Lawrence Berkeley National Laboratory

Title

Toward implementing autonomous adaptive data acquisition for scanning hyperspectral imaging of biological systems

Permalink

https://escholarship.org/uc/item/92m6f1f5

Journal

ISSN

Authors

Holman, Elizabeth A
Krishnan, Harinarayan
Holman, Derek R
et al.

Publication Date

DOI

Copyright Information

Peer reviewed

Toward implementing autonomous adaptive data acquisition for scanning hyperspectral imaging of biological systems

Elizabeth A. Holman ✉ ; Harinarayan Krishnan ; Derek R. Holman ; Hoi-Ying N. Holman ;
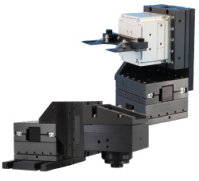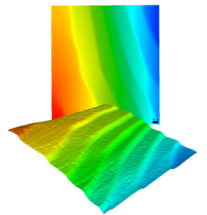Paul W. Sternberg

Check for updates
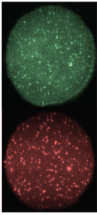
View Online   Export Citation   CrossMark

Toward implementing autonomous adaptive data acquisition for scanning hyperspectral imaging of biological systems Ⓕ ⓈⒸⒾ

View Online   Export Citation   CrossMark

Elizabeth A. Holman,[1,2,a] ⒾⒹ Harinarayan Krishnan,[2] ⒾⒹ Derek R. Holman,[3] ⒾⒹ Hoi-Ying N. Holman,[2] ⒾⒹ and Paul W. Sternberg[1] ⒾⒹ

AFFILIATIONS

[1]California Institute of Technology, Pasadena, California 91125, USA
[2]Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
[3]Stanford University, Stanford, California 94305, USA

[a]Author to whom correspondence should be addressed: eholman@alumni.caltech.edu

ABSTRACT

Autonomous experimentation is an emerging area of research, primarily related to autonomous vehicles, scientific combinatorial discovery approaches in materials science and drug discovery, and iterative research loops of planning, experimentation, and analysis. However, autonomous approaches developed in these contexts are difficult to apply to high-dimensional mapping technologies, such as scanning hyper-spectral imaging of biological systems, due to sample complexity and heterogeneity. We briefly cover the history of adaptive sampling algorithms and surrogate modeling in order to define autonomous adaptive data acquisition as an objective-based, flexible building block for future biological imaging experimentation driven by intelligent infrastructure. We subsequently summarize the recent implementations of autonomous adaptive data acquisition (AADA) for scanning hyperspectral imaging, assess how these address the difficulties of autonomous approaches in hyperspectral imaging, and highlight the AADA design variation from a goal-oriented perspective. Finally, we present a modular AADA architecture that embeds AADA-driven flexible building blocks to address the challenge of time resolution for high-dimensional scanning hyperspectral imaging of nonequilibrium dynamical systems. In our example research-driven experimental design case, we propose an AADA infrastructure for time-resolved, noninvasive, and label-free scanning hyperspectral imaging of living biological systems. This AADA infrastructure can accurately target the correct state of the system for experimental workflows that utilize subsequent expensive, high-information-content analytical techniques.

TABLE OF CONTENTS

I. INTRODUCTION

Autonomous experimentation (AE), which is critical for accelerating research and scientific discovery in hyperspectral imaging, is closely linked to original concepts of artificial intelligence (AI).[1] AE, therefore, inherits two of AI's recurring and recognized primary challenges. The first centers around distribution shifts, which occur when the data used to train the AI have a different distribution than the data encountered by the AI in either testing or deployment. The second emerges from difficulties in transfer learning—the reusing or transferring of information from previously learned tasks to new tasks.[2] There are many potential solutions that can be pursued for user-independent AE, but a simple alternative is to separate the overarching goal of AE into actionable components that can each be user-optimized for that user's specific structural and functional goals.

12 December 2023 18:07:46

Autonomous adaptive data acquisition (AADA) can be conceptualized as a flexible, tunable, "smart" building block that can eventually be embedded into built environments to create an intelligent infrastructure. In scanning hyperspectral imaging, as with any application that obtains high resolution and high precision chemical information of surfaces and materials, there are generally two major philosophies that dictate implementable AADA design strategies for the general research community. First: If the experimental system under ideal instrument performance circumstances can be described as well-characterized over the course of the researcher's study, then the design could integrate machine learning steps toward the creation and improvement of domain-oriented AI[2]—as long as an acceptable level of transfer learning can be achieved. This especially holds true when there is a large amount of accurately labeled, high quality data on experimental systems similar to the system being studied; as an example, the agricultural and materials industries use hyperspectral imaging for quality control of well-characterized products such as diseased plants[3] and silicone rubber damage.[4] Computational time associated with generating the classification model is a secondary concern;

however, this can be addressed with additional computing resources. Second: If little is known *a priori* about the experimental system except that the system will evolve over the study's duration, then increased importance is placed on sampling efficiency that balances feature discovery and characterization while minimizing computational time.

Scanning hyperspectral imaging experiments exhibit both these properties and, in this review, serve as an ideal proxy to explore addressing challenges through better AADA design principles. The broad spectrum of user inputs and varying objectives lead to differences in the methods, by which domain knowledge and human–computer interactions are integrated into the AADA process. This is especially true with respect to the availability of cluster computing resources for computationally demanding edge computing. Figure 1 depicts two practical examples of these challenges, in which hundreds to thousands of contiguous spectral bands across the electromagnetic spectrum from the x-ray to the infrared region are used to capture biochemical information in living biological samples.[5,6] Figure 1(a) exemplifies a case in which AADA could have been used to restrict sampling to the specimen, substantially decreasing data acquisition
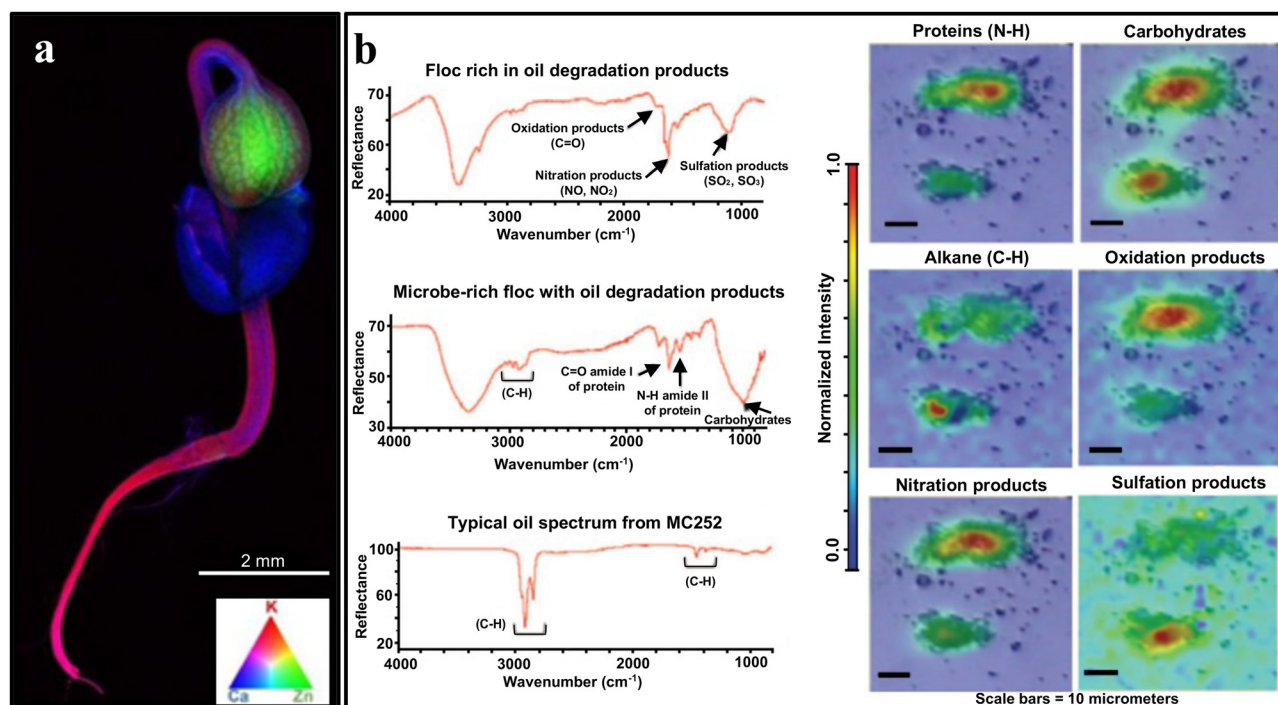

FIG. 1. AADA for addressing challenges in biological hyperspectral imaging. (a) Reproduced with permission from Kopittke *et al.*, Plant Physiol. **178**(2), 30108140 (2018). Copyright 2018 Authors licensed under a Creative Commons Attribution (CC BY) License. Synchrotron-based x-ray fluorescence microscopy (XFM) image ($\sim 5 \times 10.5$ mm^2) of calcium (blue), potassium (red), and zinc (green) distributions in a living seedling of heavy metal hyperaccumulator *Noccaea caerulescens*. Although element concentrations visibly correlate well with the maturing seedling's anatomical structures, AADA could have been used to save substantial time during data acquisition by restricting sampling points to within the seedling. Instead, considerable time was spent scanning the empty space around the seedling's irregular shape. Scale bar = 2 mm. (b) Reproduced with permission from Hazen *et al.*, Science **330**, 204 (2010). Copyright 2010 American Association for the Advancement of Science. Synchrotron-based FT-IR spectromicroscopy spectral images ($\sim 60 \times 60$ μm^2) of living microorganisms, oil, and oil degradation products inside a floc from the Gulf Horizon oil spill plume illustrate the bioremediation process. (Left) Representative spectral fingerprints for different components. (Right) Heat maps of different spectral components superimposed on the bright-field image of biological samples. Scale bar = 10 μm. Spatial heterogeneities in oxidation, nitration, and sulfation products of oil (alkane) metabolism, by bacteria in their extracellular matrix (proteins and carbohydrates), indicate metabolic specialization in regions that are not readily visibly apparent. AADA could have increased sampling and information efficiency both by avoiding empty areas and by increasing sampling density in more spatially complex regions—for example, the interface between regions producing oxidation products, nitration products, and sulfation products.

time. Figure 1(b) portrays an example in which AADA could have been used to increase sampling density in key transition regions that were revealed over the course of the experiment to be of importance. In both hyperspectral imaging studies, the data were obtained nondestructively as a two-dimensional grid of raster-scanned data points, stored digitally and visualized in false color.

In this review, we will briefly summarize how research in adaptive sampling algorithms and surrogate modeling across multiple fields shaped the research community's approach to AADA before covering recent AADA implementations for various scanning hyperspectral imaging experiments. We will discuss each AADA implementation to showcase conserved concepts along with differences in design philosophy based on the researcher's objective. We explore a potential intelligent infrastructure with embedded AADA building blocks to realistically characterize a living biological system's response in time for scientifically relevant findings. Our comprehensive understanding is often minimal in such experimental systems, which serves as a reminder that artificial intelligence for autonomous experimentation should ideally be guided by the intuitive researcher rather than logic-driven AI.[7] Therefore, this review focuses on the design and implementation of AADA for scanning hyperspectral imaging applications that work toward the development of intelligent infrastructure for AE. From this perspective, AADA architecture serves as an important tool both for optimizing efficient sampling-driven characterization of an experimental system and for refining the precision of experimental execution.

II. BRIEF HISTORY: ADAPTIVE SAMPLING ALGORITHMS AND SURROGATE MODELING FOR AADA

Sampling algorithms for autonomous adaptive data acquisition in scanning hyperspectral imaging arise from the sequential sampling theory, which addresses the challenge of scanning probe techniques only acquiring data at a single point in space at a given time. It should be noted that scanning probe techniques encompass all technologies in which data are a series of single point measurements; these experimental techniques are not limited to the presence of a physical probe such as atomic force microscopy (AFM) tips. The general theory for sequential sampling was first formulated by Wald in 1944,[8,9] although applications of statistical inference based on the sum of a sequence of random variables are attributed to Bartky in 1943.[10] However, it was not until 1947 that Wald used the foundation established in the early 1940s to propose a general theory for sequential decision functions, in which the total number of required observations depends on the outcome of the observations.[11] By 1960, Widrow proposed and evaluated an adaptive sampled data system model containing an adjustable worker and a supervisor to study the influence of performance feedback.[12] Although early adaptive sampling centered around sampling efficiency via performance feedback loops,[13] its application to pattern recognition and feature extraction,[14,15] adaptive sampling frequency,[16,17] function approximation,[18] and optimal adaptive estimation of sampled stochastic processes[19] set the stage for adaptive sampling's use in affordable model construction to replace computationally costly simulations of real-world phenomena.

The concept of constructing and implementing metamodels to replace computationally expensive simulations of real-world phenomena was introduced by Blanning in 1975, although no detailed metamodel construction methods were presented at the time.[20] This

approach became more appealing as time passed between the 1970s and 1990s, a time in which the research community already acknowledged the prohibitive computational cost and performance limitations of real-world phenomena simulations. However, the pursuit of model construction and its subsequent optimization to accurately represent a sampled data domain led to the realization that parameter estimation was not trivial.[21–24] The goal of parameter estimation with respect to model optimization was to address the effect of sampled data error on the parameter values used in the model construction. Additionally, subsequent model calibration was complicated by model characteristics relating to regions of attraction, minor local optima, roughness, sensitivity, and shape.[25] As a result, researchers focused on developing and assessing methods for spatial modeling of regional variables in the mid-1980s to address these calibration challenges. The goal of most proposed solutions was to offer statistical efficiency in describing a spatially variable system at the cost of complexity, and these approaches centered around the development of various weighted schemes.[26–28] This methodology gained more credibility with case study applications such as that from Cressie and Chan in 1989, in which they found that spatial trends identify large-scale variation in the data, whereas the variance and spatial dependence of the data captured small-scale variation.[29]

In an attempt to transfer these approaches to high-dimensional data applications, Friedman proposed using multivariate regression splines that took inspiration from recursive partitioning methods for regression. As with other leaders in the field in the early 1990s, he emphasized that a reasonable approximation for modeling of a system depends upon an ability to assess its lack of accuracy, and that the solution to approximating general functions of high dimensionality was based in adaptive computation.[18,30,31] This thought process led to the restructuring of many model calibration problems into global optimization problems in which the difference or error estimation between model simulations and corresponding observations was minimized.[32,33] This alternative solution became more practical as advancements in computational technology and its accessibility to the general research community increased by the mid-1990s, enabling the spread of metamodeling strategies to a variety of fields to serve as practical solutions to computationally costly modeling or simulation problems.

Surrogate modeling, also known as function approximation, response surface modeling, and metamodeling in other fields involve emulating a costly simulation where the model is trained via a data-driven approach. Specifically, it describes the relationship between the model's adjustable parameters termed inputs to its output, which often is evaluated performance of the model via model appraisal information and merit functions.[34–36] In 2005, Quiepo et al. provided an extensive overview of surrogate-based analysis and optimization.[37] Recently implemented AADA approaches for scanning hyperspectral imaging have utilized hybrid sequential design strategies for surrogate model construction[38] and Gaussian process (GP) based surrogate model construction[39,40] approaches to date. The commonality conserved in these approaches is that one can replace costly model simulations with cheaper surrogate model alternatives that can undergo optimization via adaptive sampling approaches, thereby decreasing the total computational burden per adaptive sampling loop iteration.

Adaptive surrogate modeling-based optimization (ASMO) method assessments were systematically evaluated in simulated

experiments in 2014 by Wang and Duan.[41] ASMO-related AADA procedures that calibrate models via optimization are similar in structure to the generalized schematic depicted in Fig. 2, which is built upon the ASMO procedures and modified to include a simplistic instrument-related performance feedback loop for practical implementation of autonomous adaptive data acquisition. Prior to the initialization of ASMO, a domain expert should assess the effects of each input parameter for the experimental system, allowing for careful parameter screening and selection to increase the accuracy of the constructed and eventually optimized surrogate model.[42] For AADA implementation in the scanning hyperspectral imaging cases, parameter selection is highly dependent upon the analytical technique to which AADA is applied as well as the experimental system that is being characterized.[38,40,43,44]

Autonomous adaptive data acquisition that incorporates ASMO framework can be summarized in several general steps. After parameter selection, a set of initial sample points are acquired according to a design of experiments (DoE) method; one such robust DoE candidate that leads to accurate convergence is the low-discrepancy quasi-Monte Carlo method.[41] These sample points are subsequently used to generate the simulation model, and at each sample point, objective function values are calculated. A surrogate model is then constructed to estimate the simulation model's response surface, which is usually built by fitting a statistical model to the performance metric of the generated

FIG. 2. General workflow of ASMO and its integration into AADA for a software-physical system. Blue boxes represent behavior embedded into the AADA design (software system aspect), whereas white boxes represent behaviors that are instrument-specific to the implementation case (physical system aspect). Resampling criteria can be set with respect to instrument performance expectations.

simulation model at the previously selected points. New sample points are selected based on the chosen adaptive sampling strategy, and these points are subsequently incorporated into the simulation model to update the surrogate model corresponding to the response surface. For each generation of the updated response surface in an implemented AADA case, a termination check is performed to determine whether instrument time should be used for further sample characterization. These termination criteria may vary greatly, being a function of many potential factors including experimental systems, techniques, and even time availability when user facility instrumentation is required. If termination criteria are not met, then a convergence check is executed on the updated response surface. If at least one pre-specified convergence criterion is met, the final response surface is used to obtain an optimal parameter set, and a new simulation model is generated with these determined optima. The new objective function value is recorded and compared to the previous objective function value. Generally, if the new value is better than the old value, then the new optimal parameters are used for model simulation step. If the new value is worse, then new adaptive sampling locations are generated until the modeled response surface performance meets the convergence criteria.

Although the general ASMO procedure is conceptually conserved, the optimal selection of surrogate model type is case-dependent, and its selection criteria are related to instrument-specific and user-specific research goals. As a practical example, the common consensus is that the Gaussian process method is computationally costly in the maximum-likelihood estimation calculation phase due to a need to compute and decompose a dense $n \times n$ covariance matrix $[\mathcal{O}(n^3)$ operations] at each iteration.[45] However, data-parallel approaches coupled with access to high performance computing (HPC) resources can reduce the computational time without lowering the computational cost. This enables an AADA user who performs scanning hyperspectral imaging to potentially construct a lower-fidelity surrogate of their sample's ground truth with high accuracy for their experimental system type,[46,47] allowing for undetermined hyperparameters to be tuned to the user's experiments. These hyperparameters may potentially be saved across the user's similar experiments when AADA incorporates appropriate machine learning approaches.[48] Recent implementations of AADA have chosen surrogate model construction methods based upon experimental system study goals, the physical limitations of techniques as the instrument interacts with the experimental system, access to high performance computing resources, and the available time period for modeling and subsequent adaptive sampling location determination.[38–40,43,49–51]

III. AUTONOMOUS ADAPTIVE DATA ACQUISITION FOR SCANNING HYPERSPECTRAL IMAGING

Recent autonomous experimentation use cases are primarily implemented in the fields of autonomous vehicles,[52,53] scientific combinatorial discovery approaches in fields similar to materials science and chemical synthesis,[54–57] and iterative research loops of planning, experimentation,[43] and analysis.[58,59] However, the first public autonomous adaptive data acquisition for scanning hyperspectral imaging was only publicly implemented for real-data acquisition in 2020.[38,43] This delay in implementation is due not only to the maturing of hardware technologies that were required to achieve acceptable signal-to-noise levels for high quality real-data generation, but also to the tailoring of autonomous adaptive data acquisition design to the
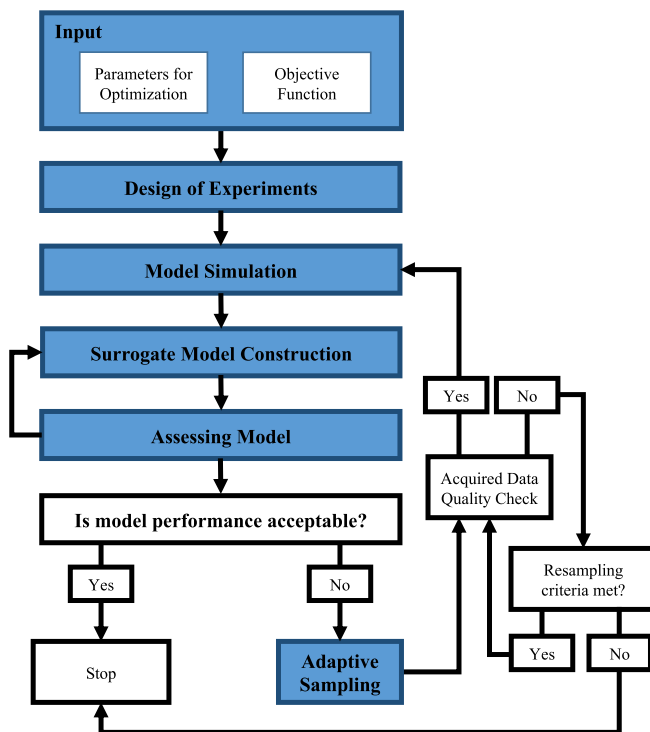
experimental techniques and to the interest of those who use the technique for scientific study.

Three major challenges were identified from the recent implementation of surrogate modeling coupled with adaptive sampling approaches in order to more efficiently create information-rich data sets that reduce redundant information storage. These challenges for practical implementation emerge from (1) instrument noise, (2) dimensionality of the parameter space, and (3) definition of the experimental goal. Roughly, analogous terms from related fields to describe these challenges would be (1) distribution shifts that are difficult to simulate *in silico*, (2) computational cost and time, and (3) complications with transfer learning or defining the scope of what should be learned per experimental run. Each recent implementation case attempted to address the relevant challenges through different methodologies, which we broadly categorize into two general discussion topics.

A. AADA implementation with high performance computing

Many techniques for constructing surrogate models exist, but historically radial basis function (RBF) and Gaussian process (GP) methods have performed well under several different modeling criteria.[60] Therefore, it is understandable that large user facilities such as synchrotrons are actively investigating GP-driven AADA design and implementation—an AADA application where instrument time rather than computational cost is the limited resource. In these cases, the GP-scaling challenge that may drive the computational cost of model construction upward to prohibitive levels can be addressed by efficient use of HPC facilities.[61,62] The first recent implementation of GP-driven autonomous adaptive data acquisition was performed by Noack *et al.* for the materials sciences. They used Gaussian priors and likelihoods to inform the decision-making process and subsequently classified their approach as a Bayesian optimization (BO) variant.[43] Their general workflow upstream of their surrogate modeling and adaptive sampling in their application to a small-angle x-ray scattering (SAXS) experiment involved:

(1) Obtaining a transmission SAXS pattern on an area detector with one intensity channel per pixel on a 1475×1679 array,
(2) Fitting the intensity profiles to extract three analysis-derived quantities of:
  (i) The degree of anisotropy for the in-plane inter-nanorod alignment,
  (ii) A factor incorporating the azimuthal angle, and
  (iii) The grain size of the ordered domains,
(3) Calculating the best-fit values and associated variances and using these parameters to construct their redefined GP models, and
(4) Using the GPR algorithms to adaptively select SAXS sampling points as a function of surface coordinates (x,y).

Notably, they proposed two modifications to their surrogate model construction to address instrument noise. They incorporate non-identically distributed observation noise by redefining their likelihood while also redefining the kernel function to incorporate the allowance of anisotropy. Although this design decision may increase computational cost in an attempt to increase model accuracy while

decreasing uncertainty, they clearly convey that this approach is viewed as necessary for practical implementation of AADA for SAXS on their specific instrument. When executing their experiment, they divided the sampling time into two phases. The first 4 h (N < 464 samples) utilized the adaptive sampling workflow described above, while the second phase of the experiment lasting 11 h ($464 \leq N \leq 1520$) aimed for the feature maximization mode. In summary, their first phase of sampling is an efficient approach to understanding the total search or mapped space, while the second phase utilized a weighted value approach to implement constraint concepts for their feature maximization mode.[43,50] For their experimental study goals, the autonomous adaptive data acquisition time was notably shorter than the gold standard acquisition method while enabling the researchers to spend the majority of their instrument time on spatially characterizing their sample's highly ordered regions.[43,51]

More recent implementations of autonomous adaptive data acquisition for materials science emphasize the concept of building a foundation for fully automated future experimentation. In an example that incorporates edge computing, Vasudevan *et al.* use a GP Bayesian optimization (GP-BO) algorithm that is implemented with a cost-effective multi-point sampling batch design for exploration of ferroelectric systems using band excitation piezoresponse spectroscopy.[40] This batch-sampling approach decreases model error per sampled region by incorporating more sample points to describe the ground truth of the sampled region while balancing the cost of moving the AFM probe to a new region for subsequent data acquisition. Conceptually, similar to the previous case, acquired multichannel data are fitted to statistical models to obtain analysis-derived quantities of amplitude, phase, frequency, and Q-factor for every sampled point.[63] These quantities were used to determine hysteresis loops, and the calculated loop areas were selected as the objective for maximization via Bayesian optimization (BO) routines.[64,65] Their approach does not adjust their kernel function like the previous publication, but they experimentally explore the dependency of hyperparameters on their instrument-specific GP-BO data acquisition. Sampling efficiency was calculated in time, where AADA took only 20% of the time required for the gold standard approach.[40] Additionally, the authors investigate a concept of prior knowledge incorporation into the sampling by using a convolutional neural network with four layers that is trained along with the GP model, and they found that physical sample drift and computational efficiency caused complications with this approach. Although both methods increase sampling efficiency in comparison to the gold standard method for their experimental technique, they concluded that sampling efficiency can be improved to an even greater degree by incorporating domain knowledge and prior physics knowledge to the sampling process.[40,66]

Vasudevan's and Kalinin's experimentally verified findings were further supported by development of active learning methods for 4D-STEM imaging that prioritized balancing the efficient data collection from regions of interest with the exploration and exploitation of characterizing observable physical phenomena.[44] Roccapriore *et al.* used a GP-BO sampling strategy for their AADA implementation, and similar to Vasudevan and Kalinin, they incorporated deep kernel learning via a deep neural network to investigate prior knowledge incorporation into their GP-BO sampling.[40,44] Importantly, Roccapriore mentions the sampling constraints of 4D-STEM imaging and potentially beam-sensitive materials—highlighting that each implementation of

AADA will likely require future tailoring to the instrument, the imaging technique, and the experimental system being studied.[44]

B. AADA for high-dimensional data acquisition with computational efficiency

For experimental techniques that can acquire data on a rapid timeframe with respect to computational time, it is important to prioritize sampling and computational efficiency in the space and time domains, respectively. This can be interpreted as placing a constraint on calculation time while emphasizing a relatively accurate spatial and spectral interpolation-based representation of the assessed experimental system. These constraints are relevant when instrument manufacturers are able to acquire more sample points per unit time using a uniform-grid scanning approach due to designed synergies between software and automated hardware protocols considerations. Hence, to avoid underperformance of implemented AADA on such systems, computational expense of the AADA-based calculations should be minimized while retaining an understanding of sequentially sampled space for subsequent adaptive data acquisition.

Holman and Fang addressed this computational cost challenge when implementing their autonomous adaptive data acquisition approach that prioritized the data acquisition efficiency with respect to computation time for high-dimensional spatiochemical hyperspectral imaging. They designed their AADA approach with computer accessibility in mind—their proof-of-concept experimental cases were performed on computer hardware that is commonly tied with available FT-IR spectral microscopes of the past decade.[38] In their instrument-specific circumstance, their analytical technique's output was a continuous infrared (IR) spectrum per spatial point in an x-y plane—each IR spectrum defined by more than 1000 multichannel intensity values that (1) are not necessarily independent from one another and (2) exhibited experimental system-specific behavior due to the physicochemical information contained in the mid-IR spectral domain.

To minimize the computational cost associated with AADA, Fang designed an alternative global surrogate modeling method called two-dimensional (2D) barycentric linear interpolation with Voronoi tessellation (LIV),[38] which was inspired by and based on previously established hybrid sequential design strategies implemented for global surrogate modeling[67] and robust error-pursuing sequential sampling approaches for global metamodeling.[68] These previous strategies share the goal of creating a model that best approximates the behavior of the simulator over the entire domain (global surrogate modeling) rather than using local models to guide an optimization algorithm toward a global optimum (local surrogate modeling). Hybrid sequential design strategies for model construction achieve this goal by defining two different criteria: one for exploration and one for exploitation.[67,69] For reference, Crombecq et al. use a Monte Carlo Voronoi approximation for their exploration criterion due to its computational efficiency and simplicity; for their exploitation criterion, they introduce an algorithm that selects sampling points by identifying those with significant deviations from a local linear approximation (LOLA) of the function.[69,70]

By drastically reducing the model construction cost, autonomous adaptive data acquisition could be driven by leave-one-out cross-validation (LOOCV)[37,71] for rapid and accurate error approximation per sampled point in the experimentally mapped space.[38,72] To briefly summarize the generalized adaptive sampling workflow (where $n = 5$ in their implemented experimental cases),

(1) Initial sample points are acquired by selected design of experiments,
(2) Operational mode selection (none vs. domain knowledge incorporation) and IR spectral processing occurs,
(3) Acquired and processed IR spectra is dimensionally reduced via principal component analysis (PCA) to n components,
(4) At each point, n components are used to construct the LIV-based surrogate models,
(5) The surrogate model's sensitivity is quantified per point through the associated and adjusted leave-one-out error at each sampled point (see described methods[38,73]),
(6) The Voronoi area or tile associated with the largest Voronoi-weighted leave-one-out error is selected for subsequent sampling,[72,74] and
(7) Steps 2–6 are iterated until termination criterion or criteria are met (e.g., pre-determined total samples, acceptable error threshold, etc.).

All operational modes of the IR spectral processing module executed rubberband baseline correction to each IR spectrum in order to reduce parameter input error—since scattering is a physical rather than chemical system phenomena.[75] When operating under the inclusion of researcher's domain knowledge, the IR spectral processing module enabled parameter (channel value) selection after a preliminary screening of reference experimental system data. This implementation of the parameter screening and selection stage of data processing can improve the quality of the constructed model. By excluding regions of the IR spectrum that are sensitive to scattering and other physical attributes of an experimental system, the researcher can reduce parametric input error by excluding known parameters that exhibit high sensitivity to noise and signal contamination while exhibiting low sensitivity to the chemical composition of the researcher's sample. This version of researcher-informed parameter screening and selection when coupled with the carefully designed AADA infrastructure allowed for rapid identification, characterization, and potential spatiochemical surveillance of regions of interest while avoiding sampling convergence to an easily characterized and potentially chemically irrelevant subspace of the spectral parameter space.[38] Holman and Fang assessed the performance of their experimental LIV-based AADA cases using instrument usage in time and their mean Voronoi-weighted leave-one-out error to quantify model accuracy. Specifically in their complex biological system case where domain knowledge was applied to confine the mapping region boundaries, LIV-based AADA efficiently obtained data over the same mapping region in 45 min, whereas the gold standard uniform-grid sampling required with ~4.9 h. They also noted that LIV-based AADA provided more comprehensive spatiochemical understanding of the total mapping region at any given time interval for their scanning hyperspectral imaging case, though they explicitly select 11 min time intervals to exhibit this concept.[38] More broadly speaking, the IR spectral processing module allows for this act-observe-update-decide cycle to function in a similar manner to a human–computer team or an eventual multi-agent system;[76] this enables implementation for high performance AADA, while the research community works toward generating specific domain-based strains of AI that realistically and robustly represent domain expert's decision-making for the future AE development.[2,77]

In 2021, Noack et al. investigated the feasibility of implementing GP-BO AADA to FT-IR spectromicroscopy, and they reported their

findings with respect to its sampling efficiency in minimizing the total number of sampled points to describe a 2D IR spectral map. Although baseline correction, parameter screening, and parameter selection were not performed, they also included dimensionality reduction using via PCA ($n = 3$) to address the high dimensionality of the technique's data output.[39] Utilizing the strengths of the GPR surrogate model, they also introduced a feature-finding capability by defining a customized acquisition function to enforce the consideration of the correlation coefficient between a given reference spectrum (feature) and the reconstructed spectrum in real time.[39] From their findings of this high-dimensional feasibility study, the authors conclude that for near-future GP-driven AADA implementation scanning hyperspectral imaging, (1) the development and use of tools for HPC-based edge computing optimization and (2) the communication of constraints for acquiring scientifically relevant data require further research and development.[39] To directly address these challenges in 2022, Noack and Krishnan proposed an algorithm that exploits the natural sparsity of GPs in their native form in order to leverage heterogeneous HPC architectures for scalable parallelism across graphics processing units (GPUs), central processing units (CPUs), and tensor processing units (TPUs) to reduce computational time.[61,65]

The importance of a computationally fast yet efficiently sampled domain is key in high-dimensional scanning hyperspectral imaging, and it serves as a potential foundation for time-resolved, noninvasive, label-free spatiochemical imaging of dynamic experimental systems. Noack and his coauthors focused on reducing computational time by efficiently harnessing HPC resources to reach computed solutions more quickly when using HPC architectures. Meanwhile, Holman's and Fang's solution to reduce computational time was achieved by decreasing the overall computational cost of their implemented AADA by carefully selecting a surrogate modeling method that met their technique's criteria[38] while incorporating their data processing module to enable flexible human–computer coordination and communication to tackle complex experimental system uncertainties.[78] Both design philosophies have their strengths and weaknesses, and perhaps a mixture of the two philosophies will offer a third variant of computational time efficiency for future high-dimensional AADA applications. In conclusion, it is important to understand the specific research community's needs and resources prior to designing and implementing goal-oriented AADA for intelligent experiment infrastructure.

IV. EXPANDING HIGH-DIMENSIONAL AADA TO BIOLOGICAL SYSTEM STUDIES

Biological systems are often composed of complex subassemblies that contain compartmentalized chemistries.[79] These systems have classically been characterized with either extremely targeted or high-information-content yet destructive approaches. Targeted methods conceptually involve either genetically encoding the reporter protein sequence or binding a target of interest with a small-molecule reporter—a common application of small-molecule fluorophores being the targeted super-resolution imaging of subcellular processes.[80–83] Recent high-information-content methods such as directed multi-omics approaches can generate comprehensive single-cell RNA, DNA, and proteomic profiles on a biological system but are limited to pseudo-temporal system characterization, in which dynamics are inferred from a series of static measurements.[84–86] However, these

now widely used approaches do not address the sensitivity of biological systems,[87,88] which is especially true in living systems that use nonequilibrium dynamics to achieve not only apparent steady states[88] such as phenotypes but also perturbation-specific behavioral evolution.

Spatiochemical imaging techniques that can directly characterize the experimental system in either its native or a largely unaltered state are important for studies ranging from the environmental sciences[6] to biomedicine.[89] As shown in Fig. 3(a), many biological studies seek to characterize their chosen experimental system's response to a variety of stimuli, such as heat, chemical, and even in some circumstances, electrical and magnetic stressors.[90,91] Mid-infrared (mid-IR) and Raman spectroscopies are at the forefront of label-free and noninvasive spatiochemical imaging, as both techniques generate molecular fingerprint vibration spectra through measurements of linear IR absorption or inelastic Raman scattering, respectively. However, due to the very small cross section of spontaneous Raman scattering, which limits the label-free bioimaging speed to tens of minutes per frame,[92] the research community has turned toward Fourier transform infrared (FT-IR) spectromicroscopy to noninvasively[93,94] capture the dynamics of living biological systems.[95–97]

FT-IR spectromicroscopy is a powerful tool for label-free and noninvasive chemical analysis that merges visible light microscopy to match physical morphology with chemical and molecular information obtained from IR absorption spectral data. In this manner, mid-IR spectra provide molecular insight into unlabeled chemical targets for a more complete understanding of the organism's evolving global biochemical landscape as it responds to its local environment.[98,99] A simplified schematic of this concept is shown in Fig. 3(b), where spectral interpretation complexity is reduced to an absolute change in specific IR absorption intensities that accurately correlate with system behavioral emergence ($t = t_i$) and the establishment of a new nonequilibrium steady state ($t = t_n$) that is distinct by some metric from its native state ($t = t_0$). As an example, previously published IR spectral monitoring reveals spatially resolved chemical and temporal changes in PC12 cell colonies responding to nerve growth factor [Fig. 3(c)].[100] From a time-resolved spatiochemical mapping perspective [Fig. 3(d)], the researcher would be interested in monitoring this change in 2D IR spectral maps over space and time with a goal of having higher spatiotemporal resolution in the regions of the system where chemical change is occurring, or when $t \in [t_i, t_n]$. Therefore, we can construct a time-resolved scanning hyperspectral imaging AADA design by decomposing the problem into two independently solvable parts.

A solution for the first solvable part should address the goal of obtaining a global understanding of the domain; in our biological system example, the domain is defined as the 2D IR spectral map that is acceptably described by some minimum number of M sample point locations that we can describe as

$$P = \left\{ (x_k, y_k) \in \mathbb{R}^2 \,|\, k \in \mathbb{N} : k \leq M \right\}.$$

For simplicity, we can determine locations P using the previously described LIV-based AADA solution[38] to avoid prohibitive HPC requirements in the construction of our tiered AADA architecture. Building upon this efficient sampling foundation, we then can calculate an optimal sampling path that minimizes total acquisition time per 2D IR spectral map, this pathing solution dependent on practical constraints of the specific instrument and its scanning probe. Optimal
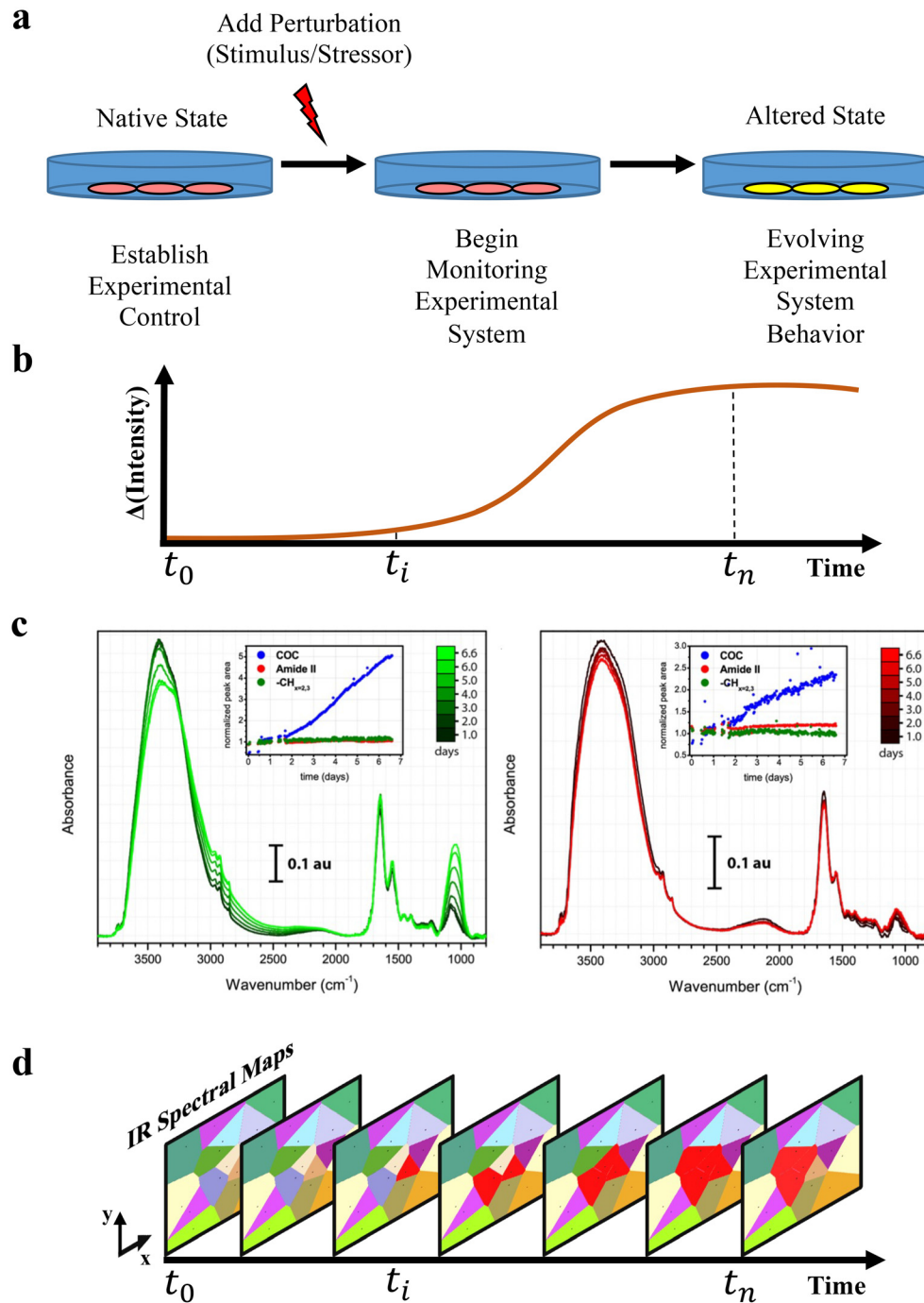
FIG. 3. Schematic of a generalized living biological system study to develop AADA for time-resolved scanning hyperspectral imaging of dynamic systems. (a) Cartoon depicting a general experimental design that monitors an experimental system for unknown behavioral response to a stimulus. (b) A simplified depiction of spectral change with respect to time of an otherwise high-dimensional space of broadband IR spectra. Representative t-values are described in the main text. (c) Reproduced with permission from Loutherback *et al.*, Anal. Chem. **87**, 4601 (2015). Copyright 2015 American Chemical Society. Multiday IR spectral progression measurements of PC12 cell colonies showing spatiochemical and spatiotemporal differences between cells in the center (left) and at edge (right) of the colony. Insets show peak area of glycogen and glycoprotein (1190–930 cm^{-1}), amide II (1580–1480 cm^{-1}), and –CH$_{2,3}$ (3000–2800 cm^{-1}). (d) Concept of temporally resolved spatiochemical mapping corresponding to the system's evolution timeframe in (a) and (b). In this case, the Voronoi areas (red) indicated spatiochemical change through a selected spectral metric detectable by models f(x, y, t) for $t \geq t_i$ when compared to the 2D IR spectral map modeled by f(x, y, t_0).

12 December 2023 18:07:46

path data acquisition (optimized map sampling) can be iterated for all Voronoi cell locations represented by P to generate 2D IR spectral maps that serve as high-fidelity surrogates of the experimental system's ground truth, since they are by definition physically based granular models[101] that preserve the main body of processes exhibited by the real experimental system.[102] By resampling at every location in P, we can establish time-course IR spectral data per Voronoi cell to identify when spatiochemical changes begin to occur in the experimental system.

A solution for the second portion of the original problem is inspired by event-sensing strategies used to monitor rare-event propagation in wireless sensor networks. Generally, their objective-oriented protocols are designed to detect, monitor, and handle specific scenarios, and these protocols can be activated upon detection of the appropriate event. As a practical example from Harrison *et al.*, burst aware protocols (BAP) respond to and handle a burst of network traffic from multiple stationary nodes when a rare event is detected by multiple sensors. In the BAP case, the goal is to avoid introducing unnecessary detection delay during these burst events, and such protocols often incorporate a carefully configured component deactivation element.[103] Similarly for our time-resolved FT-IR spectromicroscopy case, each sample point location described by P can be viewed as a stationary node that generates an IR spectrum for each time $t = \{t_0, t_1, \ldots, t_n, \ldots\}$. However, our goal is to increase the spatiochemical and spatiotemporal resolution of map regions that are represented by nodes that show an evolution of behavior over time. In living biological systems, it is common for the system to exhibit an overall response to an external stimulus, but even this response behavior can differ within a colony of the same cell type depending upon the spatial location of the individual cell [Figs. 3(c) and 4(a)].[100] A detectable event or evolution of system behavior will begin to emerge in a region of the 2D IR spectral map that can be represented by the red subset of sampling locations $U \subset P$, as conceptualized in Fig. 4(b). This spatiochemical difference between two 2D IR spectral maps can be detected by performing IR image analysis using established chemometrics[104,105] such as PC-LDA[106] and MCR-ALS[107] on each 2D IR spectral map prior to evaluating and quantifying their differences. As in Fig. 4(c), the most recently acquired 2D IR spectral map by optimized map sampling will always be compared to the initial 2D IR spectral map that closely represents the assessed experimental system's steady state $f(x, y, t_0)$. When the change is detectable in the comparison of $f(x, y, t_i)$ and $f(x, y, t_0)$, our sampling-switching architecture should prioritize a protocol that is computationally inexpensive and provides rapid, dense region of interest (ROI) sampling at locations U_{RRS} that are spatially contained within the Voronoi areas represented by U.

Uniform-grid (UG) sampling meets the criteria for our rapid ROI sampling and can quickly obtain high spatial sampling levels that are necessary for achieving diffraction-limited spatial resolution via improved image contrast.[108,109] By decreasing the computational cost with respect to the rapid sampling phase of our data acquisition, we can invest our computing resources into high-dimensional hyperspectral image analysis approaches to quantitatively compare and assess the uncertainty in our comparison of two temporally proximal 2D IR spectral maps.[110] If HPC is available, more comprehensive approaches to uncertainty modeling can be achieved either through tailored, study-specific Bayesian formulations[111] or by maintaining necessary data granularity[112] to describe the spatial and spectral uncertainty in time[113] and to avoid system under modeling in time.
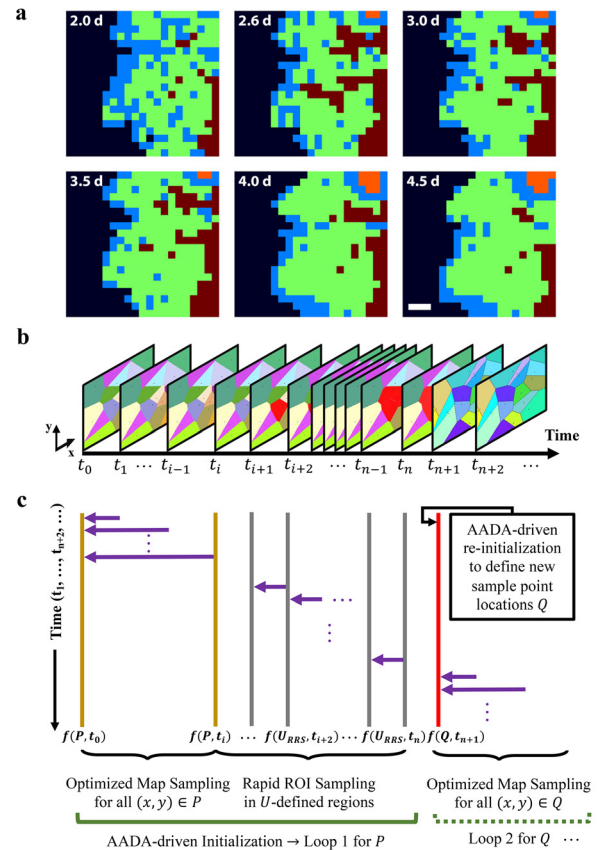

FIG. 4. Proposed sampling-switching architecture for data acquisition that builds upon established AADA for high-dimensional scanning hyperspectral imaging of biological (dynamic) systems. (a) Reproduced with permission from Loutherback *et al.*, Anal. Chem. **87**, 4609 (2015). Copyright 2015 American Chemical Society. Hierarchical cluster analysis (HCA) maps of a multiday synchrotron IR experiment on living PC12 cells as they respond to a nerve growth factor stimulus. Each pixel is colored by HCA cluster assignment, and empty pixels were filtered from the data analysis and colored dark blue. Scalebar is 30 μm. (b) A representative cartoon of time-course 2D IR spectral maps that are acquired to describe a dynamic experimental system. When t > t$_i$, Voronoi diagrams use red tiles to represent spatiochemical change from baseline IR spectral map defined by f(P, t$_0$). At time t$_n$ when detectable spatiochemical evolution stops, new sample point locations indicated with new Voronoi diagram color-coding are acquired to efficiently describe the domain of the 2D IR spectral map in spatial and spectral space. (c) Visualization of hypothetical switch-based architecture to achieve time-resolved AADA with multiple sampling tiers that can be selected modularly. Comparisons between models (purple arrows—directional relationship drawn from newly acquired model to reference model) are performed in a sampling-dependent manner. Qualitative x-axis represents evolving time-course surrogates that are generated by switch-based sampling methods (color-coded for clarity as horizontal lines). Qualitative time y-axis aids in describing the evolution of model-comparison framework as our time-resolving AADA architecture for surveillance-based scanning hyperspectral imaging operates in time.

Accurate modeling of spectral uncertainty in time is key in determining whether the spatiochemical change has occurred in the sampled experimental system over time. This implies that both experimental system and uncertainty modeling attributes directly affect the performance sensitivity of our proposed time-resolved

AADA architecture for high-dimensional scanning hyperspectral imaging. These attributes control the automated decision-making behind (1) exiting rapid ROI sampling (exploitation—loop 1, part 2) and (2) returning to a sampling mode that favors global surveillance of the experimental system (balanced exploration and exploitation—loop 1, part 1; loop 2, part 2). As shown in Fig. 4(b), this requires ending the sampling loop that was defined by sampling locations U_{RRS}, performing AADA-driven selection of new sampling locations Q that efficiently describe the spatiochemical distribution of the system's new apparent steady state, and calculating the optimal probe path to minimize total data acquisition time over the locations defined by Q. For our time-resolved AADA design case, we emphasized computational cost minimization at all steps prior to our eventual computational resource investment in the time-dependent, high-dimensional, spectral uncertainty modeling step.

Many encountered challenges of label-free noninvasive FT-IR spectromicroscopy for biological systems emerge from either the physicochemical nature of the spectral data or the sample heterogeneity exhibited by complex biological systems. However, these challenges can be minimized effectively through sample preparation, appropriate instrument settings, and IR spectral output screening, correction, and selection protocols.[38] This permits a logic-driven approach to a switch-based adaptive sampling architecture to tackle the high-dimensional data problem, which requires increasing information granulation to accurately and precisely detect minute IR spectral changes in time. Implementing a time-resolved AADA workflow for FT-IR spectromicroscopy furthers biological system study accuracy and comprehensiveness, especially those studies that generate targeted and high-information-content data on rare-event phenomena. In this case, FT-IR spectromicroscopy can serve as a noninvasive, real-time analytical surveillance tool to identify when a studied biological system starts to transition away from its apparent steady state. As soon as this transition occurs, the biological sample can be removed from the microscope stage and processed for the researcher's subsequent high-information-content experimental modalities. Some examples would be downstream spatial[114] or multi-omics[115] methods that, while capable of providing rich information about a biological system, are often destructive and, therefore, unsuitable for real-time longitudinal surveillance. Time-resolved guidance by FT-IR spectromicroscopy has the potential to drastically increase the chances of successful capturing cellular information relevant to the emerging biological event.[116,117]

V. CONCLUSION AND OUTLOOK

Autonomous adaptive data acquisition for scanning hyperspectral imaging broadly has many solutions. Only exceedingly rarely does one specific approach work as an ideal solution for all cases. Instead, conserved concepts emerge when decomposing each solution into its modular subparts and viewing each module from an objective-centric lens. Modular subcomponents can be tailored to synergistically address a plethora of instrument-, experimental system-, or study-specific constraints. These considerations include but are not limited to acquisition time per sampled point for appropriate signal-to-noise, dimensionality of the acquired data, noise and signal contamination minimization, the estimated cost of probe movement over varying distances, sample preparation variation, adjustment of AADA architecture for multi-probe parallelization, and even contextually defining efficient data acquisition, which can be both field-dependent and user-

dependent. Some of these constraints also appear in other fields such as geostatistics, engineering, and game theory, fields which have used mathematics, computation, and statistics to develop and assess the performance of many surrogate modeling methods, parameter estimation challenges relating to sampled data noise handling, and adaptive sampling strategies. This implies that as long as solutions to conserved challenges have been addressed in a similarly modular manner in other fields, these approaches can be deconstructed, modified, and transferred in a piece-wise, objective-based manner to accelerate AADA development and speciation for scanning hyperspectral imaging applications.

For discovery-based biological studies involving high-dimensional, scanning probe spectromicroscopy, AADA architecture can be seen as a surveillance and protocol-based tool—especially when applied to living biological systems. Since little is known *a priori* about the experimental system, global understanding of the mapping region is valued along with the ability to spatiotemporally capture the emergence and propagation of event-based processes. In the case of FT-IR spectromicroscopy, recent technological advancements in IR-compatible microfluidics devices for living biological system studies,[118–120] model-based correction algorithms for complex tissue-substrate systems,[121] platform-mediated hyperspectral data connection with machine learning,[122] and novel method developments to bypass IR detector limitations such as that of highly multimode quantum nonlinear interferometry[123] work toward addressing the historical hardware limitations of this analytical technique. With the development of these technologies, the incorporation of FT-IR spectromicroscopy into workflows relating to autonomous experimentation is becoming more feasible, suggesting that the research community first needs to develop various AADA architectures that can serve as building blocks for the intelligent infrastructure of the future.

A user-friendly platform that supports a visual programming language[124] to construct these AADA architectures would catalyze the development of the next generation AADA framework. This framework would be necessary for a future of intelligent infrastructure that supports not only hypothesis-driven AE but also discovery-driven AE for complex systems. By increasing the accessibility of AADA architecture development via visual programming languages, researchers can adjust and store AADA architectures that meet their user-specific and field-specific goals of experimentation. Meanwhile, those involved with AE from an AI standpoint will benefit from the platform investment since researchers will be providing workflows that represent and, therefore, communicate their decision-making process that is driven by domain expertise. By storing and accruing this information over time, trends in user-specific and system-specific decision-making can be analyzed to not only aid the researcher in improving their experimental workflows but also to generate data on researcher-driven decision-making processes—a necessary set of data to work toward AI that can optimize human–computer interactions for AE-centric intelligent infrastructure. We expect that such optimized AADA will help hyperspectral imaging reach its full potential to provide incisive analysis of biological systems.

ACKNOWLEDGMENTS

AUTHOR DECLARATIONS
Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

E.A.H. prepared, wrote, and edited the manuscript. H.K. contributed computer science domain expertise and edited the manuscript. D.R.H. reviewed, discussed, and edited the manuscript. H.-Y.N.H. contributed infrared spectroscopy of biological systems domain expertise and edited the manuscript. P.W.S. contributed biology domain expertise and edited the manuscript.

Elizabeth Anne Holman: Conceptualization (lead); Methodology (lead); Writing – original draft (lead); Writing – review & editing (equal). **Harinarayan Krishnan:** Methodology (supporting); Writing – review & editing (supporting). **Derek Rudolf Holman:** Writing – review & editing (equal). **Hoi-Ying N. Holman:** Funding acquisition (equal); Writing – review & editing (equal). **Paul W. Sternberg:** Funding acquisition (equal); Writing – review & editing (supporting).

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

[1] J. McCarthy, M. L. Minsky, and C. E. Shannon, "A proposal for the Dartmouth summer research project on artificial intelligence," AI Mag. **27**(4), 12–14 (1955).

[2] N. Shadbolt, "From so simple a beginning: Species of artificial intelligence," Daedalus **151**(2), 28–42 (2022).

[3] A. Lowe, N. Harrison, and A. P. French, "Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress," Plant Methods **13**, 80 (2017).

[4] M. Bleszynski, S. Mann, and M. Kumosa, "Visualizing polymer damage using hyperspectral imaging," Polymers **12**(9), 2071 (2020).

[5] P. M. Kopittke et al., "Synchrotron-based x-ray fluorescence microscopy as a technique for imaging of elements in plants," Plant Physiol. **178**(2), 507–523 (2018).

[6] T. C. Hazen et al., "Deep-sea oil plume enriches indigenous oil-degrading bacteria," Science **330**(6001), 204–208 (2010).

[7] K. Gal and B. J. Grosz, "Multi-agent systems: Technical and ethical challenges of functioning in a mixed group," Daedalus **151**(2), 114–126 (2022).

[8] A. Wald, "On cumulative sums of random variables," Ann. Math. Stat. **15**, 283–296 (1944).

[9] A. Wald, "Sequential tests of statistical hypotheses," Ann. Math. Stat. **16**(2), 117–186 (1945).

[10] W. Bartky, "Multiple sampling with constant probability," Ann. Math. Stat. **14**, 363–377 (1943).

[11] A. Wald, "Foundations of a general theory of sequential decision functions," Econometrica **15**(4), 279–313 (1947).

[12] B. Widrow, "Adaptive sampled-data systems," in 1st International IFAC Congress on Automatic and Remote Control, 1960.

[13] D. Ciscato and L. Mariani, "On increasing sampling efficiency by adaptive sampling," IEEE Trans. Automat. Control **Ac12**(3), 318 (1967).

[14] G. S. Sebestyen, "Pattern-recognition by an adaptive process of sample set construction," IRE Trans. Inf. Theory **8**(5), S82 (1962).

[15] Y. Chetouani, "A sequential probability ratio test (SPRT) to detect changes and process safety monitoring," Process Saf. Environ. Prot. **92**(3), 206–214 (2014).

[16] R. C. Dorf, C. A. Phillips, and M. C. Farren, "Adaptive sampling frequency for sampled-data control systems," IRE Trans. Automat. Control **Ac 7**(1), 38 (1962).

[17] R. Tomovic and G. A. Bekey, "Adaptive sampling based on amplitude sensitivity," IEEE Trans. Automat. Control **Ac11**(2), 282 (1966).

[18] R. J. Lipton and J. F. Naughton, "Query size estimation by adaptive sampling," J. Comput. Syst. Sci. **51**(1), 18–25 (1995).

[19] D. T. Magill, "Optimal adaptive estimation of sampled stochastic processes," IEEE Trans. Automat. Control **Ac10**(4), 434 (1965).

[20] R. W. Blanning, "The construction and implementation of metamodels," Simulation **24**(6), 177–184 (1975).

[21] R. P. Ibbitt and P. D. Hutchinson, "Model parameter consistency and fitting criteria," IFAC Proc. Vol. **17**(2), 3169–3173 (1984).

[22] R. P. Ibbitt, "Effects of random data errors on the parameter values for a conceptual model," Water Resour. Res. **8**(1), 70–78, https://doi.org/10.1029/WR008i001p00070 (1972).

[23] S. Sorooshian, V. K. Gupta, and J. L. Fulton, "Evaluation of maximum-likelihood parameter-estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility," Water Resour. Res. **19**(1), 251–259, https://doi.org/10.1029/WR019i001p00251 (1983).

[24] V. K. Gupta and S. Sorooshian, "The relationship between data and the precision of parameter estimates of hydrologic models," J. Hydrol. **81**(1–2), 57–77 (1985).

[25] Q. Y. Duan, S. Sorooshian, and V. Gupta, "Effective and efficient global optimization for conceptual rainfall-runoff models," Water Resour. Res. **28**(4), 1015–1031, https://doi.org/10.1029/91WR02985 (1992).

[26] N. Cressie, "Fitting variogram models by weighted least-squares," J. Int. Assoc. Math. Geol. **17**(5), 563–586 (1985).

[27] N. Cressie, "Spatial prediction and ordinary kriging," Math. Geol. **20**(4), 405–421 (1988).

[28] N. Dyn, D. Levin, and S. Rippa, "Numerical procedures for surface fitting of scattered data by radial functions," SIAM J. Sci. Stat. Comput. **7**(2), 639–659 (1986).

[29] N. Cressie and N. H. Chan, "Spatial modeling of regional variables," J. Am. Stat. Assoc. **84**(406), 393–401 (1989).

[30] J. H. Friedman, "Multivariate adaptive regression splines," Ann. Stat. **19**(1), 1–67 (1991).

[31] J. N. Morgan and J. A. Sonquist, "Problems in the analysis of survey data, and a proposal," J. Am. Stat. Assoc. **58**(302), 415–434 (1963).

[32] C. G. E. Boender et al., "A stochastic method for global optimization," Math. Program. **22**(2), 125–140 (1982).

[33] Q. Y. Duan, S. Sorooshian, and V. K. Gupta, "Optimal use of the SCE-UA global optimization method for calibrating watershed models," J. Hydrol. **158**(3–4), 265–284 (1994).

[34] A. J. Mendoza-Jasso and M. Koslowski, "Model error estimation in surrogate models of failure for composite materials," Compos. Struct. **102**, 148–153 (2013).

[35] N. V. Queipo, J. V. Goicochea, and S. Pintos, "Surrogate modeling-based optimization of SAGD processes," J. Pet. Sci. Eng. **35**(1–2), 83–93 (2002).

[36] M. J. Sasena, P. Y. Papalambros, and P. Goovaerts, *Metamodeling Sampling Criteria in a Global Optimization Framework* (American Institute of Aeronautics & Astronautics, 2000).

[37] N. V. Queipo et al., "Surrogate-based analysis and optimization," Prog. Aerosp. Sci. **41**, 1–28 (2005).

[38] E. A. Holman et al., "Autonomous adaptive data acquisition for scanning hyperspectral imaging," Commu. Biol. **3**(1), 684 (2020).

[39] M. M. Noack et al., "Gaussian processes for autonomous data acquisition at large-scale synchrotron and neutron facilities," Nat. Rev. Phys. **3**(10), 685–697 (2021).

[40] R. K. Vasudevan et al., "Autonomous experiments in scanning probe microscopy and spectroscopy: Choosing where to explore polarization dynamics in ferroelectrics," ACS Nano **15**(7), 11253–11262 (2021).

12 December 2023 18:07:46

[41]C. Wang *et al.*, "An evaluation of adaptive surrogate modeling based optimization with two benchmark problems," Environ. Modell. Software **60**, 167–179 (2014).

[42]W. Shi, J. P. C. Kleijnen, and Z. X. Liu, "Factor screening for simulation with multiple responses: Sequential bifurcation," Eur. J. Oper. Res. **237**(1), 136–147 (2014).

[43]M. M. Noack *et al.*, "Autonomous materials discovery driven by Gaussian process regression with inhomogeneous measurement noise and anisotropic kernels," Sci. Rep. **10**(1), 17663 (2020).

[44]K. M. Roccapriore *et al.*, "Automated experiment in 4D-STEM: Exploring emergent physics and structural behaviors," ACS Nano **16**(5), 7605–7614 (2022).

[45]Z. Z. Zhou *et al.*, "Combining global and local surrogate models to accelerate evolutionary optimization," IEEE Trans. Syst. Man Cybern., Part C **37**(1), 66–76 (2007).

[46]A. I. J. Forrester, A. Sobester, and A. J. Keane, "Multi-fidelity optimization via surrogate modelling," Proc. R. Soc. A **463**(2088), 3251–3269 (2007).

[47]M. C. Kennedy and A. O'Hagan, "Predicting the output from a complex computer code when fast approximations are available," Biometrika **87**(1), 1–13 (2000).

[48]Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proceedings of the 34th International Conference on Machine Learning* (PMLR, 2017), Vol. 70, pp. 1183–1192.

[49]Z. Wang *et al.*, "An adaptive surrogate-assisted endmember extraction framework based on intelligent optimization algorithms for hyperspectral remote sensing images," Remote Sens. **14**(4), 892 (2022).

[50]M. M. Noack *et al.*, "Advances in Kriging-based autonomous x-ray scattering experiments," Sci. Rep. **10**(1), 1325 (2020).

[51]M. M. Noack *et al.*, "A Kriging-based approach to autonomous experimentation with applications to x-ray scattering," Sci. Rep. **9**(1), 11809 (2019).

[52]S. Behere and M. Torngren, "A functional reference architecture for autonomous driving," Inf. Software Technol. **73**, 136–150 (2016).

[53]M. Zimmermann and F. Wotawa, "An adaptive system for autonomous driving," Software Qual. J. **28**(3), 1189–1212 (2020).

[54]E. W. McFarland and W. H. Weinberg, "Combinatorial approaches to materials discovery," Trends Biotechnol. **17**(3), 107–115 (1999).

[55]A. Ludwig, "Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods," Npj Comput. Mater. **5**, 70 (2019).

[56]D. Chen *et al.*, "Computational discovery of extremal microstructure families," Sci. Adv. **4**(1), eaao7005 (2018).

[57]Y. B. Jin *et al.*, "Intelligent on-demand design of phononic metamaterials," Nanophotonics **11**(3), 439–460 (2022).

[58]L. M. Roch *et al.*, "ChemOS: Orchestrating autonomous experimentation," Sci Rob. **3**(19), eaat5559 (2018).

[59]R. Tempke and T. Musho, "Autonomous design of new chemical reactions using a variational autoencoder," Commun. Chem. **5**(1), 40 (2022).

[60]R. Jin, W. Chen, and T. W. Simpson, "Comparative studies of metamodelling techniques under multiple modelling criteria," Struct. Multidiscip. Optim. **23**, 1–13 (2001).

[61]M. M. Noack *et al.*, "Exact Gaussian processes for massive datasets via non-stationary sparsity-discovering kernels," arXiv:2205.09070 (2022).

[62]M. Padidar *et al.*, "Scaling Gaussian processes with derivative information using variational inference," arXiv:2107.04061 (2021).

[63]S. Jesse, P. Maksymovych, and S. V. Kalinin, "Rapid multidimensional data acquisition in scanning probe microscopy applied to local polarization dynamics and voltage dependent contact mechanics," Appl. Phys. Lett. **93**(11), 112903 (2008).

[64]M. Ziatdinov (2021). "Gpim: Gaussian processes and Bayesian optimization for images and hyperspectral data," GitHub repository, https://github.com/ziatdinovmax/GPim.

[65]J. Q. Quinonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," J. Mach. Learn. Res. **6**, 1939–1959 (2005).

[66]S. V. Kalinin *et al.*, "Automated and autonomous experiments in electron and scanning probe microscopy," ACS Nano **15**, 12604 (2021).

[67]K. Crombecq *et al.*, "A novel sequential design strategy for global surrogate modeling," in 2009 Winter Simulation Conference (WSC), 2009.

[68]S. Xu *et al.*, "A robust error-pursuing sequential sampling approach for global metamodeling based on Voronoi diagram and cross validation," J. Mech. Des. **136**(7), 071009 (2014).

[69]K. Crombecq *et al.*, "A novel hybrid sequential design strategy for global surrogate modeling of computer experiments," SIAM J. Sci. Comput. **33**(4), 1948–1974 (2011).

[70]J. F. Lynch, "Analysis and application of adaptive sampling," J. Comput. Syst. Sci. **66**(1), 2–19 (2003).

[71]A. I. J. Forrester and A. J. Keane, "Recent advances in surrogate-based optimization," Prog. Aerosp. Sci. **45**(1–3), 50–79 (2009).

[72]A. Elisseeff, T. Evgeniou, and M. Pontil, "Stability of randomized learning algorithms," J. Mach. Learn. Res. **6**, 55–79 (2006).

[73]L. Ramirez-Lopez *et al.*, "Distance and similarity-search metrics for use with soil VIS–NIR spectra," Geoderma **199**, 43–53 (2013).

[74]G. Li, V. Aute, and S. Azarm, "An accumulative error based adaptive design of experiments for offline metamodeling," Struct. Multidiscip. Optim. **40**, 137 (2010).

[75]D. Deirmendjian, W. Viezee, and R. Clasen, "Mie scattering with complex index of refraction," J. Opt. Soc. Am. **51**(6), 620 (1961).

[76]J. M. Corchado, B. Lees, and N. Rees, "A multi-agent system 'test bed' for evaluating autonomous agents," in *Proceedings of the First International Conference on Autonomous Agents* (Association for Computing Machinery, 1997), pp. 386–393.

[77]B. J. Grosz and S. Kraus, "Collaborative plans for complex group action," Artif. Intell. **86**(2), 269–357 (1996).

[78]M. Tambe, "Towards flexible teamwork," J. Artif. Intell. Res. **7**, 83–124 (1997).

[79]T. L. Liu *et al.*, "Observing the cell in its native state: Imaging subcellular dynamics in multicellular organisms," Science **360**(6386), eaaq1392 (2018).

[80]M. Chalfie *et al.*, "Green fluorescent protein as a marker for gene expression," Science **263**, 802–805 (1994).

[81]L. Wang *et al.*, "Small-molecule fluorescent probes for live-cell super-resolution microscopy," J. Am. Chem. Soc. **141**(7), 2770–2781 (2019).

[82]G. Gaietta *et al.*, "Multicolor and electron microscopic imaging of connexin trafficking," Science **296**(5567), 503–507 (2002).

[83]M. J. Rust, M. Bates, and X. W. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)," Nat. Methods **3**(10), 793–795 (2006).

[84]E. Lubeck *et al.*, "Single-cell *in situ* RNA profiling by sequential hybridization," Nat. Methods **11**(4), 360–361 (2014).

[85]K. H. Chen *et al.*, "Spatially resolved, highly multiplexed RNA profiling in single cells," Science **348**(6233), aaa6090 (2015).

[86]C. Bock, M. Farlik, and N. C. Sheffield, "Multi-omics of single cells: Strategies and applications," Trends Biotechnol. **34**(8), 605–608 (2016).

[87]D. K. Kondepudi and I. Prigogine, "Sensitivity of nonequilibrium systems," Physica A **107**(1), 1–24 (1981).

[88]K. J. Rothschild *et al.*, "Nonequilibrium linear behavior of biological systems. Existence of enzyme-mediated multidimensional inflection points," Biophys. J. **30**(2), 209–230 (1980).

[89]G. Bellisola *et al.*, "Tracking infrared signatures of drugs in cancer cells by Fourier transform microspectroscopy," Analyst **135**(12), 3077–3086 (2010).

[90]N. S. Petersen, "Effects of heat and chemical stress on development," Adv. Genet. **28**, 275–296 (1990).

[91]B. R. Mcleod, A. R. Liboff, and S. D. Smith, "Biological-systems in transition: Sensitivity to extremely low-frequency fields," Electro-Magnetobiology **11**(1), 29–42 (1992).

[92]A. F. Palonpon *et al.*, "Raman and SERS microscopy for molecular imaging of live cells," Nat. Protoc. **8**(4), 677–692 (2013).

[93]M. C. Martin *et al.*, "Negligible sample heating from synchrotron infrared beam," Appl. Spectrosc. **55**(2), 111–113 (2001).

[94]H. N. Holman *et al.*, "Synchrotron infrared spectromicroscopy as a novel bioanalytical microprobe for individual living cells: Cytotoxicity considerations," J. Biomed. Opt. **7**(3), 417 (2002).

[95]H.-Y. N. Holman, M. C. Martin, and W. R. McKinney, "Tracking chemical changes in a live cell: Biomedical applications of SR-FTIR spectromicroscopy," Spectroscopy **17**, 486940 (2003), available at https://escholarship.org/uc/item/9k185794.

12 December 2023 18:07:46

[96] H. Y. Holman et al., "IR spectroscopic characteristics of cell cycle and cell death probed by synchrotron radiation based Fourier transform IR spectromicroscopy," Biopolymers **57**(6), 329–335 (2000).

[97] N. Jamin et al., "Highly resolved chemical imaging of living cells by using synchrotron infrared microspectrometry," Proc. Natl. Acad. Sci. U. S. A. **95**(9), 4837–4840 (1998).

[98] H. Y. Holman et al., "Synchrotron IR spectromicroscopy: Chemistry of living cells," Anal. Chem. **82**(21), 8757–8765 (2010).

[99] H. N. Holman et al., "Infrared spectromicroscopy: Probing live cellular responses to environmental changes," Synchrotron Radiat. News **23**(5), 12–19 (2010).

[100] K. Loutherback, L. Chen, and H. Y. Holman, "Open-channel microfluidic membrane device for long-term FT-IR spectromicroscopy of live adherent cells," Anal. Chem. **87**(9), 4601–4606 (2015).

[101] E. Wallin and M. Servin, "Data-driven model order reduction for granular media," Comput. Part. Mech. **9**(1), 15–28 (2022).

[102] S. Razavi, B. A. Tolson, and D. H. Burn, "Review of surrogate modeling in water resources," Water Resour. Res. **28**, W07401, https://doi.org/10.1029/2011WR011527 (2012).

[103] D. C. Harrison, W. K. G. Seah, and R. Rayudu, "Rare event detection and propagation in wireless sensor networks," ACM Comput. Surv. **48**(4), 58 (2016).

[104] S. Wold, "Chemometrics, why, what and where to next?," J. Pharm. Biomed. Anal. **9**(8), 589–596 (1991).

[105] H. Martens and T. Naes, Multivariate Calibration (Wiley, Chichester England; New York, 1989), Vol. XVII, p. 419.

[106] J. H. Perkins, E. J. Hasenoehrl, and P. R. Griffiths, "Expert system based on principal component analysis for the identification of molecular-structures from vapor-phase infrared-spectra. I. Theory—Identification of alcohols," Anal. Chem. **63**(17), 1738–1747 (1991).

[107] J. Felten et al., "Vibrational spectroscopic image analysis of biological material using multivariate curve resolution-alternating least squares (MCR-ALS)," Nat. Protocols **10**(2), 217–240 (2015).

[108] E. C. Mattson et al., "Restoration and spectral recovery of mid-infrared chemical images," Anal. Chem. **84**(14), 6173–6180 (2012).

[109] M. J. Nasse et al., "Multi-beam synchrotron infrared chemical imaging with high spatial resolution: Beamline realization and first reports on image restoration," Nucl. Instrum. Methods Phys. Res., Sect. A **649**(1), 172–176 (2011).

[110] D. Ami, P. Mereghetti, and S. M. Doglia, "Multivariate analysis for Fourier transform infrared spectra of complex biological systems and processes," Multivar. Anal. Manage., Eng. Sci. **2013**, 189–220.

[111] A. L. Bertozzi et al., "Uncertainty quantification in graph-based classification of high dimensional data," SIAM-ASA J. Uncertainty Quantif. **6**(2), 568–595 (2018).

[112] W. Pedrycz, "Allocation of information granularity in optimization and decision-making models: Towards building the foundations of granular computing," Eur. J. Oper. Res. **232**(1), 137–145 (2014).

[113] L. A. Zadeh, "Toward a generalized theory of uncertainty (GTU): An outline," Inf. Sci. **172**(1–2), 1–40 (2005).

[114] C. L. Eng et al., "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH," Nature **568**(7751), 235–239 (2019).

[115] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," Genome Biol. **18**, 83 (2017).

[116] Y. Takei et al., "Multiplexed dynamic imaging of genomic loci by combined CRISPR imaging and DNA sequential FISH," Biophys. J. **112**(9), 1773–1776 (2017).

[117] H. Y. Holman et al., "Real-time molecular monitoring of chemical environment in obligate anaerobes during oxygen adaptive response," Proc. Natl. Acad. Sci. U. S. A. **106**(31), 12599–12604 (2009).

[118] K. Loutherback et al., "Microfluidic approaches to synchrotron radiation-based Fourier transform infrared (SR-FTIR) spectral microscopy of living biosystems," Protein Pept. Lett. **23**(3), 273–282 (2016).

[119] G. Birarda et al., "IR-Live: Fabrication of a low-cost plastic microfluidic device for infrared spectromicroscopy of living cells," Lab Chip **16**(9), 1644–1651 (2016).

[120] L. Vaccari et al., "Infrared microspectroscopy of live cells in microfluidic devices (MD-IRMS): Toward a powerful label-free cell-based assay," Anal. Chem **84**(11), 4768–4775 (2012).

[121] A. D. Surowka et al., "Model-based correction algorithm for Fourier transform infrared microscopy measurements of complex tissue-substrate systems," Anal. Chim. Acta **1103**, 143–155 (2020).

[122] M. Toplak et al., "Infrared orange: Connecting hyperspectral data with machine learning," Synchrotron Radiat. News **30**(4), 40–45 (2017).

[123] I. Kviatkovsky et al., "Microscopy with undetected photons in the mid-infrared," Sci. Adv. **6**(42), eabd0264 (2020).

[124] J. Puigsegur and D. Robertson, "A visual syntax for logic and logic programming," J. Visual Languages Comput. **9**(4), 399–427 (1998).

12 December 2023 18:07:46