ABSTRACT

Title of Thesis: ACCELERATING MATERIAL DISCOVERY AND DESIGN THROUGH INTEGRATED ROBOTICS AND MACHINE INTELLIGENCE

Hayden C. Whitley, Master of Science, 2025

Thesis Directed By: Assistant Professor Po-Yen Chen, Department of Chemical and Biomolecular Engineering

Conventional material discovery and design is labor-, time-, and cost-intensive. To address these challenges, this work presents a robotics- and machine learning-integrated workflow designed to accelerate the material design process, maximize experimental throughput, minimize time and labor expenses, and lead to improved outcomes. This multi-stage framework replaces labor-intensive wet lab experiments with collaborative robots that can autonomously perform sample fabrication and testing, utilizes a feasibility constrained design space to minimize experimental failures, replaces inefficient trial-and-error cycles with active learning loops to efficiently navigate complex design spaces, and incorporates virtual synthesis and screening to effectively manage multi-objective optimization tasks, enabling tunable design requirements. This framework is enabled by the development and integration of autonomous robotic platforms for high-throughput sample preparation and characterization. The efficacy of this integrated approach is demonstrated through two distinct projects titled the predictive and generative modeling of mixed-dimensional aerogels with programmable properties, and the

predictive design of sustainable biobased packaging for improved postharvest preparation. This research highlights the transformative potential of combining robotics and machine intelligence to significantly accelerate the discovery and design of advanced materials with tailored property requirements.

ACCELERATING MATERIAL DISCOVERY AND DESIGN THROUGH INTEGRATED ROBOTICS AND MACHINE INTELLIGENCE

by

Hayden C. Whitley

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2025

Advisory Committee:
Assistant Professor Po-Yen Chen, Chair
Professor Peter Kofinas
Professor Shenqiang Ren

Dedication

It truly takes a village. To my family for supporting me for the past 23 years. To Dr. Po-Yen Chen for giving me an opportunity, and your unwavering support. To Tianle for your camaraderie, and willingness to teach. To all my lab mates for welcoming me into your family. To Maria for listening to my rambling and pushing me to be better every day. Thank you all.

Acknowledgements

All the work presented here is a result of the cumulative efforts from all the members of the PYC Lab. It is my opinion that the collaborative, uplifting, curious culture we have cultivated is the foundation for the great work we have accomplished, and so all of you deserve to be recognized. Thank you, Li Yang, Josh, Tianle, Haochen, Snehi, Peter, Tram, Noah, and Kameron for your unwavering support, guidance, and friendship. I enjoyed the last three years spent with you all. Every one of you has contributed to the work in this thesis, and I am incredibly grateful.

Tianle Chen, whose mentorship has been a transformative experience for me over the past four years. He has taught me more than I have learned in any classroom, and I am honored to have had the opportunity to work alongside him.

Finally, I would like to recognize Dr. Po-Yen Chen, who is the brain trust behind all these projects. His belief and trust in me fostered my curiosity, ambition, and confidence and has shaped the person and professional I am today. None of this work would have been possible without his leadership.

Table of Contents

List of Tables

List of Figures

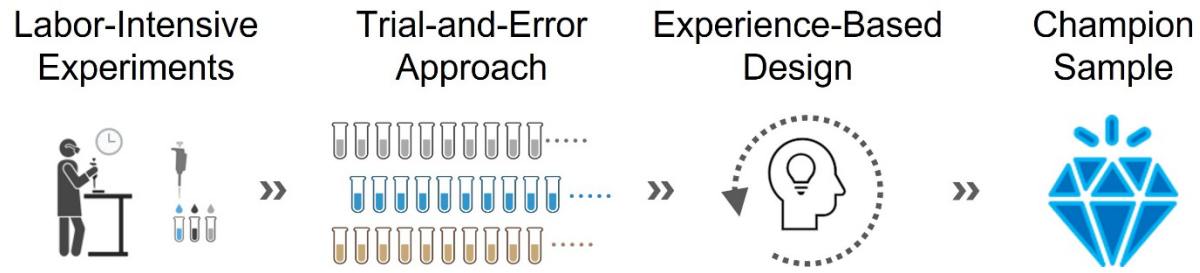
List of Equations

Chapter 1: Introduction

1.1 Conventional and machine intelligence workflow comparison

Materials discovery and design tasks are notoriously labor-, time-, and cost-intensive[1-3]. The objective is to identify the component recipe and fabrication parameters of a champion formulation, with optimal performance across targeted property requirements[4,5]. Traditional approaches (**Fig. 1**) involve labor-intensive, time-consuming wet lab experiments for sample preparation and characterization[2,3], use continuous trial-and-error cycles to navigate complex design spaces that are prone to experimental failures[6,7], and rely on experience-based design principles to perform multi-objective optimization, which does not allow for tunable property targets[7]. This conventional workflow is inefficient, resource-intensive, and often yields suboptimal results, especially when exploring large parameter spaces[8].

Figure 1. Conventional materials discovery and design workflow

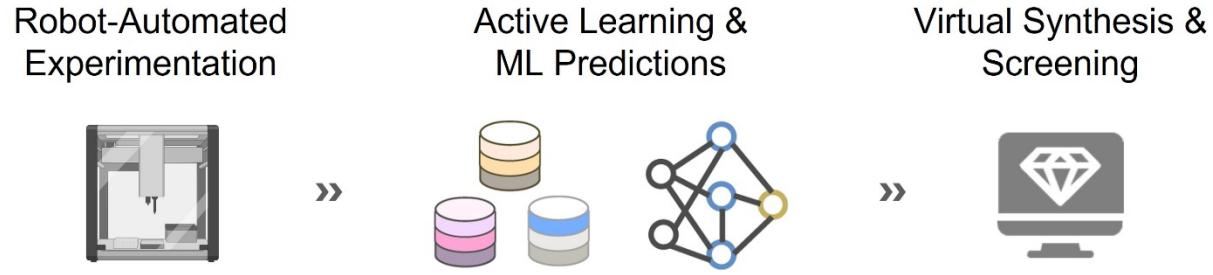


Traditional approaches to materials discovery and design tasks involve labor-intensive wet lab experiments, inefficient trial-and-error cycles, and rely on experience-based design principles to identify optimal sample formulations.

To address these challenges, we propose a robotics- and machine learning-integrated (machine intelligence) workflow (**Fig. 2**) that accelerates the design process, reduces labor costs, minimizes experimental failures, and improves experimental outcomes[8]. Rather than dedicating

human capital to labor-intensive experiments, collaborative robots autonomously perform sample fabrication and testing[9,10]. Instead of traditional trial-and-error cycles, active learning loops efficiently sample from a feasibility constrained design space, maximizing the information gain from each successive experiment, while minimizing experimental failures[5]. Finally, instead of relying on experience-based design, machine learning-enabled prediction models ensure accurate property predictions for unsampled data points that can be experimentally validated[1,3,6,8]. This enables the virtual synthesis of formulations in silico, that can be screened to effectively perform multi-objective optimization on tunable property requirements[3,11].

Figure 2. Machine intelligence materials discovery and design workflow



A robotics- and machine learning-integrated workflow automates sample preparation and characterization tasks using collaborative robots, uses machine learning-enabled prediction models and active learning loops to efficiently navigate complex design spaces, and enables virtual synthesis and screening to perform multi-objective optimization, and enable tunable design requirements.

Overall, this machine-driven approach to experimentation maximizes experimental throughput, minimizes time and labor expenses, and leads to improved outcomes compared to conventional workflows.

1.2 Multistage framework

Several challenges hinder the application of artificial intelligence and machine learning in materials discovery and design[3]. A primary issue stems from the inherent nature of experimental

design spaces: they are high-dimensional systems encompassing numerous building blocks, fabrication parameters, structural features, and chemical features[12,13]. Consequently, machine learning models require large volumes of high-quality experimental data to navigate this complexity effectively[14]. The multi-component nature of material design tasks stretches the capabilities of conventional simulation tools for generating synthetic, simulation-derived data at scale[13,15] and collecting experimental data is both time-consuming and expensive, making it difficult to construct sufficiently large, structured databases for training accurate predictive models[16,17].

Scraping literature data is a plausible approach to compiling large, structured datasets, however, several limitations exist. Literature data is often unreliable because different labs use distinct building blocks, material systems, and fabrication protocols when preparing samples[18-20]. Furthermore, publications typically only report the composition and characterizations of champion data and recipes, limiting the availability of failure data, which is crucial for training models to avoid experimental failures[5,13,21].

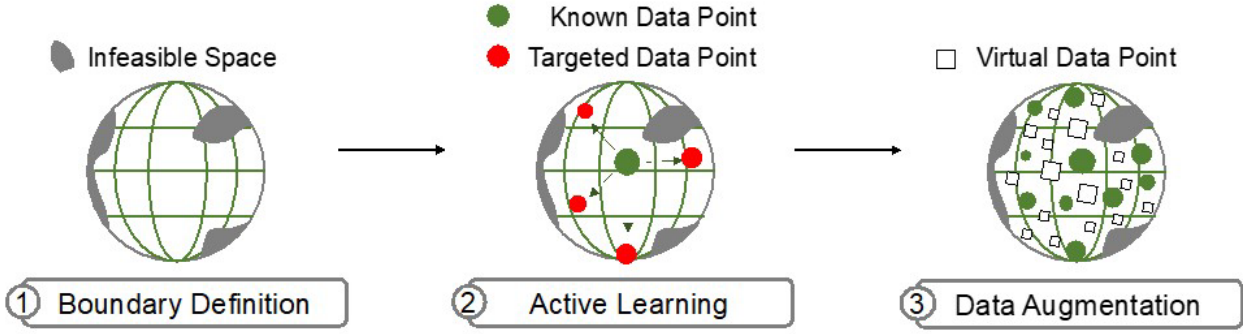
Data-scarcity is therefore an issue in material discovery and design[22]. A common challenge associated with deploying machine learning models in limited-data regimes is overfitting (**Fig 3.**), where the model's predictions conform to known data points, hindering its ability to generalize to unseen data or extrapolate beyond known regions of the design space[15,23,24].

Figure 3. Model fitting in machine learning: underfitting, overfitting, tradeoff

Underfitting occurs when a model is too simple to capture meaningful patterns in the data. Overfitting happens when a model is too complex and learns noise rather than general trends. A balanced model finds an optimal middle ground, capturing underlying patterns while maintaining generalizability.

To overcome these barriers, we propose a multi-stage framework (**Fig. 4**) that mitigates many of these issues and promotes the successful utilization and deployment of machine learning for materials discovery and design tasks[25]. Throughout these stages, we emphasize experimental throughput, producing as many samples as possible while minimizing the labor and time expenses associated with their fabrication.

Figure 4. Multi-stage framework for design space exploration[26]



A three-stage process is used for the exploration and optimization of a design space using machine learning. (1) Boundary definition establishes a feasible design boundary within the broader design space constraining the useful subregion. (2) Active learning loops iteratively select the most informative data points to synthesize structured databases. (3) Prediction models are trained using synthetic, augmented data to generalize across the design space for accurate property predictions.

First, we define a feasible design boundary. This is achieved by training a classification model (classifier) to pinpoint feasible subregions of the experimental design space. The conditions for feasibility testing vary by task, however, feasibility labeling is commonly much less time-consuming compared to other characterization tasks. During this stage, data points are initially sampled uniformly across the entire design space, after which uncertain regions are explored to minimize model uncertainty. A classifier can then be trained on this data and employed to predict formulation feasibility for further experimentation, significantly reducing experimental failures[27-29]. This process results in a continuous feasible design boundary within the broader experimental design space, confining subsequent experimentation to a much smaller subregion[30,31].

Next, we implement active learning loops to efficiently and strategically sample the most informative unlabeled formulations within the feasible design space to undergo characterization[32]. The goal is to continuously improve a navigator model's ability to accurately map formulations to targeted property labels across the entire feasible design space[33,34]. To achieve this, utility functions guide the selection process by balancing exploration, expanding knowledge of the design space, and exploitation, focusing on the most promising subregions[5]. The utility functions can be adjusted to focus on distinct property labels, enabling downstream multi-objective optimization. As a result of active learning, a well-structured database is synthesized, laying the foundation for training robust predictive models.

Finally, predictive models are trained to accurately map formulations to property predictions across the entire design space[35]. Data augmentation techniques are employed to generate synthetic data, leveraging variance in experimental measurements to create virtual data points[36,37]. This helps reduce overfitting during model training by injecting noise and introducing

uncertainty, forcing the model to learn general trends rather than memorizing direct mappings[38]. Additionally, ensemble modeling techniques are used to enhance the predictive model's ability to generalize across the dataset and reduce overfitting by averaging outputs from multiple models, each with a unique understanding of the dataset[39].

1.2.1 Discovering a feasible design boundary

Defining a feasible design boundary for material discovery and design tasks is a critical step that serves to narrow further experimentation to exclusively focus on viable formulations, reducing experimental effort and accelerating the design and property optimization process[27,28].

Unlike data-rich systems that can tolerate a higher rate of experimental failure, sample fabrication in material discovery and design experiments is often time-consuming and labor-intensive, especially when most formulations result in experimental failures (infeasible results)[29]. To address this challenge, classification models—commonly SVM classifiers or artificial neural network (ANN) classifiers—are implemented to filter out infeasible formulations before experimental validation[40,41]. This serves as a critical screening layer throughout the experimental workflow, constraining the design space and ensuring that only feasible formulations with a high likelihood of producing high-quality samples are recommended[30,31]. As a result, learning efficiency is improved, and experimental failure is reduced.

These classifiers are trained using experimental data, which label formulations as either 'feasible' or 'infeasible.' The specific labels vary depending on the design task. For example, when categorizing aerogel samples, we use a cutoff based on the retained volume of the sample after freeze-drying. For nanocomposite films, samples can be judged based on their detachability and surface uniformity[42]. Regardless of the domain, in our experience, these labels are significantly cheaper and less time-consuming to obtain than other characterization tasks.

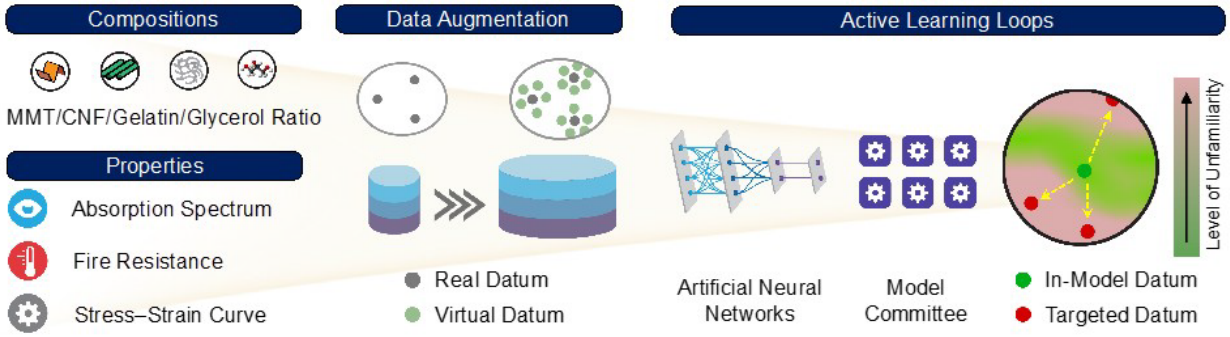
1.2.2 Active learning loops to navigate the design space

The active learning phase is a systematic and efficient approach to exploring the complex design space in material design tasks, enabling the discovery of high-performance formulations with desired properties[5].

Sample preparation and characterization is time- and labor-intensive, often limiting the number of training data points that can be collected for constructing accurate prediction models[16,17]. These challenges compounded by the vast, high-dimensional design space, which encompasses an immense number of possible formulations, make exhaustive experimental exploration impractical[12,13].

To overcome these obstacles, active learning loops (**Fig. 5**) are employed to strategically sample and explore the design space, optimizing resource allocation while improving model predictions[33]. Active learning is particularly effective in materials discovery and design tasks, as it enables the model to identify the most informative and uncertain regions for experimental validation[32]. This iterative approach enhances efficiency by prioritizing high-value experiments and minimizing redundant or less informative trials, thereby accelerating the discovery of optimal formulations[43].

Figure 5. Active learning loops for the efficient construction of structured databases[42]



The active learning workflow enables the efficient development of structured databases to train prediction models with the ability to accurately predict property labels from formulation

compositions. Data augmentation techniques augment the experimental dataset with virtually synthesized data points. Predictions are inferred by an ANN committee, which enables quantification of metrics such as uncertainty/confidence in predictions, guiding further experimentation.

During the active learning process, an ANN model is developed as the "navigator," optimizing key objectives such as maximizing specific property labels or meeting targeted property requirements[34]. To achieve these objectives, acquisition functions are defined to compute the information gain of each unsampled formulation, guiding the selection of components for subsequent experiments[5]. The most informative data points are then selected in batches for experimental validation to improve experimental throughput[32].

The experimental workflow is as follows: formulations suggested by the navigator model are prepared, and the resulting samples undergo characterization to determine their property labels. To address data scarcity and reduce the risk of overfitting by the navigator model, data augmentation techniques such as the User Input Principle (UIP) are employed[36,37]. UIP is based on the observation that the properties of specific formulations remain approximately constant when there are slight variations in their composition. Virtual data points are synthesized based on existing data, with added Gaussian noise to simulate real-world variability[44]. The experimental and virtual data points are then combined to retrain the navigator model, improving its accuracy after each round of active learning.

This process iteratively constructs a structured database of characterized formulations, which is essential for building accurate prediction models[35]. Prediction models can then be used to sample a vast number of formulations and map their predicted properties, providing a comprehensive understanding of performance trends across the design space, along with any synergistic interactions between components[8]. This approach enables the efficient identification of novel formulations with desired properties.

1.2.3 Data augmentation and ensemble modeling

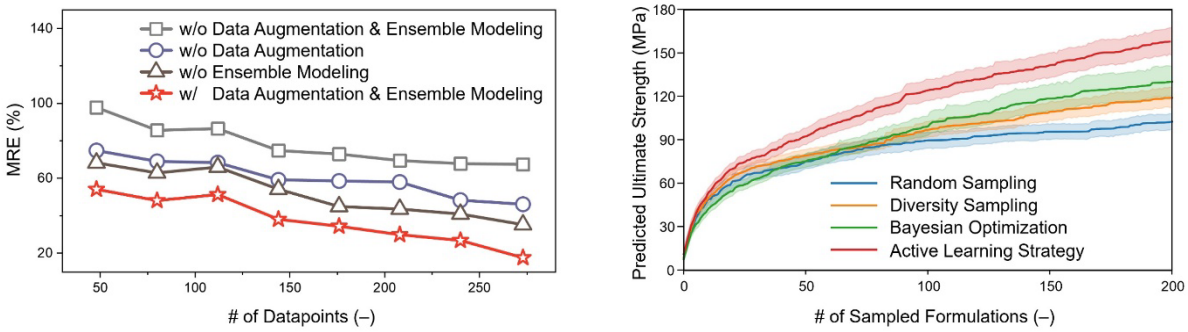
Prediction models enable accurate forecasting of material properties based on their formulations[6]. Artificial neural networks (ANNs) are particularly effective due to their ability to capture complex, non-linear relationships[45]. The goal of the model is to accurately predict property labels from sample formulations. The model is designed to, with sufficient data, eliminate the need for further experimentation by enabling the inverse design of materials with specific property requirements, which can then be experimentally validated[35]. Data augmentation and ensemble modeling techniques are employed to improve both the model's performance on experimental datasets and its generalizability[36,37].

The prediction model is trained using experimental data, systematically collected through active learning loops[5]. This data is used to build a structured database, which serves as the foundation for model development. Data augmentation complements active learning by addressing data scarcity and reducing the risk of model overfitting, particularly in small or incomplete datasets[36,37]. Techniques such as the User Input Principle (UIP) method generate virtual data points that emulate realistic variations observed in wet-lab experiments[44]. By integrating both real and synthetic data, the prediction model enhances its ability to generalize formulation-property correlations, improving prediction accuracy[38,46].

Ensemble modeling further enhances prediction performance by integrating multiple model variants into a single committee. One key advantage is its ability to mitigate and improve generalizability, yielding more reliable results across diverse formulations[39]. The prediction variance among model variants serves as an intrinsic measure of uncertainty, enabling a quantitative confidence assessment of each prediction[47].

The prediction model's performance is validated using independent testing datasets (**Fig. 6**). This validation is quantified by the mean relative error (MRE) between predicted and experimentally measured properties, where lower MRE values indicate higher prediction accuracy.

Figure 6. Performance evaluation of predictive material property modeling techniques and sampling strategies



(left) MRE values for different prediction models with and/or without data augmentation and ensemble modeling (evaluated using an independent set of testing data points). The red line (with data augmentation and ensemble modeling) shows significantly lower MRE compared to other approaches, indicating improved prediction accuracy. (right) Comparison between active learning strategy with multiple sampling methods, including random sampling, diversity sampling, and Bayesian optimization.

1.2.4 Model interpretation tools

Machine learning models, especially artificial neural networks (ANNs), are often treated as "black box" systems[48]. This opacity challenges their trustworthiness and reliability, as we lack a comprehensive understanding of how predictions are generated and how specific compositions impact the observed properties of synthesized formulations[39]. Consequently, there is a critical need for analytical tools that can (1) provide data-driven insights into model behavior, (2) enhance the interpretability of machine learning predictions, (3) illuminate complex interactions between component materials, and (4) increase overall model transparency and reliability[44,49].

These tools enhance human comprehension, allowing researchers to develop more trustworthy and explainable prediction models[50].

To uncover the complex composition-property correlations and improve the model's interpretability, SHapley Additive exPlanantions (SHAP) model interpretations can be employed over the experimental data collected during active learning loops. SHAP is a game theoretical approach to explain the output of any AI/ML model (including ensemble models)[51]. SHAP values can be computed for each datapoint in the structured dataset by iterating over complete permutations of its features. The SHAP values are calculated to approximate the contribution of each component loading to a specific property. It can explain the feature importance inside the ML models, thus enabling users to interpret the models.

1.2.5 Current challenges

There exist several challenges and limitations associated with the broad adoption of machine learning-integrated workflows for accelerating materials discovery and design tasks: (1) the lack of comprehensive, high-quality experimental datasets and (2) inefficient dissemination mechanisms that hinder collaboration among diverse stakeholders.

As previously mentioned, machine learning models require vast amounts of high-quality experimental data to function effectively. However, materials design and discovery research are not inherently data-rich, restricting the application of these technologies to highly specialized and constrained design spaces. When more building blocks, structural, and chemical features are included, the time and labor needed for constructing accurate models will be inflated.

To mitigate these growing resource demands, collaborative robotic systems can automate entire preparation and characterization processes. However, bottlenecks remain. In our experience, manual operators are still required to connect each stage of sample preparation and

characterization. While this step is the least labor-intensive, it is also the most expensive to automate.

Developing and adopting open-access, high-quality experimental datasets would be a significant advancement for the field of materials science. However, challenges persist. The quality of source components and fabrication procedures can vary between batches and across laboratories. Therefore, enforcing stringent quality controls and standardized operating procedures is essential, though not trivial.

Overcoming these challenges is critical for the successful integration of AI/ML in materials discovery and design. Addressing these issues through collaborative efforts, open-access datasets, and advanced automation will accelerate innovation and unlock new possibilities across multiple technological fields.

Chapter 2: Robotic platforms

Robotic platforms for high-throughput sample preparation and characterization enable the autonomous, unsupervised performance of otherwise labor-intensive tasks. These platforms enable precise, reliable, and reproducible outcomes by reducing any human-error in these processes. This allows researchers to focus on more abstract and high-level tasks such as design of experiments, data analysis, and model interpretation.

2.1 OT-2 robot with SID tracking and debugging interface

The OT-2 Robot is a commercially available liquid handler designed to automate the preparation of material mixtures, significantly reducing manual labor while increasing efficiency and consistency. It can be programmed to prepare aqueous mixtures in batches by precisely transferring and mixing source solutions into destination wells, enhancing experimental throughput and reproducibility in material preparation tasks.

2.1.1 OT-2 protocol generator

The OT-2 robot operates by following a sequence of commands which encompasses moving to various locations on its deck, aspirating, dispensing, picking up pipette tips, and disposing of pipette tips. This software is open source and documented to enable researchers to programmatically curate tasks. Therefore, batches are formulated, a procedure is generated, and a file is uploaded to the OT-2 which autonomously prepares mixtures with high precision, improving the reliability and reproducibility of experiments.

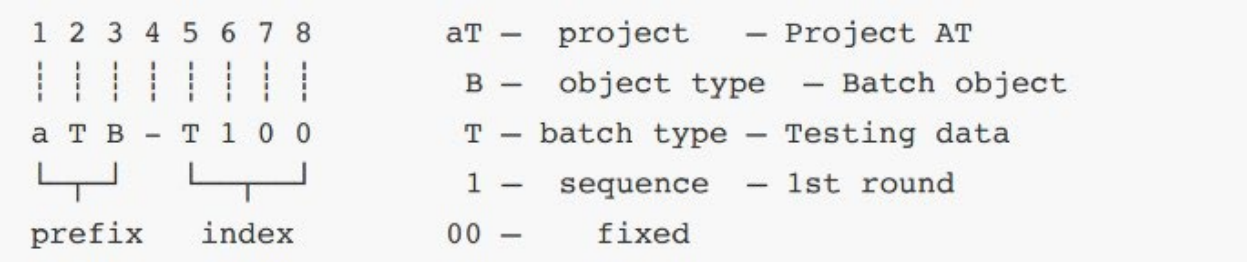
2.1.2 SID labeling system

Our novel SID labeling system generates a unique label for each sample fabricated by the OT-2. These labels are both human-readable – to maintain lab safety standards – as well as computer readable, by including a Data Matrix string data can be encoded in a computer-readable format.

The SID (String ID) is a unique string-based identification system developed for research projects. It aims to provide a standardized and efficient way to categorize and sort various objects involved in a research project, ensuring easy tracking and management. The system is designed to be user-friendly and machine-readable, making it easy for both researchers and software to understand and use the identifiers.

In the SID system, each ID has a fixed length of eight characters, divided into a "prefix" and an "index" by a hyphen (**Fig. 7**). The prefix indicates the type and scope of the object, while the index provides more specific details about the object, such as its creation date, sequence number, and associated batch number.

Figure 7. Example batch object SID



Prefixes and indices can be manipulated to represent different common lab objects and communicate information about them effectively. In this example the SID aTB-T100 represents a collection of testing samples prepared for Project AT. Another example would be the SID MMT-2801 which represents a source object, namely the first batch of the material MMT prepared during August in 2022.

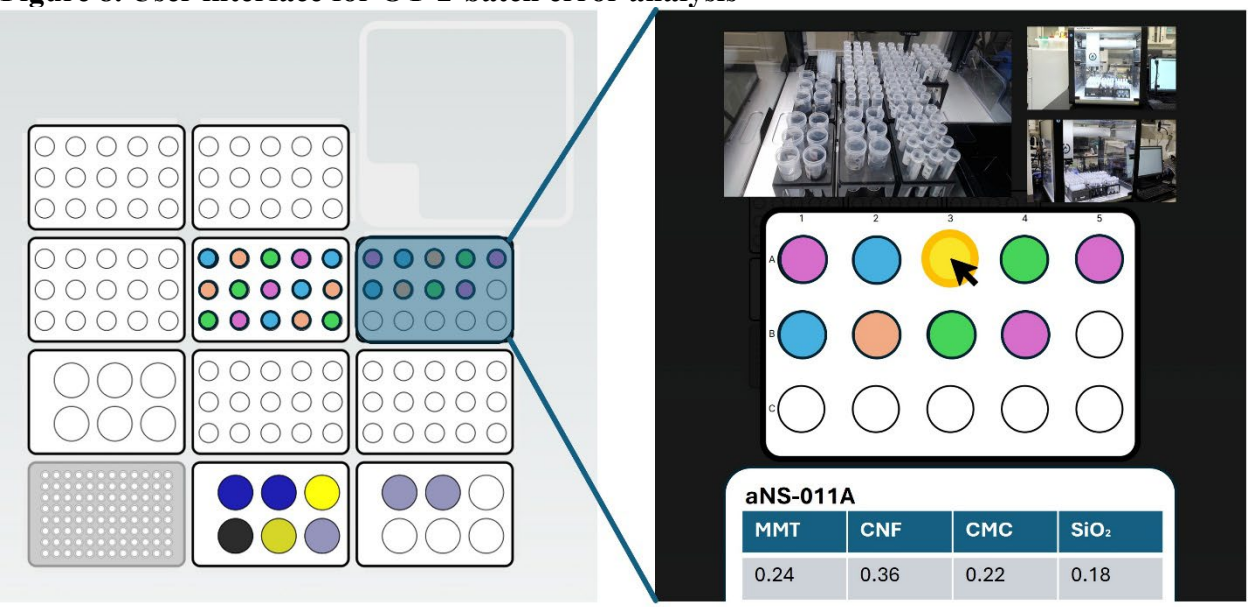
The system is flexible and adaptable, allowing for the use of project-specific codes and type codes. It was initially created for a bioplastic substitute project and has since been integrated as a core component of the OT2 Protocol Generator.

Overall, the SID system contributes to more organized and streamlined research processes, assisting in data management and ensuring accurate tracking of research projects.

2.1.3 OT-2 debugging interface

Inevitably, batches or individual samples fail. For example, a test tube may contain less solution than expected, often due to the OT-2 robot struggling to pipette highly viscous liquids. However, other failure modes also exist. While the entire OT-2 pipetting procedure is recorded, runs often exceed six hours, making manual video analysis impractical. To address this, we developed a diagnostic tool that synchronizes the timestamped OT-2 run protocol with the video recording (**Fig. 8**). This tool extracts snippets of footage showing the robot interacting with specific samples, allowing for targeted visual inspection. By identifying potential issues, it helps prevent future experimental failures.

Figure 8. User interface for OT-2-batch error analysis



This user interface provides researchers with the ability to perform targeted visual inspection on the root causes of error after experimental failures. Only video segments showing the OT-2 robot interacting with the sample of interest are presented to the user.

2.2 Automated compression testing platform

Compression testing of mixed-dimensional aerogels involves characterizing the stress–strain curves of samples (with average dimensions of 1.5 × 1.5 × 1.0 cm^3) which are measured by a mechanical testing machine (Instron 68SC-05) with a 500 N load cell for 5 cycles of loading and unloading (at 5 mm s–1 for 80% compression).

The Automated Compression Testing Platform integrates an UR5e robotic arm equipped with a Hand-E gripper to an Instron 6800 Series universal testing machine outfitted with the compression testing module to autonomously conduct mechanical characterizations, reducing the workload on human operators and enhancing data acquisition rates.

The design and implementation of this system has undergone two iterations during my time with the PYC Lab[52].

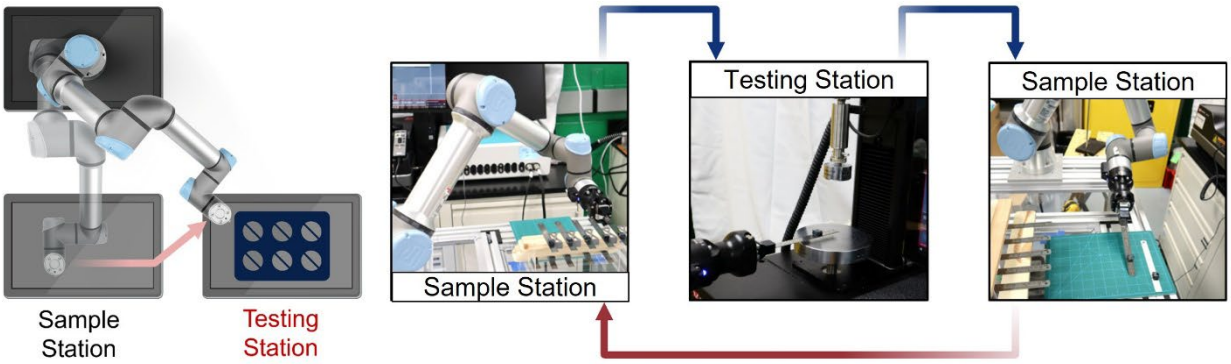
2.2.1 Autonomous batch compression testing of conductive aerogels

This project marked the inception of the Automated Compression Testing Platform, designed to perform unsupervised mechanical characterization of aerogel samples. Synchronizing the operations of a UR5e robotic arm (for sample handling) and an Instron tester (for compression testing) posed a key challenge, as the two devices lacked a shared interface and operated on different time scales; while the UR5e executed actions in fixed intervals, the Instron tester required a variable duration to complete each test.

To address this, we programmed the Instron tester to pause for a fixed interval between tests, while the UR5e awaited a signal indicating test completion before unloading the sample and loading a new one. However, direct communication was not possible since the Instron tester lacked an I/O board and the UR5e control box operated at 24V. Instead, we programmed the Instron to play an audio file at the end of each test, which was converted into a 5V analog signal. An Arduino then amplified this signal to 24V, enabling the UR5e to register it as a test completion cue.

Upon receiving the signal, the UR5e removed the tested sample, disposed of it, retrieved the next sample from the sample station, and loaded it onto the Instron tester before awaiting the next signal. The sample station featured a stage with ten slots, each holding a sample affixed to a ruler to ensure mechanical integrity during handling. The UR5e sequentially retrieved samples from these slots, conducting tests in batches of ten before requiring manual reloading for the next batch.

Figure 9. Batch autonomous compression testing by integrating an UR5e robotic arm with an Instron compression tester[52]



The UR5e robotic arm was programmed to transfer aerogel samples continuously from the sample station to the testing station. Once the UR5e arm completed placing the aerogel at the testing station, the Instron tester began the compression test. After the test was finished, the Instron tester prompted the UR5e arm, using an audio signal, to remove the aerogel, discard it, and then position a new one. More than 400 aerogels were evaluated using the autonomous testing platform during this project with a total operation time estimated to be 81 hours, averaging about 12 minutes to test one aerogel sample.

2.2.2 Autonomous compression testing of mixed-dimensional aerogels

The second iteration of the Automated Compression Testing Platform enhanced the autonomy and reliability of the platform by incorporating feedback loops, replacing the previously open-loop operation. This improvement enabled continuous operation, allowing human operators to load samples on the fly and dynamically upload sample data via SID labeling.

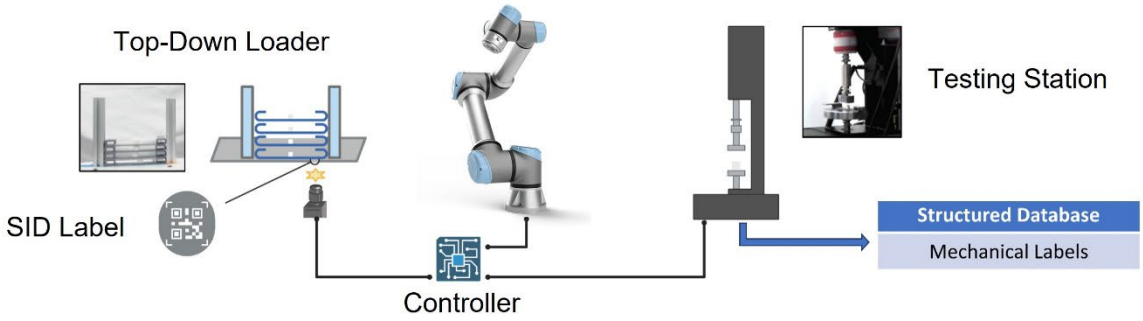
Several key features contributed to the platform's improved functionality. The sample station was redesigned with a top-down loading mechanism, ensuring each sample was consistently picked up from the same position. Samples were placed on custom holders with tabs for the UR5e robotic arm to grip, and these holders were stacked within the sample station. Labels, including data matrices encoding the sample SID, were affixed to the holders and read by a camera positioned directly beneath the station. This enabled real-time labeling, automatic data

upload from the Instron tester, and detection of sample readiness at the beginning of each test. Additionally, an IO board was integrated with the Instron tester to enable bidirectional communication, and so that a signal began the test routine upon sample loading.

Testing operated in a coordinated loop involving multiple systems. The desktop-connected camera first captured an image to determine if a primed sample was present, identified by the presence of a data matrix. If detected, the SID was decoded, stored, and a serial signal was sent to an Arduino, which activated a relay to trigger the UR5e robotic arm. The arm then retrieved the sample and loaded it into the Instron tester. A 5V signal from the Arduino confirmed successful loading, prompting the test routine to begin. Upon test completion, the Instron tester sent a signal back to the Arduino, which then instructed the UR5e to remove and discard the sample. The desktop computer retrieved the stress-strain curve via Instron's BlueHill API and stored it into a database using the corresponding SID.

The UR5e then returned to its standby position, while the Arduino signaled the computer to check for the next sample. If no sample was detected, the system periodically rechecked for up to 30 minutes, allowing operators time to load new samples.

Figure 10. Fully autonomous compression testing with integrated Instron compression tester, UR5e robotic arm, and remote data upload



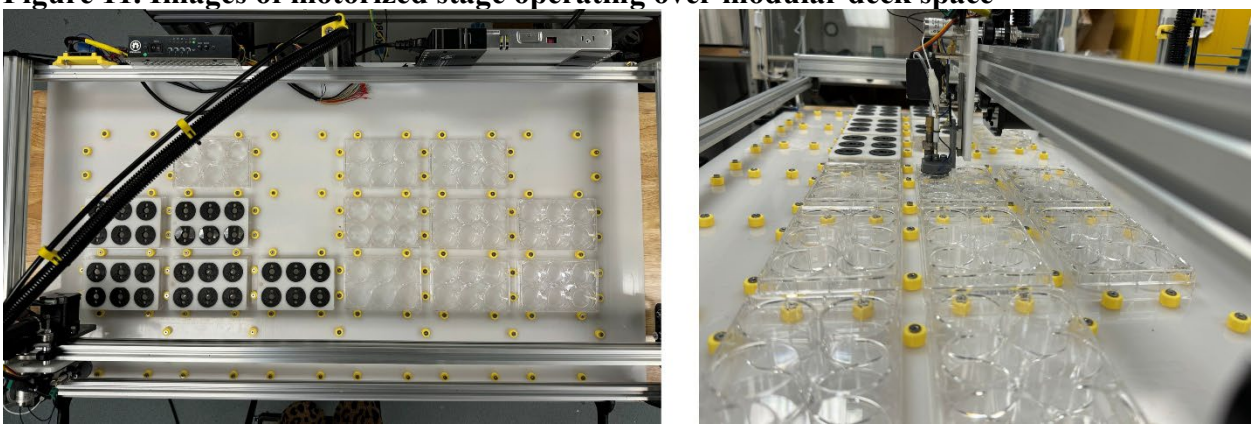
The UR5e robotic arm was integrated with an Instron compression tester to automate the testing process, reducing the reliance of labor-intensive characterization methods. The UR5e arm was programmed to continuously transfer aerogel samples from a customized sample station to an Instron compression tester. Once the UR5e arm picked up an aerogel sample for testing, another

sample automatically dropped into the same position, awaiting the next round of robotic pickup. After placing the aerogel sample in the tester, the UR5e arm signaled the Instron system to conduct compression tests and measure the compressive stress–strain curves. From the stress–strain curve of each mixed-dimensional aerogel, the compressive stress at 80% strain was extracted.

2.3 Modular autonomous electrochemical and electrical testing modules

The Modular Autonomous Electrochemical Testing Station combines a motorized stage (**Fig. 11**) with a Metrohm Autolab Electrochemical Workstation (PGSTAT302N) to enable a wide range of electrochemical and electrical characterizations. The motorized stage features an OpenBuilds Acro 510 XY router with an operational workspace of 40 inches by 20 inches. A custom workspace built using Delrin plastic is configured with slots designed to accommodate eighteen 6-well plates arranged in a 3×6 grid or any similarly sized tray.

Figure 11. Images of motorized stage operating over modular deck space

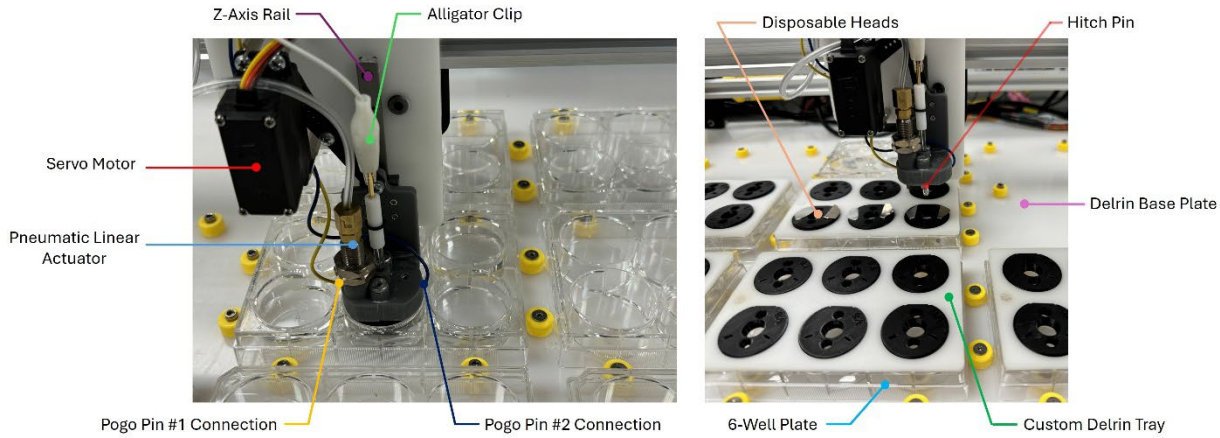


Operating workspace of motorized stage is 40” by 20” which is segmented into slots which fit a standard 6-well plate. The origin of each slot is stored, and the positions of wells are dynamically computed to enable compatibility with different lab equipment.

The platform’s versatility in electrochemical and electrical characterization tasks is achieved through a custom end effector (**Fig. 12**). This end effector incorporates a servo motor for precise control along a third (Z) axis. It features three exposed electrodes—two pogo pins and one alligator clip—that interface with probes, each equipped with external ports for connectivity

to external devices. Hitch pins secure the end effector to disposable testing heads, while a pneumatic linear actuator enables automated detachment. Digital outputs from the XY router's controller manage these actuators and facilitate communication with external devices.

Figure 12. Annotated components of motorized stage and end effector



A multicomponent end effector and modular workspace enables the motorized stage to be compatible with many electrochemical and electrical characterization tasks. The workspace contains slots for 6-well plates to be laid out in an array. Custom tops for standard 6-well plates can be manufactured to house non-standard components such as for trays of disposable heads. Disposable heads are custom manufactured to attach to the hitch pin on the end effector and be detached using the pneumatic linear actuator.

The Metrohm Autolab Electrochemical Workstation is a commercially available system capable of performing a wide range of electrochemical and electrical characterization processes. Its Nova API enables programmatic control, including data import, export, and remote signaling. Testing routines can be designed using the Nova user interface, which includes the ability to communicate with the XY router via an I/O part, enabling bidirectional signal transmission. When preparing testing routines for batches of samples, the Autolab can be configured to loop through the required number of samples automatically.

The motorized stage contains a control module (controller), which executes instructions using geometric code (G-code), a programming language commonly used to operate 3D printers

and computer numerical control (CNC) routers. G-code consists of commands that dictate motor speed and movement paths.

To accommodate both bidirectional communication between the Instron tester and the motorized stage, as well as controlling actuators on the custom end effector, several controller outputs were reassigned, repurposing standard G-code commands.

Tool Direction commands (*M3*, *M4*) were reconfigured from clockwise (CW) and counterclockwise (CCW) to, respectively, represent normal operation (*M3*) and activation of relay 2, signaling the Autolab to start the test routine (*M4*).

Servo positions (S-commands) are normally used to control tool speed, however, were adjusted to adjust the servo position via a PWM signal. For example, *S0* fully retracts the servo, while *S5000* fully extends it (to the well-plate position). Intermediate values allow for height adjustments to accommodate different apparatus configurations.

Pneumatic actuations were assigned to the *M8* and *M9* commands, which are typically assigned to coolant control. *M8* and *M9* were repurposed to, respectively, activate the solenoid controlling the pneumatic pushrod on the custom end-effector and to deactivate it. This mechanism releases disposable cartridges when testing is complete.

The *M66* code is traditionally used for input handling. We utilized an *M66 P1 L2 Q7200000* command to wait for the Autolab's signal, indicating that testing had finished, before resuming execution of the G-code program controlling the motorized stage. *P1* refers to the controller's door input pin. *L2* specifies the condition, waiting for the signal to transition from high to low. *Q7200000* sets a timeout of 2 hours which was adjusted to accommodate different testing procedures.

The G-code script also consisted of common commands such as *G4*, which pauses execution for a set duration, and *G0*, which moves the stage to a specified XY coordinates at maximum speed.

A Python script is used to generate G-code files for execution by the motorized stage. It reads a CSV file containing sample IDs (SIDs), maps the deck space layout, assigns samples to specific slots, and programmatically generates the necessary G-code to perform testing.

To ensure proper and reliable execution of characterization tasks, the motorized stage and the Autolab testing station must be synchronized and able to communicate at a few critical junctures.

When the stage is positioned to begin sample testing, the controller switches the tool direction from clockwise to counterclockwise (CCW), activating relay 2. The normally open (NO) pin is connected to the Autolab with a pulldown resistor to ground. When activated, the NO pin outputs 5V, signaling the Autolab to start the test process. The Autolab then energizes the electrodes and collects data from the potentiostat while the controller awaits a signal from the door relay. Once testing is complete, the Autolab sends a signal through the door relay to the controller, prompting the G-code to advance and return the used cartridge to its original well.

This platform has been used to perform high-throughput characterizations for two projects, both of which are currently in submission.

2.3.1 High-throughput characterization of aqueous electrolytes

While designing functional aqueous electrolytes, the soluble design space can be narrowed to satisfy the requirement of a wide electrochemical stability window (ESW), including both anodic and cathodic stability. In Zn aqueous electrolytes, linear sweep voltammetry (LSV) typically overlaps the current signals from hydrogen evolution reaction

(HER) and Zn deposition, making it difficult to quantitatively determine the electrolyte decomposition potentials. Instead, each electrolyte can be subjected to cyclic voltammetry (CV) testing of Zn deposition and dissolution to measure cathodic stability (**Fig. 13d**). The Zn electrode potential corresponding to Zn deposition and dissolution is used as the cathodic stability feature, because it reflects the reducing ability of Zn and can be tuned through the electrolyte formulation. An upward (more positive) shift indicates a lower tendency of the electrolyte to decompose during Zn deposition and dissolution.

However, CV testing is a bottleneck in this experimental workflow. Tests can last three to six hours per sample but only require a human operator to configure the testing setup: a three-electrode system consisting of reference, working, and counter electrodes. CV testing is therefore an ideal process to automate, enabling this characterization task to occur autonomously and unsupervised, significantly reducing the manual-workload from the human operator (**Fig. 13b**).

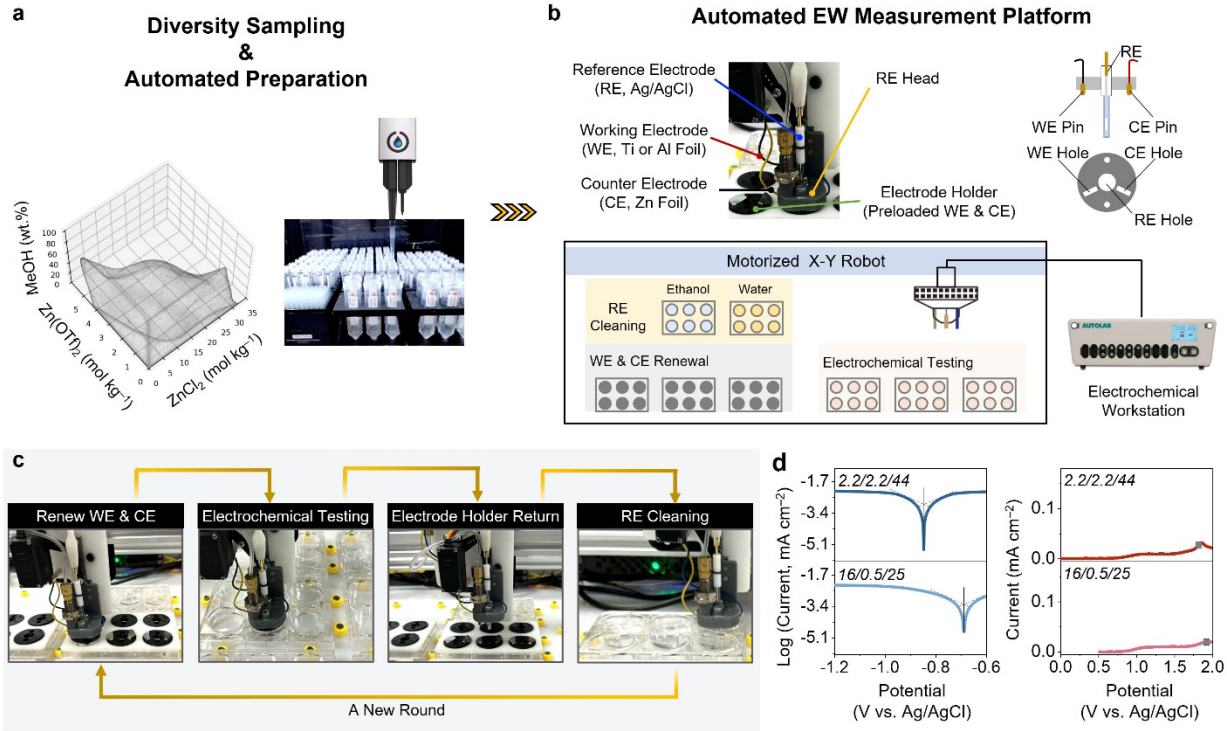
Several factors complicate this task. First, the working and counter electrodes must be replaced between tests to prevent contamination from redox reactions on their surfaces. To address this, we designed custom disposable heads with preloaded working and counter electrodes that the motorized stage's end effector can easily pick up and drop off.

Second, the disposable heads must establish a reliable connection to the Autolab testing station. This is achieved by designing slots for the working and counter electrodes in the disposable heads that allow an excess tab on the top face to contact the pogo pins on the end effector. This contact maintains a connection to the Autolab via the external ports.

Third, the reference electrode must be cleaned between tests to prevent cross-sample contamination. Unlike the working and counter electrodes, it cannot be replaced due to its high cost. To resolve this, we designed a continuously stirred washing station that fits into a deck slot,

agitating the electrode while recycling fresh water. Afterward, the electrode is dipped into an ethanol bath and air-dried. This two-step cleaning process minimizes the risk of contamination.

Figure 13. Electrochemical stability window refined feasible design space of aqueous electrolytes



(a) To build an ESW dataset, grid sampling was performed across the soluble design space, yielding 45 data points that were then prepared automatically using the OT-2 robot.
(b) To streamline electrochemical stability testing, an X-Y-Z motorized stage is integrated with an electrochemical workstation to collect electrochemical properties automatically. For each electrolyte solution, the head of the X-Y-Z motorized stage which has an Ag/AgCl reference electrode in the center, was programmed to pick up a fresh pair of working electrodes (Ti foil for cathodic and Al foil for anodic) and a count electrode (Zn foil) to conduct a three-electrode electrochemical measurement.
(c) After each test, the head returned the working and counter electrodes first, then followed by rinsing any electrolyte residue from the reference electrode in a stirred water bath. After a thorough air-drying step, the system proceeded to the next electrolyte measurement.
(d) For anodic stability, LSV testing was performed automatically on 45 sampled electrolyte solutions. In view of the large changes in viscosity among the solutions and the subjective nature of identifying oxidation potentials, we introduced the concept of a differential LSV (dLSV) value to locate the fastest oxygen evolution reaction (OER) point, thereby ensuring consistency in ESW calculations. The gray dots indicate the fastest OER process in each electrolyte, and the potential at this point was taken as the anodic stability feature.

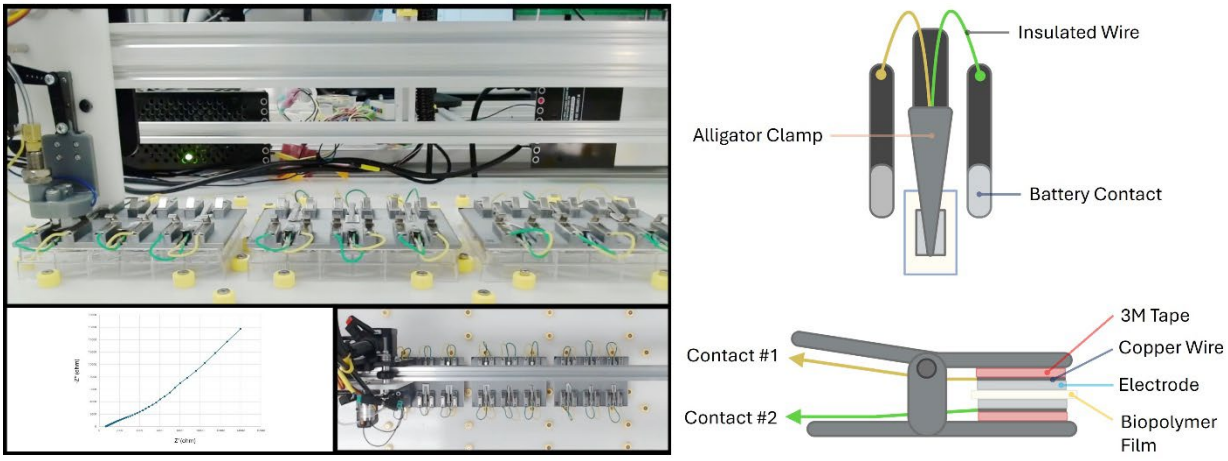
2.3.2 Autonomous impedance testing of biopolymers

Developing biopolymer-based electronics requires precise characterization of the through-plane impedance of thin films, a critical property for viable formulations. Traditionally, this test is performed manually by assembling a setup with a working and counter electrode affixed to either side of a thin film. Automating this process significantly improves experimental throughput and efficiency.

Designing an automated testing setup was straightforward. The test could be conducted using an Autolab testing station, requiring only consistent electrode contact with both planes of the film. To achieve this, we designed electrode contacts that securely held the film and interfaced with the pogo pin connections on the motorized stage's end effector. These connections were directly linked to the Autolab via external ports.

The film-holding mechanism (**Fig. 14**) consisted of an alligator clamp with insulated electrodes positioned on either side of the film. Wires connected these electrodes to battery contacts, which interfaced directly with the pogo pins. To streamline testing, trays were designed to house six of these mechanisms, allowing attachment to the tops of 6-well plates. The motorized stage was then programmed to iterate through each station, execute the testing routine, and proceed sequentially.

Figure 14. Automated film impedance testing operation and annotated design



(left) Impedance test procedures run sequentially on the Autolab test station, characterizing films in a row-major order. Data is collected and uploaded to a structured database from the Instron workstation. Trays of 2x3 testing setups are housed on 6-well plates. (right) The testing setup is housed within an alligator clamp for convenient loading and unloading of biopolymer film samples. Each face of the clamp consists of three layers: 3M tape to insulate the electrode from the conductive alligator clamp, a wire to transmit signals, and a titanium electrode to provide a smooth conductive surface is in contact with either side of the sample. The wires are soldered to battery contacts which are properly spaced to contact the pogo pins on the motorized stage end effector.

2.4 Copper leaching platform

Traditional copper extraction from low-grade slate rock relies on expensive and environmentally harmful chemicals with limited efficiency[53]. Bacterial leaching offers a more cost-effective and eco-friendlier alternative[54,55]. However, when validating the efficacy of promising leaching agents, there is a massive jump between microfluidic testing to giant columns, which is costly and time-consuming. This creates a need for an intermediate, higher throughput, testing platform.

When developing bacterial leaching agents, promising candidates are first tested on a microfluidic scale before validation in standard testing apparatus—3-meter-tall columns, several feet in diameter, that require months-long residence times and cost hundreds of thousands of

dollars to run. Our task was to create an intermediate system that appropriately scaled to the larger columns to enable higher experimental throughput at a fraction of the cost.

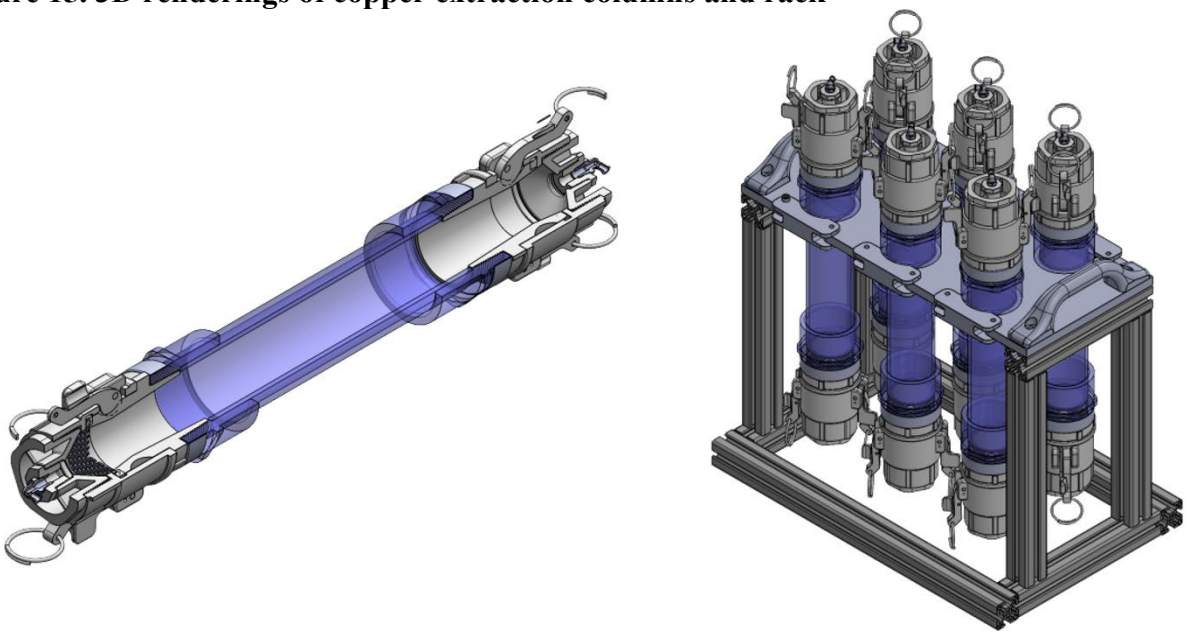
These intermediate-scale column testing apparatuses have several design requirements. One is that sampling ports were necessary for quantifying copper extraction by the microorganisms from the ore. The copper ore is housed in a 2-inch diameter, 12-inch-tall PVC pipe. Great care was taken to appropriately scale down from the larger columns while maintaining relevant test conditions, including the column height, width, as well as the particle sizes of the ore. The setup requires a constant, mild temperature, so an array of 6 testing columns were housed on a custom holder that fits within a commercial lab-grade oven (**Fig. 15**).

Media flow is maintained at 330 μL/min using a peristaltic (roller) pump. The flow rate can be adjusted by modifying the roller speed and tubing inner diameter. Media is distributed over the column, and a 5 μm filter fabric at the bottom flange retains the ore within the column. An equal flow rate of ambient air into the column applies slight positive pressure, driving media flow through the system. This also supplies oxygen to the media in the column, which is beneficial since the bacteria used are aerobic. Custom end caps were designed to securely hold the copper ore, allow controlled media flow, and maintain a pressurized seal. These caps are detachable, enabling operators to load and unload ore as needed. Tests run for two weeks, with human operators intermittently collecting media samples to analyze copper concentrations.

The liquid media is stored in a 1L flask equipped with a specialized three-port lid, allowing secure installation of inlet and outlet tubing as well as a one-way relief valve to prevent any pressure buildup. This prevents contamination from airborne particulates and minimizes acid splashing.

Materials were selected which were compatible with dilute sulfuric acid, which is produced during iron oxidation in the column. Chlorinated Polyvinyl Chloride (CPVC) was chosen due to its strength, stability, and widespread use in pilot-scale chemical processing systems. It does not leach metals that could interfere with inductively coupled plasma optical emission spectroscopy (ICP-OES) or inductively coupled plasma mass spectrometry (ICP-MS) analyses. Additionally, CPVC is relatively lightweight compared to alternative construction materials.

Figure 15. 3D renderings of copper extraction columns and rack



Copper ore is loaded into the central compartment of the column. Custom end caps form a pressurized seal with inlet and outlet ports to allow for media flow.

Chapter 3: Predictive and generative modeling of mixed-dimensional aerogels with programmable properties

3.1 Motivation and introduction

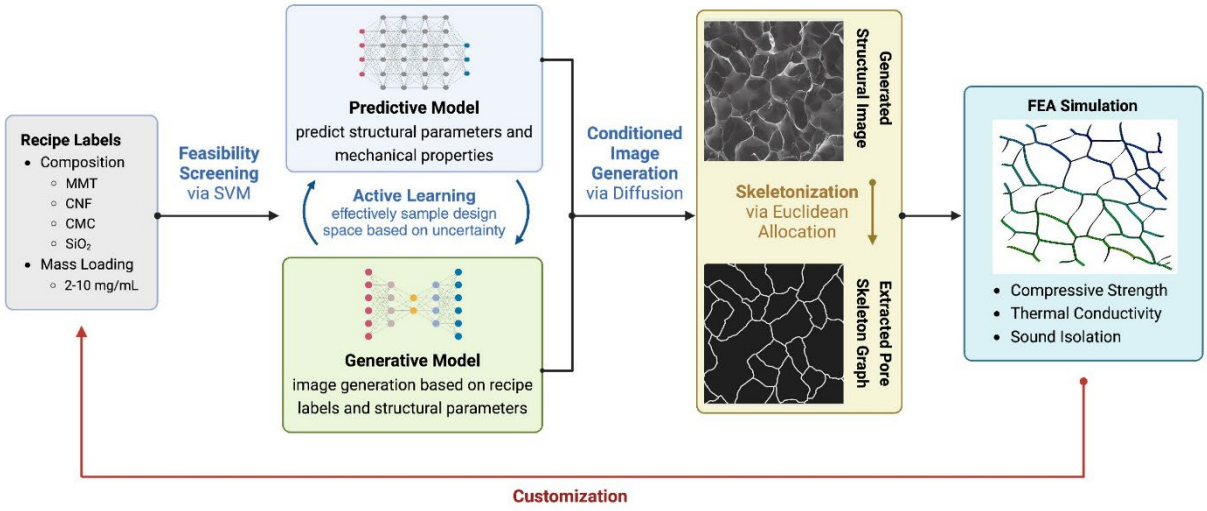
Mixed-dimensional aerogels are engineered using a combination of building blocks with varying dimensionalities—such as 0D polymers, 1D nanofibers, and 2D nanosheets—forming highly interconnected, hierarchical networks[56-58]. This multiscale architecture imparts tunable properties across mechanical resilience, thermal insulation, and acoustic absorption[59,60]. As a result, mixed-dimensional aerogels have been explored for a wide range of applications, including vibration-damping systems, thermal protection layers, and soundproofing materials[61,62]. However, discovering and optimizing aerogel compositions and structures that meet specific performance criteria remains a time-consuming and experimentally intensive challenge[56]. Several key limitations hinder progress in this field. First, current fabrication strategies rely heavily on trial-and-error experimentation, often probing a broad parameter space in a non-systematic manner[63]. The freeze-drying process, commonly employed in aerogel fabrication, further extends development timelines[59]. Conventional approaches such as one-factor-at-a-time (OFAT) and design of experiments (DoE) are labor-intensive and inadequate for efficiently navigating the expansive material design space[46]. Second, mixed-dimensional aerogels must simultaneously satisfy multiple performance requirements—such as mechanical strength, thermal insulation, and acoustic absorption—necessitating complex, iterative optimization[64]. Third, the characterization of these properties often requires samples of different

sizes, geometries, and testing conditions, adding further logistical and time-related constraints. Together, these challenges underscore the urgent need for a streamlined, data-driven framework that integrates predictive and generative modeling to minimize experimental burden, enable multi-property optimization, and dramatically accelerate the development of high-performance, application-specific aerogels.

Artificial intelligence (AI), particularly machine learning (ML), is rapidly reshaping materials research by enabling both predictive and generative modeling for complex, multi-degree-of-freedom (DOF) design problems[65]. In the field of mixed-dimensional aerogels – where the interplay between composition and structure determines performance across thermal, mechanical, and acoustic properties[66] – AI/ML presents a powerful alternative to traditional trial-and-error methods. Data-driven models can uncover hidden composition–structure–property relationships and guide formulation strategies to meet multifaceted design targets[67]. Despite this potential, significant barriers exist in applying AI/ML to aerogel design. Unlike other fields such as drug discovery[68] or catalyst development[69] – where high-throughput platforms and simulation tools can produce large volumes of standardized data – mixed-dimensional aerogel research is constrained by data scarcity[70]. The fabrication process is slow, resource-intensive, and typically involves freeze-drying and multistep assembly, limiting the rate of data acquisition[71]. Moreover, variations in component selection and experimental protocols across laboratories lead to fragmented and inconsistent datasets, undermining model accuracy and generalizability[72]. To address these limitations, a hybrid framework that integrates predictive design, generative modeling, and physics-based simulations offers a promising path forward[73,74]. This approach enables efficient exploration of the aerogel design space, reducing experimental overhead and

accelerating the discovery of high-performance aerogels with tunable, application-driven properties.

Figure 16. Construction of a prediction model via AI/ML and robotic technologies for accurate property predictions of mixed dimensional hybrid aerogels



This workflow utilizes both predictive and generative modeling to perform multi-objective optimization of directionally freeze-dried mixed dimensional hybrid aerogels. The feasibility-constrained design space if efficiently sampled through active learning loops to train prediction models on mechanical and structural labels and a generative model on surface geometries, which can be directly input to FEA simulations for high-fidelity computational characterization.

3.2 Tuning mechanical, thermal, and acoustic properties of mixed-dimensional aerogels through fabrication parameter optimization

Following a directional freeze-drying process at –80 °C and 0.3 Pa, various MMT/CNF/SNF/CMC formulations and mass loadings resulted in a diverse range of aerogels. These aerogels were composed of building blocks with different dimensions, collectively referred to as “mixed-dimensional aerogels” afterwards. Adjusting the MMT/CNF/SNF/CMC ratios and modifying solid loadings led to a non-linear variation in mechanical, thermal, and acoustic properties. These non-linear trends underscored the intricate relationships between composition, structure, and properties in mixed-dimensional aerogels[75].

To construct a comprehensive database linking fabrication parameters to the final properties of mixed-dimensional aerogels, an estimated 93,704 data points would be required, assuming a step size of 2 wt.% and four solid loadings. However, acquiring such a large dataset through conventional experimental methods is impractical due to the time-intensive and complex nature of aerogel fabrication and characterization. For instance, producing a single batch of mixed-dimensional aerogels takes an average of six days, while comprehensive characterization of their mechanical, thermal, and acoustic properties requires multiple sample dimensions and an additional five days for testing. To overcome these limitations, we developed an integrated workflow that combines robot-automated experimentation, predictive and generative modeling, and FEA simulations, aiming to accelerate experimental data acquisition, facilitate inverse aerogel design with programmable properties, and significantly reduce the need for labor-intensive aerogel characterization.

3.3 Defining a design space using an automated pipetting robot, dual-camera image analysis, and a support-vector machine (SVM) classifier

The integrated workflow for mixed-dimensional aerogels comprised four critical phases: (1) defining a design space, (2) constructing a structural database through active learning loops, (3) training predictive and generative models using a data-augmentation strategy, and (4) performing multiphysics FEA simulations.

As illustrated in **Fig. 17a**, the first phase aimed to define a design space for mixed-dimensional aerogels. The data collection process was accelerated using an OT-2 robot, a dual-camera image analysis system, and an SVM regressor. Initially, the OT-2 robot was programmed to prepare a library of aqueous mixtures with varying MMT/CNF/SNF/CMC ratios and solid

loadings (ranging from 4 to 10 mg mL–1). 224 mixed-dimensional aerogels were prepared using an interval of 20 wt.% across four solid loadings. Once prepared, these mixtures were vortexed, cast into silicone molds, and subjected to a directional freeze-drying process.

Next the volume of the 224 mixed dimensional aerogels were characterized using a dual-camera image analysis system. As shown in **Fig. 17c**, the volume retention ratio of these aerogels served as the training data points for an SVM regression model, which identified the maximal-margin hyperplanes between data points with different aerogel retention ratios. By using an independent set of testing data points to evaluate the model's prediction accuracy, the average deviation between model-predicted and actual retention ratios was quantified in terms of mean absolute error (MAE), as defined in **Equation 1**,

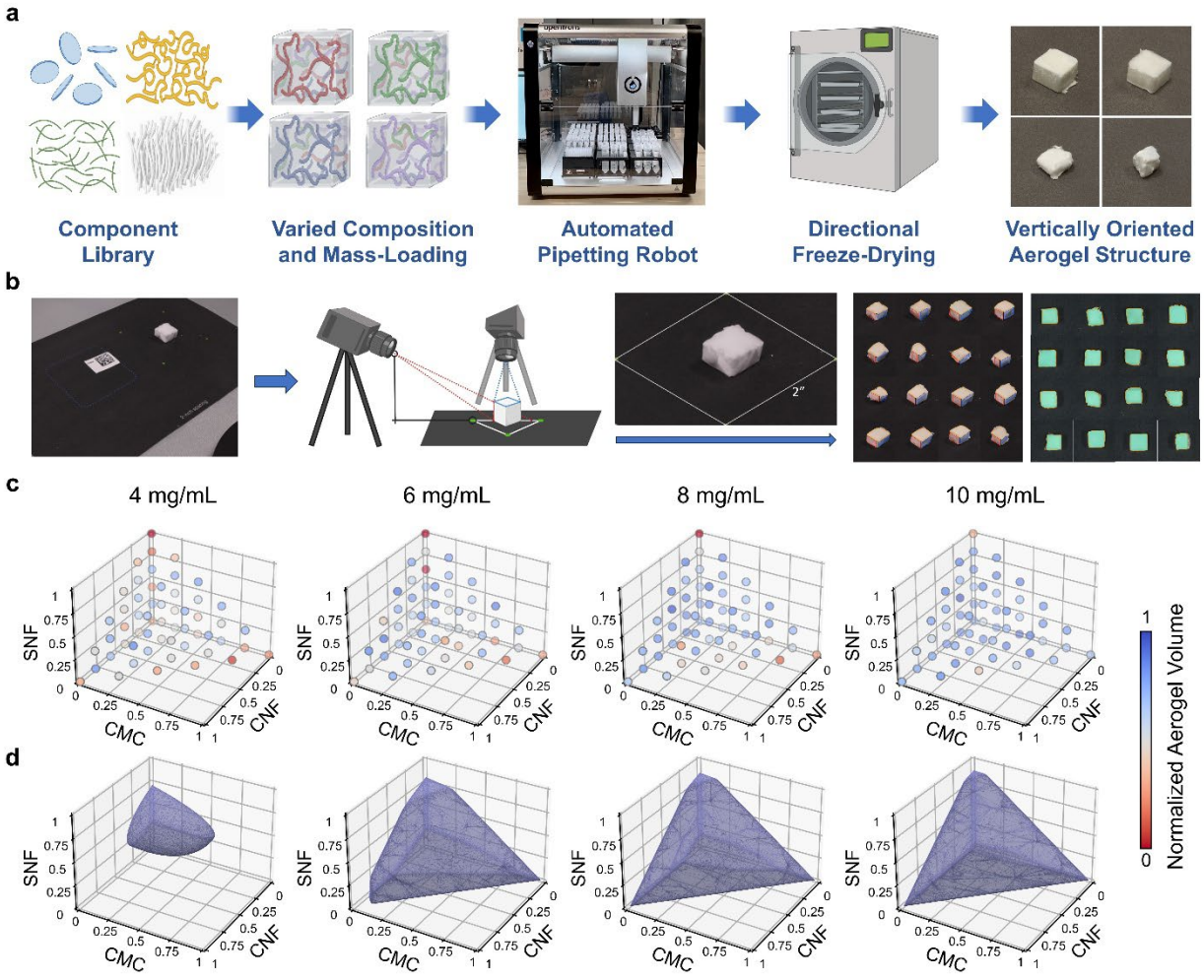
Equation 1. Mean absolute error

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|R_{predicted,i} - R_{actual,i}\right|$$

where N is the number of testing datapoints, $R_{predicted,i}$ is the volume retention ratio predicted by the SVM model based on a testing datum (i), $R_{actual,i}$ is the volume retention ratio of a testing datum (i). A smaller MAE value indicates higher prediction accuracy, and vice versa. The SVM model accurately predicted the retention ratio of mixed dimensional aerogels to a high degree of accuracy.

As shown in **Fig. 17d**, the SVM model produced three-dimensional (3D) heatmaps that represented the predicted volume retention ratios of mixed dimensional aerogels with varying MMT/CNF/CMC/SNF ratios and different solid loadings. By setting the retention ratio threshold at 80%, a design space for mixed-dimensional aerogels was established, excluding >70% of the experimental design space that contained MMT/CNF/CMC/SNF ratios and solid loadings that resulted in significant aerogel volume shrinkage.

Figure 17. Defining feasible design boundary of mixed-dimensional hybrid aerogel system using dual-camera image analysis and SVM classifier



(a) Schematic illustration of the fabrication process of mixed-dimensional hybrid aerogels accelerated by an automated pipetting robot (OT-2 robot). By adjusting the MMT/CNF/CMC/SNF ratios and mixture loadings (i.e. solid contents of aqueous mixtures), the mechanical, thermal and acoustic properties of mixed-dimensional aerogels were controlled.
(b) A dual-camera image analysis system programmatically measures volume retention rates of fabricated, directionally freeze-dried mixed-dimensional aerogel samples. A cutoff of 80% volume retention cutoff determined feasibility labels of formulations.
(c) Normalized volume retention of 224 MMT/CNF/CMC/SNF mixed-dimensional aerogels is plotted across 4 different mass loadings.
(d) 3D continuous feasible design boundary calculated using a trained SVM classifier, across 4 different mass loadings. Defined using a normalized volume retention cutoff of 70%.

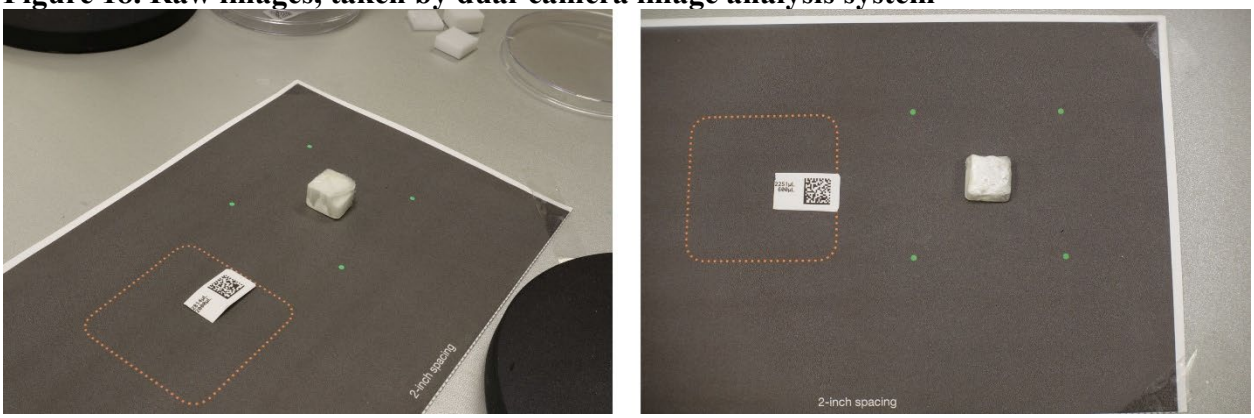
3.3.1 Collecting feasibility data using dual-camera image analysis system

The two-camera system consisted of one camera positioned directly overhead, capturing a top-down view of the aerogel, while the other was placed at an angle along the diagonal of the dot grid. A lamp was suspended above the setup to provide illumination. An illustration of this setup is presented in **Fig. 17b**.

Both cameras were focused using a white data matrix tag included in every image, enabling automatic sorting and labeling with the corresponding sample ID. The shutter time for both cameras was set to three seconds.

As depicted in **Fig. 18**, aerogel samples were placed at the center of the dot grid, which was marked by green dots forming a 2-inch by 2-inch square.

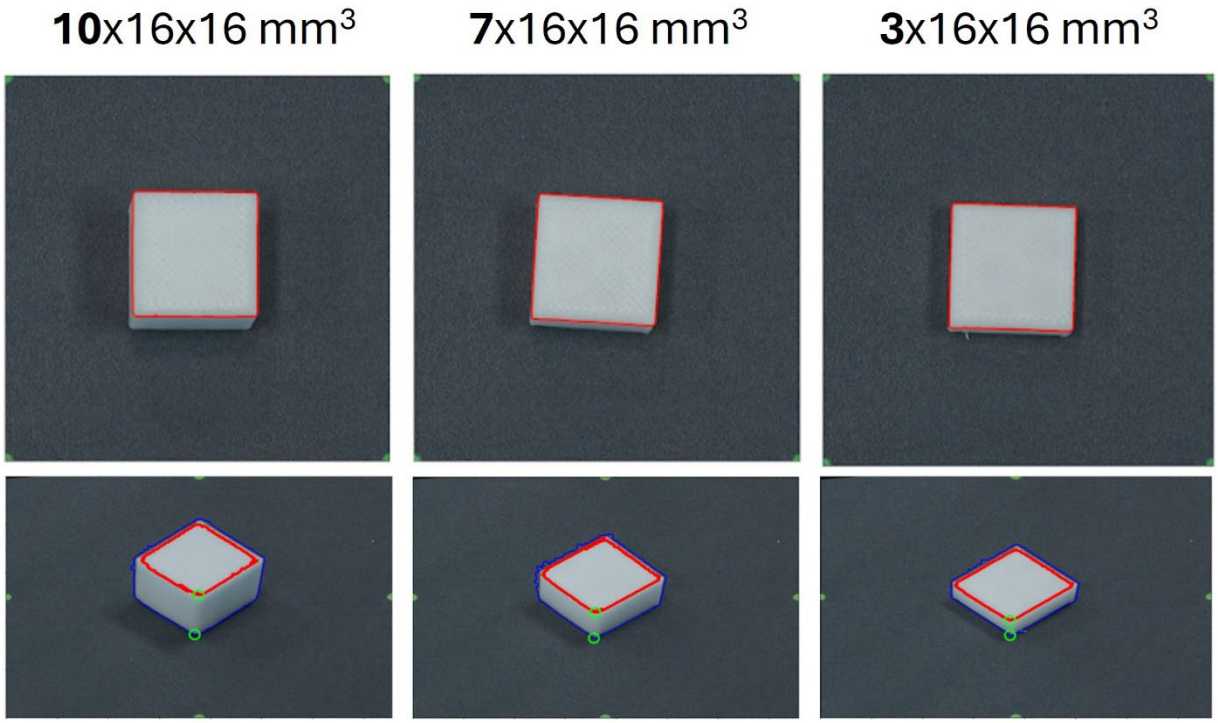
Figure 18. Raw images, taken by dual-camera image analysis system



A black backdrop was used to enhance image processing efficiency. The cameras were focused on the data matrix code to ensure accurate decoding of sample IDs and labeling of images. Aerogel samples were positioned at the center of the 2×2-inch dot grid.

Before each batch of images, the system was calibrated using three "calibration cubes" with dimensions of 3×16×16 mm^3, 7×16×16 mm^3, and 10×16×16 mm^3 (**Fig. 19**). These calibration images were used to determine scaling factors for accurate height and top-area measurements of the aerogels. Since the cameras remained stationary throughout each batch, calibration was required only once at the beginning of each session.

Figure 19. Cubes of known sizes calibrate the dual-camera system and volume estimation algorithm

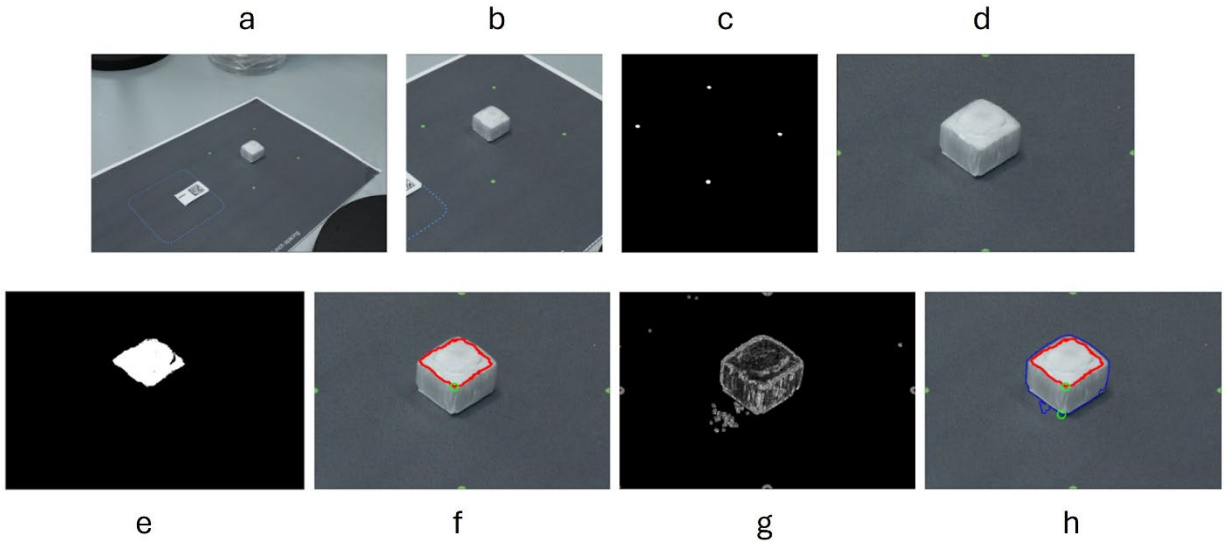


3D-printed cubes with heights of 3, 7, and 10 mm calibrated the dual-camera image analysis setup before each batch of images were taken. By first running the volume estimation algorithm on these cubes of known sizes, we solved for the camera matrix of the system and could convert from pixel to real-world length scales while computing the volume retention of aerogel samples.

3.3.2 Height estimation

First, the raw side-image is cropped to include only the sample and dot grid, removing the SID label and any extraneous elements. The green dot locations are then detected using HSV segmentation and contour detection. A perspective transformation is applied to align the image so that each green dot lies at the midpoint of its respective boundary (**Fig. 20a-d**). Since the real-world distances between the dots are known, this enables estimation of the pixel-to-millimeter conversion factor.

Figure 20. Height estimation image processing pipeline



(a) Raw side-image taken by dual-camera system.
(b) Cropped image to remove extraneous features.
(c) HSV segmenting to locate (green) dot grid.
(d) Perspective transforms places dots at center of boundaries.
(e) Binary segmentation extracts top face of aerogel sample.
(f) Top vertex location is determined as the bottom pixel on top-face contour.
(g) Sobel filter is a popular edge-finding algorithm. We use it to sharpen the sample outline and eliminate noise from extraneous foreground objects.
(h) Bottom vertex is recognized as the bottom pixel on the (whole) aerogel contour.

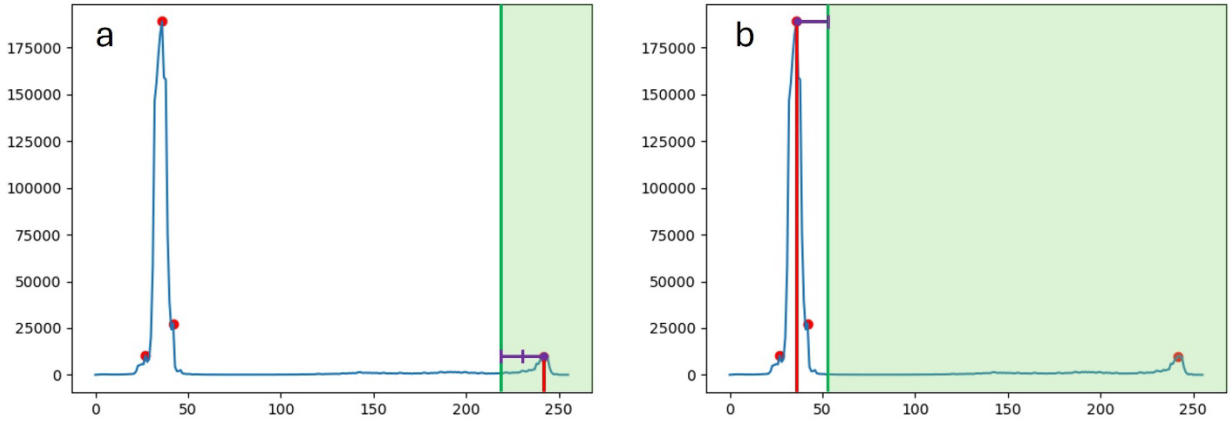
Next, the top and bottom vertices of the cubic aerogel are identified. The image undergoes preprocessing steps, including grayscale conversion, denoising, and normalization.

The top vertex is extracted by segmenting the processed image, isolating the contour of the top face (**Fig. 20e**), and identifying its lowest point along the y-axis. A dynamically selected threshold ensures only the top face is segmented. This threshold is determined by first computing the image (pixel intensity) histogram, smoothing the histogram using a Savitzky-Golay filter, and finally, identifying prominent peaks with a peak-finding algorithm.

The rightmost peak, corresponding to the lightest region in the image (illuminated by the overhead lamp), represents the aerogel's top face. The segmentation threshold is set to the peak

value minus twice the peak width to ensure complete extraction of the top face (**Fig. 21a**). The bottom-most point of the resulting contour is assigned as the aerogel's top vertex.

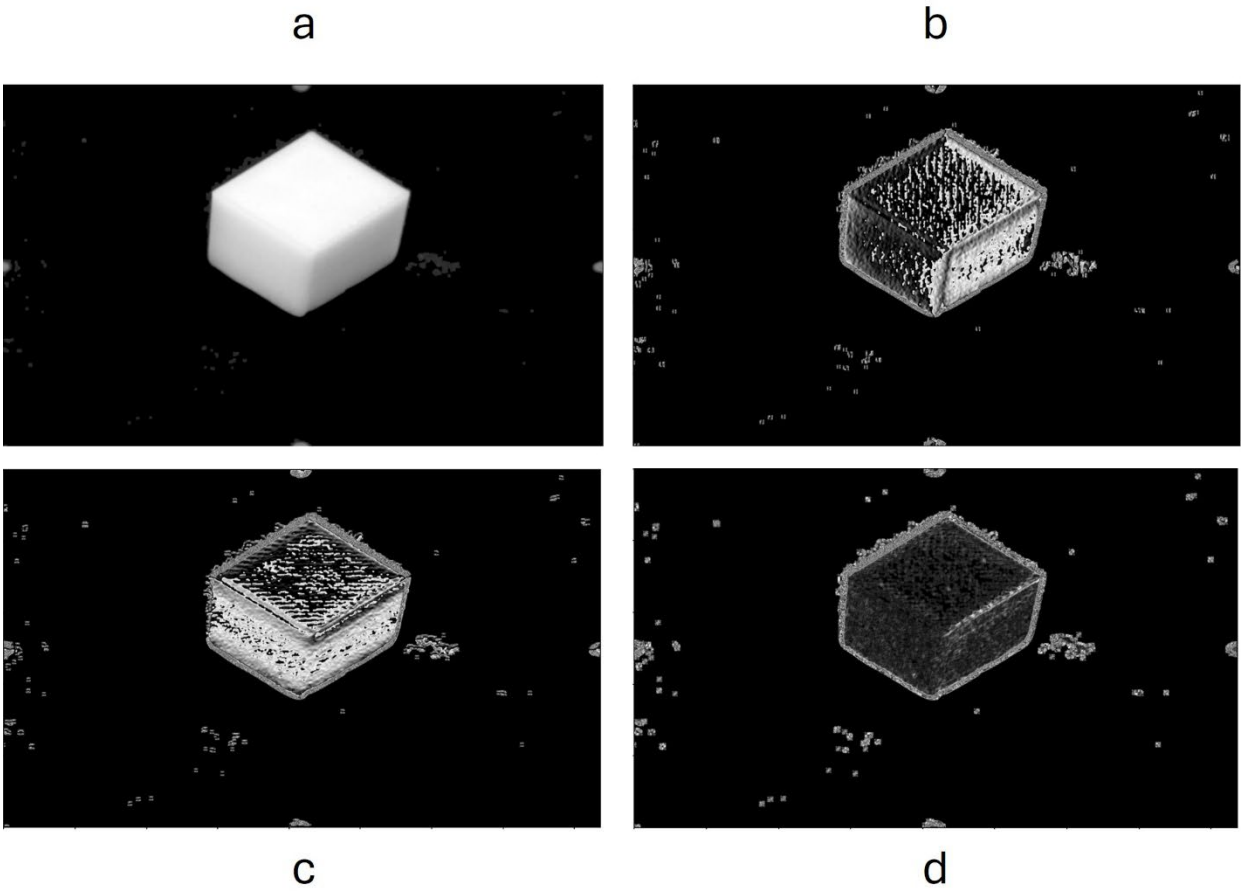
Figure 21. Locating aerogel vertices using dynamic thresholding of image histogram



(a) Dynamic calculation of threshold cutoff for top-vertex segmentation. The binary threshold cutoff set to the brightest peak value minus 2 times the peak-width. This threshold segments the top face of the aerogel sample, which is the brightest region due to the lamp illuminating the system from overhead.
(b) Dynamic calculation of threshold cutoff for bottom-vertex segmentation. To-zero threshold cutoff set to greatest magnitude peak value plus the peak's width. This threshold segments the aerogel sample, from the black background, which constitutes most pixels in the image.

To locate the bottom vertex, the preprocessed image is further refined using a bilateral filter to preserve edges while reducing noise and then performing dynamic thresholding with a to-zero threshold. The threshold is determined by extracting the image histogram, smoothing it using a Savitzky-Golay filter, and identifying the peak of greatest magnitude (representing the black background). The cutoff value is set to this peak value plus its width, ensuring the aerogel remains visible while the background is set to zero (**Fig. 21b**). We then perform morphological opening to remove small bright artifacts from the foreground (**Fig 22a**). Sobel filters are then used to sharpen the aerogel outline and eliminate noise from loose fragments or shadows (**Fig 22b-d**). Finally, we perform contour extraction, where the largest external contour is identified, and its lowest point along the y-axis is assigned as the bottom vertex.

Figure 22. Extracting sample contour using morphological openings and Sobel filters



(a) Morphological opening refines grayscale image by removing small objects (bright pixels) from the foreground.
(b) Sobel filter computed in X-direction.
(c) Sobel filter computed in Y-direction.
(d) Gradients of blended X- and Y-Sobel filters. Note that the largest and clearest contour in the image is now the aerogel sample which can be readily extracted.

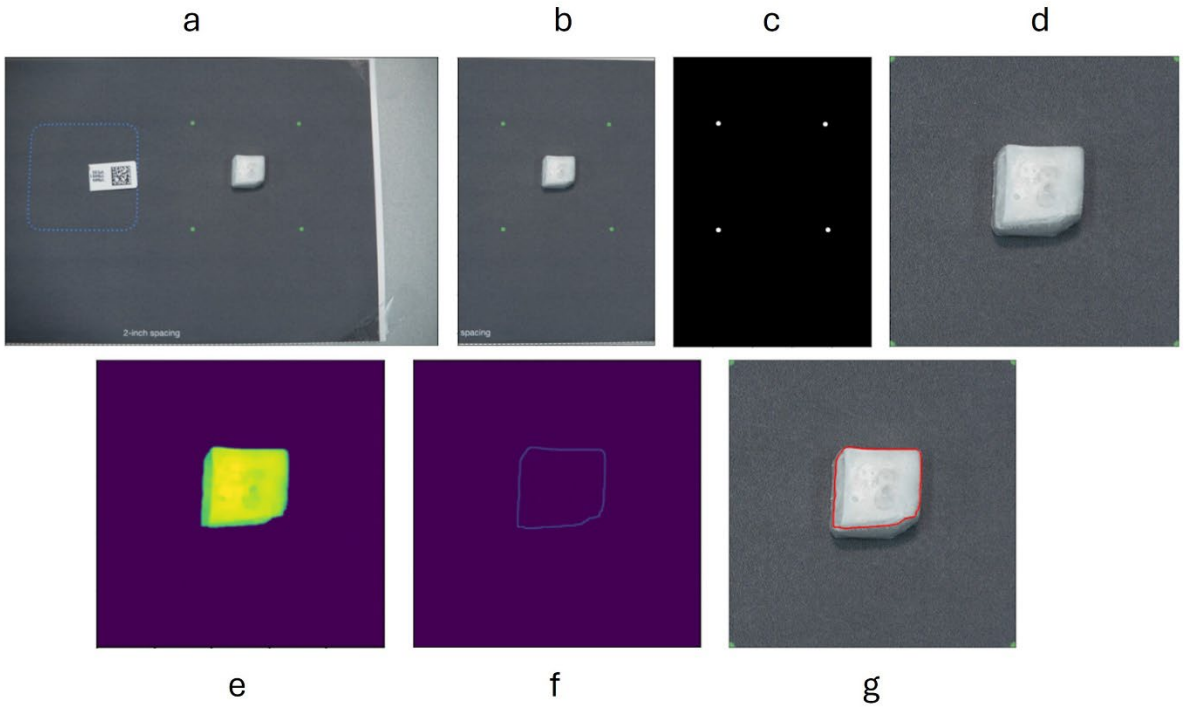
Finally, the height of the sample is calculated using the pixel difference in y-values, scaled by the previously determined pixel-to-millimeter conversion factor.

3.3.3 Top area estimation

The raw top-image is first cropped to include only the sample and dot grid, removing the SID label and any extraneous elements (**Fig. 23b**). The green dot locations are then identified

using HSV segmentation and contour detection (**Fig. 23c**). A perspective transformation is applied to align the image so that each green dot lies at the midpoint of its respective boundary (**Fig. 23d**). Since the real-world distances between the dots are known, this allows for accurate pixel-to-millimeter conversion.

Figure 23. Top-area estimation image processing pipeline



(a) Raw top-down image taken by dual-camera system.
(b) Cropped image to remove extraneous features.
(c) HSV segmenting to locate (green) dot grid.
(d) Perspective transforms places dots at corners of image.
(e) Fixed binary threshold segments the top-face of the aerogel sample.
(f) Laplacian filter extracts major boundaries, highlighting the edges of the top-face of the sample by detecting sudden changes in intensity.
(g) Contour of top-face area is extracted, and pixel area is calculated. This value is adjusted by a scaling factor determined during calibration and determined by the height of the cube.

Next, the image is converted to grayscale and blurred using a bilateral filter to preserve edges. Segmentation is performed using a fixed threshold, as the well-lit top surface of the

sample maintains consistent bright pixel intensity. A morphological opening is then applied to remove small bright artifacts from the foreground.

To extract the edges of the top face, a Laplacian filter is used to detect sudden intensity transitions, highlighting the sample's boundaries (**Fig. 23f**). The contour of this outline is then extracted, and the top-face area is calculated using the pixel-to-millimeter conversion factor, determined from the perspective transformation and known distances in the dot grid.

Since perspective effects can cause variations in perceived area due to the sample's distance from the camera lens, the calculated area is adjusted using a scaling factor proportional to the sample's height—derived from calibration cube data—to ensure accuracy.

3.3.4 Constructing SVM classifier to perform feasibility predictions

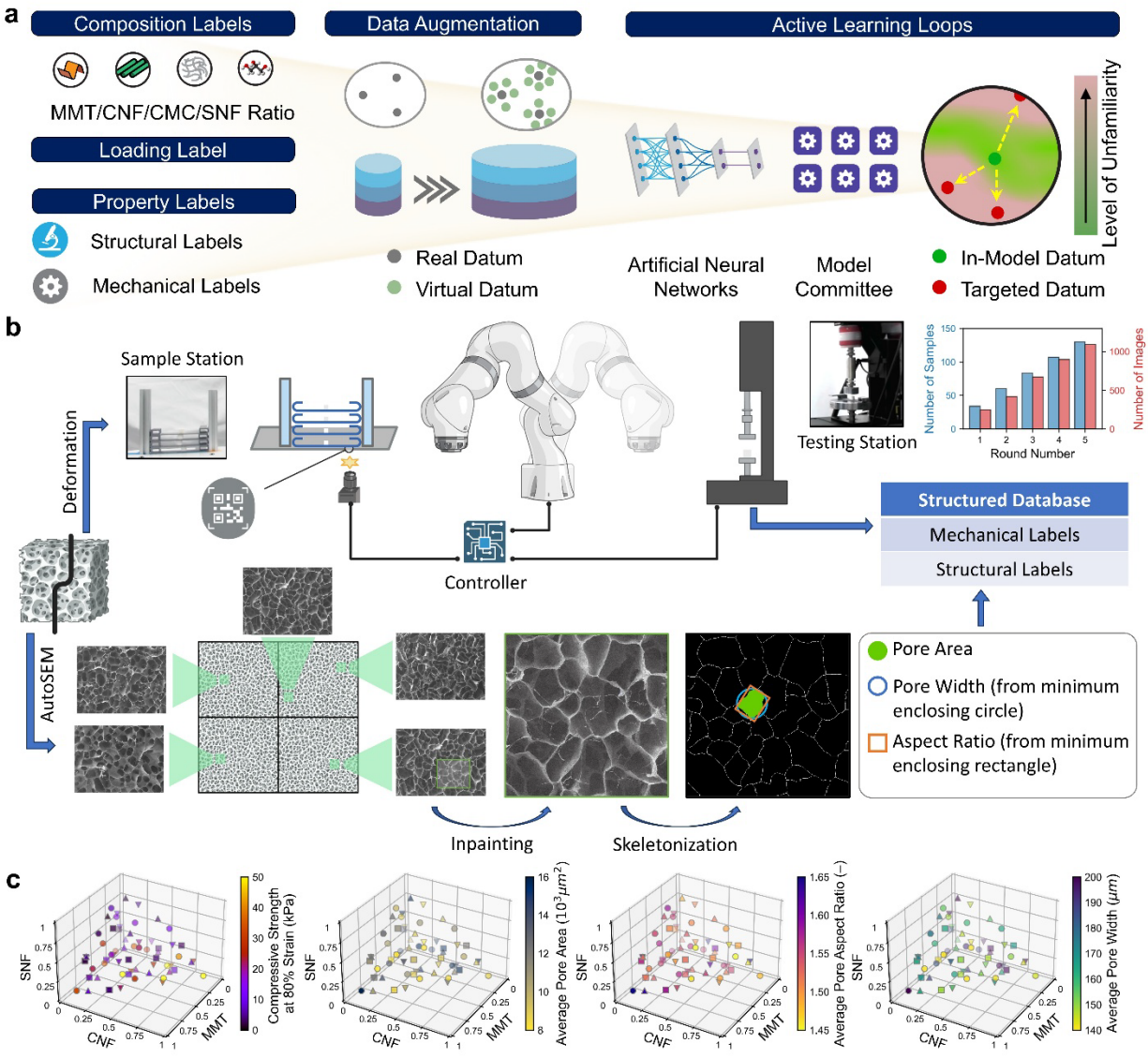
The SVM regression model served as a critical screening mechanism, directing active learning loops to operate exclusively within the design space and produce aerogel samples with high volume retention and structural integrity. Specifically, the SVM model restricted the prediction model to suggest only MMT/CNF/CMC/SNF ratios and solid loadings that yielded mixed-dimensional aerogels with high volume retention (>80%) at a success rate exceeding 93%. Unlike data-rich systems that could tolerate experimental failures, re-fabricating mixed-dimensional aerogels was a time-exhausting process, particularly when the suggested ratios were near the edges of the discrete data diagram. Consequently, transforming the discrete data diagram into a probabilistic heat map using the SVM regression model was essential.

3.4 Constructing a structural database via active learning loops and robot-automated experimentation platforms

Within the design space, active learning loops were utilized to guide the sampling of representative data points, enabling more efficient data collection and constructing a structured database. Multiple robot-automated experimentation platforms were programmed to streamline the fabrication, characterization, and analysis processes of mixed-dimensional aerogels. Specifically, three robotic/automated platforms – an OT-2 robot, a UR5e-automated compression tester, and a scanning electron microscopy (SEM) system with automated imaging – were integrated to minimize human workload and accelerate data acquisition.

As illustrated in **Fig. 24a**, the active learning loops were initiated by commanding the OT-2 robot to prepare 30 aqueous mixtures at random MMT/CNF/CMC/SNF ratios and solid loadings. Once vortexed, cast in silicone molds, and freeze-dried, these aqueous mixtures formed the first batch of mixed-dimensional aerogels. The MMT/CNF/CMC/SNF ratios were recorded as "composition" labels, while the corresponding solid loadings were documented as "loading" labels. Next, the porous microstructures of mixed-dimensional aerogels were analyzed using a SEM system with automated imaging functions. As shown in **Fig. 24b**, four aerogel samples were positioned on a stage with their surfaces oriented upward to expose the openings of vertical pores for imaging. Four imaging positions were defined to capture the structural information of aerogel samples, and SEM images were acquired using automatic contrast, brightness, and focus functions. For each position, images were taken at five different magnifications (ranging from 30x to 500x). For each SEM operation, four aerogel samples were characterized, and around 80 SEM images were collected (20 SEM images per aerogel sample across 5 locations).

Figure 24. Constructing a structured database of mechanical and structural labels of mixed-dimensional aerogels through active learning loops



(a) Active learning loops efficiently sample from high-dimensional design spaces, accelerating design space exploration, and enabling multi-objective optimization. Data augmentation and ensemble modeling strategies yield accurate prediction models that ensure accurate property predictions from formulation inputs.
(b) Fabricated mixed-dimensional aerogels undergo both structural and mechanical characterizations using collaborative robots. (top) A UR5e robotic arm and Instron tester integrated system autonomously measure the stress-strain curves of samples. (bottom) A tabletop SEM collects 80 images per batch of 4 fabricated samples. Structural parameters are then algorithmically extracted directly from the SEM images.
(c) Mechanical and structural property labels of mixed-dimensional samples at 10mg/mL mass loading.

After acquiring the SEM images of mixed-dimensional aerogels, we developed an image analysis software tool to automatically process the images and extract multiple "structural" labels. The software applied an image skeletonization algorithm to delineate the structural framework of the aerogel's microporous network. This skeletonization process enhanced the visibility of pore boundaries, enabling precise quantification of key structural labels, including pore area, pore width, pore height, and pore aspect ratio. Pore area was quantified by measuring the total surface area enclosed by the pore outline. Pore width was defined as the maximum distance between any two points on the pore outline. Pore height was determined by summing the two maximum perpendicular distances from the width-axis to any point on the pore outline, on both sides of the width-axis. Pore aspect ratio was calculated as the ratio of pore width to pore height. These structural labels provided critical insights into the microstructure of mixed-dimensional aerogels, facilitating deeper correlations between composition, solid loading, and aerogel structure.

Next, the mixed-dimensional aerogels were sent to an autonomous testing platform to characterize their compressive strengths. As shown in **Fig. 24c**, a UR5e robotic arm was integrated with an Instron compression tester to automate the testing process, reducing the reliance of labor-intensive characterization methods. The UR5e arm was programmed to continuously transfer aerogel samples from a customized sample station to an Instron compression tester. Once the UR5e arm picked up an aerogel sample for testing, another sample automatically dropped into the same position, awaiting the next round of robotic pickup. After placing the aerogel sample in the tester, the UR5e arm signaled the Instron system to conduct compression tests and measure the compressive stress–strain curves. From the stress–strain curve of each mixed-dimensional aerogel, the compressive stress at 80% strain (abbreviated as σ_{80})

was extracted, and the average σ_{80} value from at least three replicates were designated as the "mechanical" label.

Upon test completion, the Instron system signaled the UR5e arm to remove the tested sample and position a new one. During the active learning loops, more than 90 mixed-dimensional aerogels (with 2–3 replicates per data point) were evaluated using this autonomous testing platform. In short, each kind of mixed-dimensional aerogel resulted in one data point, which included four "composition" labels, one "loading" label, four "structural" labels, and one "mechanical" label.

To improve model learning efficiency and counteract potential model overfitting, the User Input Principle (UIP) method was adopted to synthesize virtual data points. The creation of virtual data points took place in the vicinity of collected real data points. To synthesize virtual data points, Gaussian noises were introduced in the proximity of the composition labels based on the observed composition insensitivity. The optimal virtual-to-real data ratio was determined to be 100:1, maximizing the model's learning efficiency while maintaining the training time of under three days per loop. It is important to note that no virtual data points were synthesized around the structural labels, ensuring that all analyzed data was derived directly from SEM observations. Subsequently, both real and virtual data points were used as training data for an ANN-based model employing 5-fold cross-validation. The top five ANN models, selected based on validation metrics, were assembled into a model committee. This ANN committee functioned as a navigator, guiding the exploration of the design space and identifying unlabeled data points with high information value for further evaluation.

In the next active learning loop, the navigator model assessed the unfamiliarity level of each data point within the design space using a hybrid acquisition function termed A Score[26], as represented by **Eqn. 2**,

Equation 2. A-score

$$A = \hat{L} \cdot \hat{\sigma}$$

where $\hat{L}$ denotes the Euclidean distance between in-model and model-targeted data points and $\hat{\sigma}$ denotes the ANN model's prediction variance on the mechanical label. The mechanical label was chosen as an indicator of model uncertainty because it serves as a singular macroscale property that captures potential structural variations across the design space. By reflecting aggregated effects from microstructural features (e.g., pore area, pore width), it provides a comprehensive measure of structural influence on mechanical performance. The data points with the highest A Scores were the least familiar to the model and pinpointed for experimental validation in the next loop.

By extracting the composition and loading labels of sampled data points, the OT-2 robot was activated to prepare a new set of MMT/CNF/CMC/SNF mixtures. Once vortexed, cast, and freeze-dried, a new batch of mixed-dimensional aerogels was produced. Similarly, the structural and mechanical labels of as-fabricated aerogels were characterized using the SEM system with automated imaging and the autonomous testing platform, respectively. Based on these real data points, virtual data points were synthesized using the UIP method. With both real and virtual data points as inputs, the ANN model was retrained, A-scores were reassessed, and a new set of composition and loading labels was suggested for the next active learning cycle.

With the operation of two collaborative robots, the active learning loops were significantly streamlined. Each loop took an average of 7 days, including 2 hours for OT-2

pipetting, 120 hours for freeze-drying, 8 hours for autonomous testing, and 4 hours for model training. In total, 5 active learning loops were conducted, resulting in the stagewise production of 130 kinds of mixed-dimensional aerogels. Over the course of this study, we collected 130 real data points during these 5 active learning loops, constructing a structural database for mixed-dimensional aerogels.

3.5 Data-augmented, ensemble modeling strategy to develop a prediction model for mixed-dimensional aerogels

After active learning loops were completed, experimental data were used to construct a prediction model by sequentially applying data augmentation and ensemble modeling techniques. Similarly, the UIP data augmentation method was applied to synthesize virtual data points. Using this augmented dataset, an ensemble modeling approach was implemented, where multiple ANN variants – each with different hidden layers, node configurations, and activation functions – were trained. These models underwent 5-fold cross-validation, and their validation errors were ranked. The top-performing ANN models were then selected to form an ensemble committee, where their outputs were averaged to generate more generalized predictions. Model accuracy was evaluated using an independent test set (excluded from the training process) and quantified in terms of MAE. As shown on **Fig. 25e**, after 5 learning loops, the MAE decreased to approximately 18%, approaching the observed measurement variation (~15% for structural labels). Compared to models without data augmentation or ensemble learning, this data-augmented, ensemble modeling strategy demonstrated superior efficiency and lower testing error. Data augmentation and ensemble modeling strategies are particularly effective for small experimental datasets, where data limitations often hinder model performance.

3.6 Diffusion-based strategy to develop a generative model for mixed-dimensional aerogels

To complement the predictive modeling described above, we developed a diffusion-based generative model to synthesize realistic microporous structures of mixed-dimensional aerogels under user-specified composition, loading, and structural label constraints. This generative framework was designed to (1) produce high-fidelity aerogel images with programmable pore shapes and sizes and (2) expand the structural database beyond experimentally collected SEM images, enriching subsequent simulations and design explorations.

From the structural database acquired during active learning loops, we extracted a training dataset consisting of SEM images (2,560 × 1,920 pixels) that captured top-view pore morphologies of aerogels at multiple magnifications. Before training, each image was randomized through rotation and cropping, center-cropped, and resized to 128 × 128 pixels to normalize input dimensions. To guide the diffusion model, we trained it with conditioning inputs, including composition labels (e.g., MMT/CNF/SNF/CMC ratios), loading labels (e.g., solid loading), and structural labels (e.g., pore area, width, height, and aspect ratio). The composition and loading labels provided style guidance, capturing appearance differences resulting from distinct material ratios, while the structural labels offered physical guidance, enabling the model to generate pore structures with realistic geometries.

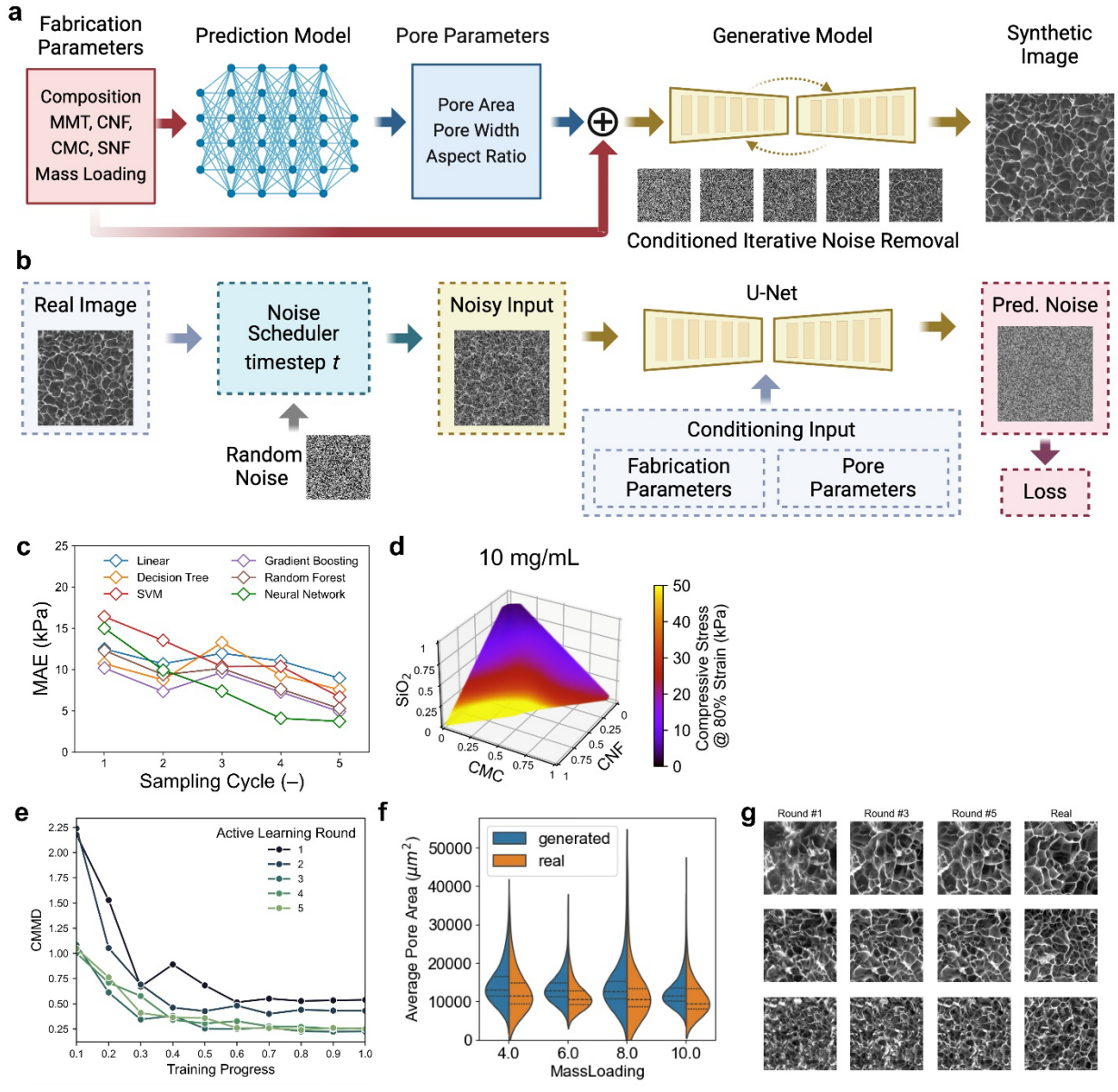
Next, a UNet model architecture was adopted[76], coupled with a DDPM noise scheduler configured for 1,000 diffusion steps. During inference, or image generation, noisy images were progressively denoised by the UNet model across 1,000 timesteps, with each denoising step guided by the specified composition, loading, and structural inputs. In each training iteration, a forward diffusion process first corrupted the training SEM images with Gaussian noise proportional to a timestep-dependent schedule. The generative model then attempted to predict

and remove this added noise, conditioned on the composition, loading, and structural inputs. The MSE between predicted and actual noises served as the training loss. Training was performed in mini batches, with the augmented dataset shuffled to enhance generalizability. At periodic intervals, the model was evaluated on a fixed subset of reference conditioned inputs, generating sample images that were visually inspected for stability and fidelity. Once the MSE loss plateaued and the generated samples converged to the desired structural characteristics, training was terminated. After each active learning loop, the generative model was updated by appending newly acquired SEM images to the training dataset, and training was restarted until convergence stabilized once again.

Each trained model was evaluated both qualitatively and quantitatively. Fréchet Inception Distance (FID) scores provided a style-focused comparative measure between generated and real SEM images. Lower FID values indicated higher perceptual similarity, demonstrating the effectiveness of composition and loading conditioning in capturing surface-level textures and characteristic pore features. As shown in **Fig. 25e**, the FID evolution was tracked across five active learning rounds. To validate the physical accuracy of the generated images, both real and model-synthesized SEMs were processed using the same skeletonization and pore-analysis algorithms applied elsewhere in this study. The resulting distributions of average pore area, width, height, and aspect ratio were compared using violin plots (**Fig. 25f**). The generated pore distributions followed the same broad trends and ranges as real aerogels at each mass loading. Minor offsets between real and generated distributions were attributed to inherent variability in the training data or model smoothing of extreme outliers. However, the closeness of interquartile ranges and median values confirmed the diffusion model's strong capability to replicate crucial structural attributes of mixed-dimensional aerogels. Through these qualitative and quantitative

evaluations, each diffusion model iteration was assessed for both stylistic realism (FID) and physical fidelity (pore shape metrics). In subsequent sections, these generative models are leveraged to expand design space exploration and guide inverse design tasks, offering a rapid means to preview aerogel morphologies without requiring exhaustive physical fabrication.

Figure 25. Construction of a generative model for aerogel surface microstructure generation



(a) A conditioned U-Net received inputs including fabrication parameters and structural parameters and generates high-resolution synthetic images. Uses diffusion algorithm to generate

images by iteratively removing noise from Gaussian image. This is a stochastic process and so no two generated images are alike.
(b) The diffusion model is trained by adding noise to real sample images and evaluating the model's ability to remove that noise to recover the original image. MAE loss metric can be used to evaluate model performance in this fashion.
(c) Neural network algorithm outperforms other common machine learning implementations. This is due to the algorithm's ability to capture non-linear features.
(d) Heat map visualizing the trend of compressive strength across the feasible design space of mixed dimensional aerogel samples at a mass loading of 10 mg/mL.
(e) Image generation performance improves after each iteration of active learning. This is because more data was collected with each iteration of active learning loops.
(f) Comparison between distributions of pore areas between real and generated surface microstructures.
(g) Visual depiction of diffusion model generating improved, more realistic, outputs after each active learning loop.

Chapter 4: Predictive design of sustainable biobased packaging via machine intelligence for improved postharvest preservation

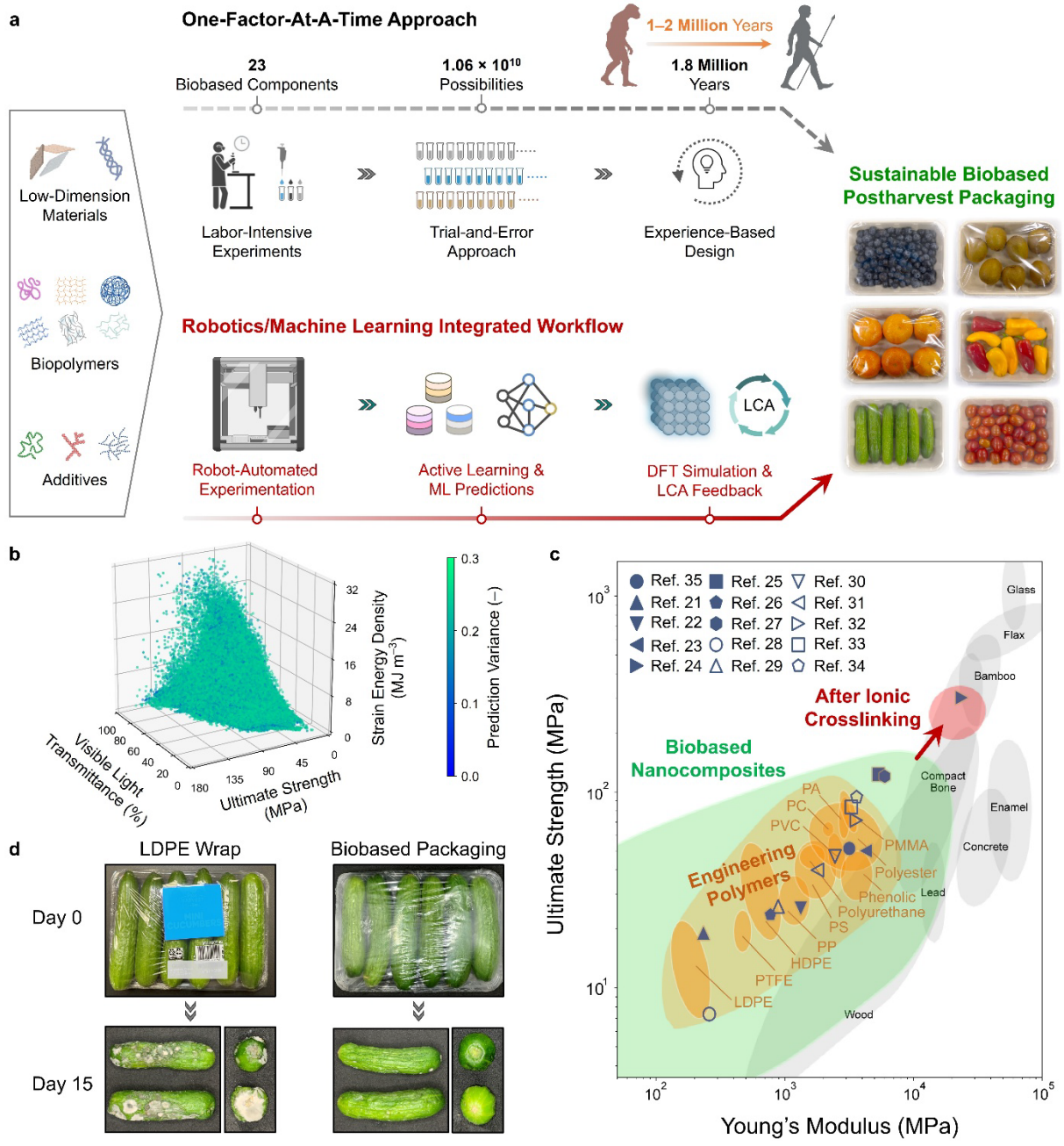
4.1 Motivation and introduction

The food packaging industry has long relied on engineered polymers, such as polystyrene (PS), polyvinyl chloride (PVC), low-density polyethylene (LDPE), high-density polyethylene (HDPE), and layered composites for their transparency, flexibility, processability, hydrostability, and oxygen barrier properties[77,78]. However, their production is highly energy-intensive, significantly contributing to greenhouse gas emissions and environmental concerns[79,80]. In 2020, PVC manufacturing in the United States alone generated approximately 18 million metric tons of CO_2[77]. Post-consumer plastic waste further exacerbates the environmental crisis beyond the manufacturing stage, with over 85% accumulating in landfills and oceans, polluting terrestrial and marine ecosystems, and driving microplastic contamination[81-83]. Compounding these challenges, conventional plastic packaging lacks inherent antimicrobial properties, limiting its ability to prevent microbial growth and foodborne contamination during storage[84,85]. Incorporating antimicrobial functionality into food packaging presents a promising strategy to extend shelf life, reduce food waste, and enhance supply chain security and safety[86,87].

Biobased materials have emerged as sustainable alternatives in the food packaging industry, offering abundance, renewability, home compostability, and functional advantages, such as moisture absorption and inherent antimicrobial properties that actively suppress microbial growth[88-91]. The transition from petrochemical-based plastics to biobased alternatives

presents a viable strategy to reduce the carbon footprint of packaging production, curb plastic pollution, and mitigate microplastic contamination[92,93]. However, discovering and designing biobased packaging with optimized properties remains challenging and time-consuming due to three key obstacles. First, the vast material library, populated with diverse natural and Generally Recognized As Safe (GRAS) components, poses a significant challenge for systematic exploration. Traditional methods, such as one-factor-at-a-time (OFAT) approaches and design of experiments (DoE), are labor-intensive and limit formulation design to only a few component combinations, hindering efficient optimization of biobased packaging with targeted properties[91,94]. Second, biobased packaging must meet multiple performance criteria – including tunable transparency, mechanical resilience, compostability, and scalable production – necessitating iterative optimization and adding substantial design complexity[84,91]. Third, conventional simulation tools (e.g., molecular dynamics) require extensive computational resources to accurately model complex nanocomposite structures, limiting comprehensive formulation screening and multi-property optimization[95-97]. These challenges underscore the urgent need for a streamlined, data-driven workflow that can accelerate nanocomposite design, minimize reliance on trial-and-error experimentation, and enable efficient multi-property optimization, ultimately advancing the development of sustainable biobased packaging solutions.

Figure 26. Integrated workflow to accelerate the discovery of sustainable biobased packaging for enhanced postharvest preservation with reduced environmental footprint



(a) Schematic illustration to compare the conventional OFAT approach and the robotics/ML integrated workflow for discovering sustainable biobased packaging. Conventional OFAT approach relies heavily on labor-intensive experiments with trial-and-error learning, making it 8 inefficient and time-consuming. For instance, with 23 natural/GRAS components in the materials library, exploring all possible formulations (with a 2 wt.% step size) would necessitate 1.06 × 1010 experiments and require about 1.8 million years (assuming 30 minutes per experiment and working 8 hours a day), a timescale close to the 1.8 million years of human evolution from

Homo species. In contrast, this work integrates robot-automated experimentation, active learning, ML-enabled predictions, DFT simulations, and LCA-informed feedback to efficiently develop sustainable biobased packaging. Right inset shows the digital photograph of various postharvest produce (including cucumbers, blueberries, mandarin oranges, cherry tomatoes, peppers, and kiwifruits) wrapped in the model-discovered biobased packaging films.
(b) A 3D heatmap visualizes the model-predicted T_{Vis}, σ_u, and SED values of ~0.15 billion formulations of biobased nanocomposites. The color gradient represents the prediction variance within the model committee, with green colors indicating higher prediction variance and blue colors indicating lower variance. All predicted properties exhibit low variances (<30%), ensuring reliable predictions.
(c) Ashby diagram ($\sigma\sigma uu$ vs. EE) for engineering polymers and our nanocomposite system. The prediction model can suggest a library of biobased nanocomposites with programmable $\sigma\sigma uu$ and EE values, which well cover and outperform the mechanical properties of conventional food packaging materials, such as PS, PVC, LDPE, and HDPE.
(d) Comparative photographs of postharvest cucumbers stored at 4 °C for 15 days, with conventional LDPE wraps and biobased packaging films. The biobased films exhibit superior antimicrobial properties and preservation efficacy, effectively suppressing mold and bacterial growth compared to LDPE wraps.

4.2 Defining the design space of biobased nanocomposites via an automated pipetting robot and an ANN classifier

The robotics/ML-integrated workflow consisted of four key phases: (1) defining the design space of biobased nanocomposites, (2) exploring the design space through active learning loops, (3) constructing a prediction model using a data-augmented, ensemble modeling strategy, and (4) suggesting biobased nanocomposite formulations via reverse design processes. By incorporating active learning, data augmentation, and ensemble modeling, this workflow enhances predictive accuracy and addresses challenges associated with small experimental datasets.

In the first phase, an automated pipetting robot (OT-2 robot) was programmed to prepare 2,120 different aqueous mixtures using a diversity sampling strategy. Each mixture contained any five components selected from a materials library of 23 building blocks at varying ratios. The OT-2 robot is both precise and efficient, and successfully prepared 1,060 mixtures within 64

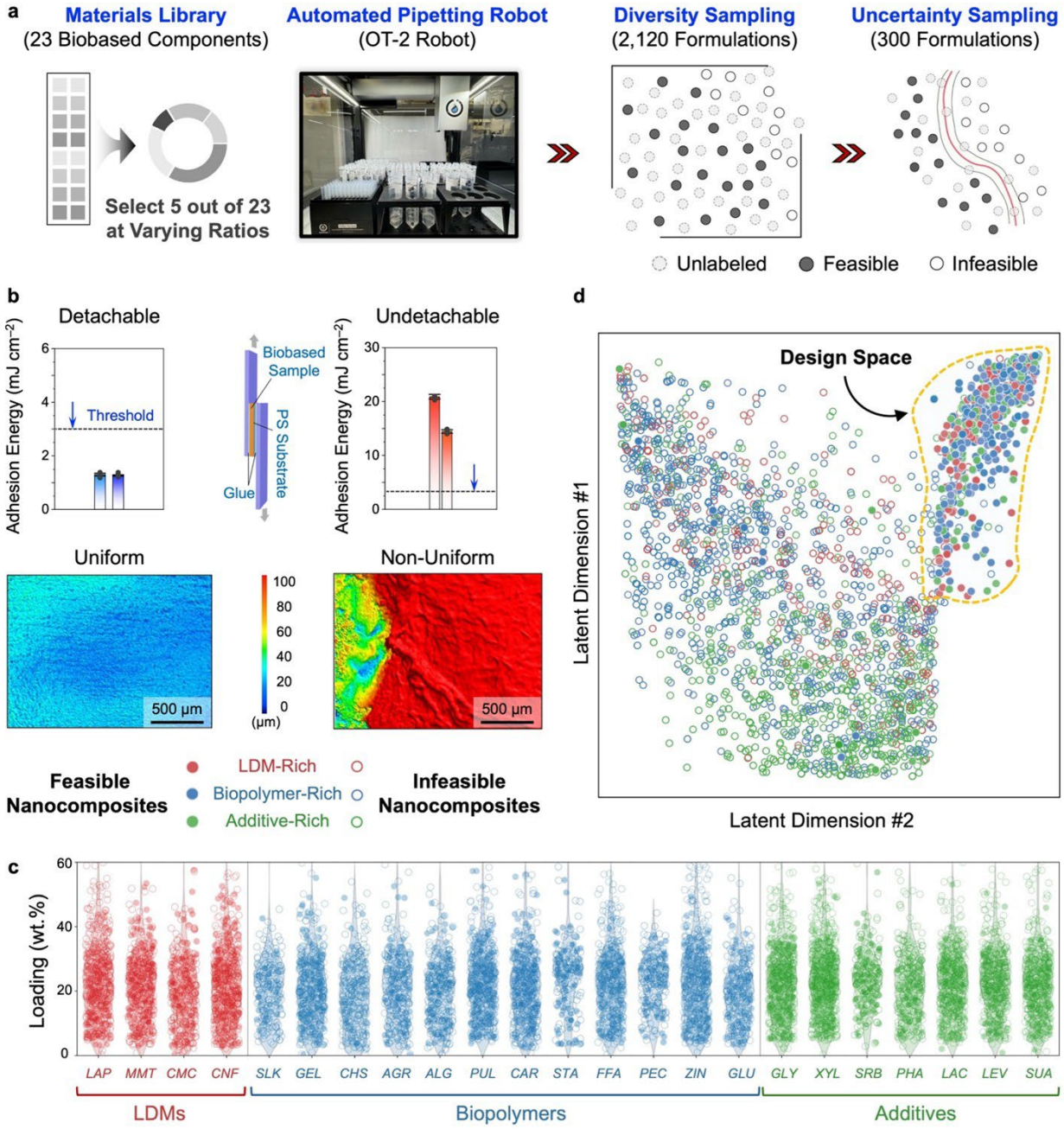
machine hours for 7 days. Then these mixtures were drop-cast into hydrophobic PS boats and air-dried overnight at 40 ºC. These nanocomposites then underwent film detachability and surface uniformity evaluation. Nanocomposites with low adhesion energy to PS substrates (<3 mJ cm^{-2}) and smooth surfaces (roughness <10 μm) were classified as "feasible," while those failing to meet these criteria were categorized as "infeasible." Through diversity sampling, 465 feasible and 1,655 infeasible nanocomposites were identified.

Using this experimental dataset, a preliminary ANN classifier was trained via 5-fold cross-validation. The classifier then conducted uncertainty sampling, identifying the formulations near the high-variance regions where their classification labels were ambiguous. The OT-2 robot then prepared an additional 300 mixtures, which underwent the same casting, drying, and film evaluation processes. Though uncertainty sampling, 126 formulations were classified as feasible and 174 were infeasible, which were added into the dataset. The ANN classifier was then re-trained to improve its prediction accuracy in distinguishing feasible and infeasible formulations. When evaluated on an independent testing dataset, the ANN classifier achieved a low testing error of 17.5%.

Fig. 27c presents the distribution profile of 2,240 biobased nanocomposites across 23 components, collected through diversity and uncertainty sampling. Each component was incorporated into ~700 samples of different nanocomposites across a wide range of loadings from 0% to 70%, with the even distribution of sample numbers across 23 components, demonstrating an extensive and systematic exploration of possible formulations. As shown in **Fig. 27d**, the ANN classifier generated a dimension-reduction visualization map that effectively clustered 2,420 nanocomposite samples into two distinct subregions. The subregion containing the majority of feasible samples was defined as a design space, which contained formulations

predicted to consistently yield high-quality (detachable and uniform) nanocomposite films. This map was generated from the ANN classifier by extracting outputs from the penultimate hidden layer of the ANN classifier, allowing high-dimensional formulation data projected into a 2D latent space. This approach enabled a clear visual representation of nanocomposite feasibility. Unlike data-rich systems that could tolerate experimental failures, re-designing and fabricating biobased nanocomposites was a time- and labor-intensive process, especially when suggested formulations yielded undetachable or rough films. Therefore, implementing an ANN classifier as a screening layer was crucial for filtering out infeasible formulations before experimental validation. For subsequent active learning loops, we exclusively focused on sampling within the design space.

Figure 27. Defining the design space of biobased nanocomposites via an automated pipetting robot and an ANN classifier



(a) Schematic illustration of the robot-automated preparation of biobased nanocomposites by selecting any five components from a materials library of 23 building blocks (their abbreviations listed in **Table 1**) and mixing them at varying ratios. Through diversity and uncertainty sampling processes, 2,420 biobased nanocomposites are fabricated. After drop-casting and air drying, these nanocomposites are evaluated for film detachability and surface uniformity. Based on these assessments, 2,420 biobased nanocomposites are classified as either “feasible” or “infeasible.”

These discrete feasibility data are used to construct an ANN classifier, defining the design space of biobased nanocomposites.
(b) Representation of feasible and infeasible biobased nanocomposites. Feasible nanocomposites exhibit low adhesion low adhesion energy (<3 mJ cm^{-2}) and low surface roughness (<10 μm), whereas infeasible ones show high adhesion energy (>3 mJ cm^{-2}) or high surface roughness (>10 μm).
(c) Distribution profile of 2,420 biobased nanocomposites across 23 components. The plot illustrates the comprehensive investigation of possible formulations, with component loadings varying from 0% to 70% and even distribution of sample numbers across 23 components.
(d) Dimension-reduction visualization map clustering 591 feasible (solid dots) and 1,829 infeasible (empty dots) biobased nanocomposites into a 2D latent space. The dot color represents the predominant component category in each nanocomposite formulation: red indicates an LDM-rich formulation, blue indicates a biopolymer-rich formulation, and green indicates an additive-rich formulation. The yellow-enclosed sub-region defines a design space of biobased nanocomposites, containing formulations predicted to consistently yield high-quality films.

Table 1. Twenty-three natural and Generally Regarded As Safe components as building blocks for sustainable biobased packaging

LDMs	**Biopolymers**			**Additives**	
Laponite (LAP)	Silk (SLK)	Sodium Alginate (ALG)	Furfural (FFA)	Glycerol (GLY)	Lactic Acid (LAC)
Montmorillonite (MMT)	Gelatin (GEL)	Pullulan (PUL)	Pectin (PEC)	Xylitol (XYL)	Levulinic Acid (LEV)
Sodium Carboxymethyl Cellulose (CMC)	Chitosan (CHS)	Carrageenan (CAR)	Zein (ZIN)	Sorbitol (SRB)	Succinic Acid (SUA)
Cellulose Nanofiber (CNF)	Agarose (AGR)	Starch (STA)	Gluten (GLU)	Phytic Acid (PHA)	

4.3 Exploring the design space of biobased nanocomposites through active learning loops

In the second phase (Fig. 28a), multiple active learning loops were conducted to iteratively sample, fabricate, and characterize biobased nanocomposite films within the design space. This iterative approach enabled the progressive development of a structured database, which served as the foundation for training ML-enabled prediction models. During the active learning loops, an ANN model was developed as the navigator, optimizing two key objectives:

maximizing σu and SED. To achieve these objectives, two acquisition functions were defined to compute the information indices for unlabeled data points within the design space, as shown in **Eqn. 3** and **Eqn. 4**,

Equation 3. Information index for ultimate strength

$$\hat{L}^n \times \widehat{\sigma_u}^m$$

Equation 4. Information index for strain energy density

$$\hat{L}^n \times \widehat{SED}^m$$

where $\hat{L}$ was the Euclidean distance between in-model and model-targeted formulation labels, $\widehat{\sigma_u}$ and $\widehat{SED}$ was the model-predicted σ_u and SED values, respectively, and the parameters *n* and *m* were dynamically adjusted across different active learning loops. In each loop, the navigator model used their information indices to rank unlabeled data points within the design space, prioritizing those with the highest rankings for experimental validation. Based on the model-suggested formulation labels, the OT-2 robot was then activated to prepare a new batch of aqueous mixtures.

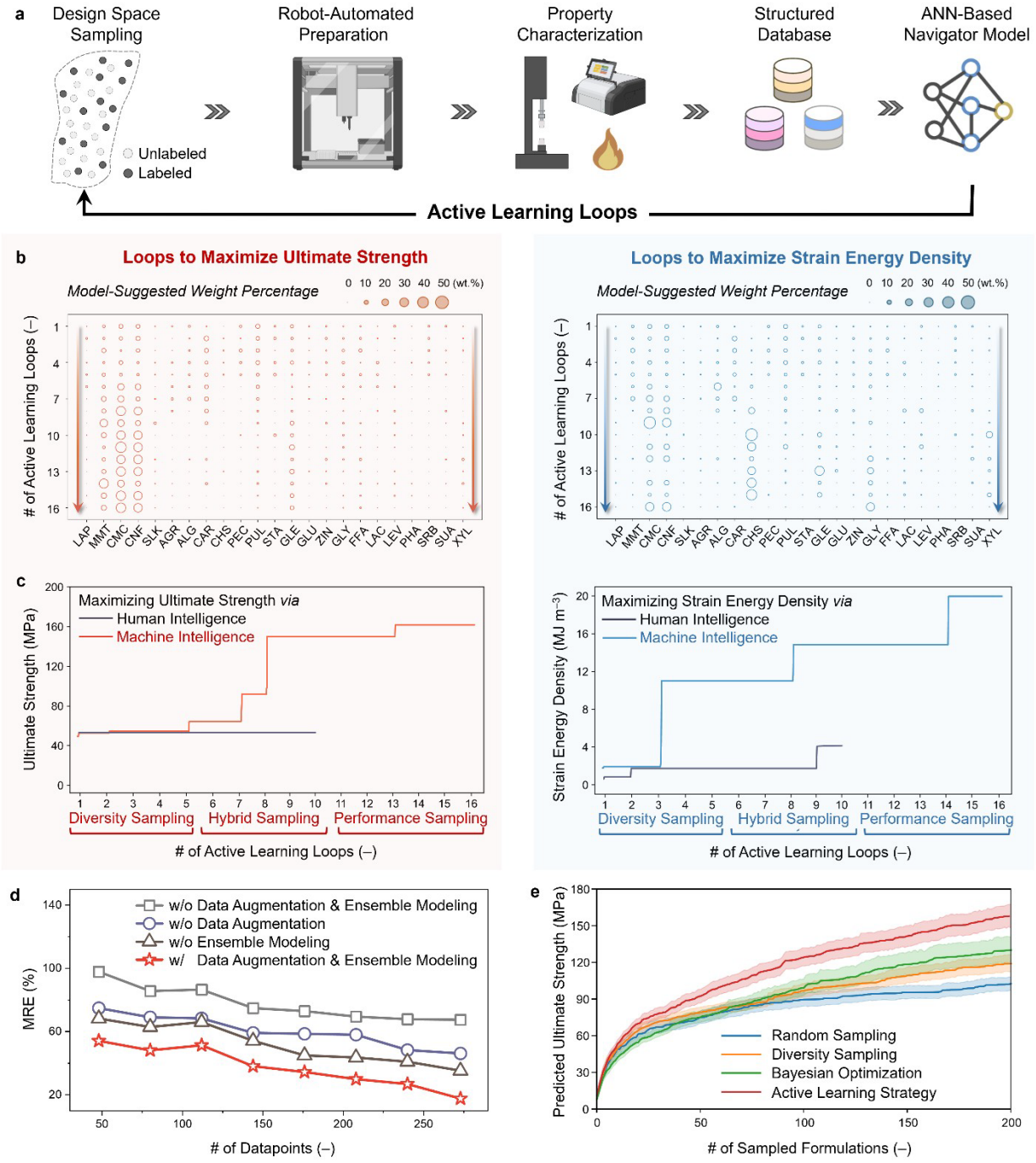
After cast drying, the resulting nanocomposite films underwent a series of mechanical, optical, and fire-resistant characterizations. Mechanical properties were assessed through tensile testing, extracting four key parameters from stress-strain curves as the "mechanical" labels - σ_u, fracture strain (ε_f), E, and SED. Optical properties were evaluated by measuring the light transmittance value at 365, 550, and 950 nm, recorded as the "spectral" labels – T_{UV}, T_{Vis}, and T_{IR}, respectively. Fire resistance was characterized using a modified ASTM D6413 fire test, with the pre- and post-test area ratio documented as the "fire" label – RR. Each nanocomposite film contributed one data point, comprising formulation labels (independent parameters), along with four mechanical, three spectral, and one fire label (dependent parameters).

By dynamically adjusting the exponent values (*n* and *m*) in the acquisition functions, the navigator model dynamically adjusted its sampling strategies to explore the design space. In the initial phase (first to fifth loops), *n* and *m* were set to 1 and 0, respectively, prioritizing diversity sampling to ensure an even distribution of sample numbers across all 23 components. In the second phase (sixth to tenth loops), both *n* and *m* were set to 1, enabling the navigator model to adopt a hybrid strategy that balance diversity sampling with σ_u and SED maximization. In the final phase (eleventh to sixteenth loops), *n* and *m* were set to 0.25 and 1, respectively, shifting the optimization focus toward maximizing the σ_u and SED of biobased nanocomposites. Over 16 loops, 343 nanocomposite films were fabricated and characterized.

Fig. 28b presents two distribution maps showing the average weight percentages for 23 components suggested by the navigator model across 16 active learning loops. The left map represents the active learning loops to maximize the σ_u of biobased nanocomposites, while the right map depicts the loops to maximize SED. In these maps, circle diameters correspond to the model-suggested weight percentages, with larger diameters indicating higher recommended loadings. To maximize σ_u, the navigator model identified a synergistic interaction between MMT and CNF, and sodium carboxymethyl cellulose (CMC). In the later active learning loops, the model consistently recommended the nanocomposite formulations with cumulative MMT/CNF/CMC loading exceeding 70% wt.%. In contrast, to maximize SED, the model identified two distinct formulation types. The first was a chitosan (CHS)-rich formulation (>60 wt.%), supplemented with small amounts of gelatin (GEL) and glycerol (GLY). The second was a CNF/CMC-based formulation (cumulative CNF/CMC loadings exceeding 50 wt.%), and GLY was incorporated to enhance the mechanical resilience in biobased nanocomposites.

Fig. 38c illustrates the characterized σ_u and SED values of nanocomposite films fabricated during active learning loops. From the first to fifth loops (diversity sampling), the navigator model prioritized exploring a broad range of component combinations, and the σ_u and SED values were limited to 54.6 MPa and 11.2 MJ m^{-3}, respectively. From sixth to tenth loops (hybrid sampling), the navigator model began to identify several formulations with synergistic component combinations, significantly improving σ_u to 150.2 MPa and SED to 14.5 MJ m^{-3}. From eleventh to sixteenth loops (σ_u and SED optimization), the navigator model focused on fine-tuning the component ratios within high-performing formulations, further elevation σ_u to 161.7 MPa and SED to 20.0 MJ m^{-3}.

Figure 28. Developing a prediction model for biobased nanocomposites using active learning, data augmentation, and ensemble modeling



(a) Schematic illustration of active learning loops to sample the design space of biobased nanocomposites and collect the experimental data points systematically, establishing a structured database as the foundation for prediction model development. The navigator model aims to maximize the σ_u and SED of biobased nanocomposites using two distinct acquisition functions. Over 16 active learning loops, 343 nanocomposite films are fabricated with the OT-2 robot, and

their mechanical, optical, and fire-resistant properties are characterized. Using data augmentation and ensemble modeling, a prediction model comprising a committee of ANNs is established to accelerate the discovery of biobased nanocomposites with programmable properties.
(b) Mean weight percentages of 23 natural/GRAS components recommended by the navigator model during 16 active learning loops, with the objectives to maximize the σ_u (left inset) and SED values (right inset) of biobased nanocomposites. A larger circle represents a higher weight percentage of a specific component suggested by the navigator model, while a smaller circle represents a lower recommended component loading.
(c) Solid dots represent the characterized σ_u and SED values of biobased nanocomposites at different phases of active learning loops. Solid lines represent the highest σ_u and SED values of biobased nanocomposites within a structured databased across active learning loops.
(d) MRE values for different prediction models with and/or without data augmentation and ensemble modeling (evaluated using an independent set of testing data points).
(e) Comparison between active learning strategy with multiple sampling methods, including random sampling, diversity sampling, and Bayesian optimization.

4.4 Constructing a prediction model using data augmentation and ensemble modeling

After active learning loops, experimental data were used to construct a prediction model by sequentially integrating data augmentation and ensemble modeling techniques. To enhance prediction accuracy and minimize potential overfitting, an in silico data augmentation method, known as the User Input Principle (UIP), was applied, which posits that minor formulation adjustments result in similar mechanical, spectral, and fire-retardant properties. This is consistent with experimental measurements which show variations of ~5% for spectral labels, ~3% for fire-retardant labels, and ~15% for mechanical labels. To account for these variabilities, gaussian noises were introduced around the formulation, mechanical, spectral, and fire-retardant labels, synthesizing virtual data points to expand the dataset. The optimal virtual-to-real data ratio was determined to be 100:1, maximizing the model's learning efficiency while maintaining the model training time of under three days per loop.

Next, using the augmented dataset, an ensemble modeling strategy was implemented to train a pool of ANN variants with differing hidden layer numbers, node configurations, and activation functions. These ANN variants were evaluated through 5-fold cross-validation, and

their validation errors were used to rank prediction accuracy. The five top-performing ANN variants were selected to form an ensemble committee, which average their outputs to generate more generalized predictions. The accuracy of the prediction model was assessed using an independent testing dataset, which was not included in the training process. The deviation between model-predicted and actual property values was quantified in terms of MRE, as defined in **Eqn. 5**,

Equation 5. Mean relative error (MRE)

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{output^i - E^i}{E^i}\right|$$

where N is the total number of testing data, $output^i$ is the model-predicted property labels for a given testing datum (i), E_i is the corresponding actual property values for a given testing datum (i). A lower MRE value indicated higher prediction accuracy, while a higher MRE value signified a lower accuracy. As shown in **Fig. 28d**, after 16 active learning loops, the MRE value was reduce to ~16%, close to the observed measurement variations (e.g., ~15% for the mechanical labels). Compared to other prediction models lacking data augmentation and/or ensemble modeling, this approach exhibited higher learning efficiency, developing a prediction model with the lowest testing error. These results highlight the effectiveness of integrating data augmentation with ensemble modeling in developing an accurate prediction model, particularly for small experimental datasets, where limited data often constrains model performance.

Fig. 28e compares the active learning strategy with multiple sampling methods, including random sampling, diversity sampling, and Bayesian optimization. A high-accuracy prediction model with low testing errors served as an independent oracle, providing model-predicted σ_u values for each selected formulation. Each sampling process began with a randomly selected set of 40 formulations, followed by subsequent formulation selections based on the model-predicted

σ_u values. Each sampling method was limited to 200 formulations per process and iterated 256 times using different initial formulation sets to ensure a fair performance assessment. The solid curves in **Fig. 28e** represent the average σ_u values from 256 iterations, while the shaded regions indicate the variability in predicted averages across 256 iterations. Under the constraint of 200 formulations, random and diversity sampling methods exhibited the slowest improvements in σ_u, plateauing around 100 MPa and 120 MPa respectively. Bayesian optimization was similarly limited, with an average σ_u value around 130 MPa, likely due to the vast design space (with 23 independent parameters) and the relatively small initial dataset (40 formulations). These conditions restricted the ability of Bayesian optimization to effectively identify global maxima and minima, and similar limitations have been reported in the literature, particularly in complex, high-dimensional design spaces[98]. In contrast, the active learning strategy, employing dynamic acquisition functions, consistently outperformed all other sampling methods, achieving the highest σ_u value of 157.9 MPa based on 200 samples, closely aligning with experimental results. This superior performance was attributed to a versatile sampling strategy, which initially prioritized diversity sampling to explore the large design space, followed by property optimization to ensure the global maxima of targeted properties were not overlooked.

4.5 Development of an open-access data-sharing platform for sustainable biobased nanocomposites

The widespread adoption of sustainable biobased nanocomposites faces several key challenges including (1) the lack of comprehensive, high-quality experimental datasets, (2) inefficient dissemination mechanisms that limit collaboration among diverse stakeholders, and (3) absence of accessible data platforms and user-friendly visualization tools. To address these

barriers, we established a public-access platform that has compiled approximately 1 billion formulations of biobased nanocomposites, along with their model-predicted properties, including film quality T_{Vis}, σ_u, and SED.

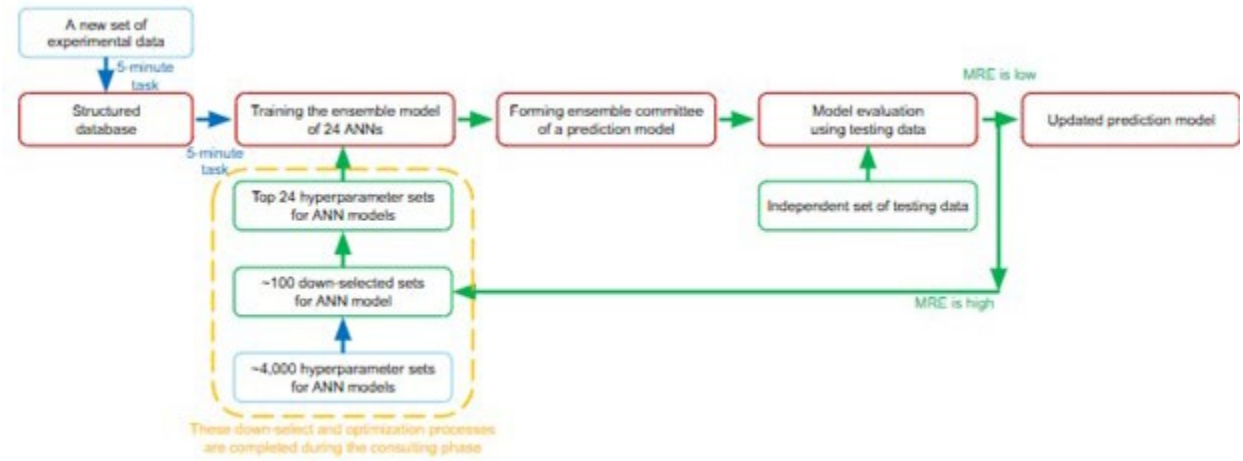
The data-sharing platform features two key functionalities: forward design and reverse engineering. In the forward design section (**Fig. 29a**), users can select a set of any five components at varying ratios, and the platform will utilize its embedded prediction model to generate predictions for film quality, T_{Vis}, σ_u, and SED values when the film quality is predicted to be high and the prediction variance is below 30%, ensuring reliable outputs. In the reverse engineering section (**Fig. 29b**), users can define targeted property criteria (T_{Vis}, σ_u, and/or SED), prompting the platform to perform cluster analyses using the embedded prediction model. The platform will suggest multiple most suitable formulations, allowing users to interactively optimize formulations within the recommended loading range for each selected component. The platform's backend is currently deployed on a local server, integrating the ML-enabled prediction model with an expandable database. This setup facilitates seamless real-time model training and updates, ensuring continuous improvement in predictive accuracy. As the community grows, the platform can be migrated to a secure, scalable cloud-based infrastructure, supporting continuous data expansion, iterative model enhancements, and broader accessibility.

Figure 29. User interface of open-access data-sharing platform

(a) Open-access platform fosters a data-driven materials science community, accelerating the adoption of ML/AI technologies and driving innovation. ML models are hosted on backend servers which embed model-driven capabilities onto the front-end platform. In this way users can be presented with an intuitive and user-friendly application and have the power of model inference without having to understand or worry about the intricacies of developing, implementing, and deploying these models.
(b) The platform offers two main views, Forward Prediction and Inverse Design. In the forward prediction feature, users can specify a formulation consisting of 4-5 components and request property labels using the embedded model. Using the inverse design feature, based on targeted property requirements and selected materials, formulations can be proposed that exhibit those targeted properties (subject to experimental validation).

The open-access data-sharing platform was built using the Next.js React framework. It relies on remote servers to interact with machine learning models and render two primary views: (1) forward prediction and (2) inverse design. This design choice ensures that the front end remains lightweight, without incorporating any computationally intensive processes. Page state is saved across the platform using a Redux store. The platform uses the Redux Persist library to save the Redux store to the client's local storage to preserve user data between sessions. A Persist Gate component delays rendering the UI until the persisted state has been retrieved and rehydrated into the Redux store, ensuring a seamless user experience.

Figure 30. Platform model training data flow



This flow diagram demonstrates the current data flow from the backend system when model retraining is enacted. New experimental data is fetched and concatenated to a structured training database. 24 sets of model hyperparameters have been previously cached as "champion configuration" on the backend, filtered down from over 4,000 configurations during the model training process. These configurations will vary in layer number, hidden layer sizes, learning rate, regularization parameters, and random initialization (random seed). An ensemble of 5-10 of these models is then selected using k-fold cross-validation, by selecting a champion model on each fold of the training dataset from the original 24. This ensemble forms a voting committee, with improved prediction accuracy and generalizability compared to any single ANN network. The updated ensemble prediction model is then evaluated against an independent testing dataset, which is separate from the training data and was not used during the training phase. If the mean relative error (MRE) or mean absolute error (MAE) exceed acceptable thresholds, the process loops back to the committee down selection phase, requiring retraining. If the updated ensemble meets the required MRE and MAE criteria, its weights are hot swapped into the backend servers for inference.

The first feature the platform offers is Forward Prediction. Users can directly run inference on prediction models trained on high-quality, structured, experimental databases containing a diverse-component library. The platform balances ease of use with the flexibility to conduct large-scale experiments while ensuring data is presented in a way that minimizes misuse or misinterpretation (e.g., avoiding high-uncertainty formulations).

The Forward Prediction interface features an intuitive data input layout designed for nanocomposite formulation prediction. Users can make single-point predictions, receiving outputs from the trained model based on their input data. Additionally, all prediction results are logged, allowing users to track and record their past queries.

The second feature the platform offers is Inverse Design. The platform generates large synthetic datasets using trained prediction models, enabling users to identify formulations that meet specific property requirements. These datasets are constructed by sampling input features from a Gaussian distribution while ensuring constraints are met (e.g., material compositions summing to 100%). Output labels are then generated using the trained prediction model.

Since the dataset reflects the state of the trained model, it must be resampled or reconstructed whenever the model is retrained or updated. Through the platform interface, users can specify up to three required materials and define target property ranges. These inputs are translated into a SQL query, filtering viable samples from an optimized SQL database hosted on a remote ClickHouse server. The platform then presents trends among the identified formulations.

As shown in **Fig. 31**, both features are accessible via the platform's frontend using strictly typed and validated API endpoints. Communication with backend servers occurs through HTTP requests and responses.

Figure 31. Propagation of updated prediction model to the frontend platform

When the state of the trained prediction model changes (e.g., due to retraining), several tasks are initiated to ensure consistency across the platform. Static statistical analyses, such as computing SHAP values and generating violin plots to visualize feature distributions, must be recomputed and published. The prediction endpoint is updated with the new model parameters and is then ready for use. To update the cluster endpoint, formulations must be resampled and stored on the backend using the updated model. Once this process is complete, the cluster endpoint can resume communication with the frontend.

Chapter 5: Conclusion and perspective

5.1 Fundamental advancements to the field

Through my work, I have introduced several novel concepts and methodologies to the fields of chemical engineering and materials science.

First, I developed an experimental framework that enables the integration of machine learning and artificial intelligence with materials discovery and design research. This includes implementing sampling strategies such as active learning to efficiently sample from unlabeled formulations, maximizing the experimental efficiency of resource-intensive experiments. Additionally, I have compiled a suite of tools and model-training techniques that enable the successful implementation of machine learning models, both predictive and generative, in a data-scarce domain.

Second, I have contributed to the development of autonomous characterization techniques, reducing the manual workload for researchers by delegating tasks to collaborative robots. These systems enhance experimental throughput while ensuring the reliability and reproducibility of data. The successful implementation of such platforms demands expertise beyond traditional disciplines like transport phenomena, kinetics, and thermodynamics, emphasizing the value of cross-disciplinary researchers and students.

Finally, I have contributed to a data-driven ecosystem within the materials science community by developing open-access data-sharing platforms. For researchers to buy into machine learning, they need accessible tools that demonstrate its value. While the availability of comprehensive experimental datasets could significantly advance the field, this will not happen

until there is collective buy-in throughout the research community, which I have done my small part to promote.

5.2 Further work

I believe the next step for this work is to develop foundation models for nanocomposite design that are material-agnostic and not confined to specific design spaces. This requires creating informative material embeddings—a novel way to represent material recipes without relying on direct composition input. Several methods, such as Mol2Vec, ChemBERT, molecular fingerprinting algorithms, as well as natural language embeddings offer potential solutions, and we need to systematically evaluate them to determine the most effective approach for representing complex nanomaterial building blocks and sample preparation methods.

To assess the viability of these representations, we must establish a robust testing framework with two key criteria. First, the new approach should perform comparably to existing artificial neural network-based modeling approaches, which perform direct composition-based predictions. Second, it must demonstrate the capability to extrapolate beyond its training data, maintaining accuracy even when certain materials are excluded from the training set but included in the test set.

Our current modeling approaches lack the granularity to predict system changes at a more atomic level, for instance, to predict the performance impact of functionalizing chitosan within our system. A more advanced representation framework would allow us to directly model these fine-scale alterations, significantly enhancing our ability to design new materials.

Another application of this approach is in driving the adoption of biobased plastic substitutes, an area that I am particularly passionate about. A current challenge is the high cost of

material building blocks, which limits their ability to compete with petrochemical staples. A model capable of identifying cost-effective replacements for expensive components in champion formulations without degrading performance would help address this issue.

By advancing material representation and predictive capabilities, we can drive the development of scalable, cost-effective nanocomposites with broader applications.

5.3 Recommendations

Machine learning is poised to revolutionize the chemical engineering and material science fields. As I have demonstrated, even without deep domain expertise or an intrinsic understanding of atomic synergies in nanomaterial assembly, I can programmatically curate champion formulations with fewer experiments and in less time.

These benefits extend far beyond nanomaterial composites. Machine learning will redefine how process engineers design heat exchangers, how control engineers optimize reactions, and how clinical studies are conducted and approved.

Integrating machine learning fundamentals into the undergraduate chemical engineering curriculum at the University of Maryland is essential. This doesn't mean building models with toy data—it requires strengthening core foundations, particularly Linear Algebra, which is currently absent from the curriculum. Additionally, replacing MATLAB and MathCAD with more functional programming languages, in particular Python, will provide future chemical engineers with a significant competitive edge.

In this new era, chemical engineers who can quickly grasp machine learning concepts will be invaluable.

References

1 Liu, Y., Zhao, T., Ju, W. & Shi, S. Materials discovery and design using machine learning. *Journal of Materiomics* **3**, 159, (2017).

2 Tabor, D. P. *et al.* Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature Reviews Materials* **3**, 5, (2018).

3 Pyzer-Knapp, E. O. *et al.* Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials* **8**, 84, (2022).

4 Rajan, K. Materials informatics. *Materials Today* **8**, 38, (2005).

5 Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials* **5**, 21, (2019).

6 Fang, J. *et al.* Machine learning accelerates the materials discovery. *Materials Today Communications* **33**, 104900, (2022).

7 Juan, Y., Dai, Y., Yang, Y. & Zhang, J. Accelerating materials discovery using machine learning. *Journal of Materials Science & Technology* **79**, 178, (2021).

8 Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624**, 80, (2023).

9 Shevlin, M. Practical high-throughput experimentation for chemists. *ACS Medicinal Chemistry Letters* **8**, 601, (2017).

10 Szymanski, N. J. *et al.* An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86, (2023).

11 Jain, A., Shin, Y. & Persson, K. A. Computational predictions of energy materials using density functional theory. *Nature Reviews Materials* **1**, 15004, (2016).

12 Curtarolo, S. *et al.* Aflowlib.Org: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science* **58**, 227, (2012).

13 Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547, (2018).

14 Hestness, J. *et al.* Deep learning scaling is predictable, empirically. (2017).

15 Hoffmann, J. *et al.* Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets. *Science Advances* **5**, eaau6792.

16 Baghel, R. S., Reddy, C. R. K. & Singh, R. P. Seaweed-based cellulose: Applications, and future perspectives. *Carbohydrate Polymers* **267**, 118241, (2021).

17 Tsang, Y. F. *et al.* Production of bioplastic through food waste valorization. *Environment International* **127**, 625, (2019).

18 Fourches, D., Muratov, E. & Tropsha, A. Trust, but verify: On the importance of chemical structure curation in cheminformatics and qsar modeling research. *Journal of Chemical Information and Modeling* **50**, 1189, (2010).

19 Xia, Q. *et al.* A strong, biodegradable and recyclable lignocellulosic bioplastic. *Nature Sustainability* **4**, (2021).

20 Jiang, B. *et al.* Lignin as a wood-inspired binder enabled strong, water stable, and biodegradable paper for plastic replacement. *Advanced Functional Materials*, (2019).

21 Ward, L. *et al.* Strategies for accelerating the adoption of materials informatics. *MRS Bulletin* **43**, 683, (2018).

22 Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: Recent applications and prospects. *npj Computational Materials* **3**, 54, (2017).
23 Althnian, A. *et al.* Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences* **11** (2021).
24 Ying, X. An overview of overfitting and its solutions. *Journal of Physics: Conference Series* **1168**, 022022, (2019).
25 Li, J. *et al.* Ai applications through the whole life cycle of material discovery. *Matter* **3**, 393, (2020).
26 Yang, H. *et al.* Automatic strain sensor design via active learning and data augmentation for soft machines. *Nature Machine Intelligence* **4**, 84, (2022).
27 Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nature Materials* **12**, 191, (2013).
28 Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials* **15**, 1120, (2016).
29 MacLeod, B. P. *et al.* Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances* **6**, eaaz8867, (2020).
30 Gómez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science* **4**, 268, (2018).
31 Lavecchia, A. & Di Giovanni, C. Virtual screening strategies in drug discovery: A critical review. *Current Medicinal Chemistry* **20**, 2839, (2013).
32 Cole, J. M. A design-to-device pipeline for data-driven materials discovery. *Accounts of Chemical Research* **53**, 599, (2020).
33 Stegle, O., Payet, L., Mergny, J.-L., MacKay, D. J. C. & Huppert, J. L. Predicting and understanding the stability of g-quadruplexes. *Bioinformatics* **25**, i374, (2009).
34 Casciato, M. J., Kim, S., Lu, J. C., Hess, D. W. & Grover, M. A. Optimization of a carbon dioxide-assisted nanoparticle deposition process using sequential experimental design with adaptive design space. *Industrial & Engineering Chemistry Research* **51**, 4363, (2012).
35 Eyke, N. S., Green, W. H. & Jensen, K. F. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *Reaction Chemistry & Engineering* **5**, 1963, (2020).
36 Pan, Y., Jing, Y., Wu, T. & Kong, X. Knowledge-based data augmentation of small samples for oil condition prediction. *Reliability Engineering & System Safety* **217**, 108114, (2022).
37 Oviedo, F. *et al.* Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials* **5**, 60, (2019).
38 Mumuni, A. & Mumuni, F. Data augmentation: A comprehensive survey of modern approaches. *Array* **16**, 100258, (2022).
39 Qavi, I. & Tan, G. Harnessing interpretable and ensemble machine learning techniques for precision fabrication of aligned micro-fibers. *Manufacturing Letters* **41**, 364, (2024).
40 Noble, W. S. What is a support vector machine? *Nature Biotechnology* **24**, 1565, (2006).
41 Mueller, T., Kusne, A. & Ramprasad, R. 186 (2016).

42 Chen, T. *et al.* Machine intelligence-accelerated discovery of all-natural plastic substitutes. *Nature Nanotechnology* **19**, 782, (2024).
43 Murphy, R. F. An active role for machine learning in drug development. *Nature Chemical Biology* **7**, 327, (2011).
44 Fu, X., Cheng, W., Wan, G., Yang, Z. & Tee, B. C. K. Toward an ai era: Advances in electronic skins. *Chemical Reviews* **124**, 9899, (2024).
45 Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *International Journal of Quantum Chemistry* **115**, 1032, (2015).
46 Cao, B. *et al.* How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics. *ACS Nano* **12**, 7434, (2018).
47 Chen, L.-Y. & Li, Y.-P. Uncertainty quantification with graph neural networks for efficient molecular design. *Nature Communications* **16**, 3262, (2025).
48 A, S. & R, S. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal* **7**, 100230, (2023).
49 Christoph, M. Interpretable machine learning: A guide for making black box models explainable. (2020).
50 Ghosh, A. Towards physics-informed explainable machine learning and causal models for materials research. *Computational Materials Science* **233**, 112740, (2024).
51 Lundberg, S. M. & Lee, S.-I. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768 (Curran Associates Inc., Long Beach, California, USA, 2017).
52 Shrestha, S. *et al.* Machine intelligence accelerated design of conductive mxene aerogels with programmable properties. *Nature Communications* **15**, 4685, (2024).
53 Saldaña, M. *et al.* Dynamic of mining systems: Impact of cl− ion concentration on heap copper leaching process at industrial scale. *Metals* **13** (2023).
54 Sun, J. *et al.* Bioleaching of copper-containing electroplating sludge. *Journal of Environmental Management* **285**, 112133, (2021).
55 Ren, Z. *et al.* Catalytic effect of ethylene thiourea on the leaching of chalcopyrite. *Hydrometallurgy* **196**, 105410, (2020).
56 Zhao, S. *et al.* Additive manufacturing of silica aerogels. *Nature* **584**, 387, (2020).
57 Liu, H., Liu, J. a., Tian, Y., Wu, X. & Li, Z. Investigation of high temperature thermal insulation performance of fiber-reinforced silica aerogel composites. *International Journal of Thermal Sciences* **183**, 107827, (2023).
58 Zimmermann, M. V. G. & Zattera, A. J. Silica aerogel reinforced with cellulose nanofibers. *Journal of Porous Materials* **28**, 1325, (2021).
59 Chen, Y. *et al.* Recent progress on nanocellulose aerogels: Preparation, modification, composite fabrication, applications. *Advanced Materials* **33**, e2005569, (2021).
60 Takeshita, S. & Yoda, S. Chitosan aerogels: Transparent, flexible thermal insulators. *Chemistry of Materials* **27**, 7569, (2015).
61 Tang, X. *et al.* Generalized 3d printing of graphene-based mixed-dimensional hybrid aerogels. *ACS Nano* **12**, 3502, (2018).
62 Peng, F. *et al.* Thermally insulating, fiber-reinforced alumina–silica aerogel composites with ultra-low shrinkage up to 1500 °c. *Chemical Engineering Journal* **411**, 128402, (2021).

63 Pandit, P., Abdusalamov, R., Itskov, M. & Rege, A. Deep reinforcement learning for microstructural optimisation of silica aerogels. *Scientific Reports* **14**, 1511, (2024).
64 Hu, D., Jiang, P. & Huang, X. Mixed-dimensional engineering of 3d mxene ultralight hybrid aerogel for anticorrosive and microwave absorption applications. *Composites Part A: Applied Science and Manufacturing* **156**, 106865, (2022).
65 Badini, S., Regondi, S. & Pugliese, R. Unleashing the power of artificial intelligence in materials design. *Materials (Basel)* **16**, (2023).
66 Yu, D., Liu, M., Xu, F., Kong, Y. & Shen, X. Structure tailoring and thermal performances of water glass-derived silica aerogel composite with high specific surface area and enhanced thermal stability. *Journal of Non-Crystalline Solids* **630**, 122889, (2024).
67 Tao, L., Byrnes, J., Varshney, V. & Li, Y. Machine learning strategies for the structure-property relationship of copolymers. *iScience* **25**, 104585, (2022).
68 Blanco-González, A. *et al.* The role of ai in drug discovery: Challenges, opportunities, and strategies. *Pharmaceuticals (Basel)* **16**, (2023).
69 Mazheika, A. *et al.* Artificial-intelligence-driven discovery of catalyst genes with application to co2 activation on semiconductor oxides. *Nature Communications* **13**, 419, (2022).
70 Li, J. *et al.* Research progress on the application of aerogels in atmospheric water harvesting. *Separation and Purification Technology* **363**, 132074, (2025).
71 Smirnova, I. & Gurikov, P. Aerogel production: Current status, research directions, and future opportunities. *The Journal of Supercritical Fluids* **134**, 228, (2018).
72 Malfait, W. J. *et al.* The poor reliability of thermal conductivity data in the aerogel literature: A call to action! *Journal of Sol-Gel Science and Technology* **109**, 569, (2024).
73 Walker, R. C., Hyer, A. P., Guo, H. & Ferri, J. K. Silica aerogel synthesis/process–property predictions by machine learning. *Chemistry of Materials* **35**, 4897, (2023).
74 Patil, S. P., Parale, V. G., Park, H.-H. & Markert, B. Mechanical modeling and simulation of aerogels: A review. *Ceramics International* **47**, 2981, (2021).
75 Zhao, Y. *et al.* Structural engineering of hierarchical aerogels comprised of multi-dimensional gradient carbon nanoarchitectures for highly efficient microwave absorption. *Nano-Micro Letters* **13**, 144, (2021).
76 Ronneberger, O., Fischer, P. & Brox, T. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* (eds Nassir Navab, Joachim Hornegger, William M. Wells, & Alejandro F. Frangi) 234 (Springer International Publishing).
77 MacLeod, M., Arp, H. P. H., Tekman, M. B. & Jahnke, A. The global threat from plastic pollution. *Science* **373**, 61, (2021).
78 Stegmann, P., Daioglou, V., Londo, M., van Vuuren, D. P. & Junginger, M. Plastic futures and their co2 emissions. *Nature* **612**, 272, (2022).
79 The future of petrochemicals. (IEA, 2018).
80 Geyer, R., Jambeck, J. R. & Law, K. L. Production, use, and fate of all plastics ever made. *Science Advances* **3**, e1700782.
81 Redondo-Hasselerharm, P. E., Rico, A. & Koelmans, A. A. Risk assessment of microplastics in freshwater sediments guided by strict quality criteria and data alignment methods. *Journal of Hazardous Materials* **441**, 129814, (2023).
82 Law, K. L. Plastics in the marine environment. *Annual Review of Marine Science* **9**, 205, (2017).

83 Malhotra, B., Keshwani, A. & Kharkwal, H. Antimicrobial food packaging: Potential and pitfalls. *Front Microbiol* **6**, 611, (2015).
84 Han, J.-W., Ruiz-Garcia, L., Qian, J.-P. & Yang, X.-T. Food packaging: A comprehensive review and future trends. *Comprehensive Reviews in Food Science and Food Safety* **17**, 860, (2018).
85 Xu, Z., Da-Wen, S., Xin-An, Z., Dan, L. & and Pu, H. Research developments in methods to reduce the carbon footprint of the food system: A review. *Critical Reviews in Food Science and Nutrition* **55**, 1270, (2015).
86 Ahmed, S. *et al.* Research progress on antimicrobial materials for food packaging. *Critical Reviews in Food Science and Nutrition* **62**, 3088, (2022).
87 Motelica, L. *et al.* Biodegradable antimicrobial food packaging: Trends and perspectives. *Foods* **9** (2020).
88 Upadhyay, A. *et al.* Bio-based smart packaging: Fundamentals and functions in sustainable food systems. *Trends in Food Science & Technology* **145**, 104369, (2024).
89 Tardy, B. L. *et al.* Advancing bio-based materials for sustainable solutions to food packaging. *Nature Sustainability* **6**, 360, (2023).
90 Sid, S., Mor, R., Kishore, A. & Sharanagat, V. Bio-sourced polymers as alternatives to conventional food packaging materials: A review. *Trends in Food Science & Technology* **115**, (2021).
91 Otoni, C. G. *et al.* The food–materials nexus: Next generation bioplastics and advanced materials from agri-food residues. *Advanced Materials* **33**, 2102520, (2021).
92 Law, K. L. & Narayan, R. Reducing environmental plastic pollution by designing polymer materials for managed end-of-life. *Nature Reviews Materials* **7**, 104, (2022).
93 Fei, X. *et al.* The long-term fates of land-disposed plastic waste. *Nature Reviews Earth & Environment* **3**, 733, (2022).
94 Zhao, X. *et al.* Sustainable bioplastics derived from renewable natural resources for food packaging. *Matter* **6**, 97, (2023).
95 Lau, D., Jian, W., Yu, Z. & Hui, D. Nano-engineering of construction materials using molecular dynamics simulations: Prospects and challenges. *Composites Part B: Engineering* **143**, 282, (2018).
96 Shukla, R. & Tripathi, T. in *Innovations and implementations of computer aided drug discovery strategies in rational drug design* (ed Sanjeev Kumar Singh) 295 (Springer Singapore, 2021).
97 Rezić, I. & Somogyi Škoc, M. Computational methodologies in synthesis, preparation and application of antimicrobial polymers, biomolecules, and nanocomposites. *Polymers* **16** (2024).
98 Binois, M. & Wycoff, N. A survey on high-dimensional gaussian process modeling with application to bayesian optimization. *ACM Transactions on Evolutionary Learning and Optimization* **2**, Article 8, (2022).