information would be extremely useful but is not available in a programmatically accessible format. Machine learning methods promise rapid

access to predictions of some engineering quantities, but issues with generalizability and reliability make their blanket application difficult.

A2.5 References

(1) Moosavi, S. M.; Jablonka, K. M.; Smit, B. The Role of Machine Learning in the Understanding and Design of Materials. *J. Am. Chem. Soc.* **2020**, *142* (48), 20273–20287. https://doi.org/10.1021/jacs.0c09105.

(2) Maier, W. F.; Stöwe, K.; Sieg, S. Combinatorial and High-Throughput Materials Science. *Angew. Chem. Int. Ed.* **2007**, *46* (32), 6016–6067. https://doi.org/10.1002/anie.200603675.

(3) Leardi, R. Experimental Design in Chemistry: A Tutorial. *Anal. Chim. Acta* **2009**, *652* (1), 161–172. https://doi.org/10.1016/j.aca.2009.06.015.

(4) Kumar, A.; Kini, S. G.; Rathi, E. A Recent Appraisal of Artificial Intelligence and In Silico ADMET Prediction in the Early Stages of Drug Discovery. *Mini-Rev. Med. Chem.* **2021**, *21* (18), 2788–2800. https://doi.org/10.2174/1389557521666210401091147.

(5) Rodriguez, J.; Politi, M.; Adler, S.; Beck, D.; Pozzo, L. High-Throughput and Data Driven Strategies for the Design of Deep-Eutectic Solvent Electrolytes. *Mol. Syst. Des. Eng.* **2022**, *7* (8), 933–949. https://doi.org/10.1039/D2ME00050D.

(6) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951.

(7) Kim, S.; Thiessen, P. A.; Cheng, T.; Yu, B.; Bolton, E. E. An Update on PUG-REST: RESTful Interface for Programmatic Access to PubChem. *Nucleic Acids Res.* **2018**, *46* (W1), W563–W570. https://doi.org/10.1093/nar/gky294.

(8) *Python API Reference — The RDKit 2023.03.1 documentation*. https://www.rdkit.org/docs/api-docs.html (accessed 2023-09-21).

(9) Verslycke, T.; Reid, K.; Bowers, T.; Thakali, S.; Lewis, A.; Sanders, J.; Tuck, D. The Chemistry Scoring Index (CSI): A Hazard-Based Scoring and Ranking Tool for Chemicals and Products Used in the Oil and Gas Industry. *Sustainability* **2014**, *6* (7), 3993–4009. https://doi.org/10.3390/su6073993.

(10) Dehghan-Manshadi, B.; Mahmudi, H.; Abedian, A.; Mahmudi, R. A Novel Method for Materials Selection in Mechanical Design: Combination of Non-Linear Normalization and a Modified Digital Logic Method. *Mater. Des.* **2007**, *28* (1), 8–15. https://doi.org/10.1016/j.matdes.2005.06.023.

(11) Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9* (11).

(12) *Streamlit • A faster way to build and share data apps*. https://streamlit.io/ (accessed 2024-02-05).

(13) Meng, D.; Zheng, R.; Zhao, Y.; Zhang, E.; Dou, L.; Yang, Y. Near-Infrared Materials: The Turning Point of Organic Photovoltaics. *Adv. Mater.* **2022**, *34* (10), 2107330. https://doi.org/10.1002/adma.202107330.

(14) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic Accessibility Score (RAscore) – Rapid Machine Learned Synthesizability Classification from AI Driven Retrosynthetic Planning. *Chem. Sci.* **2021**, *12* (9), 3339–3349. https://doi.org/10.1039/D0SC05401A.

(15) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminformatics* **2009**, *1* (1), 8. https://doi.org/10.1186/1758-2946-1-8.
(16) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58* (2), 252–261. https://doi.org/10.1021/acs.jcim.7b00622.
(17) Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian Estimation of Synthetic Accessibility of Organic Compounds. *J. Cheminformatics* **2020**, *12* (1), 35. https://doi.org/10.1186/s13321-020-00439-2.
(18) Zeitler, S. M.; Chakma, P.; Golder, M. R. Diaryliodonium Salts Facilitate Metal-Free Mechanoredox Free Radical Polymerizations. *Chem. Sci.* **2022**, *13* (14), 4131–4138.
(19) Chakma, P.; Zeitler, S. M.; Baum, F.; Yu, J.; Shindy, W.; Pozzo, L. D.; Golder, M. R. Mechanoredox Catalysis Enables a Sustainable and Versatile Reversible Addition-Fragmentation Chain Transfer Polymerization Process. *Angew. Chem. Int. Ed.* **2023**, *62* (2), e202215733. https://doi.org/10.1002/anie.202215733.
(20) Nothling, M. D.; Daniels, J. E.; Vo, Y.; Johan, I.; Stenzel, M. H. Mechanically Activated Solid-State Radical Polymerization and Cross-Linking via Piezocatalysis. *Angew. Chem.* **2023**, *135* (20), e202218955. https://doi.org/10.1002/ange.202218955.
(21) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360–365. https://doi.org/10.1126/science.aat2663.
(22) *Frank Westheimer - Wikiquote*. https://en.wikiquote.org/wiki/Frank_Westheimer (accessed 2024-02-05).

Bibliography

Abdel-Latif, K.; Epps, R. W.; Bateni, F.; Han, S.; Reyes, K. G.; Abolhasani, M. Self-Driven Multistep Quantum Dot Synthesis Enabled by Autonomous Robotic Experimentation in Flow. Adv. Intell. Syst. 2021, 3 (2), 2000245. https://doi.org/10.1002/aisy.202000245.

Abed, J.; Bai, Y.; Persaud, D.; Kim, J.; Witt, J.; Hattrick-Simpers, J.; H. Sargent, E. AMPERE: Automated Modular Platform for Expedited and Reproducible Electrochemical Testing. Digit. Discov. 2024, 3 (11), 2265–2274. https://doi.org/10.1039/D4DD00203B.

Abolhasani, M.; Kumacheva, E. The Rise of Self-Driving Labs in Chemical and Materials Sciences. Nat. Synth. 2023, 1–10. https://doi.org/10.1038/s44160-022-00231-0.

Acceleration Consortium. https://acceleration.utoronto.ca/ (accessed 2024-01-29).

Akepati, S. V. R.; Gupta, N.; Jayaraman, A. Computational Reverse Engineering Analysis of the Scattering Experiment Method for Interpretation of 2D Small-Angle Scattering Profiles (CREASE-2D). JACS Au 2024, 4 (4), 1570–1582. https://doi.org/10.1021/jacsau.4c00068.

Allan, D.; Caswell, T.; Campbell, S.; Rakitin, M. Bluesky's Ahead: A Multi-Facility Collaboration for an a La Carte Software Project for Data Acquisition and Management. Synchrotron Radiat. News 2019, 32 (3), 19–22. https://doi.org/10.1080/08940886.2019.1608121.

Ament, S.; Daulton, S.; Eriksson, D.; Balandat, M.; Bakshy, E. Unexpected Improvements to Expected Improvement for Bayesian Optimization. arXiv January 7, 2025. https://doi.org/10.48550/arXiv.2310.20708.

An Update on Research Security: Streamlining Disclosure Standards to Enhance Clarity, Transparency, and Equity | OSTP. The White House. https://www.whitehouse.gov/ostp/news-updates/2022/08/31/an-update-on-research-securitystreamlining-disclosure-standards-to-enhance-clarity-transparency-and-equity/ (accessed 2023-02-09).

Andersen, C. W.; Armiento, R.; Blokhin, E.; Conduit, G. J.; Dwaraknath, S.; Evans, M. L.; Fekete, Á.; Gopakumar, A.; Gražulis, S.; Merkys, A.; Mohamed, F.; Oses, C.; Pizzi, G.; Rignanese, G.-M.; Scheidgen, M.; Talirz, L.; Toher, C.; Winston, D.; Aversa, R.; Choudhary, K.; Colinet, P.; Curtarolo, S.; Di Stefano, D.; Draxl, C.; Er, S.; Esters, M.; Fornari, M.; Giantomassi, M.; Govoni, M.; Hautier, G.; Hegde, V.; Horton, M. K.; Huck, P.; Huhs, G.; Hummelshøj, J.; Kariryaa, A.; Kozinsky, B.; Kumbhar, S.; Liu, M.; Marzari, N.; Morris, A. J.; Mostofi, A. A.; Persson, K. A.; Petretto, G.; Purcell, T.; Ricci, F.; Rose, F.; Scheffler, M.; Speckhard, D.; Uhrin, M.; Vaitkus, A.; Villars, P.; Waroquiers, D.; Wolverton, C.; Wu, M.; Yang, X. OPTIMADE, an API for Exchanging Materials Data. Sci. Data 2021, 8 (1), 217. https://doi.org/10.1038/s41597-021-00974-z.

Archibald, R. K.; Doucet, M.; Johnston, T.; Young, S. R.; Yang, E.; Heller, W. T. Classifying and Analyzing Small-Angle Scattering Data Using Weighted k Nearest Neighbors Machine

Learning Techniques. J. Appl. Crystallogr. 2020, 53 (2), 326–334. https://doi.org/10.1107/S1600576720000552.

Argento, N. Institutional ELN/LIMS Deployment. EMBO Rep. 2020, 21 (3), e49862. https://doi.org/10.15252/embr.201949862.

Ashraf, C.; Joshi, N.; Beck, D. A. C.; Pfaendtner, J. Data Science in Chemical Engineering: Applications to Molecular Science. Annu. Rev. Chem. Biomol. Eng. 2021, 12 (1), 15–37. https://doi.org/10.1146/annurev-chembioeng-101220-102232.

Automation-Hardware/Cartridge Sample Holder for SAS Experiments/SAXS-USAXS Liquids 48 well plate holder at master · pozzo-research-group/Automation-Hardware. https://github.com/pozzo-research-group/Automation-Hardware/tree/master/Cartridge%20Sample%20Holder%20for%20SAS%20Experiments/SAXS-USAXS%20Liquids%2048%20well%20plate%20holder (accessed 2025-05-05).

Automation-Hardware/Cartridge Sample Holder for SAS Experiments/SAXS-USAXS_AntonPaar_FlowCellHolder/USAXS_flow cell holder v5.stl at master · pozzo-research-group/Automation-Hardware. GitHub. https://github.com/pozzo-research-group/Automation-Hardware/blob/master/Cartridge%20Sample%20Holder%20for%20SAS%20Experiments/SAXS-USAXS_AntonPaar_FlowCellHolder/USAXS_flow%20cell%20holder%20v5.stl (accessed 2025-05-05).

Automation-Hardware/Vial Holders/2mLdram_44_wellplate_withRetainer at master · pozzo-research-group/Automation-Hardware. https://github.com/pozzo-research-group/Automation-Hardware/tree/master/Vial%20Holders/2mLdram_44_wellplate_withRetainer (accessed 2025-06-05).

Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. Science 2021, 373 (6557), 871–876. https://doi.org/10.1126/science.abj8754.

Bai, J.; Cao, L.; Mosbach, S.; Akroyd, J.; Lapkin, A. A.; Kraft, M. From Platform to Knowledge Graph: Evolution of Laboratory Automation. JACS Au 2022, 2 (2), 292–309. https://doi.org/10.1021/jacsau.1c00438.

Baird, S. G. Sparks-Baird/Self-Driving-Lab-Demo: V0.8.4, 2023. https://doi.org/10.5281/ZENODO.7855492.

Balandat, M.; Karrer, B.; Jiang, D.; Daulton, S.; Letham, B.; Wilson, A. G.; Bakshy, E. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In Advances in Neural Information Processing Systems; Curran Associates, Inc., 2020; Vol. 33, pp 21524–21538.

Bayerlein, B.; Hanke, T.; Muth, T.; Riedel, J.; Schilling, M.; Schweizer, C.; Skrotzki, B.; Todor, A.; Moreno Torres, B.; Unger, J. F.; Völker, C.; Olbricht, J. A Perspective on Digital Knowledge Representation in Materials Science and Engineering. Adv. Eng. Mater. 2022, 24 (6), 2101176. https://doi.org/10.1002/adem.202101176.

Beaucage, P. A.; Martin, T. B. The Autonomous Formulation Laboratory: An Open Liquid Handling Platform for Formulation Discovery Using X-ray and Neutron Scattering. Chem. Mater. 2023, 35 (3), 846–852. https://doi.org/10.1021/acs.chemmater.2c03118.

Beck, D. A.; Carothers, J. M.; Subramanian, V. R.; Pfaendtner, J. Data Science: Accelerating Innovation and Discovery in Chemical Engineering; 2016; Vol. 62, pp 1402–1416.

Bennett, J. A.; Orouji, N.; Khan, M.; Sadeghi, S.; Rodgers, J.; Abolhasani, M. Autonomous Reaction Pareto-Front Mapping with a Self-Driving Catalysis Laboratory. Nat. Chem. Eng. 2024, 1 (3), 240–250. https://doi.org/10.1038/s44286-024-00033-5.

Bik, E. M.; Casadevall, A.; Fang, F. C. The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications. mBio 2016, 7 (3), e00809-16. https://doi.org/10.1128/mBio.00809-16.

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

Borges, J. L. The Library of Babel; 1941.

Bosman, L.; Garcia-Bravo, J. Lessons Learned: Research Benefits and Beyond Associated with Participating in the NSF I-CorpsTM Customer Discovery Program. Technol. Innov. 2021, 22 (1), 41–54. https://doi.org/10.21300/21.4.2021.5.

botorch.acquisition — BoTorch documentation. https://botorch.readthedocs.io/en/latest/acquisition.html#ament2023logei (accessed 2025-05-15).

Bouchoucha, M.; Côté, M.-F.; C.-Gaudreault, R.; Fortin, M.-A.; Kleitz, F. Size-Controlled Functionalized Mesoporous Silica Nanoparticles for Tunable Drug Release and Enhanced Anti-Tumoral Activity. Chem. Mater. 2016, 28 (12), 4243–4258. https://doi.org/10.1021/acs.chemmater.6b00877.

Brendel, A.; Dorfmüller, F.; Liebscher, A.; Kraus, P.; Kress, K.; Oehme, H.; Arnold, M.; Koschitzki, R. Laboratory and Analytical Device Standard (LADS): A Communication Standard Based on OPC UA for Networked Laboratories. In Smart Biolabs of the Future; Beutel, S., Lenk, F., Eds.; Advances in Biochemical Engineering/Biotechnology; Springer International Publishing: Cham, 2022; pp 175–194. https://doi.org/10.1007/10_2022_209.

brendenpelkie/usaxs_processing: Run processing for USAXS data for silica nanoparticle optimization campaign. https://github.com/brendenpelkie/usaxs_processing (accessed 2025-05-10).

Bressler, I.; Pauw, B. R.; Thünemann, A. F. McSAS: Software for the Retrieval of Model Parameter Distributions from Scattering Patterns. J. Appl. Crystallogr. 2015, 48 (3), 962–969. https://doi.org/10.1107/S1600576715007347.

Brinson, L. C.; Bartolo, L. M.; Blaiszik, B.; Elbert, D.; Foster, I.; Strachan, A.; Voorhees, P. W. FAIR Data Will Fuel a Revolution in Materials Research. arXiv April 6, 2022. https://doi.org/10.48550/arXiv.2204.02881.

Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A Mobile Robotic Chemist. Nature 2020, 583 (7815), 237–241. https://doi.org/10.1038/s41586-020-2442-2.

Burley, S. K.; Berman, H. M.; Christie, C.; Duarte, J. M.; Feng, Z.; Westbrook, J.; Young, J.; Zardecki, C. RCSB Protein Data Bank: Sustaining a Living Digital Data Resource That Enables Breakthroughs in Scientific Research and Biomedical Education. Protein Sci. Publ. Protein Soc. 2018, 27 (1), 316–330. https://doi.org/10.1002/pro.3331.

Candela-Noguera, V.; Alfonso, M.; Amorós, P.; Aznar, E.; Marcos, M. D.; Martínez-Máñez, R. In-Depth Study of Factors Affecting the Formation of MCM-41-Type Mesoporous Silica Nanoparticles. Microporous Mesoporous Mater. 2024, 363, 112840. https://doi.org/10.1016/j.micromeso.2023.112840.

canSAS.org. https://www.cansas.org/ (accessed 2022-11-22).

Chakma, P.; Zeitler, S. M.; Baum, F.; Yu, J.; Shindy, W.; Pozzo, L. D.; Golder, M. R. Mechanoredox Catalysis Enables a Sustainable and Versatile Reversible Addition-Fragmentation Chain Transfer Polymerization Process. Angew. Chem. Int. Ed. 2023, 62 (2), e202215733. https://doi.org/10.1002/anie.202215733.

Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; Palizhati, A.; Sriram, A.; Wood, B.; Yoon, J.; Parikh, D.; Zitnick, C. L.; Ulissi, Z. Open Catalyst 2020 (OC20) Dataset and Community Challenges. ACS Catal. 2021, 2020, 6059–6072. https://doi.org/10.1021/acscatal.0c04525.

Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; KDD '16; Association for Computing Machinery: New York, NY, USA, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.

Chowdhury, A. J.; Yang, W.; Walker, E.; Mamun, O.; Heyden, A.; Terejanu, G. A. Prediction of Adsorption Energies for Chemical Species on Metal Catalyst Surfaces Using Machine Learning. J. Phys. Chem. C 2018, 122 (49), 28142–28150. https://doi.org/10.1021/acs.jpcc.8b09284.

Christensen, M.; E. Yunker, L. P.; Shiri, P.; Zepel, T.; L. Prieto, P.; Grunert, S.; Bork, F.; E. Hein, J. Automation Isn't Automatic. Chem. Sci. 2021, 12 (47), 15473–15490. https://doi.org/10.1039/D1SC04588A.

Cloud-based platform for biotech R&D | Benchling. https://www.benchling.com/ (accessed 2022-11-21).

Colbert, D. T.; Miller, W. H. A Novel Discrete Variable Representation for Quantum Mechanical Reactive Scattering via the Sâ□• matrix Kohn Method. J. Chem. Phys. 1992, 96 (3), 1982–1991. https://doi.org/10.1063/1.462100.

Cole, J. M. The Chemistry of Errors. Nat. Chem. 2022, 14 (9), 973–975. https://doi.org/10.1038/s41557-022-01028-6.

Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. J. Chem. Inf. Model. 2018, 58 (2), 252–261. https://doi.org/10.1021/acs.jcim.7b00622.

Collins, C. R.; Gordon, G. J.; Von Lilienfeld, O. A.; Yaron, D. J. Constant Size Descriptors for Accurate Machine Learning Models of Molecular Properties. J. Chem. Phys. 2018, 148 (24). https://doi.org/10.1063/1.5020441.

Data Management. Citrine Informatics. https://citrine.io/product/what-is-the-citrine-platform/data-management/ (accessed 2022-11-21).

Dehghan-Manshadi, B.; Mahmudi, H.; Abedian, A.; Mahmudi, R. A Novel Method for Materials Selection in Mechanical Design: Combination of Non-Linear Normalization and a Modified Digital Logic Method. Mater. Des. 2007, 28 (1), 8–15. https://doi.org/10.1016/j.matdes.2005.06.023.

Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine Learning Methods for Small Data Challenges in Molecular Science. Chem. Rev. 2023, 123 (13), 8736–8780. https://doi.org/10.1021/acs.chemrev.3c00189.

Draxl, C.; Scheffler, M. The NOMAD Laboratory: From Data Sharing to Artificial Intelligence. J. Phys. Mater. 2019, 2 (3), 036001. https://doi.org/10.1088/2515-7639/ab13bb.

Dryad | Our mission. https://datadryad.org/stash/our_mission (accessed 2023-02-02).

Duet3D · Duet3D. Duet3D. https://www.duet3d.com/ (accessed 2024-02-05).

Duke, R.; Bhat, V.; Risko, C. Data Storage Architectures to Accelerate Chemical Discovery: Data Accessibility for Individual Laboratories and the Community. Chem. Sci. 2022. https://doi.org/10.1039/D2SC05142G.

Electronic Lab Notebook (ELN). Labfolder. https://labfolder.com/ (accessed 2022-12-23).

Emerald Cloud Lab: Remote Controlled Life Sciences Lab. https://www.emeraldcloudlab.com/ (accessed 2025-05-14).

Empty Rhetoric over Data Sharing Slows Science. Nature 2017, 546 (7658), 327–327. https://doi.org/10.1038/546327a.

Entity Registration | Dotmatics. Entity Registration | Dotmatics. https://www.dotmatics.com/capabilities/entity-registration (accessed 2022-11-21).

Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. J. Cheminformatics 2009, 1 (1), 8. https://doi.org/10.1186/1758-2946-1-8.

ESAMP: Event-Sourced Architecture for Materials Provenance Management and Application to Accelerated Materials Discovery | Materials Chemistry | ChemRxiv | Cambridge Open Engage. https://chemrxiv.org/engage/chemrxiv/article-details/60c73cbf842e650956db1678 (accessed 2022-10-04).

Excipients. https://grace.com/industries/pharmaceutical-solutions/purification-formulation-delivery/excipients/, https://grace.com/industries/pharmaceutical-solutions/purification-formulation-delivery/excipients/ (accessed 2025-05-06).

Export Administration Regulations; Vol. 15.734.8.

Fei, Y.; Rendy, B.; Kumar, R.; Dartsi, O.; Sahasrabuddhe, H. P.; McDermott, M. J.; Wang, Z.; Szymanski, N. J.; Walters, L. N.; Milsted, D.; Zeng, Y.; Jain, A.; Ceder, G. AlabOS: A Python-Based Reconfigurable Workflow Management Framework for Autonomous Laboratories. Digit. Discov. 2024, 3 (11), 2275–2288. https://doi.org/10.1039/D4DD00129J.

Flynn, E. J.; Keane, D. A.; Tabari, P. M.; Morris, M. A. Pervaporation Performance Enhancement through the Incorporation of Mesoporous Silica Spheres into PVA Membranes. Sep. Purif. Technol. 2013, 118, 73–80. https://doi.org/10.1016/j.seppur.2013.06.034.

Frank Westheimer - Wikiquote. https://en.wikiquote.org/wiki/Frank_Westheimer (accessed 2024-02-05).

Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-Based Machine Learning Representations for Predicting the Enantioselectivity of Organocatalysts. Chem. Sci. 2021, 12 (20), 6879–6889. https://doi.org/10.1039/d1sc00482d.

Gallis, K. W.; Araujo, J. T.; Duff, K. J.; Moore, J. G.; Landry, C. C. The Use of Mesoporous Silica in Liquid Chromatography. Adv. Mater. 1999, 11 (17), 1452–1455. https://doi.org/10.1002/(SICI)1521-4095(199912)11:17<1452::AID-ADMA1452>3.0.CO;2-R.

Garnett, R. Bayesian Optimization. https://www.cse.wustl.edu/~garnett/cse515t/spring_2019/files/lecture_notes/12.pdf (accessed 2024-02-05).

GEMD Documentation. https://citrineinformatics.github.io/gemd-docs/ (accessed 2022-11-18).

Ghimire, P. P.; Jaroniec, M. Renaissance of Stöber Method for Synthesis of Colloidal Particles: New Developments and Opportunities. J. Colloid Interface Sci. 2021, 584, 838–865. https://doi.org/10.1016/j.jcis.2020.10.014.

Giacomelli, C.; Borsali, R. Disordered Phase and Self-Organization of Block Copolymer Systems. In Soft Matter Characterization; Borsali, R., Pecora, R., Eds.; Springer Netherlands: Dordrecht, 2008; pp 133–189. https://doi.org/10.1007/978-1-4020-4465-6_3.

Glatter, O. Chapter 3 - The Inverse Scattering Problem. In Neutrons, X-rays, and Light (Second Edition); Lindner, P., Oberdisse, J., Eds.; Elsevier, 2025; pp 61–90. https://doi.org/10.1016/B978-0-443-29116-6.00004-7.

Green, M. L.; Choi, C. L.; Hattrick-Simpers, J. R.; Joshi, A. M.; Takeuchi, I.; Barron, S. C.; Campo, E.; Chiang, T.; Empedocles, S.; Gregoire, J. M.; Kusne, A. G.; Martin, J.; Mehta, A.; Persson, K.; Trautt, Z.; Van Duren, J.; Zakutayev, A. Fulfilling the Promise of the Materials Genome Initiative with High-Throughput Experimental Methodologies. Appl. Phys. Rev. 2017, 4 (1), 011105. https://doi.org/10.1063/1.4977487.

Hamley, I. W.; Castelletto, V. Small-Angle Scattering of Block Copolymers: In the Melt, Solution and Crystal States. Prog. Polym. Sci. 2004, 29 (9), 909–948. https://doi.org/10.1016/j.progpolymsci.2004.06.001.

Hardware Interface Packages — bluesky 1.10.0.post14+gfc4204d4 documentation. https://blueskyproject.io/bluesky/hardware-interfaces.html (accessed 2022-11-17).

High-throughput and data driven strategies for the design of deep-eutectic solvent electrolytes - Molecular Systems Design & Engineering (RSC Publishing) DOI:10.1039/D2ME00050D. https://pubs.rsc.org/en/content/articlehtml/2022/me/d2me00050d (accessed 2023-02-15).

HIPAA Compliance with Google Workspace and Cloud Identity - Google Workspace Admin Help. https://support.google.com/a/answer/3407054?hl=en (accessed 2023-02-15).

Hitzler, P. A Review of the Semantic Web Field. Commun. ACM 2021, 64 (2), 76–83. https://doi.org/10.1145/3397512.

Hollamby, M. J.; Borisova, D.; Brown, P.; Eastoe, J.; Grillo, I.; Shchukin, D. Growth of Mesoporous Silica Nanoparticles Monitored by Time-Resolved Small-Angle Neutron Scattering. Langmuir 2012, 28 (9), 4425–4433. https://doi.org/10.1021/la203097x.

Home - Gathering for Open Science Hardware. https://openhardware.science/ (accessed 2024-02-05).

Home | ICSD. https://icsd.products.fiz-karlsruhe.de/ (accessed 2022-11-22).

Horton, M. K.; Woods-Robinson, R. Addressing the Critical Need for Open Experimental Databases in Materials Science. Patterns 2021, 2 (12), 100411. https://doi.org/10.1016/j.patter.2021.100411.

Houston, P. L.; Nandi, A.; Bowman, J. M. A Machine Learning Approach for Prediction of Rate Constants. J. Phys. Chem. Lett. 2019, 10 (17), 5250–5258. https://doi.org/10.1021/acs.jpclett.9b01810.

Hänggi, P.; Talkner, P.; Borkovec, M. Reaction-Rate Theory: Fifty Years after Kramers. Rev. Mod. Phys. 1990, 62 (2), 251–341. https://doi.org/10.1103/RevModPhys.62.251.

Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Aspuru-Guzik, A. Gryffin: An Algorithm for Bayesian Optimization of Categorical Variables Informed by Expert Knowledge. Appl. Phys. Rev. 2021, 8 (3), 031406. https://doi.org/10.1063/5.0048164.

Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving Laboratories. Trends Chem. 2019, 1 (3), 282–291. https://doi.org/10.1016/j.trechm.2019.02.007.

Ilavsky, J.; Zhang, F.; Andrews, R. N.; Kuzmenko, I.; Jemian, P. R.; Levine, L. E.; Allen, A. J. Development of Combined Microstructure and Structure Characterization Facility for in Situ and Operando Studies at the Advanced Photon Source. J. Appl. Crystallogr. 2018, 51 (3), 867–882. https://doi.org/10.1107/S160057671800643X.

Illustration of prior and posterior Gaussian process for different kernels. scikit-learn. https://scikit-learn/stable/auto_examples/gaussian_process/plot_gpr_prior_posterior.html (accessed 2024-01-30).

Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification. Federal Register. https://www.federalregister.gov/documents/2022/10/13/2022-21658/implementation-of-additional-export-controls-certain-advanced-computing-and-semiconductor (accessed 2023-02-09).

Inc, B. Laboratory Information Management System | LIMS | Labguru. https://www.labguru.com/lims (accessed 2022-11-21).

Inc, L. Automate Your Laboratory with the Global Leader for LIMS and ELN. https://www.labware.com (accessed 2022-11-21).

Jacobsson, T. J.; Hultqvist, A.; García-Fernández, A.; Anand, A.; Al-Ashouri, A.; Hagfeldt, A.; Crovetto, A.; Abate, A.; Ricciardulli, A. G.; Vijayan, A.; Kulkarni, A.; Anderson, A. Y.; Darwich, B. P.; Yang, B.; Coles, B. L.; Perini, C. A. R.; Rehermann, C.; Ramirez, D.; Fairen-Jimenez, D.; Di Girolamo, D.; Jia, D.; Avila, E.; Juarez-Perez, E. J.; Baumann, F.; Mathies, F.; González, G. S. A.; Boschloo, G.; Nasti, G.; Paramasivam, G.; Martínez-Denegri, G.; Näsström, H.; Michaels, H.; Köbler, H.; Wu, H.; Benesperi, I.; Dar, M. I.; Bayrak Pehlivan, I.; Gould, I. E.; Vagott, J. N.; Dagar, J.; Kettle, J.; Yang, J.; Li, J.; Smith, J. A.; Pascual, J.; Jerónimo-Rendón, J. J.; Montoya, J. F.; Correa-Baena, J.-P.; Qiu, J.; Wang, J.; Sveinbjörnsson, K.; Hirselandt, K.; Dey, K.; Frohna, K.; Mathies, L.; Castriotta, L. A.; Aldamasy, M. H.; Vasquez-Montoya, M.; Ruiz-Preciado, M. A.; Flatken, M. A.; Khenkin, M. V.; Grischek, M.; Kedia, M.; Saliba, M.; Anaya, M.; Veldhoen, M.; Arora,

N.; Shargaieva, O.; Maus, O.; Game, O. S.; Yudilevich, O.; Fassl, P.; Zhou, Q.; Betancur, R.; Munir, R.; Patidar, R.; Stranks, S. D.; Alam, S.; Kar, S.; Unold, T.; Abzieher, T.; Edvinsson, T.; David, T. W.; Paetzold, U. W.; Zia, W.; Fu, W.; Zuo, W.; Schröder, V. R. F.; Tress, W.; Zhang, X.; Chiang, Y.-H.; Iqbal, Z.; Xie, Z.; Unger, E. An Open-Access Database and Analysis Tool for Perovskite Solar Cells Based on the FAIR Data Principles. Nat. Energy 2022, 7 (1), 107–115. https://doi.org/10.1038/s41560-021-00941-3.

Johansson, F. Arb: Efficient Arbitrary-Precision Midpoint-Radius Interval Arithmetic. IEEE Trans. Comput. 2017, 66 (8), 1281–1292. https://doi.org/10.1109/TC.2017.2690633.

Johansson, F. Mpmath: A Python Library for Arbitrary-Precision Floating-Point Arithmetic (Version 0.18), December 2013. URL Httpmpmath Org 2013.

Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P. O. Organic Reactivity from Mechanism to Machine Learning. Nat. Rev. Chem. 2021, 5 (4), 240–255. https://doi.org/10.1038/s41570-021-00260-x.

Jubilee. https://jubilee3d.com/index.php?title=Main_Page (accessed 2025-05-13).

Juchli, D. SiLA 2: The Next Generation Lab Automation Standard. In Smart Biolabs of the Future; Beutel, S., Lenk, F., Eds.; Advances in Biochemical Engineering/Biotechnology; Springer International Publishing: Cham, 2022; pp 147–174. https://doi.org/10.1007/10_2022_204.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. Nature 2021, 596 (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kanza, S.; Willoughby, C.; Gibbins, N.; Whitby, R.; Frey, J. G.; Erjavec, J.; Zupančič, K.; Hren, M.; Kovač, K. Electronic Lab Notebooks: Can They Replace Paper? J. Cheminformatics 2017, 9 (1), 31. https://doi.org/10.1186/s13321-017-0221-3.

Kim, C.; Yoon, S.; Lee, J. H. Facile Large-Scale Synthesis of Mesoporous Silica Nanoparticles at Room Temperature in a Monophasic System with Fine Size Control. Microporous Mesoporous Mater. 2019, 288, 109595. https://doi.org/10.1016/j.micromeso.2019.109595.

Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. Nucleic Acids Res. 2016, 44 (D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951.

Kim, S.; Thiessen, P. A.; Cheng, T.; Yu, B.; Bolton, E. E. An Update on PUG-REST: RESTful Interface for Programmatic Access to PubChem. Nucleic Acids Res. 2018, 46 (W1), W563–W570. https://doi.org/10.1093/nar/gky294.

Kim, T.-W.; Chung, P.-W.; Lin, V. S.-Y. Facile Synthesis of Monodisperse Spherical MCM-48 Mesoporous Silica Nanoparticles with Controlled Particle Size. Chem. Mater. 2010, 22 (17), 5093–5104. https://doi.org/10.1021/cm1017344.

Komp, E.; Valleau, S. Low-Cost Prediction of Molecular and Transition State Partition Functions via Machine Learning. Chem. Sci. 2022, 13 (26), 7900–7906. https://doi.org/10.1039/D2SC01334G.

Komp, E.; Valleau, S. Machine Learning Quantum Reaction Rate Constants. J. Phys. Chem. A 2020, 124 (41), 8607–8613.

Kresge, C. T.; Leonowicz, M. E.; Roth, W. J.; Vartuli, J. C.; Beck, J. S. Ordered Mesoporous Molecular Sieves Synthesized by a Liquid-Crystal Template Mechanism. Nature 1992, 359 (6397), 710–712. https://doi.org/10.1038/359710a0.

Kumar, A.; Kini, S. G.; Rathi, E. A Recent Appraisal of Artificial Intelligence and In Silico ADMET Prediction in the Early Stages of Drug Discovery. Mini-Rev. Med. Chem. 2021, 21 (18), 2788–2800. https://doi.org/10.2174/1389557521666210401091147.

Kusne, A. G.; Gao, T.; Mehta, A.; Ke, L.; Nguyen, M. C.; Ho, K.-M.; Antropov, V.; Wang, C.-Z.; Kramer, M. J.; Long, C.; Takeuchi, I. On-the-Fly Machine-Learning for High-Throughput Experiments: Search for Rare-Earth-Free Permanent Magnets. Sci. Rep. 2014, 4 (1), 6367. https://doi.org/10.1038/srep06367.

Lai, C.-Y.; Trewyn, B. G.; Jeftinija, D. M.; Jeftinija, K.; Xu, S.; Jeftinija, S.; Lin, V. S.-Y. A Mesoporous Silica Nanosphere-Based Carrier System with Chemically Removable CdS Nanoparticle Caps for Stimuli-Responsive Controlled Release of Neurotransmitters and Drug Molecules. J. Am. Chem. Soc. 2003, 125 (15), 4451–4459. https://doi.org/10.1021/ja028650l.

Langner, S.; Häse, F.; Perea, J. D.; Stubhan, T.; Hauch, J.; Roch, L. M.; Heumueller, T.; Aspuru-Guzik, A.; Brabec, C. J. Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems. Adv. Mater. 2020, 32 (14), 1907801. https://doi.org/10.1002/adma.201907801.

Leardi, R. Experimental Design in Chemistry: A Tutorial. Anal. Chim. Acta 2009, 652 (1), 161–172. https://doi.org/10.1016/j.aca.2009.06.015.

Leong, C. J.; Low, K. Y. A.; Recatala-Gomez, J.; Quijano Velasco, P.; Vissol-Gaudin, E.; Tan, J. D.; Ramalingam, B.; I Made, R.; Pethe, S. D.; Sebastian, S.; Lim, Y.-F.; Khoo, Z. H. J.; Bai, Y.; Cheng, J. J. W.; Hippalgaonkar, K. An Object-Oriented Framework to Enable Workflow Evolution across Materials Acceleration Platforms. Matter 2022, 5 (10), 3124–3134. https://doi.org/10.1016/j.matt.2022.08.017.

Lewis-Atwell, T.; Townsend, P. A.; Grayson, M. N. Machine Learning Activation Energies of Chemical Reactions. WIREs Comput. Mol. Sci. 2022, 12 (4), e1593. https://doi.org/10.1002/wcms.1593.

Light, J. C.; Hamilton, I. P.; Lill, J. V. Generalized Discrete Variable Approximation in Quantum Mechanics. J. Chem. Phys. 1985, 82 (3), 1400–1409. https://doi.org/10.1063/1.448462.

LIMS- Laboratory Information Management Systems - US. https://www.thermofisher.com/us/en/home/digital-solutions/lab-informatics/lab-information-management-systems-lims.html (accessed 2022-11-21).

Lindner, P. (Peter). Neutrons, X-Rays and Light : Scattering Methods Applied to Soft Condensed Matter. No Title.

Lo, S.; G. Baird, S.; Schrier, J.; Blaiszik, B.; Carson, N.; Foster, I.; Aguilar-Granda, A.; V. Kalinin, S.; Maruyama, B.; Politi, M.; Tran, H.; D. Sparks, T.; Aspuru-Guzik, A. Review of Low-Cost Self-Driving Laboratories in Chemistry and Materials Science: The "Frugal Twin" Concept. Digit. Discov. 2024, 3 (5), 842–868. https://doi.org/10.1039/D3DD00223C.

Lu, Y.; Fan, H.; Stump, A.; Ward, T. L.; Rieker, T.; Brinker, C. J. Aerosol-Assisted Self-Assembly of Mesostructured Spherical Nanoparticles. Nature 1999, 398 (6724), 223–226. https://doi.org/10.1038/18410.

Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems; Curran Associates, Inc., 2017; Vol. 30.

López, V.; Villegas, M. R.; Rodríguez, V.; Villaverde, G.; Lozano, D.; Baeza, A.; Vallet-Regí, M. Janus Mesoporous Silica Nanoparticles for Dual Targeting of Tumor Cells and Mitochondria. ACS Appl. Mater. Interfaces 2017, 9 (32), 26697–26706. https://doi.org/10.1021/acsami.7b06906.

M, S. I. The Distribution of Points in a Cube and the Approximate Evaluation of Integrals. USSR Comput. Math. Math. Phys. 1967, 7, 86–112.

M. Weigandt, K.; C. Pozzo, D.; Porcar, L. Structure of High Density Fibrin Networks Probed with Neutron Scattering and Rheology. Soft Matter 2009, 5 (21), 4321–4330. https://doi.org/10.1039/B906256D.

Machineagency/Science_jubilee, 2023. https://github.com/machineagency/science_jubilee (accessed 2024-02-05).

MacLeod, B. P.; Parlane, F. G. L.; Brown, A. K.; Hein, J. E.; Berlinguette, C. P. Flexible Automation Accelerates Materials Discovery. Nat. Mater. 2022, 21 (7), 722–726. https://doi.org/10.1038/s41563-021-01156-3.

Maier, W. F.; Stöwe, K.; Sieg, S. Combinatorial and High-Throughput Materials Science. Angew. Chem. Int. Ed. 2007, 46 (32), 6016–6067. https://doi.org/10.1002/anie.200603675.

Martin, T. B.; Sutherland, D. R.; McDannald, A.; Kusne, A. G.; Beaucage, P. A. Autonomous Small-Angle Scattering for Accelerated Soft Material Formulation Optimization. arXiv March 14, 2025. https://doi.org/10.48550/arXiv.2503.11859.

Materials Project - Community. Materials Project. https://materialsproject.org/community (accessed 2023-02-14).

Materials Project - Materials Explorer. Materials Project. https://materialsproject.org/materials (accessed 2023-02-14).

Medina, J.; Ziaullah, A. W.; Park, H.; Castelli, I. E.; Shaon, A.; Bensmail, H.; El-Mellouhi, F. Accelerating the Adoption of Research Data Management Strategies. Matter 2022, 5 (11), 3614–3642. https://doi.org/10.1016/j.matt.2022.10.007.

Mehr, S. H. M.; Craven, M.; Leonov, A. I.; Keenan, G.; Cronin, L. A Universal System for Digitization and Automatic Execution of the Chemical Synthesis Literature. Science 2020, 370 (6512), 101–108. https://doi.org/10.1126/science.abc2986.

Meier, M.; Ungerer, J.; Klinge, M.; Nirschl, H. Synthesis of Nanometric Silica Particles via a Modified Stöber Synthesis Route. Colloids Surf. Physicochem. Eng. Asp. 2018, 538, 559–564. https://doi.org/10.1016/j.colsurfa.2017.11.047.

Mekki-Berrada, F.; Ren, Z.; Huang, T.; Wong, W. K.; Zheng, F.; Xie, J.; Tian, I. P. S.; Jayavelu, S.; Mahfoud, Z.; Bash, D.; Hippalgaonkar, K.; Khan, S.; Buonassisi, T.; Li, Q.; Wang, X. Two-Step Machine Learning Enables Optimized Nanoparticle Synthesis. Npj Comput. Mater. 2021, 7 (1), 1–10. https://doi.org/10.1038/s41524-021-00520-w.

Meng, D.; Zheng, R.; Zhao, Y.; Zhang, E.; Dou, L.; Yang, Y. Near-Infrared Materials: The Turning Point of Organic Photovoltaics. Adv. Mater. 2022, 34 (10), 2107330. https://doi.org/10.1002/adma.202107330.

Meyer, B.; Sawatlon, B.; Heinen, S.; Lilienfeld, O. A. von; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. Chem. Sci. 2018, 9 (35), 7069–7077. https://doi.org/10.1039/C8SC01949E.

Miller, W. H.; Schwartz, S. D.; Tromp, J. W. Quantum Mechanical Rate Constants for Bimolecular Reactions. J. Chem. Phys. 1983, 79 (10), 4889–4898. https://doi.org/10.1063/1.445581.

Mills, G.; Jacobsen, K. W. CHAPTER 16 Nudged Elastic Band Method for Nding Minimum En-Ergy Paths of Transitions. Class. Quantum Dyn. Condens. Phase Simul. 1997.

Mohan, A.; Jaison, A.; Lee, Y.-C. Emerging Trends in Mesoporous Silica Nanoparticle-Based Catalysts for CO2 Utilization Reactions. Inorg. Chem. Front. 2023, 10 (11), 3171–3194. https://doi.org/10.1039/D3QI00378G.

Monge, N. Development of a Machine Learning Methodology for Nanoparticle Classification from Small-Angle X-ray Scattering (SAXS) Data. phdthesis, Université Grenoble Alpes [2020-....], 2024. https://theses.hal.science/tel-05039356 (accessed 2025-05-01).

Moosavi, S. M.; Jablonka, K. M.; Smit, B. The Role of Machine Learning in the Understanding and Design of Materials. J. Am. Chem. Soc. 2020, 142 (48), 20273–20287. https://doi.org/10.1021/jacs.0c09105.

Nandi, A.; Bowman, J. M.; Houston, P. A Machine Learning Approach for Rate Constants. II. Clustering, Training, and Predictions for the O(3P) + HCl → OH + Cl Reaction. J. Phys. Chem. A 2020, 124 (28), 5746–5755. https://doi.org/10.1021/acs.jpca.0c04348.

Narayan, R.; Nayak, U. Y.; Raichur, A. M.; Garg, S. Mesoporous Silica Nanoparticles: A Comprehensive Review on Synthesis and Recent Advances. Pharmaceutics 2018, 10 (3), 118. https://doi.org/10.3390/pharmaceutics10030118.

National Academies of Sciences, E. NSF Efforts to Achieve the Nation's Vision for the Materials Genome Initiative: Designing Materials to Revolutionize and Engineer Our Future (DMREF); 2022. https://doi.org/10.17226/26723.

Networked laboratory equipment. SPECTARIS - Deutscher Industrieverband für Optik, Photonik, Analysen- und Medizintechnik. https://www.spectaris.de/en/association/thespectarisindustries/networked-laboratory-equipment/ (accessed 2023-02-01).

Nikolaev, P.; Hooper, D.; Webber, F.; Rao, R.; Decker, K.; Krein, M.; Poleski, J.; Barto, R.; Maruyama, B. Autonomy in Materials Research: A Case Study in Carbon Nanotube Growth. Npj Comput. Mater. 2016, 2 (1), 1–6. https://doi.org/10.1038/npjcompumats.2016.31.

Nnakwe, C. C.; Cooch, N.; Huang-Saad, A. Investing in Academic Technology Innovation and Entrepreneurship: Moving Beyond Research Funding through the NSF I-CORPSTM Program. Technol. Innov. 2018, 19 (4), 773–786. https://doi.org/10.21300/19.4.2018.773.

Nothling, M. D.; Daniels, J. E.; Vo, Y.; Johan, I.; Stenzel, M. H. Mechanically Activated Solid-State Radical Polymerization and Cross-Linking via Piezocatalysis. Angew. Chem. 2023, 135 (20), e202218955. https://doi.org/10.1002/ange.202218955.

Noureddine, A.; Maestas-Olguin, A.; Tang, L.; Corman-Hijar, J. I.; Olewine, M.; Krawchuck, J. A.; Tsala Ebode, J.; Edeh, C.; Dang, C.; Negrete, O. A.; Watt, J.; Howard, T.; Coker, E. N.; Guo, J.; Brinker, C. J. Future of Mesoporous Silica Nanoparticles in Nanomedicine: Protocol for Reproducible Synthesis, Characterization, Lipid Coating, and Loading of Therapeutics (Chemotherapeutic, Proteins, siRNA and mRNA). ACS Nano 2023, 17 (17), 16308–16325. https://doi.org/10.1021/acsnano.3c07621.

Oellermann, M.; Jolles, J. W.; Ortiz, D.; Seabra, R.; Wenzel, T.; Wilson, H.; Tanner, R. L. Open Hardware in Science: The Benefits of Open Electronics. Integr. Comp. Biol. 2022, 62 (4), 1061–1075. https://doi.org/10.1093/icb/icac043.

Open Source Hardware Definition. https://oshwa.org/resources/open-source-hardware-definition/ (accessed 2025-05-14).

OpenAI. GPT-4 Technical Report. arXiv March 16, 2023. https://doi.org/10.48550/arXiv.2303.08774.

OPTIMADE. materials-consortia.github.io. https://optimade.org/ (accessed 2023-02-22).

Pal, M.; Ganesan, V. Zinc Phthalocyanine and Silver/Gold Nanoparticles Incorporated MCM-41 Type Materials as Electrode Modifiers. Langmuir 2009, 25 (22), 13264–13272. https://doi.org/10.1021/la901792b.

Pauw, B. R. Everything SAXS: Small-Angle Scattering Pattern Collection and Correction. J. Phys. Condens. Matter 2013, 25 (38), 383201. https://doi.org/10.1088/0953-8984/25/38/383201.

Pauw, B. R.; Pedersen, J. S.; Tardif, S.; Takata, M.; Iversen, B. B. Improvements and Considerations for Size Distribution Retrieval from Small-Angle Scattering Data by Monte Carlo Methods. J. Appl. Crystallogr. 2013, 46 (2), 365–371. https://doi.org/10.1107/S0021889813001295.

Pelkie, B.G; Pozzo, L. D. The Laboratory of Babel: Highlighting Community Needs for Integrated Materials Data Management. Digit. Discov. 2023, 2 (3), 544–556. https://doi.org/10.1039/D3DD00022B.

Pelkie, B.; Baird, S.; Aissi, E.; Aspuru-Takata, K.; Cao, Y.; Chang, J. H.; Gambhir, K.; Hale, W. S.; Hao, L.; Hattrick, C.; Hein, J.; Luo, D.; Melville, O.; Ngan, M.; Nyeland, L. L. B.; Peek, N.; Politi, M.; Rajkumar, E. E.; Siemenn, A.; Subbaraman, B.; Vasquez, S.; Watchorn, J.; Zhang, W.; Ziskason, R.; Pozzo, L.; Buonassisi, T.; Vegge, T. Democratizing Self-Driving Labs through User-Developed Automation Infrastructure. ChemRxiv February 12, 2025. https://doi.org/10.26434/chemrxiv-2025-zhkrf.

Pendleton, I. M.; Cattabriga, G.; Li, Z.; Najeeb, M. A.; Friedler, S. A.; Norquist, A. J.; Chan, E. M.; Schrier, J. Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): A Software Pipeline for Automated Chemical Experimentation and Data Management. MRS Commun. 2019, 9 (3), 846–859. https://doi.org/10.1557/mrc.2019.72.

Peters, B. Reaction Rate Theory and Rare Events; Elsevier, 2017.

PhasIR: An Instrumentation and Analysis Software for High-throughput Phase Transition Temperature Measurements. https://openhardware.metajnl.com/articles/10.5334/joh.39/ (accessed 2022-11-17).

Politi, M.; Baum, F.; Vaddi, K.; Antonio, E.; Vasquez, S.; Bishop, B. P.; Peek, N.; Holmberg, V. C.; Pozzo, L. D. A High-Throughput Workflow for the Synthesis of CdSe Nanocrystals Using a Sonochemical Materials Acceleration Platform. Digit. Discov. 2023.

pozzo-research-group/automated-MSN-synthesis: Documentation and example code for forthcoming automated mesoporous silica nanoparticle synthesis paper. https://github.com/pozzo-research-group/automated-MSN-synthesis (accessed 2025-06-02).

Pozzo-Research-Group/Saxs_data_processing, 2024. https://github.com/pozzo-research-group/saxs_data_processing (accessed 2024-02-06).

pozzo-research-group/silica-np-synthesis. https://github.com/pozzo-research-group/silica-np-synthesis (accessed 2025-06-05).

PubChem. Ammonia. https://pubchem.ncbi.nlm.nih.gov/compound/222 (accessed 2025-05-04).

PubChem. Monoethanolamine. https://pubchem.ncbi.nlm.nih.gov/compound/700 (accessed 2025-05-04).

Python API Reference — The RDKit 2023.03.1 documentation. https://www.rdkit.org/docs/api-docs.html (accessed 2023-09-21).

Pörtner, H.-O.; Roberts, D. C.; Adams, H.; Adelekan, I.; Adler, C.; Adrian, R.; Aldunce, P.; Ali, E.; Begum, R. A.; Bednar-Friedl, B.; Bezner Kerr, R.; Biesbroek, R.; Birkmann, J.; Bowen, K.; Caretta, M. A.; Carnicer, J.; Castellanos, E.; Cheong, T. S.; Chow, W.; Cissé, G.; Clayton, S.; Constable, A.; Cooley, S.; Costello, M. J.; Craig, M.; Cramer, W.; Dawson, R.; Dodman, D.; Efitre, J.; Garschagen, M.; Gilmore, E.; Glavovic, B.; Gutzler, D.; Haasnoot, M.; Harper, S.; Hasegawa, T.; Hayward, B.; Hicke, J. A.; Hirabayashi, Y.; Huang, C.; Kalaba, K.; Kiessling, W.; Kitoh, A.; Lasco, R.; Lawrence, J.; Lemos, M. F.; Lempert, R.; Lennard, C.; Ley, D.; Lissner, T.; Liu, Q.; Liwenga, E.; Lluch-Cota, S.; Löschke, S.; Lucatello, S.; Luo, Y.; Mackey, B.; Mintenbeck, K.; Mirzabaev, A.; Möller, V.; Vale, M. M.; Morecroft, M. D.; Mortsch, L.; Mukherji, A.; Mustonen, T.; Mycoo, M.; Nalau, J.; New, M.; Okem, A.; Ometto, J. P.; O'Neill, B.; Pandey, R.; Parmesan, C.; Pelling, M.; Pinho, P. F.; Pinnegar, J.; Poloczanska, E. S.; Prakash, A.; Preston, B.; Racault, M.-F.; Reckien, D.; Revi, A.; Rose, S. K.; Schipper, E. L. F.; Schmidt, D. N.; Schoeman, D.; Shaw, R.; Simpson, N. P.; Singh, C.; Solecki, W.; Stringer, L.; Totin, E.; Trisos, C.; Trisurat, Y.; Aalst, M. van; Viner, D.; Wairu, M.; Warren, R.; Wester, P.; Wrathall, D.; Ibrahim, Z. Z. Technical Summary. In Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change; Pörtner, H.-O., Roberts, D. C., Tignor, M. M. B., Poloczanska, E. S., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., Rama, B., Eds.; Cambridge University Press, 2022.

RepRapFirmware.org. https://www.reprapfirmware.org/ (accessed 2024-02-05).

Roberts, G.; Nieh, M.-P.; W.K. Ma, A.; Yang, Q. Automated Structural Analysis of Small Angle Scattering Data from Common Nanoparticles via Machine Learning. Digit. Discov. 2025. https://doi.org/10.1039/D5DD00059A.

Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: An Orchestration Software to Democratize Autonomous Discovery. PLOS ONE 2020, 15 (4), e0229862. https://doi.org/10.1371/journal.pone.0229862.

Rodriguez, J.; Politi, M.; Adler, S.; Beck, D.; Pozzo, L. High-Throughput and Data Driven Strategies for the Design of Deep-Eutectic Solvent Electrolytes. Mol. Syst. Des. Eng. 2022, 7 (8), 933–949. https://doi.org/10.1039/D2ME00050D.

Rupnow, C. C.; MacLeod, B. P.; Mokhtari, M.; Ocean, K.; Dettelbach, K. E.; Lin, D.; Parlane, F. G. L.; Chiu, H. N.; Rooney, M. B.; Waizenegger, C. E. B.; Hoog, E. I. de; Soni, A.; Berlinguette, C. P. A Self-Driving Laboratory Optimizes a Scalable Process for Making Functional Coatings. Cell Rep. Phys. Sci. 2023, 4 (5). https://doi.org/10.1016/j.xcrp.2023.101411.

Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. Phys. Rev. Lett. 2012, 108 (5), 058301. https://doi.org/10.1103/PhysRevLett.108.058301.

Salley, D.; Keenan, G.; Grizou, J.; Sharma, A.; Martín, S.; Cronin, L. A Nanomaterials Discovery Robot for the Darwinian Evolution of Shape Programmable Gold Nanoparticles. Nat. Commun. 2020, 11 (1), 2771. https://doi.org/10.1038/s41467-020-16501-4.

Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. Science 2018, 361 (6400), 360–365. https://doi.org/10.1126/science.aat2663.

SasView. SasView. https://sasview.github.io/ (accessed 2025-05-12).

SasView/Sasmodels, 2025. https://github.com/SasView/sasmodels (accessed 2025-05-15).

savgol_filter — SciPy v1.15.3 Manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.savgol_filter.html (accessed 2025-06-12).

Savitzky, Abraham.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 1964, 36 (8), 1627–1639. https://doi.org/10.1021/ac60214a047.

Scheffler, M.; Aeschlimann, M.; Albrecht, M.; Bereau, T.; Bungartz, H.-J.; Felser, C.; Greiner, M.; Groß, A.; Koch, C. T.; Kremer, K.; Nagel, W. E.; Scheidgen, M.; Wöll, C.; Draxl, C. FAIR Data Enabling New Horizons for Materials Research. Nature 2022, 604 (7907), 635–642. https://doi.org/10.1038/s41586-022-04501-x.

Scheurer, C.; Reuter, K. Role of the Human-in-the-Loop in Emerging Self-Driving Laboratories for Heterogeneous Catalysis. Nat. Catal. 2025, 8 (1), 13–19. https://doi.org/10.1038/s41929-024-01275-5.

Schlexer Lamoureux, P.; Winther, K. T.; Garrido Torres, J. A.; Streibel, V.; Zhao, M.; Bajdich, M.; Abild-Pedersen, F.; Bligaard, T. Machine Learning for Computational Heterogeneous Catalysis. ChemCatChem 2019, 11 (16), 3581–3601. https://doi.org/10.1002/cctc.201900595.

Schwarz, N.; Veseli, S.; Jarosz, D. Data Management at the Advanced Photon Source. Synchrotron Radiat. News 2019, 32 (3), 13–18. https://doi.org/10.1080/08940886.2019.1608120.

Schüth, F. The Evolution of Ordered Mesoporous Materials. In Studies in Surface Science and Catalysis; Terasaki, O., Ed.; Mesoporous Crystals and Related Nano-Structured Materials; Elsevier, 2004; Vol. 148, pp 1–13. https://doi.org/10.1016/S0167-2991(04)80190-3.

SDL for KIDS. https://sites.google.com/matterhorn.studio/sdl4kids/home (accessed 2024-02-05).

Search Citrination. https://citrination.com/search/simple?searchMatchOption=fuzzyMatch (accessed 2022-11-22).

Seidel, S.; Cruz-Bournazou, M. N.; Groß, S.; Schollmeyer, J. K.; Kurreck, A.; Krauss, S.; Neubauer, P. A Comprehensive IT Infrastructure for an Enzymatic Product Development in a Digitalized Biotechnological Laboratory. In Smart Biolabs of the Future; Beutel, S.,

Lenk, F., Eds.; Advances in Biochemical Engineering/Biotechnology; Springer International Publishing: Cham, 2022; pp 61–82. https://doi.org/10.1007/10_2022_207.

Seifrid, M.; Hattrick-Simpers, J.; Aspuru-Guzik, A.; Kalil, T.; Cranford, S. Reaching Critical MASS: Crowdsourcing Designs for the next Generation of Materials Acceleration Platforms. Matter 2022, 5 (7), 1972–1976. https://doi.org/10.1016/j.matt.2022.05.035.

Seifrid, M.; Pollice, R.; Aguilar-Granda, A.; Morgan Chan, Z.; Hotta, K.; Ser, C. T.; Vestfrid, J.; Wu, T. C.; Aspuru-Guzik, A. Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab. Acc. Chem. Res. 2022. https://doi.org/10.1021/acs.accounts.2c00220.

SENAITE · Enterprise Open Source Laboratory System. https://github.com/senaite/senaite.github.io/ (accessed 2022-11-21).

Shankar, K. Order from Chaos: The Poetics and Pragmatics of Scientific Recordkeeping. J. Am. Soc. Inf. Sci. Technol. 2007, 58 (10), 1457–1466. https://doi.org/10.1002/asi.20625.

Shen, D.; Yang, J.; Li, X.; Zhou, L.; Zhang, R.; Li, W.; Chen, L.; Wang, R.; Zhang, F.; Zhao, D. Biphase Stratification Approach to Three-Dimensional Dendritic Biodegradable Mesoporous Silica Nanospheres. Nano Lett. 2014, 14 (2), 923–932. https://doi.org/10.1021/nl404316v.

Sim, M.; Vakili, M. G.; Strieth-Kalthoff, F.; Hao, H.; Hickman, R. J.; Miret, S.; Pablo-García, S.; Aspuru-Guzik, A. ChemOS 2.0: An Orchestration Architecture for Chemical Self-Driving Laboratories. Matter 2024, 7 (9), 2959–2977. https://doi.org/10.1016/j.matt.2024.04.022.

Sivia, D. S. Elementary Scattering Theory: For X-Ray and Neutron Users; Oxford University Press, 2011. https://doi.org/10.1093/acprof:oso/9780199228676.001.0001.

Sobol — SciPy v1.15.3 Manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.qmc.Sobol.html#re15be05a07a0-1 (accessed 2025-06-02).

Soedarmadji, E.; Stein, H. S.; Suram, S. K.; Guevarra, D.; Gregoire, J. M. Tracking Materials Science Data Lineage to Manage Millions of Materials Experiments and Analyses. Npj Comput. Mater. 2019, 5 (1), 1–9. https://doi.org/10.1038/s41524-019-0216-x.

Stach, E.; DeCost, B.; Kusne, A. G.; Hattrick-Simpers, J.; Brown, K. A.; Reyes, K. G.; Schrier, J.; Billinge, S.; Buonassisi, T.; Foster, I.; Gomes, C. P.; Gregoire, J. M.; Mehta, A.; Montoya, J.; Olivetti, E.; Park, C.; Rotenberg, E.; Saikin, S. K.; Smullin, S.; Stanev, V.; Maruyama, B. Autonomous Experimentation Systems for Materials Development: A Community Perspective. Matter 2021, 4 (9), 2702–2726. https://doi.org/10.1016/j.matt.2021.06.036.

Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; Cronin, L. Organic Synthesis in a Modular Robotic System Driven by a Chemical Programming Language. Science 2019, 363 (6423), eaav2211. https://doi.org/10.1126/science.aav2211.

Streamlit • A faster way to build and share data apps. https://streamlit.io/ (accessed 2024-02-05).

Strieth-Kalthoff, F.; Hao, H.; Rathore, V.; Derasp, J.; Gaudin, T.; Angello, N. H.; Seifrid, M.; Trushina, E.; Guy, M.; Liu, J.; Tang, X.; Mamada, M.; Wang, W.; Tsagaantsooj, T.; Lavigne, C.; Pollice, R.; Wu, T. C.; Hotta, K.; Bodo, L.; Li, S.; Haddadnia, M.; Wołos, A.; Roszak, R.; Ser, C. T.; Bozal-Ginesta, C.; Hickman, R. J.; Vestfrid, J.; Aguilar-Granda, A.; Klimareva, E. L.; Sigerson, R. C.; Hou, W.; Gahler, D.; Lach, S.; Warzybok, A.; Borodin, O.; Rohrbach, S.; Sanchez-Lengeling, B.; Adachi, C.; Grzybowski, B. A.; Cronin, L.; Hein, J. E.; Burke, M. D.; Aspuru-Guzik, A. Delocalized, Asynchronous, Closed-Loop Discovery of Organic Laser Emitters. Science 2024, 384 (6697), eadk9227. https://doi.org/10.1126/science.adk9227.

Stöber, W.; Fink, A.; Bohn, E. Controlled Growth of Monodisperse Silica Spheres in the Micron Size Range. J. Colloid Interface Sci. 1968, 26 (1), 62–69. https://doi.org/10.1016/0021-9797(68)90272-5.

Subbaraman, B.; de Lange, O.; Ferguson, S.; Peek, N. The Duckbot: A System for Automated Imaging and Manipulation of Duckweed. PloS One 2024, 19 (1), e0296717. https://doi.org/10.1371/journal.pone.0296717.

Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. J. Chem. Inf. Model. 2016, 56 (10), 1894–1904. https://doi.org/10.1021/acs.jcim.6b00207.

Szymanski, N. J.; Rendy, B.; Fei, Y.; Kumar, R. E.; He, T.; Milsted, D.; McDermott, M. J.; Gallant, M.; Cubuk, E. D.; Merchant, A.; Kim, H.; Jain, A.; Bartel, C. J.; Persson, K.; Zeng, Y.; Ceder, G. An Autonomous Laboratory for the Accelerated Synthesis of Novel Materials. Nature 2023, 624 (7990), 86–91. https://doi.org/10.1038/s41586-023-06734-w.

Talley, K. R.; White, R.; Wunder, N.; Eash, M.; Schwarting, M.; Evenson, D.; Perkins, J. D.; Tumas, W.; Munch, K.; Phillips, C.; Zakutayev, A. Research Data Infrastructure for High-Throughput Experimental Materials Science. Patterns 2021, 2 (12), 100373. https://doi.org/10.1016/j.patter.2021.100373.

Tao, H.; Wu, T.; Kheiri, S.; Aldeghi, M.; Aspuru-Guzik, A.; Kumacheva, E. Self-Driving Platform for Metal Nanoparticle Synthesis: Combining Microfluidics and Machine Learning. Adv. Funct. Mater. 2021, 31 (51), 2106725. https://doi.org/10.1002/adfm.202106725.

Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic Accessibility Score (RAscore) – Rapid Machine Learned Synthesizability Classification from AI Driven Retrosynthetic Planning. Chem. Sci. 2021, 12 (9), 3339–3349. https://doi.org/10.1039/D0SC05401A.

Tiled — tiled 0.1.0a87 documentation. https://blueskyproject.io/tiled/ (accessed 2023-03-20).

Tom, G.; Schmid, S. P.; Baird, S. G.; Cao, Y.; Darvish, K.; Hao, H.; Lo, S.; Pablo-García, S.; Rajaonson, E. M.; Skreta, M.; Yoshikawa, N.; Corapi, S.; Akkoc, G. D.; Strieth-Kalthoff, F.; Seifrid, M.; Aspuru-Guzik, A. Self-Driving Laboratories for Chemistry and Materials Science. Chem. Rev. 2024, 124 (16), 9633–9732. https://doi.org/10.1021/acs.chemrev.4c00055.

Tomaszewski, P.; Yu, S.; Borg, M.; Rönnols, J. Machine Learning-Assisted Analysis of Small Angle X-ray Scattering. In 2021 Swedish Workshop on Data Science (SweDS); 2021; pp 1–6. https://doi.org/10.1109/SweDS53855.2021.9638297.

Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. I. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. ACS Catal. 2020, 10 (3), 2260–2297. https://doi.org/10.1021/acscatal.9b04186.

Usnistgov/AFL-Automation, 2025. https://github.com/usnistgov/AFL-automation (accessed 2025-05-13).

Usnistgov/AFL-Hardware, 2024. https://github.com/usnistgov/AFL-hardware (accessed 2025-05-13).

Vaddi, K. Kiranvad/Amplitude-Phase-Distance, 2025. https://github.com/kiranvad/Amplitude-Phase-Distance (accessed 2025-06-05).

Vaddi, K.; Chiang, H. T.; Pozzo, L. D. Autonomous Retrosynthesis of Gold Nanoparticles via Spectral Shape Matching. Digit. Discov. 2022, 1 (4), 502–510.

Vallet-Regi, M.; Rámila, A.; del Real, R. P.; Pérez-Pariente, J. A New Property of MCM-41: Drug Delivery System. Chem. Mater. 2001, 13 (2), 308–311. https://doi.org/10.1021/cm0011559.

Vallet-Regí, M.; Schüth, F.; Lozano, D.; Colilla, M.; Manzano, M. Engineering Mesoporous Silica Nanoparticles for Drug Delivery: Where Are We after Two Decades? Chem. Soc. Rev. 2022, 51 (13), 5365–5451. https://doi.org/10.1039/D1CS00659B.

Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. J. Mach. Learn. Res. 2008, 9 (11).

Vasquez, S.; Twigg-Smith, H.; Tran O'Leary, J.; Peek, N. Jubilee: An Extensible Machine for Multi-Tool Fabrication. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; CHI '20; Association for Computing Machinery: New York, NY, USA, 2020; pp 1–13. https://doi.org/10.1145/3313831.3376425.

Venugopal, V.; Sahoo, S.; Zaki, M.; Agarwal, M.; Gosvami, N. N.; Krishnan, N. M. A. Looking through Glass: Knowledge Discovery from Materials Science Literature Using Natural Language Processing. Patterns 2021, 2 (7), 100290. https://doi.org/10.1016/j.patter.2021.100290.

Verslycke, T.; Reid, K.; Bowers, T.; Thakali, S.; Lewis, A.; Sanders, J.; Tuck, D. The Chemistry Scoring Index (CSI): A Hazard-Based Scoring and Ranking Tool for Chemicals and Products Used in the Oil and Gas Industry. Sustainability 2014, 6 (7), 3993–4009. https://doi.org/10.3390/su6073993.

Vescovi, R.; Ginsburg, T.; Hippe, K.; Ozgulbas, D.; Stone, C.; Stroka, A.; Butler, R.; Blaiszik, B.; Brettin, T.; Chard, K.; Hereld, M.; Ramanathan, A.; Stevens, R.; Vriza, A.; Xu, J.; Zhang, Q.; Foster, I. Towards a Modular Architecture for Science Factories. Digit. Discov. 2023, 2 (6), 1980–1998. https://doi.org/10.1039/D3DD00142C.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat. Methods 2020, 17 (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Volk, A. A.; Abolhasani, M. Performance Metrics to Unleash the Power of Self-Driving Labs in Chemistry and Materials Science. Nat. Commun. 2024, 15 (1), 1378. https://doi.org/10.1038/s41467-024-45569-5.

Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian Estimation of Synthetic Accessibility of Organic Compounds. J. Cheminformatics 2020, 12 (1), 35. https://doi.org/10.1186/s13321-020-00439-2.

Walsh, D.; Zou, W.; Schneider, L.; Mello, R.; Deagen, M.; Mysona, J.; Lin, T.-S.; Pablo, J. de; Jensen, K.; Audus, D.; Olsen, B. CRIPT: A Scalable Polymer Material Data Structure. 2022. https://doi.org/10.26434/chemrxiv-2022-xpz37.

Wang, D.; Chen, X.; Feng, J.; Sun, M. Recent Advances of Ordered Mesoporous Silica Materials for Solid-Phase Extraction. J. Chromatogr. A 2022, 1675, 463157. https://doi.org/10.1016/j.chroma.2022.463157.

Wang, X.; Xuan, X.; Wang, Y.; Li, X.; Huang, H.; Zhang, X.; Du, X. Nano-Au-Modified TiO2 Grown on Dendritic Porous Silica Particles for Enhanced CO2 Photoreduction. Microporous Mesoporous Mater. 2021, 310, 110635. https://doi.org/10.1016/j.micromeso.2020.110635.

Wang, X.; Zhu, L.; Zhuo, Y.; Zhu, Y.; Wang, S. Enhancement of CO2 Methanation over La-Modified Ni/SBA-15 Catalysts Prepared by Different Doping Methods. ACS Sustain. Chem. Eng. 2019, 7 (17), 14647–14660. https://doi.org/10.1021/acssuschemeng.9b02563.

What is the Z-average? https://www.malvernpanalytical.com/en/learn/knowledge-center/faqs/faq0015averagediameter (accessed 2025-05-14).

Wierenga, R. P.; Golas, S. M.; Ho, W.; Coley, C. W.; Esvelt, K. M. PyLabRobot: An Open-Source, Hardware-Agnostic Interface for Liquid-Handling Robots and Accessories. Device 2023, 1 (4). https://doi.org/10.1016/j.device.2023.100111.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der

Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci. Data 2016, 3 (1), 160018. https://doi.org/10.1038/sdata.2016.18.

Williams, C. K.; Rasmussen, C. E. Gaussian Processes for Machine Learning; MIT press Cambridge, MA, 2006; Vol. 2.

Winther, K. T.; Hoffmann, M. J.; Boes, J. R.; Mamun, O.; Bajdich, M.; Bligaard, T. Catalysis-Hub.Org, an Open Electronic Structure Database for Surface Reactions. Sci. Data 2019, 6 (1), 1–10. https://doi.org/10.1038/s41597-019-0081-y.

Xiong, L.; Du, X.; Kleitz, F.; Qiao, S. Z. Cancer-Cell-Specific Nuclear-Targeted Drug Delivery by Dual-Ligand-Modified Mesoporous Silica Nanoparticles. Small Weinh. Bergstr. Ger. 2015, 11 (44), 5919–5926. https://doi.org/10.1002/smll.201501056.

Yan, R.; Jiang, X.; Wang, W.; Dang, D.; Su, Y. Materials Information Extraction via Automatically Generated Corpus. Sci. Data 2022, 9 (1), 401. https://doi.org/10.1038/s41597-022-01492-2.

Ye, Z.; Wu, Z.; Jayaraman, A. Computational Reverse Engineering Analysis for Scattering Experiments (CREASE) on Vesicles Assembled from Amphiphilic Macromolecular Solutions. JACS Au 2021, 1 (11), 1925–1936. https://doi.org/10.1021/jacsau.1c00305.

Yoshikawa, N.; Darvish, K.; Vakili, M. G.; Garg, A.; Aspuru-Guzik, A. Digital Pipette: Open Hardware for Liquid Transfer in Self-Driving Laboratories. Digit. Discov. 2023, 2 (6), 1745–1751. https://doi.org/10.1039/D3DD00115F.

Yu, X.; T. Williams, C. Recent Advances in the Applications of Mesoporous Silica in Heterogeneous Catalysis. Catal. Sci. Technol. 2022, 12 (19), 5765–5794. https://doi.org/10.1039/D2CY00001F.

Zaki, M.; Prinz, C.; Ruehle, B. A Self-Driving Lab for Nano- and Advanced Materials Synthesis. ACS Nano 2025, 19 (9), 9029–9041. https://doi.org/10.1021/acsnano.4c17504.

Zakutayev, A.; Perkins, J.; Schwarting, M.; White, R.; Munch, K.; Tumas, W.; Wunder, N.; Phillips, C. High Throughput Experimental Materials Database, 2017, 2 files. https://doi.org/10.7799/1407128.

Zeitler, S. M.; Chakma, P.; Golder, M. R. Diaryliodonium Salts Facilitate Metal-Free Mechanoredox Free Radical Polymerizations. Chem. Sci. 2022, 13 (14), 4131–4138.

Zenodo - Research. Shared. https://zenodo.org/ (accessed 2023-02-02).

Zhao, D.; Feng, J.; Huo, Q.; Melosh, N.; Fredrickson, G. H.; Chmelka, B. F.; Stucky, G. D. Triblock Copolymer Syntheses of Mesoporous Silica with Periodic 50 to 300 Angstrom Pores. Science 1998, 279 (5350), 548–552. https://doi.org/10.1126/science.279.5350.548.

🔬🧪 Science Jubilee ⚡⚙️ — Science Jubilee 0.3.2.post1.dev170+g0ee6bd1 documentation. https://science-jubilee.readthedocs.io/en/latest/index.html (accessed 2025-06-02).