ACS central science

http://pubs.acs.org/journal/acscii

Outlook

Synergizing Chemical and AI Communities for Advancing Laboratories of the Future

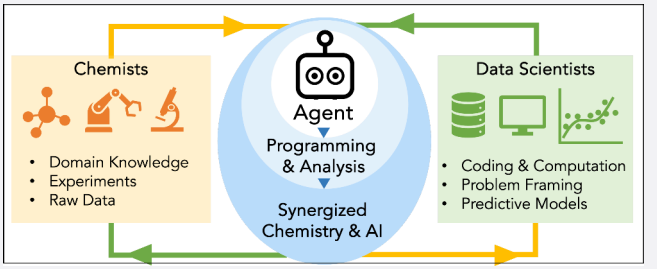
Saejin Oh,△ Xinyi Fang,△ I-Hsin Lin, Paris Dee, Christopher S. Dunham, Stacy M. Copp,* Abigail G. Doyle,* Javier Read de Alaniz,* and Mengyang Gu*

Cite This: https://doi.org/10.1021/acscentsci.5c01994

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The development of automated experimental facilities and the digitization of experimental data have introduced numerous opportunities to radically advance chemical laboratories. As many laboratory tasks involve predicting and understanding previously unknown chemical relationships, machine learning (ML) approaches trained on experimental data can substantially accelerate the conventional design-build-test-learn process. This outlook article aims to help chemists understand and begin to adopt ML predictive models for a variety of laboratory tasks, including experimental design, synthesis optimization, and materials characterization. Furthermore, this article introduces how artificial intelligence (AI) agents based on large language models can help researchers acquire background knowledge in chemical or data science and accelerate various aspects of the discovery process. We present three case studies in distinct areas to illustrate how ML models and AI agents can be leveraged to reduce time-consuming experiments and manual data analysis. Finally, we highlight existing challenges that require continued synergistic effort from both experimental and computational communities to address.



INTRODUCTION

Laboratory experiments are one of the most critical conduits to advance basic science and technology. In recent years, the field of chemistry has experienced numerous significant milestones in accelerating laboratory experiments with the introduction of critical techniques, including laboratory automation, high-performance computing, machine learning (ML) algorithms, and artificial intelligence (AI) agents based on large language models (LLMs). These advancements automate various laboratory processes, ranging from synthesis and purification to characterization and data analysis with minimal human intervention, stimulating the transition toward self-driving laboratories.[1−8]

Figure 1 shows a timeline of the introduction of selective high-throughput experimentation (HTE)/lab automation approaches, ML/AI algorithms, and LLMs over the past three decades. Although automated and self-driving laboratories are a relatively new concept, tools for tracking and cataloging data for experimentation, such as laboratory information management systems (LIMS)[9] and electronic laboratory notebooks (ELNs),[10,11] were conceptualized 30−40 years ago. As data acquisition and processing became increasingly multistep and time-consuming, automated and parallel operations of HT experiments have evolved in different areas.[2,12,13]

HTE has been developed through the standardization and automation of laboratory workflows. In the 1980s, the integration of robotics and analytical tools enabled high-throughput screening in pharmaceutical research.[14,15] The first commercial LIMS was released around the same time.[9] Also, the wide adoption of the 96-well plate in the early 1990s established a practical standard for parallel assays and screening, enabling reproducible experimentation in large sample volumes.[16] By the mid-1990s, the concept of ELN began to emerge[17] that enable researchers to capture their scientific observations and laboratory data electronically. In the early 2000s, automation expanded beyond screening, with reports of end-to-end systems integrating synthesis, purification, and characterization, marking a shift toward autonomous materials discovery workflows.[18] Additionally, the integration of active learning and Bayesian optimization into the process of Design of Experiments (DoE),[19,20] and the emergence of lab automation hardware companies improved the efficiency in exploration of multidimensional parameter spaces. Finally,

Received: October 17, 2025
Revised: January 12, 2026
Accepted: January 13, 2026

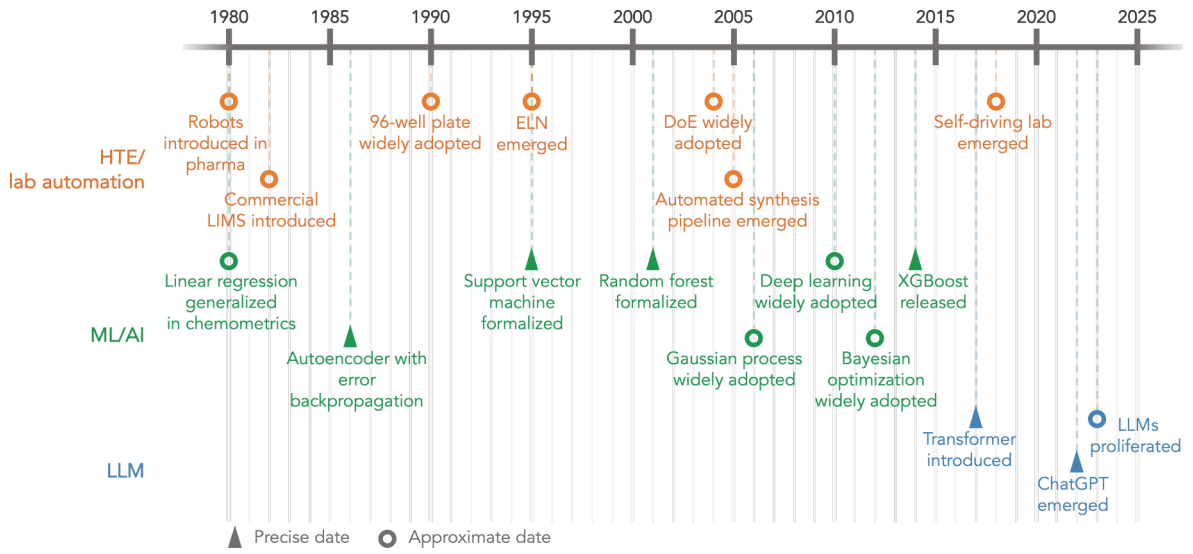


Figure 1. A brief timeline for the major developmental milestones of HTE/lab automation, ML/AI algorithms, and LLMs for the laboratories of the future.

from the late 2010s, closed-loop, self-driving laboratories emerged, in which machine-learning models iteratively proposed and executed experiments with minimal human intervention.[21,22]

The hardware of laboratory research has evolved along with the computational tools capable of powering the feedback loops that guide operations. Linear regression or linear models, for instance, have been widely used as a tool of chemometrics in 1980s.[23] Algorithms, such as backpropagation,[24] one of the most useful approaches to optimize artificial neural networks,[25] were formed in the 1970s and formally introduced in the early 1980s for building an autoencoder, a fundamental approach to represent latent information between inputs and outputs.[26] The 1990s saw the development of kernel methods, such as support vector machines (SVMs),[27] to encode similarities of multidimensional inputs by kernels for classification. The early 2000s saw the formalization of ensemble tree techniques, such as random forests, and probabilistic models, including Gaussian processes, for nonlinear regression and classification problems with small to moderate data sizes.[28−32] With the arrival of massive data collections of text and images on the Internet, different architectures of neural networks, such as convolutional neural networks and recurrent neural networks, were developed and evolved to be more flexible and accurate for tasks such as image classification and segmentation.[33−35] Furthermore, probabilistic data reduction tools and generative models, including variational autoencoder[36] and denoising diffusion probabilistic models[37,38] have been applied for protein structure prediction and design.[39,40] The development of neural network architectures[25,41] and their profound impacts in predicting protein structures[40,42,43] was awarded the 2024 Nobel Prizes in Physics and Chemistry, respectively. Trained by simulated or experimental data, ML methods can be routinely used as models for predicting untested inputs,[44,45] which can facilitate operations in almost all areas of laboratory science, including experimental design, synthesis optimization, and materials characterization.[46,47] However, significant challenges remains for applying ML approaches in accelerating laboratory research. For instance, the performance of the ML approach is highly dependent on the quality and quantity of the training data, and predictions may be unreliable when data is scarce or systematically biased. Furthermore, some complex ML models, such as neural networks and their variants, lack interpretability and require manual tuning of hyperparameters. Consequently, it is difficult for researchers to understand their underlying chemical mechanisms and reliably use them in different applications. Thus, high-quality data acquisition and understanding different ML model assumptions are key elements for successfully selecting and applying ML approaches to accelerate chemical laboratory research.The development of high-throughput experimentation and data digitization has facilitated the deployment of AI/ML approaches to replace labor-intensive tasks in chemical laboratories by automated design and discovery processes.

Over the past decade, LLMs based on transformer architecture[48] have gained tremendous attention across the world for text generation, and have opened up a new era of scientific research. The transformer, a neural network architecture for training LLMs, for instance, inspired the development of the Generative Pretrained Transformer (GPT),[49,50] and other LLM models, such as Claude, Gemini, Llama, Qwen, and DeepSeek.[51−55] The versatility of LLMs for use in a variety of operations, ranging from literature summary to computer code generation, reduces barriers to learning new disciplines and facilitates interdisciplinary collaboration, which has started to transform the paradigm in chemical laboratory research.[56,57] Despite the remarkable capabilities, LLMs also have notable limitations. These models can produce information that appears plausible yet is actually incorrect,

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

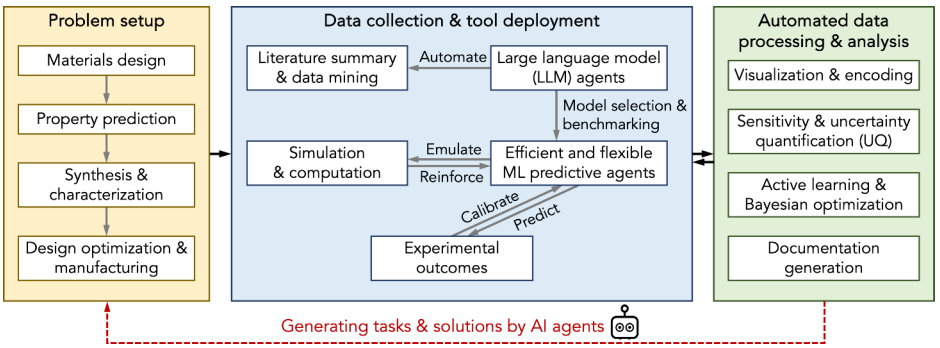


Figure 2. Laboratory workflows automated and accelerated by agentic AI.

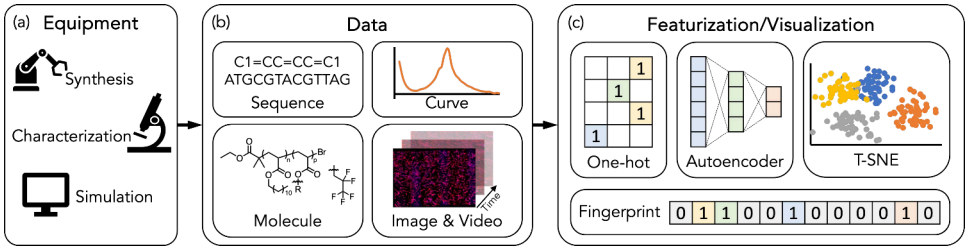


Figure 3. (a–c) Data collection, processing, and featurization in chemical research.

and their performance remains inferior to domain experts when addressing complex chemical problems that require specialized knowledge. Additionally, LLM outputs exhibit randomness, which necessitates careful validation.

Today, we stand at a pivotal moment for radically transforming laboratory research and education. Traditional chemical laboratories require significant human labor for manual experimental designs, product screening, and data analysis, which can be substantially accelerated by robotic systems and AI agents, illustrated by the workflows in Figure 2. The LLMs and ML predictive models can encode multiscale, cross-disciplinary information, enabling scalable and accurate prediction for a large number of test samples, thereby substantially reducing the experimental cost and time. However, many researchers, particularly in experimental science, are unsure where to begin and what ML methods they should use to minimize deployment effort and cost. Although research tasks can be drastically different between chemical science communities, many involve forming, predicting, and understanding chemical relationships, i.e. f: $\mathbf{x} \rightarrow f(\mathbf{x})$, where $\mathbf{x}$ can be descriptors of molecules, chemicals, experimental conditions or experimental outcomes, such as microscopy images and scattering curves, and f is a function that maps the input to system properties, such as conductivity, chemical reaction yields, structural and mechanical properties of the materials. Our modern world is built upon the discovery of maps that accurately predict previously unknown relationships. In the past, however, to discover the underlying principles of a new system, chemists often relied on time-consuming lab experiments and manual analysis of data in a traditional lab.

Two critical advances have paved the way for data-driven discovery of unknown relationships in chemical science. First, experimental and simulation data have gradually become digitalized, enabling the use of fundamental statistical learning principles, such as Bayes' theorem, to automatically update rules from the *status quo*, or prior distribution, to a new paradigm, or posterior distribution, by conditioning on new data. Second, ML models have advanced over the years to learn complex relationships from data, such as numerical values, texts, and sequences, which can substantially reduce time and computational cost for analyzing complex data. Through the lens of these changes, this outlook article will assess the current status of chemical laboratory research, highlight existing gaps, and suggest a path for uniting experimental and computational communities to accelerate progress.

■ ACCELERATING DATA COLLECTION AND PROCESSING

Data Acquisition

Materials synthesis, characterization, and simulation are three main sources of chemical data, shown in Figure 3(a), which produce, for instance, molecular sequences, curves, images, and videos (Figure 3(b)). The key goals are to accelerate and automate data collection, processing, and featurization (Figure 3(c)) for guiding the process of learning chemical relationships.

First, advances in automation are transforming the way materials are synthesized and fabricated for downstream analysis.[2,13,58] Robotic platforms can be flexibly programmed to perform a range of chemical reactions and formulations with high precision and reproducibility, enabling parallel experimentation in multiwell plate formats.[2,5,59] Flow chemistry further extends automation by providing continuous control over reaction conditions, incorporating in-line characterization tools for real-time monitoring, and improving safety when handling hazardous compounds.[60−62] Once reactions are complete, automated flash purification systems and preparative high-performance liquid chromatography[63] streamline isolation of small molecules and can be adapted to generate well-defined polymer libraries with minimal human intervention.[64,65] Beyond producing physical samples, these automated platforms generate distinct types of records, including molecular

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

structures, reaction conditions, and experimental procedures, which can be digitized into machine-compatible formats. For instance, information on molecular structures can be converted into SMILES and SELFIES strings.[66−68] Furthermore, efforts are being made to standardize experimental procedures, such as the Open Reaction Database[69] and Chemical Description Language,[70] for training ML models to optimize synthesis and reaction conditions. Commonly used methods to represent discrete inputs include one-hot encoding, which expresses discrete inputs by sequences of '0' and '1', and molecular fingerprints by numerical vectors.[71−73] Encoding these methods helps bridge synthesis outputs with machine learning models that can analyze reaction trends and accelerate discovery.

Second, a wide range of materials characterization tools, including microscopy, rheology, spectrometry, scattering, and spectroscopy, have been developed. These tools generate images, time-series data, spectra, or other quantitative values in chemical laboratories. Data processing tools, such as image segmentation and particle tracking,[74] have been developed for extracting and linking data from microscopy images. These data processing tools have been implemented into software packages, such as ImageJ and Fiji,[75,76] which contain easy-to-use graphical user interfaces (GUIs), empowering users to view and analyze large quantities of data, particularly useful for biochemical research.[77] The availability of a high volume of labeled data enables the development of more accurate supervised learning tools, such as Cellpose,[78] which utilizes a large database of labeled data to train U-Net,[35] a convolutional neural network for segmenting cells from microscopy images. For more challenging scenarios, such as capturing optically dense systems and fast dynamics, Fourier-based tools, e.g. differential dynamic microscopy (DDM),[79,80] remove the need to segment particles to extract system information, e.g. mean squared displacement of the particles, that determine the mechanical properties (storage, loss modulus).[81,82] Building upon existing tools, it is possible to construct probabilistic generative models and automated estimators for existing data processing methods, such as by removing manual selection of the Fourier range in DDM[83] which otherwise needs to be chosen on a case-by-case manner.[84−87]

Third, computational simulations from distinct space-time length scales can provide scientific insights and a pathway to explore chemical systems before conducting chemical experiments.[88−90] These simulations can reveal mechanistic insights prior to experimentation but are often limited by large computational and/or storage costs, and the need for accurate model calibration, such as determining the form of observed model parameters.[91−93] To address this challenge, Meta FAIR has released Open Molecules 2025 (OMol25), a large-scale open-source data set comprising over 100 million density functional theory (DFT) calculations. It aims to provide high-accuracy quantum chemical data to support the development of machine learning models in molecular chemistry.[94] The past decade witnessed the success of ML surrogate models[95−102] for predicting outcomes of expensive simulations, such as the potential energy, force field, and particle density at untested inputs from nanoscale to bulk environment. For example, neural network potentials and Gaussian process regression have been used to accelerate molecular dynamics and DFT calculations.[45,103,104] Integrating ML-accelerated simulations into laboratory workflows can reduce the number of experiments in laboratories and guide synthesis toward the most promising targets. Realizing this vision requires closer collaboration between experimental and computational communities, ensuring that simulation-informed predictions are seamlessly incorporated into automated experimentation and data-driven discovery workflows.

As the tools used to inform laboratory operations have expanded and evolved, so has the need to record and manage data from these systems. Software, such as LIMS and ELNs, is capable of providing mechanisms for researchers to catalog and record key experimental data in ways that are searchable, labeled, uniquely identified, and accessible in machine-readable formats. Additionally, digital representations of laboratory protocols and associated data can simplify sharing and enable greater collaboration between researchers. The information in an ELN can be utilized to provide training data to update data-driven methods for prediction and optimization. Because of these advantages, physical notebooks of laboratories are gradually being replaced by ELNs.[11,105] Furthermore, data from an ELN can be stored in or connected to a LIMS to enable comprehensive lab data management.[106−108] Together, ELN and LIMS serve as tools that can foster open access data for researchers to retrieve, review, and analyze.

Input Featurization and Visualization

As the input or descriptor $\mathbf{x}$ is not often available to learn chemical relationships $f(\mathbf{x})$, domain knowledge, cheminformatics, and simulation are often used to generate feature sets that capture underlying chemical structures. Representative cheminformatics packages, including OpenBabel, RDKit, and CDK, have been integrated with popular programming languages (Table 1),[109] which enables processing scientific data to obtain meaningful input features for a wide range of problems.

Table 1. Examples of Typical Cheminformatics Packages

Cheminformatics package	Languages	Strength
OpenBabel	C++, Python, Java	Format conversion, Structure search
RDKit	C++, Python	Molecular analysis, ML
CDK	Java	Computational chemistry, Bioinformatics

Featurization of molecules and materials depends on available data, and they are often problem-specific. For applications involving categorical variables, such as catalyst type or solvent selection, one-hot encoding can represent each category as a binary vector. Yet this approach significantly increases the feature dimension when dealing with a large number of categories. To predict molecular properties, molecular fingerprints are widely used to encode molecules as fixed-length binary vectors, where each entry indicates the presence of a specific substructure,[110] This approach can rapidly screen similarity across large molecular libraries and it is widely used in pharmaceutical virtual screening and drug similarity searches. However, most molecular fingerprints provide 2-dimensional structural similarity and do not capture 3-dimensional geometry or electronic effects. To accurately capture 3-dimensional spatial configurations, electronic effects, and physicochemical interactions, computational tools, such as DFT calculations and molecular dynamics simulations, provide quantitative characterization of systems, such as partial charges and activation energies, which can be used as descriptors or

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

features.[111,112] A drawback of these computational models is their high computational cost, and the assumptions of the computational models may not be satisfied in some real-world systems. To reduce the high computational cost, ML surrogate models have been developed to predict system properties, including potential energy, electronic density, and atomic forces, from atomic features, such as inverse pairwise distances between atomic coordinates.[96,97,101,113−115] Furthermore, the discrepancies between the computational models and reality can be modeled by Gaussian processes or their variants through model calibration approaches.[116−118]

> Accelerated data collection, processing, and featurization through computational and data-driven approaches enabled constructing advanced predictive models that encode multiscale, cross-disciplinary information to rapidly screen huge input spaces and substantially reduce experimental cost and time.

Furthermore, exploratory data analysis tools are commonly used for visualization and featurization.[119] A common challenging scenario for featurization involves high-dimensional data, including curves, images, or videos, and discrete inputs such as molecular sequences and graphs. Unsupervised dimension reduction tools, such as principal component analysis (PCA),[120] t-distributed stochastic neighbor embedding (t-SNE),[121] uniform manifold projection and reduction (UMAP),[122] dynamic mode decomposition,[123] autoencoders and decoders,[26,36] are developed for extracting features of high-dimensional data. These methods can be used to visualize the high-dimensional data sets, and the vectors with a reduced dimension can be input as the features for ML models. As some information on the data sets will be inevitably lost in data reduction processes,[124−126] it is crucial to understand the underlying assumptions of these data-driven approaches. For instance, the linear subspace of PCA is equivalent to the maximum marginal likelihood estimator of a probabilistic latent factor model,[127] and one can examine whether the probabilistic model is suitable to represent the data sets for dimension reduction. Domain knowledge, such as physical and chemical principles, can also be used to reduce the dimension of data and improve the accuracy of noisy experimental data. For instance, for classifying phases of block copolymers by small-angle X-ray scattering (SAXS) data, using several features relevant to the location, width, and curvature of the primary peaks of the X-ray curves substantially improves the predictive accuracy of ML models compared to using the entire curve as input in ML models.[65] Furthermore, scattering measurements were used to estimate the micelle structure of block copolymer solutions inversely,[128] and ML surrogate models can improve the inverse estimation by learning the map from reduced-dimensional features of micelle structural parameters to scattering patterns.[129]

The overarching goal of featurization is to inform the similarity of chemical candidates in terms of their system properties. A common challenge of featurization involves discrete or categorical inputs, such as different types of atoms, molecules, and chemical bonds. Compared with numerical inputs, discrete inputs are more challenging to model due to the lack of ordering between the inputs. ML models have achieved success for predicting discrete sequences in some applications, including transformers in LLMs that predict the next text token given the context,[48] and AlphaFold that maps amino acids to protein spatial structure.[42] These advanced ML approaches are capable of learning latent features embedded in a continuous space from a large number of samples. Large opportunities exist to emulate the success in these examples in other areas of chemical and materials sciences, by identifying objective functions, constructing standardized data sets, and developing novel ML architectures for discovering complex relationships from high-dimensional discrete inputs or mixed discrete and continuous inputs.

LEARNING CHEMICAL RELATIONSHIPS BY PREDICTIVE MODELS

Predictive Models

A predictive model, sometimes referred to as statistical methods of chemometrics by chemists,[130] is an indispensable component for learning chemical relationships. A common goal is to predict system properties for a given input or condition and quantify the uncertainty of the prediction, a process typically involving training a data-driven predictive

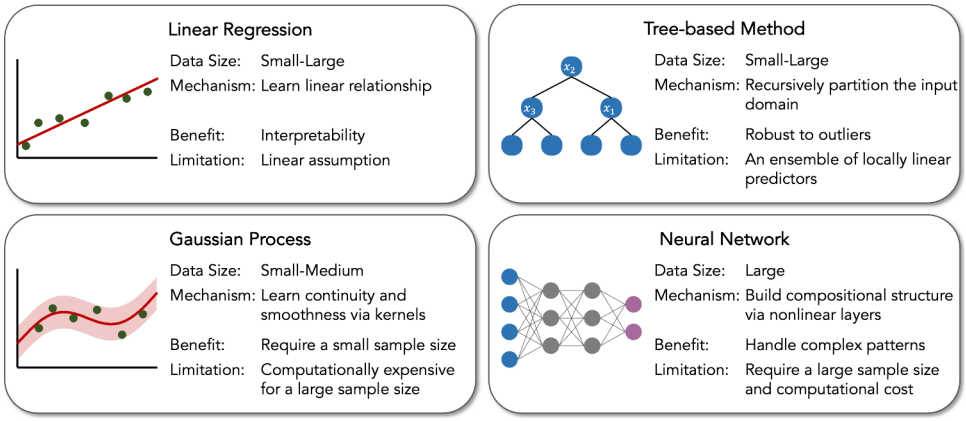


Figure 4. Data-driven predictive models for chemical research.

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

model. We will start from regression, in which a continuous-valued outcome is typically modeled by

$$y = f(\mathbf{x}) + \epsilon \quad (1)$$

where $f(\mathbf{x})$ that maps the input vector $\mathbf{x} = (x_1, ..., x_p)^T$ to system properties, and ϵ is an independent noise, often assumed to follow a zero-mean Gaussian distribution. We introduce four classes of widely used predictive models listed in Figure 4. All these models can be generalized to predict categorical data and counts (often called classification), which utilizes a link function to convert continuous numerical predictions into probabilities. For example, a logistic function[131] can map any real number to a probability value between 0 and 1.

A linear model or linear regression is potentially the oldest and most widely used benchmark model. In linear regression, the relationship is assumed to be linear: $f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} x_j\beta_j$, where β_0 is the coefficient of an intercept, x_j and β_j are the jth input variable and coefficient, respectively. Statistical theory has been well established for estimating the coefficient of linear regression, along with their uncertainties, for noisy observations. Due to the assumption of linearity, linear regression typically does not require large amounts of data to estimate the parameters. However, when processing chemical measurements (such as spectral data), extremely high dimensionality and multicollinearity- of the data could cause traditional regression approaches to fail. Partial least-squares regression extends linear models to high-dimensional spaces,[23] which integrates the dimension reduction idea of PCA. In addition to prediction, linear methods offer a rigorous framework for statistical inference, hypothesis testing, and variable selection for automating model construction.[132,133] Therefore, though the predictive power of a linear model is constrained by its restrictive assumption, the interpretability and the ease of fitting the linear model make it a suitable benchmark model to estimate unknown chemical relationships.

Tree-based methods[134] assume locally constant relationships through partitioning the input variable or feature space: $f(\mathbf{x}) = \sum_{m=1}^{M} c_m 1_{\mathbf{x}\in\mathcal{R}_m}$, where $\mathcal{R}_m$ is an input subspace and c_m is a fitting value of the subspace.[135] Ensembles of trees, such as random forests[28,30,136] and gradient-boosted trees,[31] construct multiple trees for prediction or classification, and they are widely used for their robustness and ability to model nonlinear relationships. Random forests, for instance, construct multiple decision trees in parallel, each trained on a bootstrap sample generated by randomly sampling the observed input-output pairs with replacement. Predictions are obtained by aggregating across all trees, via the majority vote for classification or averaging for regression, which reduces variance and mitigating overfitting. In contrast, gradient-boosted trees are built sequentially, with each new tree focusing on correcting the residuals or errors of the previous model. These methods naturally handle both numerical and categorical inputs, are insensitive to feature scaling, and are computationally efficient. In addition, tree-based methods and ensembles of trees provide feature importance metrics to quantify the contribution of each feature to prediction accuracy, which allows researchers to identify key features that dominate the properties of molecules or materials. Tree-based methods, however, may not be efficient to capture smooth functional relationships, as they are constrained by the assumptions that outputs can be approximated by the ensembles of local predictors of the input subspaces.

Gaussian process regression (GPR) utilizes a kernel function to characterize the distance between the input space.[32] Any latent functions evaluated at n input values, $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))^T$, are assumed to follow a multivariate normal distribution in GPR: $\mathbf{f} \sim \mathcal{MN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is a n-vector of mean often assumed to be constant, and $\boldsymbol{\Sigma}$ is an $n \times n$ covariance matrix that often parametrized by a kernel function. When $f(\cdot)$ is a continuous function, the two outcome values will be similar if the associated two inputs are close to each other. The distance between inputs is characterized by kernel functions in GPR, to measure how close two sets of input conditions, such as experimental conditions and molecular descriptors, are to each other. Commonly used kernel functions include power exponential kernels and Matérn kernels,[137] and chemical information and physical symmetries can be integrated into kernel functions for prediction.[96,138] Conditioning on a set of observations, the predictive mean of GPR provides point predictions, equivalent to solution of kernel ridge regression,[139] and the predictive intervals of GPR assess the confidence of a prediction, which identifies which experimental conditions are worth further exploration. Compared to linear models and tree-based models, Gaussian processes often require less training data to learn nonlinear relationships, when the underlying map is smooth. The high efficiency with respect to small samples and availability of uncertainty make the Gaussian process a suitable candidate for predictions and design optimization in chemical studies.[45] When the sample size is large, approximation methods[140,141] are often required for GPR due to the computational expense that involves inversion of a large covariance matrix, yet the predictive accuracy can deteriorate.

Artificial neural networks (NNs) are capable of learning intricate patterns from large data sets by stacking multiple layers of nonlinear transformation.[142,143] Each layer applies a nonlinear activation function $\sigma(\cdot)$ after a linear transformation of the input with weight matrix $\mathbf{W}^{(l)}$ and bias vector $\mathbf{b}^{(l)}$, allowing the network to approximate arbitrarily complex functions by combining nonlinear building blocks. Mathematically, a feedforward NN can be represented as $f(\mathbf{x}) = f^{(L)}(f^{(L-1)}(...f^{(1)}(\mathbf{x})))$, where each layer follows $f^{(l)}(\mathbf{x}^{(l-1)}) = \sigma(\mathbf{W}^{(l)}\mathbf{x}^{(l-1)} + \mathbf{b}^{(l)})$ with $\sigma(\cdot)$ acting on each element of the input. The large number of parameters enables neural networks to effectively learn nonlinear input-output mappings, including those relationships that are difficult to model using traditional methods. In recent years, many NN architectures,[144] such as convolutional neural networks[34] and recurrent neural networks,[33] have found great success, particularly for image analysis such as image classification,[145] segmentation,[35] generation and inpainting.[37,38] As the neural network models often require a large amount of data to train, they are suitable for certain scenarios, such as learning potential energy and atomic forces from simulation,[115,146] and segmenting cells from microscopy images.[78]

Examples of the Python and R packages for the four classes of predictive models are given in Table 2. These approaches have been widely used for predicting experimental outcomes[44,160] or as a surrogate model for approximating computationally expensive simulations.[98] In practice, it is also

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

Table 2. Examples of Python and R Packages for Predictive Models

Predictive models	Python packages	R packages
Linear regression	scikit-learn[147]	stats,[148] glmnet[149]
Tree-based models	scikit-learn, XGBoost[150]	randomForest,[136] xgboost[151]
Gaussian processes	scikit-learn, GPyTorch[152]	RobustGaSP,[153] GpGp[154]
Neural networks	PyTorch,[155] TensorFlow,[156] Keras[157]	torch,[158] keras[159]

critical to have reliable uncertainty quantification of the predictions, expressed as predictive intervals, for optimizing experimental designs[161] and controlling predictive error.[162] The predictive intervals of linear regression and Gaussian processes are often used to quantify uncertainty of the prediction, and they can be easily computed. The predictive uncertainty of tree-based methods or ensembles of trees can be computed through Bayesian inference, quantile regression, or asymptotic analysis.[163−165] Assessing the uncertainty of neural network approaches is still an open area of research, with various methods been developed, including dropout, ensemble samples, and conformal inference.[166−170]

Experimental Design Optimization

Leveraging the predictive power from simulation and ML methods enables the efficient design of experiments to understand an enormous space of molecules and materials. A primary goal of efficient materials design can be mathematically formulated as an optimization problem: $\mathbf{x}^* = \arg\max_{\mathbf{x}} g(\mathbf{x})$, where $g(\mathbf{x})$ is the objective function, representing the system property to be maximized or minimized (such as reaction yield) for given input $\mathbf{x}$ (such as materials and experimental conditions). DoE offers powerful tools for exploring multivariate variable spaces, particularly for factorized inputs, which can substantially accelerate the conventional process to tune one variable at a time (OVAT).[171] As experimental data contains noise, and design input space can be enormous, applying traditional optimization methods such as a quasi-Newton method[172] can be prohibitive, as they typically require gradient information, noise-free outcomes of the objective functions, and a relatively large number of evaluations. To overcome these challenges, a predictive model, such as a Gaussian process, can be used as a probabilistic proxy to sequentially design the next experiments that give the most valuable experimental outcome through an acquisition function (a strategy that balances exploration of new regions with exploitation of known good regions). This process is often referred to as Bayesian optimization or active learning.[173] The quantified uncertainty from the predictions is crucial to strike a balance between exploration and exploitation for making better predictions and improving the gain function, respectively.[174]

FILLING THE GAPS BY LLM AGENTS

Advancing laboratory research involves a large set of tools and techniques. The rise of LLMs, such as ChatGPT, offers a promising path forward in connecting distinct domains to accelerate learning and problem formulation processes, where LLMs act as the agent at the interface between chemists and data scientists, enabling researchers to quickly learn knowledge in other disciplines. Figure 5 illustrates several potential applications of LLMs, including generating computer code to perform data analysis for chemists and helping computational experts better understand concepts in chemistry which leads to developing new computational tools for predicting polymer phases that will be introduced in the first case study.[65] By accelerating learning processes and reducing communication barriers, LLMs can serve as helpful mediators to facilitate collaborations between distinct communities.

Several recent studies have explored the use of LLMs in chemical research, including assisting with coding and framing scientific questions using chemical data.[175−177] LLMs offer an accessible entry point for novices lacking computational skills, enabling efficient data processing, high-quality visualization,[178] and generating computer codes with little prior programming experience.[179,180] In surveys conducted after introducing LLMs as learning tools, users reported notable improvements in their coding skills, demonstrating that LLMs can accelerate learning with minimal barriers.[179] Beyond basic use, LLMs can support general chemistry problem-solving,[181] and they can be fine-tuned for domain-specific tasks to further enhance output quality.[182] However, the responses from LLMs may not be accurate, and they sometimes may hallucinate. Consequently,

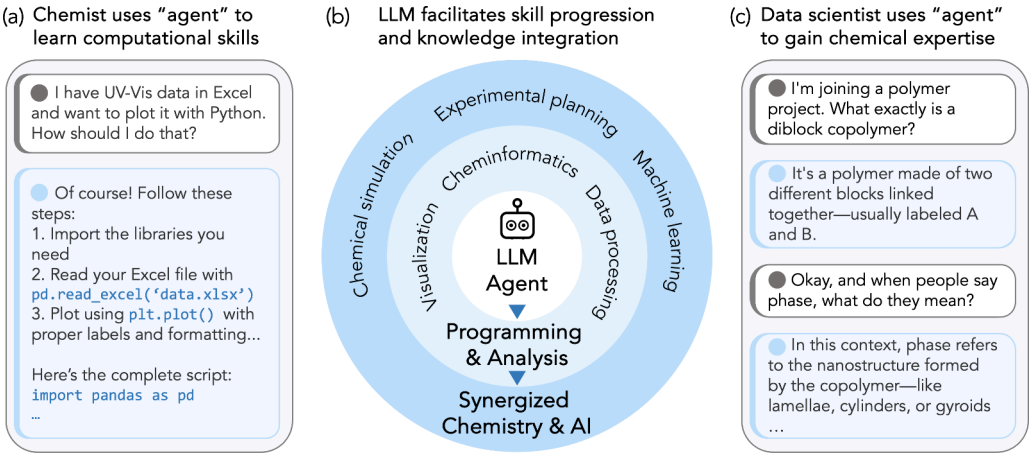


Figure 5. LLM agents facilitate cross-disciplinary collaboration and skill development in chemical research. (a) Example dialogue of chemists acquiring Python programming skills for data analysis. (b) Skill progression framework from basic computational tools to advanced chemistry-AI applications. (c) Example dialogue of LLM agents helping explain chemical concepts.

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

LLMs cannot replace experts in many disciplines, and their answers should be verified by domain experts.

Figure 6 provides examples of distinct expertise from chemists and data scientists for building a collaborative workflow with the aid of LLM agents. Conversely, the expertise can contribute to enhancing LLM agents, as LLMs are essentially trained on text sequences, including dialogues, publications, and computer code. To amplify the utilities of LLMs, it is imperative to educate students and researchers on framing the laboratory research tasks as properly defined mathematical problems. Furthermore, As the LLM agents largely remove the barriers of learning and programming, the existing curriculum of chemical science can include more components of statistical machine learning and data analysis with the assistance of LLM agents.

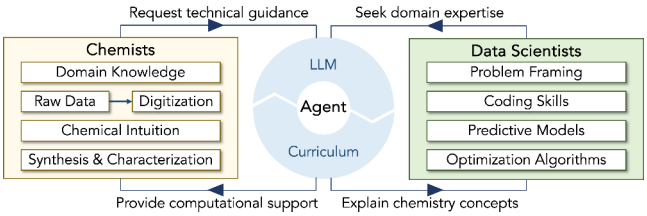


Figure 6. Collaborative workflows between chemists and data scientists facilitated by LLM agents.

> LLM agents can serve as an interface between chemists and data scientists, to assist in a variety of tasks, such as lowering the barriers for accessing the domain knowledge and performing data analysis, but their answers should be validated carefully.

CASE STUDIES

Physics-Informed Machine Learning for Automated Block Copolymer Phase Identification

Nature has long mastered the synthesis and use of well-defined macromolecules in biology. While this level of structural specificity remains out of reach with most synthetic polymers, significant progress has been made in preparing precise polymers and developing new strategies to access well-defined materials in high-throughput.[183−187] When these methods leverage common laboratory equipment that is simple to use and broadly available, it can facilitate widespread use in answering fundamental questions or carefully tailoring structure−property relationships for a specific application.[64] For example, Hawker and co-workers have recently demonstrated the use of automated chromatography to rapidly generate block copolymer libraries.[188] Block copolymers are an important class of materials that self-assemble into a rich array of nanoscale morphologies.[189] Key to applications, such as advanced separation membranes, thermoplastic elastomers, photonic crystals, microelectronics, and drug delivery, is the ability to tune self-assembly through synthetic handles, including block chemistry, block sequence, composition, molecular weight, and dispersity using controlled polymerization techniques.[190−193] This long list of structural variables illustrates the difficulty in navigating and controlling a multidimensional design space. Traditional methods of constructing even an incomplete block copolymer phase diagram involve iterative synthesis followed by multiple purification and isolation steps, which are time-consuming and labor-intensive. The repetitive synthesis of multiple block copolymers is also complicated by slight variations in reaction conditions and/or purification that led to undesired differences among samples and the presence of variable amounts of homopolymer impurities.

This process can be substantially accelerated and automated by leveraging the advances of techniques and predictive models shown in Figure 7. For example, a library of 20 well-defined diblock copolymers, spanning a broad range of compositions, was readily prepared in 1 h from a single parent block copolymer and used to prepare an enhanced phase diagram.[188,194,195] Because automated chromatography accelerates polymer library construction so significantly, it is essential to pair it with more efficient methods for mapping phase diagrams of diverse block copolymer chemistries. SAXS

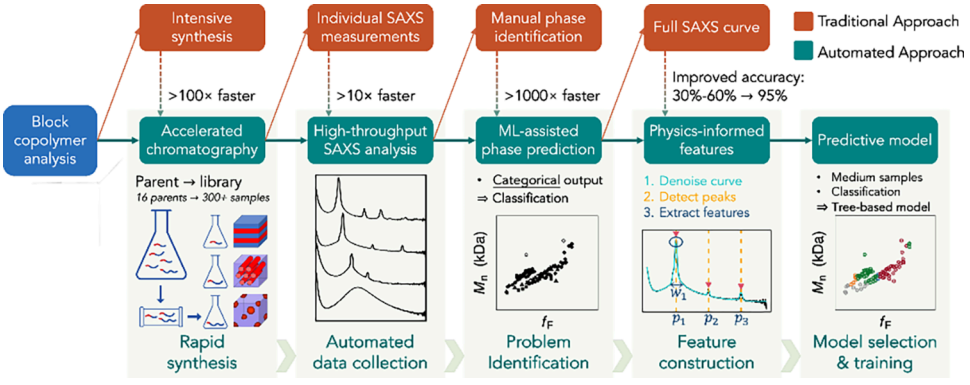


Figure 7. Accelerated workflow for block copolymer phase identification comparing traditional (red) and automated (green) approaches. At each decision point, automated approaches reduce time or improve accuracy compared to conventional methods. Adapted with permission from ref 65. Copyright 2025 Wiley.

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

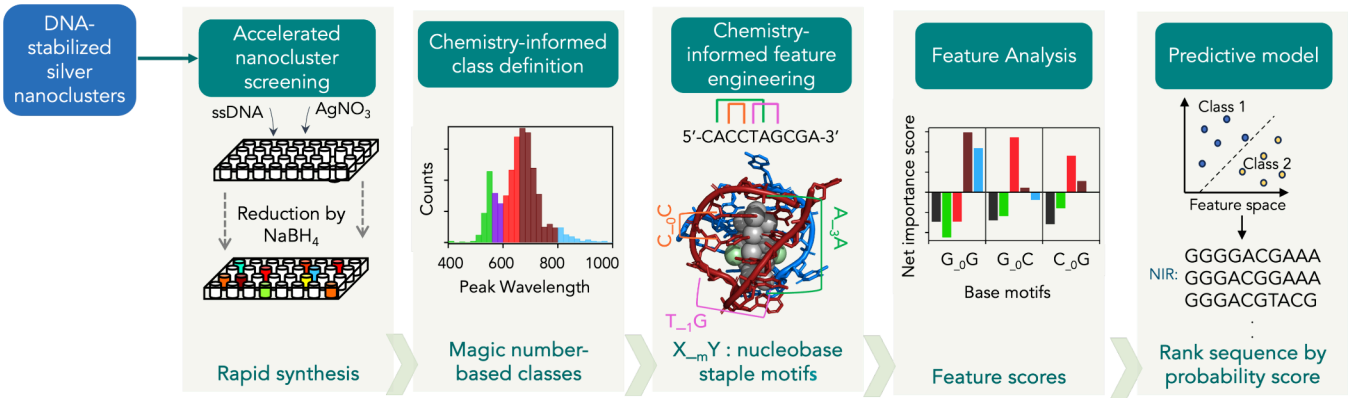


Figure 8. Workflow for ML-enabled DNA-Ag_N discovery. Experimental DNA-Ag_N synthesis is performed on 10^3 DNA oligomers with different sequences, and automated fluorimetry is used to generate training data for ML models. Chemical information guides the choice of the ML problem definition and feature engineering, enabling predictive ML with limited experimental training data and interpretation of sequence-to-property relationships learned by the model. Adapted with permission from ref 209. Copyright 2022 American Chemical Society.

can determine the polymer phases of these samples, yet it requires an expert to manually identify the phase of the polymer by interpreting SAXS curves, which is time-consuming. This problem This problem was addressed by BioPACIFIC MIP team through the development of a physics-informed predictive model to automate polymer phase identification from SAXS.[65] Instead of inputting the entire SAXS data into ML models for classifying polymer phases, the authors extend the Kalman filter[196] for automated peak detection to extract physics-informed morphological features (PIMF), including the peak locations, width, and sharpness of the peaks. These features are used to construct a random forest model,[30] suitable for classification problems with a small to medium number of training samples. Identifying the phases of hundreds of samples using the random forest model takes less than a second on a desktop computer, and it can be executed without the help of a computational expert.

The PIMF from SAXS curves substantially improved the predictive accuracy, achieving around 95% out-of-sample accuracy even for predicting new monomers with different volume fractions not in the database for training ML models.[65] The substantial improvement comes from the integration of polymer theory for featurization in machine learning algorithms for determining polymer phases, which dramatically reduces the dimension of the input space in predictions. Furthermore, the maximum prediction probability from a machine learning model, such as a random forest classifier, can be used for quantifying the uncertainty of the prediction. The assessed uncertainty enables reinspecting a small subset of the samples with maximum prediction probability lower than a prespecified threshold, to achieve near 100% accuracy for polymer phase identification. Furthermore, the authors found 3 samples that were mislabeled by the expert but predicted correctly by the ML model.

As polymer phase identification is a new problem for the data scientists, the LLM was used to efficiently acquire domain-specific knowledge about block copolymer behavior and SAXS curves, as illustrated in Figure 5(c). This LLM-assisted process accelerates the learning process required in interdisciplinary collaboration. This example illustrates how the integration of advanced experimental approaches, data-driven predictive models, and domain expertise expedites the characterization of structure−property relationships.

ML-Guided Experimental Screening for Discovery of DNA-Stabilized Silver Nanocluster Fluorophores

DNA-stabilized silver nanoclusters (DNA-Ag_N) are ultrasmall fluorescent nanoparticles with highly tunable properties. First reported in 2004, DNA-Ag_N contains only 10 to 30 silver atoms stabilized by one to three single-stranded DNA oligomers.[197−199] DNA-Ag_N are attractive for their sequence-tuned excitation and emission wavelengths that can be tuned from blue to near-infrared (NIR) by the DNA template sequence.[200,201] Together with high quantum yields and extinction coefficients, these properties make DNA-Ag_N promising emitters for biosensing, bioimaging, and nano-photonics.[202,203] For example, emerging NIR-emitting DNA-Ag_N could enable deep tissue imaging within the NIR tissue transparency window, where biological tissues and fluids are highly transparent to electromagnetic radiation.[204]

The unique sequence-programmed nature of DNA-Ag_N presents opportunities to engineer these emitters precisely for specific applications, but DNA-Ag_N design is highly challenged by the large number of possible templating DNA sequences. Most sequences do not yield useful fluorescent DNA-Ag_N, and the rules connecting DNA sequence to DNA-Ag_N properties are complex.[205] Moreover, very few X-ray crystal structures of DNA-Ag_N have been reported, and first-principles computational modeling is currently intractable for DNA-Ag_N design.[200,206−208]

Copp, Bogdanov, and coauthors have developed approaches that combine high-throughput experimental synthesis and characterization with ML models[209−213] to significantly increase DNA-Ag_N design efficacy, using the workflow in Figure 8. First, automated liquid handling is used to synthesize DNA-Ag_N on 10^3 different DNA oligomers in well plates, with one oligomer sequence per well. The fluorescence spectrum of each sample is then collected using automated fluorimetry with a well plate reader; universal UV excitation via the nucleobases is employed to excite all DNA-Ag_N with a single wavelength for rapid fluorimetry. Finally, automated spectral fitting is used to determine the spectral peak parameters for each DNA sequence, thereby generating a large data library that connects DNA sequences to DNA-Ag_N fluorescence.

This data set has been leveraged to train chemistry-informed classification models, due to the quantized "magic number" properties of nanoclusters, which naturally yield certain DNA-Ag_N sizes.[205] Chemically informed featurization has been

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

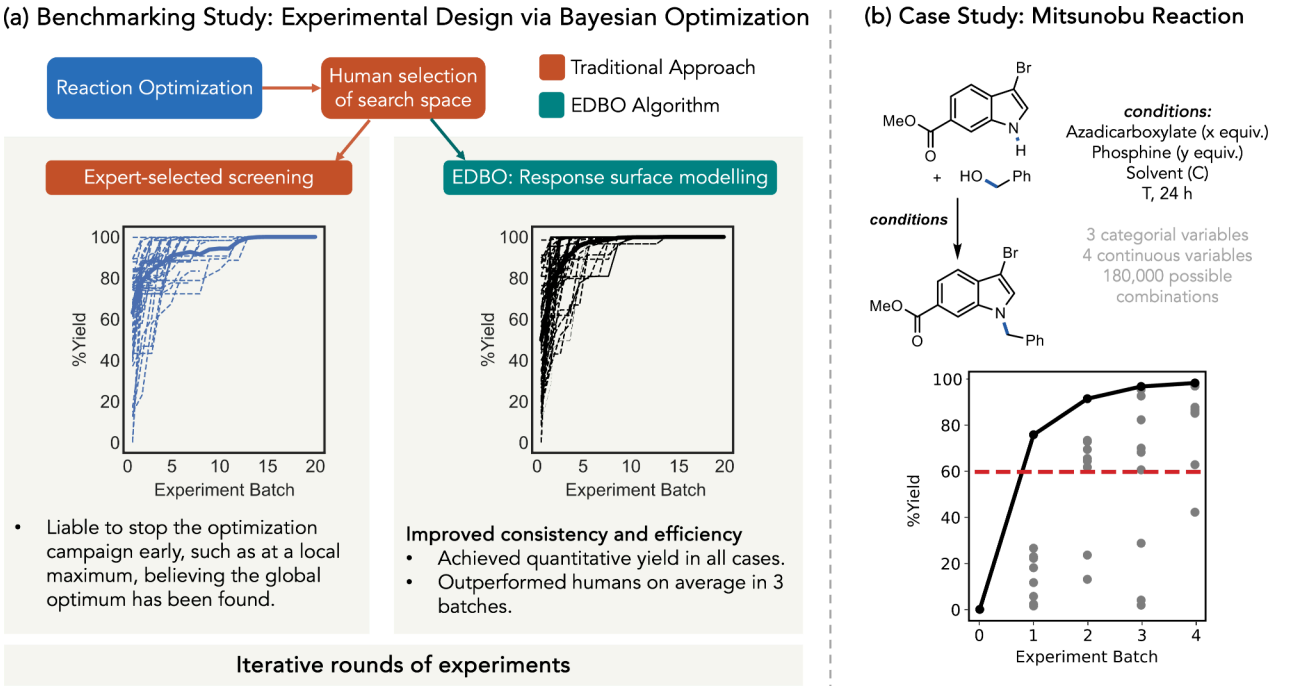


Figure 9. Experimental design via Bayesian optimization. (a) Validation of Bayesian reaction optimization via direct comparison between human performance (left) and machine learning performance (right); optimization curves for individual players and optimizer runs (dashed) and average (solid) as a function of experiment batch (size: 5). (b) Optimization of a Mitsunobu reaction via EDBO: cumulative best observed yield (black) and individual experiment outcomes (gray) as a function of experiment batch (batch size: 10), yield for standard reaction conditions (red dashed). Adapted with permission from ref 47. Copyright 2021 Springer Nature.

essential for ML classifiers to learn sequence-to-color relationships, rather than using simple methods such as one-hot encoding. For example, by featurizing DNA sequence using nucleobase "staple" motifs inspired by DNA-Ag_N crystal structure,[207] support vector machines[27] were trained to predict the emission color class of a DNA-Ag_N given input DNA sequence.[209] To ensure that the ML models are effective for predicting nanocluster properties, it is important to mitigate overfitting and data imbalance issues commonly encountered in experimental nanocluster data sets, which are often limited in size. An imbalanced data set is one that contains relatively few examples of some types of materials while containing more examples of others, e.g., the histogram in Figure 8 panel 2 with very few data points with wavelengths >900 nm as compared to other wavelengths. ML models trained on imbalanced data sets can perform poorly in regions where little training data is available. Regularization techniques by adding a penalty term to the loss function in estimating parameters are effective in reducing overfitting in SVMs,[214,215] and classifier ensembles can be used together with data balancing strategies such as subsampling to achieve predictive ML models with limited experimental data sets. More recently, deep learning models that perform automatic feature extraction and enable continuous property design were introduced and demonstrated for DNA-Ag_N.[210,211] Beyond prediction, ML models can provide valuable chemical insights into how DNA sequence influences DNA-Ag_N color through interpretability analysis using feature analysis tools such as BorutaSHAP.[216]

Experiments have verified the efficacy of ML-guided design approaches for DNA-Ag_N. One of the most notable findings is the discovery of NIR-emitting DNA-Ag_N, which are rare in training data libraries, yet can be designed at a 12.3 times enhanced success rate using ML-guided sequence selection.[209] This strategy illustrates the strength of integrating domain knowledge (DNA-Ag_N chemistry) and ML algorithms to facilitate the systematic discovery of materials and to enhance fundamental chemical understanding in ways that are not achievable using conventional methods.

Open-Source Bayesian Optimization Tool for Reaction Development in Small-Molecule Organic Synthesis

Experimental optimization is ubiquitous in small-molecule organic synthesis. These optimization problems are usually high-dimensional, with reaction spaces defined by both categorical variables (e.g., reagent and solvent identities) and continuous variables (e.g., catalyst loading and temperature). A synthetic chemist selects the initial reaction space to explore based on successful conditions for similar reactions, mechanistic reasoning, and chemical intuition, then iteratively performs rounds of experiments with varied conditions to seek the optimum. The most common conventional strategy for exploration of this space, namely OVAT testing, has proven effective, but is inefficient for exploring a large number of variables and overlooks interactions between variables.

Bayesian optimization (BO) is well-suited to reaction optimization, as it can suggest multiple experiments by utilizing the quantified uncertainty of a predictive model in a search space defined by both categorical and continuous parameters, to ultimately identify the global optimum in a low-data regime.[47] In this setting, a successful BO algorithm could substantially reduce the number of experiments necessary to complete optimization.

In 2021, the Doyle group developed Experimental Design via Bayesian Optimization (EDBO), an open-source Python package for reaction development.[47] The algorithm was tested with featurization at three different levels of chemical information: one-hot encoding, wherein the algorithm is

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

given only the name of each reaction component and its binary presence or absence in each set of conditions; Mordred cheminformatics descriptors,[217] which provide basic information about reaction components such as polarizability, molecular weight, and number of aromatic rings; and properties that require DFT calculation, such as the charges at the coordinating atoms of potential ligands in their minimum energy conformers, which are the most computationally expensive but also provide the richest information. When tuned for each level of featurization using experimental data from the chemical literature, the algorithm performed similarly on average during the optimization process. However, DFT-level encoding gave the most consistent results, with its worst optimization trials still converging on a yield within 5% of the optimum. DFT encoding was thus selected for use in the remainder of the study; these results suggest that users with similar search spaces may still find the algorithm effective even with only simple one-hot encoding. Additionally, optimizer tuning revealed the best performance using a Gaussian process surrogate model[152] and parallel expected improvement[218] as the acquisition function. This acquisition function suggests batches of experiments that maximize expected utility until the objective is optimized or the reaction space is explored sufficiently that the probability of finding an improved condition is low. This platform can be used in diverse settings for any parametrizable reaction, including everyday bench-scale experimentation and automated systems, making it widely applicable for modern chemical laboratories.

To benchmark the EDBO algorithm's performance against the choices of human experts, Doyle and co-workers developed a computer game that asked the player to find the highest-yielding conditions for a Pd-catalyzed C−H arylation reaction within a search space of 1,728 possible reaction conditions, defined by three categorical variables (solvent, ligand, and base identity) and two continuous variables (temperature and concentration). To mimic a real laboratory, the resource budget was limited: players chose 5 experiments to run "per workday" and had 20 "workdays" to maximize the yield of the reaction. The experimental outcomes supplied to the players were real, with the yield data for every possible reaction being collected beforehand via HTE.

For performance comparison, 50 expert chemists were asked to play the benchmarking game and the EDBO algorithm was asked to play it a corresponding 50 times (Figure 9a). While human experts selected higher-yielding conditions on average for the first round of experiments, the optimizer's average performance surpassed humans' average performance in only three "workdays" and typically achieved quantitative yield within the first ten. In addition to EDBO's greater efficiency, it displays improved consistency: the optimizer identified the optimal conditions every time it played the game, while many human participants concluded they had identified the best conditions before achieving quantitative yield and stopped optimization early.

To demonstrate the platform's ability to optimize real-world reactions used in pharmaceutical development, Doyle and co-workers applied EDBO to a test case of the Mitsunobu reaction.[47] This reaction was selected because it is used frequently in synthesis, but tends to deliver moderate yields under standard conditions. Methyl 3-bromo-1*H*-indole-6-carboxylate and benzyl alcohol were chosen as substrates. These substrates afforded a moderate 60% yield of the desired product under the standard conditions used at Bristol Myers Squibb. Seven total categorical and continuous reaction parameters were selected to define the reaction space: the identity and equivalents of the azadicarboxylate reagent, the identity and equivalents of the phosphine reagent, the identity and concentration of the solvent, and the temperature. Chemical information about the reagents and solvent was encoded in the form of DFT-computed descriptors. With 6 azadicarboxylates, 12 phosphines, 5 equivalencies for each reagent, 5 solvents, 4 concentrations, and 5 temperatures, the full reaction space consists of 180,000 possible combinations.

With the search space in hand, EDBO was initialized with conditions chosen at random. Ten reactions were run in parallel per experiment batch. The optimizer identified three conditions that delivered the product in nearly quantitative yield (99%) in only four rounds, totaling 40 experiments (Figure 9b). EDBO's ability to deliver a suite of distinct optimized conditions is advantageous, as it enables chemists to choose between several options based on additional factors such as cost and operational convenience.

In 2022, the Doyle group expanded the utility of EDBO with the release of EDBO+.[219] The upgraded platform accommodates multiobjective optimization and allows the user to modify the reaction space during the optimization campaign. These improvements adapt the system well to common use-cases in organic synthesis, where multiple objectives (e.g., yield, selectivity, cost) are often in play and condition space is routinely updated as the system is better understood. In addition to its availability as an open-source software package, EDBO+ can be used via a web-based application with a step-by-step graphical user interface designed for users who have little to no coding knowledge, which helps bridge the gap between data scientists and experimental chemists. Furthermore, the integration of EDBO+ as a decision-making tool with other data-driven technologies is already showing promise: the year after its release, EDBO+ proved effective for the optimization of a pyridinium salt synthesis via continuous flow with semiautomated low-resolution data processing,[220] which is gaining popularity for automated reaction development.[221,222]

SUMMARY AND OUTLOOK

Chemical lab research has been transformed by the availability of large volumes of digital data generated by high-throughput experimental facilities that are increasingly automated. These data offer unique opportunities to develop new approaches and algorithms to substantially accelerate the discovery process. A key step to advance lab research is to formulate lab tasks as mathematical questions, which is crucial to leveraging progress in machine learning algorithms and AI tools. As many chemical tasks involve identifying unknown relationships, a suitable predictive model can open doors for numerous applications, including accelerating experimental design, processing, and optimization of material properties. To bridge the knowledge gap between distinct areas, LLM agents can help chemical scientists select suitable predictive models, provide standard computer code, and assist computational experts in understanding domain knowledge for developing algorithms to facilitate the discovery process. Furthermore, the answers from LLM agents may inspire new ideas and facilitate the discovery process. Yet LLM agents may generate inaccurate responses and can fabricate or hallucinate information about nonexistent theorems or references, which may lead to unsafe experiments, such as providing access to synthesis information that poses

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

security issues. Prompt engineering, including providing contexts and examples, breaking large research questions into smaller pieces, and integrating coscientists specializing in different domains, can guide LLMs to generate more accurate solutions.[223] Some of these strategies require not only domain knowledge, but also more understanding of data science. Thus, integration of more statistical thinking and machine learning concepts into the pedagogy of chemical science can assist chemists in better interacting with LLM agents and ensuring the correctness of LLM-derived solutions.

Overcoming several other common challenges can lead to fruitful outcomes in advancing lab research. First, many experimental characterization tools produce data that are, to varying degrees, closed-source, meaning that access to the data is restricted to an ecosystem supported only by the vendor. Recent efforts have been made to facilitate connections between closed-source vendor ecosystems and external software (e.g., LIMS, ELN, or analysis tools) by gaining access to application programming interfaces (APIs) directly from the vendors. For example, a software development kit in a common programming language (Python) was developed and released to consume the API for the HTE instrument, thereby providing greater access to system commands.[59] Efforts to convert proprietary data into standard formats and share them in an open-source repository can cultivate community efforts. The availability of a standard format of data has driven, for instance, the progress in LLMs and accurate protein structure prediction tools, such as Alphafold.[42] Furthermore, there is a vast need to develop standard software that can be easily plug-in into daily experimental tasks, including automating data processing, making reliable predictions of chemical relationships, generating interpretable analysis of experiments, and suggesting solutions for experimental challenges. These tools need to overcome several challenges, including the limited number of training samples in experiments, automating model training processes, enabling uncertainty assessment, and assimilation to integrate different types of data. On the other hand, a deeper understanding of the assumptions behind these tools enables chemists to better deploy them in suitable scenarios, identify the reasons when ML tools do not work well, and resolve problems more quickly when interacting with AI agents. Together, the joint efforts in experimental and computational fields can substantially accelerate the discovery process in chemical science.

AUTHOR INFORMATION

Corresponding Authors

Stacy M. Copp − *Department of Materials Science and Engineering, Department of Chemical and Biomolecular Engineering, Department of Physics and Astronomy, and Department of Chemistry, University of California, Irvine, California 92697, United States*; Email: copps@uci.edu

Abigail G. Doyle − *Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California 90095, United States*; orcid.org/0000-0002-6641-0833; Email: abigaildoyle@g.ucla.edu

Javier Read de Alaniz − *Department of Chemistry and Biochemistry, University of California, Santa Barbara, Santa Barbara, California 93106, United States*; orcid.org/0000-0003-2770-9477; Email: jalaniz@ucsb.edu

Mengyang Gu − *Department of Statistics and Applied Probability, University of California, Santa Barbara, Santa Barbara, California 93106, United States*; orcid.org/0000-0002-3959-8965; Email: mengyang@pstat.ucsb.edu

Authors

Saejin Oh − *BioPACIFIC Materials Innovation Platform, University of California, Santa Barbara, Santa Barbara, California 93106, United States*

Xinyi Fang − *Department of Statistics and Applied Probability, University of California, Santa Barbara, Santa Barbara, California 93106, United States*

I-Hsin Lin − *Department of Materials Science and Engineering, University of California, Irvine, California 92697, United States*

Paris Dee − *Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California 90095, United States*

Christopher S. Dunham − *BioPACIFIC Materials Innovation Platform, University of California, Santa Barbara, Santa Barbara, California 93106, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acscentsci.5c01994

Author Contributions

△S.O. and X.F. contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is supported by the BioPACIFIC Materials Innovation Platform of the National Science Foundation (NSF BioPACIFIC MIP) under Award No. DMR-2445868, the NSF Cyberinfrastructure for Sustained Scientific Innovation Program under Award No. OAC-2411043, and Multicampus Research Programs and Initiatives from the University of California under Award No. M23PL5990. We thank Rose Yu from UC San Diego and the anonymous referees for helpful suggestions.

REFERENCES

(1) Abolhasani, M.; Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth.* **2023**, *2*, 483−492.

(2) Tom, G.; Schmid, S. P.; Baird, S. G.; Cao, Y.; Darvish, K.; Hao, H.; Lo, S.; Pablo-García, S.; Rajaonson, E. M.; Skreta, M.; et al. Self-driving laboratories for chemistry and materials science. *Chem. Rev.* **2024**, *124*, 9633−9732.

(3) Vriza, A.; Chan, H.; Xu, J. Self-driving laboratory for polymer electronics. *Chem. Mater.* **2023**, *35*, 3046−3056.

(4) Wang, C.; Kim, Y.-J.; Vriza, A.; Batra, R.; Baskaran, A.; Shan, N.; Li, N.; Darancet, P.; Ward, L.; Liu, Y.; Chan, M. K. Y.; Sankaranarayanan, S. K.R.S.; Fry, H. C.; Miller, C. S.; Chan, H.; Xu, J. Autonomous platform for solution processing of electronic polymers. *Nat. Commun.* **2025**, *16*, 1498.

(5) Seifrid, M.; Pollice, R.; Aguilar-Granda, A.; Morgan Chan, Z.; Hotta, K.; Ser, C. T.; Vestfrid, J.; Wu, T. C.; Aspuru-Guzik, A. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Acc. Chem. Res.* **2022**, *55*, 2454−2466.

(6) MacLeod, B. P.; Parlane, F. G. L.; Morrissey, T. D.; Hase, F.; Roch, L. M.; Dettelbach, K. E.; Moreira, R.; Yunker, L. P. E.; Rooney, M. B.; Deeth, J. R.; Lai, V.; Ng, G. J.; Situ, H.; Zhang, R. H.; Elliott, M. S.; Haley, T. H.; Dvorak, D. J.; Aspuru-Guzik, A.; Hein, J. E.; Berlinguette, C. P. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **2020**, *6*, No. eaaz8867.

(7) Zhang, W.; Guy, M. A.; Yang, J.; Hao, L.; Liu, J.; Hawkins, J. M.; Mustakis, J.; Monfette, S.; Hein, J. E. Leveraging GPT-4 to transform

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

chemistry from paper to practice. *Digit. Discovery* **2024**, *3*, 2367−2376.
(8) Wu, T. C; Aguilar-Granda, A.; Hotta, K.; Yazdani, S. A.; Pollice, R.; Vestfrid, J.; Hao, H.; Lavigne, C.; Seifrid, M.; Angello, N.; Bencheikh, F.; Hein, J. E.; Burke, M.; Adachi, C.; Aspuru-Guzik, A. A materials acceleration platform for organic laser discovery. *Adv. Mater.* **2023**, *35*, 2207070.
(9) Gibbon, G. A. A brief history of LIMS. *Lab. Autom. Inf. Manag.* **1996**, *32*, 1−5.
(10) Du, P.; Kofman, J. A. Electronic laboratory notebooks in pharmaceutical R&D: on the road to maturity. *J. Lab. Autom.* **2007**, *12*, 157−165.
(11) Rubacha, M.; Rattan, A. K.; Hosselet, S. C. A review of electronic laboratory notebooks available in the market today. *JALA* **2011**, *16*, 90−98.
(12) Dai, T.; Vijayakrishnan, S.; Szczypiński, F. T.; Ayme, J.-F.; Simaei, E.; Fellowes, T.; Clowes, R.; Kotopanov, L.; Shields, C. E.; Zhou, Z.; et al. Autonomous mobile robots for exploratory synthetic chemistry. *Nature* **2024**, *635*, 890−897.
(13) Ha, T.; Lee, D.; Kwon, Y.; Park, M. S.; Lee, S.; Jang, J.; Choi, B.; Jeon, H.; Kim, J.; Choi, H.; Seo, H.-T.; Choi, W.; Hong, W.; Park, Y. J.; Jang, J.; Cho, J.; Kim, B.; Kwon, H.; Kim, G.; Oh, W. S.; Kim, J. W.; Choi, J.; Min, M.; Jeon, A.; Jung, Y.; Kim, E.; Lee, H.; Choi, Y.-S. AI-driven robotic chemist for autonomous synthesis of organic molecules. *Sci. Adv.* **2023**, *9*, No. eadj0461.
(14) Saharan, V. A. *Computer Aided Pharmaceutics and Drug Delivery: An Application Guide for Students and Researchers of Pharmaceutical Sciences*; Springer, 2022; pp 381−414.
(15) North, N. Robotics and laboratory automation in pharmaceuticals analysis *Encyclopedia of Analytical Chemistry*; Wiley, 2000. DOI: 10.1002/9780470027318.a1919.
(16) Wells, D. E.; Price, P. J. Simple rapid methods for freezing hybridomas in 96-well microculture plates. *J. Immunol. Methods* **1983**, *59*, 49−52.
(17) Borman, S. Electronic laboratory notebooks may revolutionize research record keeping. *Chem. Eng. News* **1994**, *72* (21), 6.
(18) Weber, A.; Von Roedern, E.; Stilz, H. U. SynCar: an approach to automated synthesis. *J. Comb. Chem.* **2005**, *7*, 178−184.
(19) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **2016**, *104*, 148−175.
(20) Mills, J. E. Design of experiments in pharmaceutical process research and development. *Drug Dev. Ind. Pharm.* **2003**, *43*, 889−901.
(21) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chem. Sci.* **2018**, *9*, 7642−7655.
(22) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-generation experimentation with self-driving laboratories. *Trends Chem.* **2019**, *1*, 282−291.
(23) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(24) Linnainmaa, S. *The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors.* Ph.D. thesis, Master's thesis (in Finnish), Univ. Helsinki, 1970.
(25) Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 2554−2558.
(26) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Learning Internal Representations by Error Propagation*; MIT Press, 1986; pp 318 − 362.
(27) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273−297.
(28) Ho, T. K. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*; IEEE, 1995; pp 278−282.
(29) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123−140.
(30) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.
(31) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189−1232.
(32) Rasmussen, C. E. *Gaussian Processes for Machine Learning*; MIT Press, 2006.
(33) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput* **1997**, *9*, 1735−1780.
(34) LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278−2324.
(35) Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention−MICCAI 2015:18th International Conference, Munich, Germany, October 5−9, 2015, Proceedings, Part III* **2015**, *18* (9351), 234−241.
(36) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. International Conference on Learning Representation; *arXiv*, arXiv:1412.6980, 2014. DOI: 10.48550/arXiv.1312.6114.
(37) Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840−6851.
(38) Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv*, arXiv 2011.13456, 2020. DOI: 10.48550/arXiv.2011.13456.
(39) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; et al. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089−1100.
(40) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493−500.
(41) Ackley, D. H.; Hinton, G. E.; Sejnowski, T. J. A learning algorithm for Boltzmann machines. *Cogn. Sci.* **1985**, *9*, 147−169.
(42) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(43) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.
(44) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C−N cross-coupling using machine learning. *Science* **2018**, *360*, 186−190.
(45) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian process regression for materials and molecules. *Chem. Rev.* **2021**, *121*, 10073−10141.
(46) Zhang, Y.; Apley, D. W.; Chen, W. Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Sci. Rep.* **2020**, *10*, 1−13.
(47) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **2021**, *590*, 89−96.
(48) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.*, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017; *30*.
(49) Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. Improving language understanding by generative pre-training. *Preprint. Work in progress*, 2018.
(50) Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* 2023, arXiv:2303.08774. DOI: 10.48550/arXiv.2303.08774.
(51) Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A comprehensive overview of

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

large language models. *arXiv*, 2023, arXiv:2307.06435. DOI: 10.48550/arXiv.2307.06435.
(52) Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv*, 2023, arXiv:2302.13971. DOI: 10.48550/arXiv.2302.13971.
(53) Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C., et al. Qwen3 technical report. *arXiv*, arXiv preprint arXiv:2505.09388, 2025. DOI: 10.48550/arXiv.2505.09388.
(54) Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C., et al. DeepSeek-V3 technical report. *arXiv*, arXiv preprint arXiv:2412.19437, 2024. DOI: 10.48550/arXiv.2412.19437.
(55) Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **2025**, *645*, 633−638.
(56) Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570−578.
(57) Ramos, M. C.; Collison, C. J.; White, A. D. A review of large language models and autonomous agents in chemistry. *Chem. Sci.* **2025**, *16*, 2514−2572.
(58) Caramelli, D.; Granda, J. M.; Mehr, S. H. M.; Cambié, D.; Henson, A. B.; Cronin, L. Discovering new chemistry with an autonomous robotic platform driven by a reactivity-seeking neural network. *ACS Cent. Sci.* **2021**, *7*, 1821−1830.
(59) Seifrid, M.; Strieth-Kalthoff, F.; Haddadnia, M.; Wu, T. C.; Alca, E.; Bodo, L.; Arellano-Rubach, S.; Yoshikawa, N.; Skreta, M.; Keunen, R.; et al. Chemspyd: an open-source python interface for Chemspeed robotic chemistry and materials platforms. *Digit. Discovery* **2024**, *3*, 1319−1326.
(60) Plutschack, M. B.; Pieber, B.; Gilmore, K.; Seeberger, P. H. The hitchhiker's guide to flow chemistry. *Chem. Rev.* **2017**, *117*, 11796−11893.
(61) Guidi, M.; Seeberger, P. H.; Gilmore, K. How to approach flow chemistry. *Chem. Soc. Rev.* **2020**, *49*, 8910−8932.
(62) Capaldo, L.; Wen, Z.; Noël, T. A field guide to flow chemistry for synthetic organic chemists. *Chem. Sci.* **2023**, *14*, 4230−4247.
(63) Still, W. C.; Kahn, M.; Mitra, A. Rapid chromatographic technique for preparative separations with moderate resolution. *J. Org. Chem.* **1978**, *43*, 2923−2925.
(64) Murphy, E. A.; Zhang, C.; Bates, C. M.; Hawker, C. J. Chromatographic separation: A versatile strategy to prepare discrete and well-defined polymer libraries. *Acc. Chem. Res.* **2024**, *57*, 1202−1213.
(65) Fang, X.; Murphy, E. A.; Kohl, P. A.; Li, Y.; Hawker, C. J.; Bates, C. M.; Gu, M. Universal Phase Identification of Block Copolymers From Physics-Informed Machine Learning. *J. Polym. Sci.* **2025**, *63*, 1433−1440.
(66) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(67) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; et al. BigSMILES: a structurally-based line notation for describing macromolecules. *ACS Cent. Sci.* **2019**, *5*, 1523−1531.
(68) Krenn, M.; Ai, Q.; Barthel, S.; Carson, N.; Frei, A.; Frey, N. C.; Friederich, P.; Gaudin, T.; Gayle, A. A.; Jablonka, K. M.; Lameiro, R. F.; Lemm, D.; Lo, A.; Moosavi, S. M.; Napoles-Duarte, J. M.; Nigam, A.; Pollice, R.; Rajan, K.; Schatzschneider, U.; Schwaller, P.; Skreta, M.; Smit, B.; Strieth-Kalthoff, F.; Sun, C.; Tom, G.; Falk von Rudorff, G.; Wang, A.; White, A. D.; Young, A.; Yu, R.; Aspuru-Guzik, A. SELFIES and the future of molecular string representations. *Patterns* **2022**, *3*, 100588.
(69) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The open reaction database. *J. Am. Chem. Soc.* **2021**, *143*, 18820−18826.
(70) Mehr, S. H. M.; Craven, M.; Leonov, A. I.; Keenan, G.; Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **2020**, *370*, 101−108.
(71) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; John Wiley & Sons, 2008.
(72) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(73) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem.* **2020**, *6*, 1379−1390.
(74) Crocker, J. C.; Grier, D. G. Methods of digital video microscopy for colloidal studies. *J. Colloid Interface Sci.* **1996**, *179*, 298−310.
(75) Schneider, C. A.; Rasband, W. S.; Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **2012**, *9*, 671−675.
(76) Schindelin, J.; Arganda-Carreras, I.; Frise, E.; Kaynig, V.; Longair, M.; Pietzsch, T.; Preibisch, S.; Rueden, C.; Saalfeld, S.; Schmid, B.; et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **2012**, *9*, 676−682.
(77) Luo, Y.; Gu, M.; Park, M.; Fang, X.; Kwon, Y.; Urueña, J. M.; Read de Alaniz, J.; Helgeson, M. E.; Marchetti, C. M.; Valentine, M. T. Molecular-scale substrate anisotropy and crowding drive long-range nematic order of cell monolayers. *J. R. Soc. Interface* **2023**, *20*, 20230160.
(78) Stringer, C.; Wang, T.; Michaelos, M.; Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **2021**, *18*, 100−106.
(79) Cerbino, R.; Trappe, V. Differential dynamic microscopy: probing wave vector dependent dynamics with a microscope. *Phys. Rev. Lett.* **2008**, *100*, 188102.
(80) Giavazzi, F.; Brogioli, D.; Trappe, V.; Bellini, T.; Cerbino, R. Scattering information obtained by optical microscopy: differential dynamic microscopy and beyond. *Phys. Rev. E* **2009**, *80*, 031403.
(81) Bayles, A. V.; Squires, T. M.; Helgeson, M. E. Probe microrheology without particle tracking by differential dynamic microscopy. *Rheol. Acta* **2017**, *56*, 863−869.
(82) Gu, M.; Luo, Y.; He, Y.; Helgeson, M. E.; Valentine, M. T. Uncertainty quantification and estimation in differential dynamic microscopy. *Phys. Rev. E* **2021**, *104*, 034610.
(83) Gu, M.; He, Y.; Liu, X.; Luo, Y. Ab initio uncertainty quantification in scattering analysis of microscopy. *Phys. Rev. E* **2024**, *110*, 034601.
(84) Martinez, V. A.; Besseling, R.; Croze, O. A.; Tailleur, J.; Reufer, M.; Schwarz-Linek, J.; Wilson, L. G.; Bees, M. A.; Poon, W. C. Differential dynamic microscopy: A high-throughput method for characterizing the motility of microorganisms. *Biophys. J.* **2012**, *103*, 1637−1647.
(85) Lu, P. J.; Giavazzi, F.; Angelini, T. E.; Zaccarelli, E.; Jargstorff, F.; Schofield, A. B.; Wilking, J. N.; Romanowsky, M. B.; Weitz, D. A.; Cerbino, R. Characterizing concentrated, multiply scattering, and actively driven fluorescent systems with confocal differential dynamic microscopy. *Phys. Rev. Lett.* **2012**, *108*, 218103.
(86) Bayles, A. V.; Squires, T. M.; Helgeson, M. E. Dark-field differential dynamic microscopy. *Soft Matter* **2016**, *12*, 2440−2452.
(87) Guidolin, C.; Heim, C.; Adams, N. B.; Baaske, P.; Rondelli, V.; Cerbino, R.; Giavazzi, F. Protein Sizing with Differential Dynamic Microscopy. *Macromolecules* **2023**, *56*, 8290−8297.
(88) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules; International Series of Monographs on Chemistry*; Oxford University Press: New York Oxford England, 1989; pp x, 333 p.
(89) Rapaport, D. C.; Rapaport, D. C. R. *The Art of Molecular Dynamics Simulation*; Cambridge University Press, 2004.
(90) Fredrickson, G. *The Equilibrium Theory of Inhomogeneous Polymers*; Oxford University Press, 2006.
(91) Sholl, D. S.; Steckel, J. A. *Density Functional Theory: A Practical Introduction*; John Wiley & Sons, 2022.
(92) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; John Wiley & Sons, 2015.

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

(93) Matsen, M. W. Self-consistent field theory and its applications. *Soft Matter* **2005**, *1*, 87.
(94) Levine, D. S. et al. The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models. *arXiv* ,2025. DOI: 10.48550/arXiv.2505.08762.
(95) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(96) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.
(97) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **2017**, *8*, 1−10.
(98) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, No. e1603015.
(99) Li, H.; Zhou, M.; Sebastian, J.; Wu, J.; Gu, M. Efficient force field and energy emulation through partition of permutationally equivalent atoms. *J. Chem. Phys.* **2022**, *156*, 184304.
(100) Lu, D.; Wang, H.; Chen, M.; Lin, L.; Car, R.; Weinan, E.; Jia, W.; Zhang, L. 86 PFLOPS Deep Potential Molecular Dynamics simulation of 100 million atoms with ab initio accuracy. *Comput. Phys. Commun.* **2021**, *259*, 107624.
(101) Behler, J. Four generations of high-dimensional neural network potentials. *Chem. Rev.* **2021**, *121*, 10037−10072.
(102) Sammüller, F.; Hermann, S.; de Las Heras, D.; Schmidt, M. Neural functional theory for inhomogeneous fluids: Fundamentals and applications. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120*, No. e2312484120.
(103) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(104) Duignan, T. T. The potential of neural network potentials. *ACS Phys. Chem. Au.* **2024**, *4*, 232−241.
(105) Machina, H. K.; Wild, D. J. Electronic laboratory notebooks progress and challenges in implementation. *J. Lab. Autom.* **2013**, *18*, 264−268.
(106) Paszko, C.; Turner, E. *Laboratory Information Management Systems*; CRC Press, 2018.
(107) Skobelev, D.; Zaytseva, T.; Kozlov, A.; Perepelitsa, V.; Makarova, A. Laboratory information management systems in the work of the analytic laboratory. *Meas. Technol.* **2011**, *53*, 1182−1189.
(108) Prasad, P. J.; Bodhe, G. Trends in laboratory information management system. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 187−192.
(109) O'Boyle, N. M.; Hutchison, G. R. Cinfony−combining Open Source cheminformatics toolkits behind a common interface. *Chem. Cent. J.* **2008**, *2*, 24.
(110) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(111) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Muller, K.-R. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.
(112) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **2020**, *4*, 347−358.
(113) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(114) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887.
(115) Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.
(116) Kennedy, M. C.; O'Hagan, A. Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2001**, *63*, 425−464.
(117) Bayarri, M. J.; Berger, J. O.; Paulo, R.; Sacks, J.; Cafeo, J. A.; Cavendish, J.; Lin, C.-H.; Tu, J. A framework for validation of computer models. *Technometrics* **2007**, *49*, 138−154.
(118) Gu, M.; Wang, L. Scaled Gaussian stochastic process for computer model calibration and prediction. *SIAM/ASA J. Uncertain. Quantif.* **2018**, *6*, 1555−1583.
(119) Tukey, J. W. *Exploratory Data Analysis*; Addision-Wesley, 1977; Vol. 2.
(120) Jolliffe, I. T. *Principal Component Analysis for Special Types of Data*; Springer, 2002.
(121) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*.2579.
(122) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 2018, arXiv:1802.03426. DOI: 10.48550/arXiv.1802.03426.
(123) Schmid, P. J. Dynamic mode decomposition of numerical and experimental data. *J. Fluid Mech.* **2010**, *656*, 5−28.
(124) Wattenberg, M.; Viégas, F.; Johnson, I. How to use t-SNE effectively. *Distill* **2016**, *1*, No. e2.
(125) Kobak, D.; Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **2019**, *10*, 5416.
(126) Chari, T.; Pachter, L. The specious art of single-cell genomics. *PLOS Comput. Biol.* **2023**, *19*, No. e1011288.
(127) Tipping, M. E.; Bishop, C. M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 611−622.
(128) Beltran-Villegas, D. J.; Wessels, M. G.; Lee, J. Y.; Song, Y.; Wooley, K. L.; Pochan, D. J.; Jayaraman, A. Computational reverse-engineering analysis for scattering experiments on amphiphilic block polymer solutions. *J. Am. Chem. Soc.* **2019**, *141*, 14916−14930.
(129) Heil, C. M.; Patil, A.; Dhinojwala, A.; Jayaraman, A. Computational reverse-engineering analysis for scattering experiments (CREASE) with machine learning enhancement to determine structure of nanoparticle mixtures and solutions. *ACS Cent. Sci.* **2022**, *8*, 996−1007.
(130) Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V. The evolution of data-driven modeling in organic chemistry. *ACS Cent. Sci.* **2021**, *7*, 1622−1637.
(131) Hastie, T.; Tibshirani, R.; Friedman, J. H.; Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer, 2009; Vol. 2.
(132) Luo, Y.; Gu, M.; Edwards, C. E.; Valentine, M. T.; Helgeson, M. E. High-throughput microscopy to determine morphology, microrheology, and phase boundaries applied to phase separating coacervates. *Soft Matter* **2022**, *18*, 3063−3075.
(133) Gu, M.; Fang, X.; Luo, Y. Data-driven model construction for anisotropic dynamics of active matter. *PRX Life* **2023**, *1*, 013009.
(134) Breiman, L.; Friedman, J.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Routledge, 2017.
(135) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*; Springer, 2013.
(136) Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18−22.
(137) Gu, M.; Wang, X.; Berger, J. O. Robust Gaussian stochastic process emulation. *Ann. Stat.* **2018**, *46*, 3038−3066.
(138) Chmiela, S.; Sauceda, H. E.; Poltavsky, I.; Müller, K.-R.; Tkatchenko, A. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. *Comput. Phys. Commun.* **2019**, *240*, 38−45.
(139) Kanagawa, M.; Hennig, P.; Sejdinovic, D.; Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv*, arXiv preprint arXiv:1807.02582, 2018. DOI: 10.48550/arXiv.1807.02582.
(140) Vecchia, A. V. Estimation and model identification for continuous spatial processes. *J. R. Stat. Soc. Ser. B Methodol.* **1988**, *50*, 297−312.
(141) Snelson, E.; Ghahramani, Z. Sparse Gaussian processes using pseudo-inputs. *Adv. Neural Inf. Process. Syst.* **2006**, *18*, 1257.

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

(142) Murphy, K. P. *Probabilistic Machine Learning: An Introduction*; MIT Press, 2022.
(143) Bishop, C. M.; Bishop, H. *Deep Learning: Foundations and Concepts*; Springer Nature, 2023.
(144) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436−444.
(145) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*..
(146) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(147) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(148) R Core Team, R.: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2024.
(149) Friedman, J. H.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1−22.
(150) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016; pp 785−794.
(151) Chen, T. et al. xgboost: Extreme Gradient Boosting. 2025; R package version 1.7.11.1.
(152) Gardner, J.; Pleiss, G.; Weinberger, K. Q.; Bindel, D.; Wilson, A. G. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.7587.
(153) Gu, M.; Palomo, J.; Berger, J. O. RobustGaSP: Robust Gaussian Stochastic Process Emulation in R. *R J.* **2019**, *11*, 112−136.
(154) Guinness, J.; Katzfuss, M.; Fahmy, Y. GpGp: Fast Gaussian Process Computation Using Vecchia's Approximation, R package version 0.5.1, 2024.
(155) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. DOI: 10.48550/arXiv.1912.01703.
(156) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M., et al. TensorFlow: a system for Large-Scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016; pp 265−283.
(157) Chollet, F.; et al. Keras. https://keras.io, 2015.
(158) Falbel, D.; Luraschi, J. torch: Tensors and Neural Networks with 'GPU' Acceleration; *R package*, version 0.15.1, 2025.
(159) Chollet, F.; Allaire, J., et al. R Interface to Keras. https://github.com/rstudio/keras, 2017.
(160) Luo, Y.; Chen, J.; Gu, M.; Luo, Y. Optimizing gelation time for cell shape control through active learning. *Soft Matter* **2025**, *21*, 970−981.
(161) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst.* **2012**, *25*.
(162) Fang, X.; Gu, M.; Wu, J. Reliable emulation of complex functionals by active learning with error control. *J. Chem. Phys.* **2022**, *157*, 214109.
(163) Chipman, H. A.; George, E. I.; McCulloch, R. E. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **2010**, *4*, 266−298.
(164) Meinshausen, N. Quantile regression forests. *J. Mach. Learn. Res.* **2006**, *7*.983.
(165) Mentch, L.; Hooker, G. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **2016**, *17*, 1−41.
(166) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.
(167) Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 2016; pp 1050−1059.
(168) Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
(169) Kristiadi, A.; Hein, M.; Hennig, P. Being Bayesian, even just a bit, fixes overconfidence in ReLU networks. *International Conference on Machine Learning*, 2020; pp 5436−5446.
(170) Fontana, M.; Zeni, G.; Vantini, S. Conformal prediction: a unified review of theory and new challenges. *Bernoulli* **2023**, *29*, 1−23.
(171) Williamson, E. M.; Sun, Z.; Mora-Tamez, L.; Brutchey, R. L. Design of experiments for nanocrystal syntheses: a how-to guide for proper implementation. *Chem. Mater.* **2022**, *34*, 9823−9835.
(172) Nocedal, J.; Wright, S. J. *Numerical Optimization*; Springer, 1999.
(173) Frazier, P. I. A tutorial on Bayesian optimization. *arXiv* 2018, arXiv:1807.02811. DOI: 10.48550/arXiv.1807.02811.
(174) Wang, X.; Jin, Y.; Schmitt, S.; Olhofer, M. Recent advances in Bayesian optimization. *ACM Comput. Surv.* **2023**, *55*, 1−36.
(175) White, A. D. The future of chemistry is language. *Nat. Rev. Chem.* **2023**, *7*, 457−458.
(176) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digit. Discovery* **2023**, *2*, 1233−1250.
(177) Griffen, E. J.; Dossetter, A. G.; Leach, A. G. Chemists: AI is here; unite to get the benefits. *J. Med. Chem.* **2020**, *63*, 8695−8704.
(178) Subasinghe, S. S.; Gersib, S. G.; Mankad, N. P. Large Language Models (LLMs) as Graphing Tools for Advanced Chemistry Education and Research. *J. Chem. Educ.* **2025**, *102*, 1563−1571.
(179) Hare, P. M. Coding with AI in the physical chemistry laboratory. *J. Chem. Educ.* **2024**, *101*, 3869−3874.
(180) Nam, D.; Macvean, A.; Hellendoorn, V.; Vasilescu, B.; Myers, B. Using an llm to help with code understanding. *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*; IEEE, 2024; pp 1−13.
(181) White, A. D.; Hocky, G. M.; Gandhi, H. A.; Ansari, M.; Cox, S.; Wellawatte, G. P.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; et al. Assessment of chemistry knowledge in large language models that generate code. *Digit. Discovery* **2023**, *2*, 368−376.
(182) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **2024**, *6*, 161−169.
(183) DeStefano, A. J.; Segalman, R. A.; Davidson, E. C. Where biology and traditional polymers meet: the potential of associating sequence-defined polymers for materials science. *JACS Au* **2021**, *1*, 1556−1571.
(184) Barnes, J. C.; Ehrlich, D. J.; Gao, A. X.; Leibfarth, F. A.; Jiang, Y.; Zhou, E.; Jamison, T. F.; Johnson, J. A. Iterative exponential growth of stereo-and sequence-controlled polymers. *Nat. Chem.* **2015**, *7*, 810−815.
(185) Jiang, Y.; Golder, M. R.; Nguyen, H. V.-T.; Wang, Y.; Zhong, M.; Barnes, J. C.; Ehrlich, D. J.; Johnson, J. A. Iterative exponential growth synthesis and assembly of uniform diblock copolymers. *J. Am. Chem. Soc.* **2016**, *138*, 9369−9372.
(186) Al Ouahabi, A.; Charles, L.; Lutz, J.-F. Synthesis of non-natural sequence-encoded polymers using phosphoramidite chemistry. *J. Am. Chem. Soc.* **2015**, *137*, 5629−5635.
(187) Coin, I.; Beyermann, M.; Bienert, M. Solid-phase peptide synthesis: from standard procedures to the synthesis of difficult sequences. *Nat. Protoc.* **2007**, *2*, 3247−3256.
(188) Zhang, C.; Bates, M. W.; Geng, Z.; Levi, A. E.; Vigil, D.; Barbon, S. M.; Loman, T.; Delaney, K. T.; Fredrickson, G. H.; Bates, C. M.; Whittaker, A. K.; Hawker, C. J. Rapid generation of block copolymer libraries using automated chromatographic separation. *J. Am. Chem. Soc.* **2020**, *142*, 9843−9849.

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX

(189) Bates, C. M.; Bates, F. S. 50th Anniversary Perspective: Block Polymers Pure Potential. *Macromolecules* **2017**, *50*, 3−22.
(190) Wang, X.; Zhang, C.; Sawczyk, M.; Sun, J.; Yuan, Q.; Chen, F.; Mendes, T. C.; Howlett, P. C.; Fu, C.; Wang, Y.; et al. Ultra-stable all-solid-state sodium metal batteries enabled by perfluoropolyether-based electrolytes. *Nat. Mater.* **2022**, *21*, 1057−1065.
(191) Zhang, C.; Yan, K.; Fu, C.; Peng, H.; Hawker, C. J.; Whittaker, A. K. Biological utility of fluorinated compounds: from materials design to molecular imaging, therapeutics and environmental remediation. *Chem. Rev.* **2022**, *122*, 167−208.
(192) Maji, P.; Naskar, K. Styrenic block copolymer-based thermoplastic elastomers in smart applications: Advances in synthesis, microstructure, and structure−property relationships—A review. *J. Appl. Polym. Sci.* **2022**, *139*, No. e52942.
(193) Moon, J. D.; Freeman, B. D.; Hawker, C. J.; Segalman, R. A. Can self-assembly address the permeability/selectivity trade-offs in polymer membranes? *Macromolecules* **2020**, *53*, 5649−5654.
(194) Murphy, E. A.; Skala, S. J.; Kottage, D.; Kohl, P. A.; Li, Y.; Zhang, C.; Hawker, C. J.; Bates, C. M. Accelerated discovery and mapping of block copolymer phase diagrams. *Phys. Rev. Mater.* **2024**, *8*, 015602.
(195) Murphy, E. A.; Roth, K. G.; Bates, M. W.; Murphy, M. C.; Edmund, J.; Bates, C. M.; Hawker, C. J. High-Throughput Generation of Block Copolymer Libraries via Click Chemistry and Automated Chromatography. *Macromolecules* **2025**, *58*, 8369−8376.
(196) Kalman, R. E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **1960**, *82*, 35−45.
(197) Petty, J. T.; Zheng, J.; Hud, N. V.; Dickson, R. M. DNA-templated Ag nanocluster formation. *J. Am. Chem. Soc.* **2004**, *126*, 5207−5212.
(198) Schultz, D.; Gwinn, E. G. Silver atom and strand numbers in fluorescent and dark Ag: DNAs. *Chem. Commun.* **2012**, *48*, 5748−5750.
(199) Guha, R.; Gonzàlez-Rosell, A.; Rafik, M.; Arevalos, N.; Katz, B. B.; Copp, S. M. Electron count and ligand composition influence the optical and chiroptical signatures of far-red and NIR-emissive DNA-stabilized silver nanoclusters. *Chem. Sci.* **2023**, *14*, 11340−11350.
(200) Huard, D. J.; Demissie, A.; Kim, D.; Lewis, D.; Dickson, R. M.; Petty, J. T.; Lieberman, R. L. Atomic structure of a fluorescent Ag8 cluster templated by a multistranded DNA scaffold. *J. Am. Chem. Soc.* **2019**, *141*, 11465−11470.
(201) Swasey, S. M.; Copp, S. M.; Nicholson, H. C.; Gorovits, A.; Bogdanov, P.; Gwinn, E. G. High throughput near infrared screening discovers DNA-templated silver clusters with peak fluorescence beyond 950 nm. *Nanoscale* **2018**, *10*, 19701−19705.
(202) Neacşu, V. A.; Cerretani, C.; Liisberg, M. B.; Swasey, S. M.; Gwinn, E. G.; Copp, S. M.; Vosch, T. Unusually large fluorescence quantum yield for a near-infrared emitting DNA-stabilized silver nanocluster. *Chem. Commun.* **2020**, *56*, 6384−6387.
(203) Mastracco, P.; Copp, S. M. Beyond nature's base pairs: machine learning-enabled design of DNA-stabilized silver nanoclusters. *Chem. Commun.* **2023**, *59*, 10360−10375.
(204) Hong, G.; Antaris, A. L.; Dai, H. Near-infrared fluorophores for biomedical imaging. *Nat. Biomed. Eng.* **2017**, *1*, 0010.
(205) Copp, S. M.; Schultz, D.; Swasey, S.; Pavlovich, J.; Debord, M.; Chiu, A.; Olsson, K.; Gwinn, E. Magic numbers in DNA-stabilized fluorescent silver clusters lead to magic colors. *J. Phys. Chem. Lett.* **2014**, *5*, 959−963.
(206) Sapnik, A. F.; Romolini, G.; Cerretani, C.; Vosch, T.; Jensen, K. M. Structure of a DNA-Stabilized Ag16Cl2 Nanocluster in Solution. *Angew. Chem., Int. Ed.* **2025**, *64*, No. e202422432.
(207) Cerretani, C.; Kanazawa, H.; Vosch, T.; Kondo, J. Crystal structure of a NIR-emitting DNA-stabilized Ag16 nanocluster. *Angew. Chem., Int. Ed.* **2019**, *58*, 17153−17157.
(208) Romolini, G.; Kanazawa, H.; Mollerup, C. B.; Liisberg, M. B.; Lind, S. W.; Huang, Z.; Cerretani, C.; Kondo, J.; Vosch, T. Shining Bright at 960 nm: A 28-Silver-Atom Nanorod Stabilized by DNA. *Small Struct.* **2025**, *6*, No. e202500022.
(209) Mastracco, P.; Gonzàlez-Rosell, A.; Evans, J.; Bogdanov, P.; Copp, S. M. Chemistry-informed machine learning enables discovery of DNA-stabilized silver nanoclusters with near-infrared fluorescence. *ACS Nano* **2022**, *16*, 16322−16331.
(210) Sadeghi, E.; Mastracco, P.; Gonzàlez-Rosell, A.; Copp, S. M.; Bogdanov, P. Multi-objective design of DNA-stabilized nanoclusters using variational autoencoders with automatic feature extraction. *ACS Nano* **2024**, *18*, 26997−27008.
(211) Moomtaheen, F.; Killeen, M.; Oswald, J.; Gonzàlez-Rosell, A.; Mastracco, P.; Gorovits, A.; Copp, S. M.; Bogdanov, P. DNA-Stabilized Silver Nanocluster Design via Regularized Variational Autoencoders. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery, 2022; pp 3593−3602.
(212) Copp, S. M.; Bogdanov, P.; Debord, M.; Singh, A.; Gwinn, E. Base motif recognition and design of DNA templates for fluorescent silver clusters by machine learning. *Adv. Mater.* **2014**, *26*, 5839−5845.
(213) Copp, S. M.; Gorovits, A.; Swasey, S. M.; Gudibandi, S.; Bogdanov, P.; Gwinn, E. G. Fluorescence color by data-driven design of genomic silver clusters. *ACS Nano* **2018**, *12*, 8240−8247.
(214) Evgeniou, T.; Pontil, M.; Poggio, T. Regularization networks and support vector machines. *Adv. Comput. Math.* **2000**, *13*, 1−50.
(215) Xu, H.; Caramanis, C.; Mannor, S. Robustness and regularization of support vector machines. *J. Machine Learning Res.* **2009**, *10*.1485.
(216) Kursa, M. B.; Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1−13.
(217) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **2018**, *10*, 4.
(218) Mockus, J. On the Bayes methods for seeking the extremal point. *IFAC Proc.* **1975**, *8* (8), 428−431.
(219) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A multi-objective active learning platform and web app for reaction optimization. *J. Am. Chem. Soc.* **2022**, *144*, 19999−20007.
(220) Dunlap, J. H.; Ethier, J. G.; Putnam-Neeb, A. A.; Iyer, S.; Luo, S.-X. L.; Feng, H.; Garrido Torres, J. A.; Doyle, A. G.; Swager, T. M.; Vaia, R. A.; Mirau, P.; Crouse, C. A.; Baldwin, L. A. Continuous flow synthesis of pyridinium salts accelerated by multi-objective Bayesian optimization with active learning. *Chem. Sci.* **2023**, *14*, 8061−8069.
(221) Sans, V.; Porwol, L.; Dragone, V.; Cronin, L. A self optimizing synthetic organic reactor system using real-time in-line NMR spectroscopy. *Chem. Sci.* **2015**, *6*, 1258−1264.
(222) Cortés-Borda, D.; Wimmer, E.; Gouilleux, B.; Barré, E.; Oger, N.; Goulamaly, L.; Peault, L.; Charrier, B.; Truchet, C.; Giraudeau, P.; et al. An autonomous self-optimizing flow reactor for the synthesis of natural product carpanone. *J. Org. Chem.* **2018**, *83*, 14286−14299.
(223) Luo, F.; Zhang, J.; Wang, Q.; Yang, C. Leveraging prompt engineering in large language models for accelerating chemical research. *ACS Cent. Sci.* **2025**, *11*, 511−519.

https://doi.org/10.1021/acscentsci.5c01994
ACS Cent. Sci. XXXX, XXX, XXX−XXX