Volume 4
Number 12
December 2025
Pages 3415-3830

Digital Discovery

rsc.li/digitaldiscovery

ISSN 2635-098X

ROYAL SOCIETY OF CHEMISTRY

PAPER
Olexandr Isayev *et al.*
Machine learning anomaly detection of automated HPLC experiments in the cloud laboratory

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 3445

Machine learning anomaly detection of automated HPLC experiments in the cloud laboratory

Filipp Gusev,[ab] Benjamin C. Kline,[c] Ryan Quinn,[d] Anqin Xu,[c] Ben Smith,[c] Brian Frezza[cd] and Olexandr Isayev[*ab]

Automation of experiments in cloud laboratories promises to revolutionize scientific research by enabling remote experimentation and improving reproducibility. However, maintaining quality control without constant human oversight remains a critical challenge. Here, we present a novel machine learning framework for automated anomaly detection in High-Performance Liquid Chromatography (HPLC) experiments conducted in a cloud lab. Our system specifically targets air bubble contamination—a common yet challenging issue that typically requires expert analytical chemists to detect and resolve. By leveraging active learning combined with human-in-the-loop annotation, we trained a binary classifier on approximately 25 000 HPLC traces. Prospective validation demonstrated robust performance, with an accuracy of 0.96 and an F1 score of 0.92, suitable for real-world applications. Beyond anomaly detection, we show that the system can serve as a sensitive indicator of instrument health, outperforming traditional periodic qualification tests in identifying systematic issues. The framework is protocol-agnostic, instrument-agnostic, and, in principle, vendor-neutral, making it adaptable to various laboratory settings. This work represents a significant step toward fully autonomous laboratories by enabling continuous quality control, reducing the expertise barrier for complex analytical techniques, and facilitating proactive maintenance of scientific instrumentation. The approach can be extended to detect other types of experimental anomalies, potentially transforming how quality control is implemented in self-driving laboratories (SDLs) across diverse scientific disciplines.

Received 6th June 2025
Accepted 23rd October 2025

DOI: 10.1039/d5dd00253b

rsc.li/digitaldiscovery

1 Introduction

Laboratory automation refers to the integration of scientific instrumentation, software, and processes to streamline laboratory workflows and enhance efficiency, reproducibility, and throughput while minimizing human error.[1] Laboratory autonomy—an advancement beyond basic automation—entails integrating artificial intelligence (AI) and self-driven systems to conduct experiments in a closed-loop manner. In such systems, data are continuously collected, analyzed, and used to plan subsequent experiments, all with minimal (if any) human intervention.[2–7] Fields such as drug discovery,[8,9] materials science,[10–14] and synthetic biology[15]—which routinely require the rapid screening and testing of hundreds to thousands of samples—are among the early adopters harnessing this paradigm shift.

The emergence of cloud laboratories is transforming autonomous experimentation by enabling remote execution of complex biological and chemical research with enhanced reproducibility, scalability, and accessibility. These facilities integrate robotic automation and networked control systems to conduct experiments continuously and in parallel, significantly reducing physical and logistical constraints. The pioneering work by CMU alumni through Emerald Cloud Lab (ECL) provides researchers with a suite of instruments for biological and chemical experimentation at scale. While cloud laboratories hold great promise for democratizing access to sophisticated experimental infrastructure, they also introduce challenges related to remote troubleshooting, real-time experimental adaptability, and standardization across diverse research domains. As these platforms evolve, they present an opportunity to accelerate scientific discovery while necessitating new frameworks for data integrity, automation-driven research methodologies, and integration into traditional experimental workflows.

This study focuses on improving High-Performance Liquid Chromatography (HPLC) in the Cloud Lab. HPLC is an essential analytical technique used across various scientific disciplines, from pharmaceuticals and biotechnology to environmental studies, making it a prime target for automation in a Cloud Lab environment. In a traditional lab, modern HPLC instruments incorporate basic automation (*e.g.*, for liquid handling, sample

[a]*Department of Chemistry, Mellon College of Science, Carnegie Mellon University, 4400 Fifth Ave, Pittsburgh, PA 15213, USA. E-mail: olexandr@olexandrisayev.com*
[b]*Ray and Stephanie Lane Computational Biology Department, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA*
[c]*Emerald Cloud Lab, 15500 Wells Port Dr, Austin, TX 78728, USA*
[d]*Emerald Therapeutics, 15500 Wells Port Dr, Austin, TX 78728, USA*

collection, and executing predefined protocols); however, a scientist is still often required to be present to monitor the data readout and ensure its validity as well as proper functioning of the instrument. This manual oversight enables experts to catch issues such as pressure fluctuations due to air bubbles, clogs, leaks, empty mobile phase bottles, and other unexpected system behaviors. In contrast, in autonomous closed-loop systems like Cloud Lab, such real-time human intervention is impractical given the high-throughput use of instruments and the need for a fully closed data flow in the Design-Make-Test-Analyze (DMTA) cycle.

Many integrated data analysis systems for HPLC instruments, for example those performing peak assignment, operate on the assumption that the recorded signal is valid, which is true in most cases. In a field of data-driven research, it is common to 'trust but verify' historically accumulated data,[16,17] or re-generate it *de novo* to minimize discrepancy among data sources or avoid implicit biases. The classic 'garbage in, garbage out'[18] principle can undermine any data-driven system; closed-loop experiments like Bayesian optimization are among the most vulnerable. Bayesian optimization algorithms, once misdirected by a false signal, will require several observations (or rounds in batch execution) of data acquisition to self-correct at the cost of time and resources at best and fully degrade at worst. Currently, ensuring correct execution, a common step in computer science, is overlooked among target metrics for the evaluation of self-driving laboratories.[19]

HPLC chromatogram peaks, as frequently monitored by absorbance, can be negatively affected by many variables, including column health and age, purity of the sample, and—germane for this investigation—air bubbles. Although most modern instruments have some ways to detect common and expected pitfalls (*e.g.* by qualification/controlled experiments), complications arise when rare, stochastic events occur during large-scale experimental campaigns. Air bubbles—one such pitfall—can disrupt an HPLC experiment: when air enters the buffer tubing, it will eventually reach the column, where the chemical separation occurs. These intermittent pockets of air alter the interactions between analytes and the stationary phase, often leading to unpredictable retention times (Fig. 1A), distorted peak shapes (Fig. 1B), loss of a peak (Fig. 1C) or even an HPLC chromatogram that is indiscernible to the scientist. Moreover, the presence of air bubbles may be especially problematic for preparative HPLC experiments, where the entire source sample is used up during the experiment and repeating the protocol is not always an option.

Several user behaviors or instrument shortcomings can lead to the introduction of air bubbles and resultant pressure fluctuations in an HPLC run. Air bubbles in HPLC systems are most commonly introduced when mobile phases are not adequately degassed, allowing dissolved gases to come out of solution under pump pressure. Temperature fluctuations between different parts of the system can also reduce gas solubility and trigger bubble formation. In addition, leaks at pump seals, fittings, or inlet lines can draw air into the system, while insufficient priming after solvent changes may leave residual air

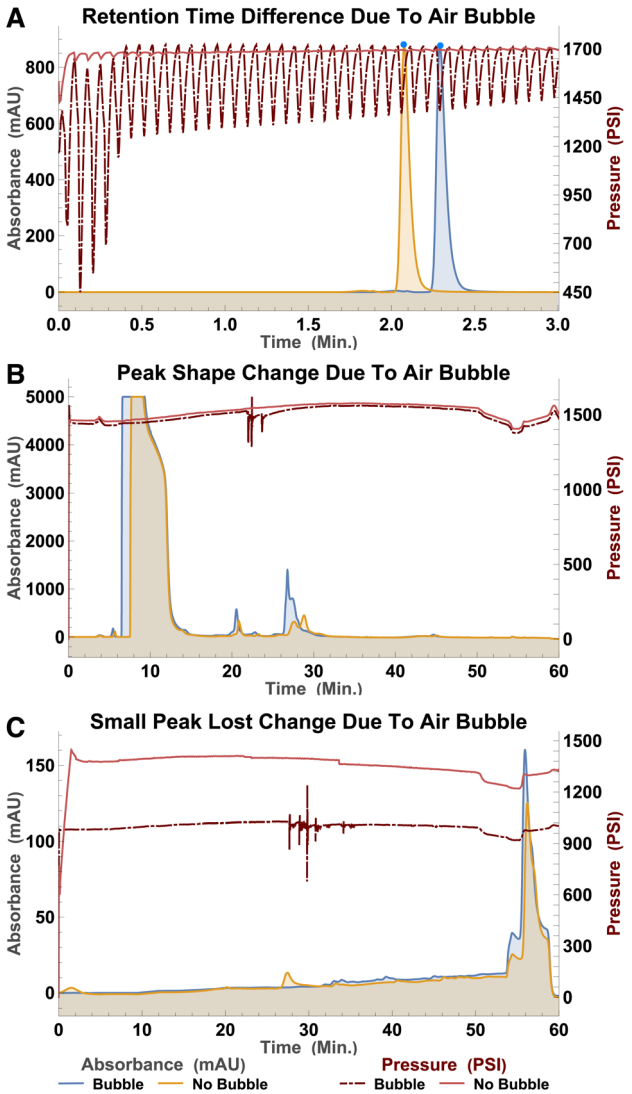

Fig. 1 Representative examples of air bubble presence on absorbance and pressure traces for same sample and HPLC protocol. (A) Effect on retention time; (B) effect on peak shape; (C) effect on peak presence.

pockets in the tubing. Together, these factors represent the primary sources of bubble formation in HPLC.[20]

Several factors can influence peak shape and retention time in HPLC, including column age or identity, mobile phase composition, temperature, sample concentration, and flow rate variations. A major advantage of a cloud laboratory is that all collected data are linked within a central database, enabling rapid root-cause analysis of problems and anomalies. The representative data shown in Fig. 1A–C were selected such that the other variables affecting peak shape were held constant, with the main difference being the pressure trace during the run. Column age or health is the most likely alternative cause; to mitigate this, standards are routinely run on all columns to ensure they are not used beyond their effective lifetime.

We designed our automatic anomaly detection system for HPLC experiments that operates on-the-fly and without human intervention. The system is based on a binary classifier: the ML

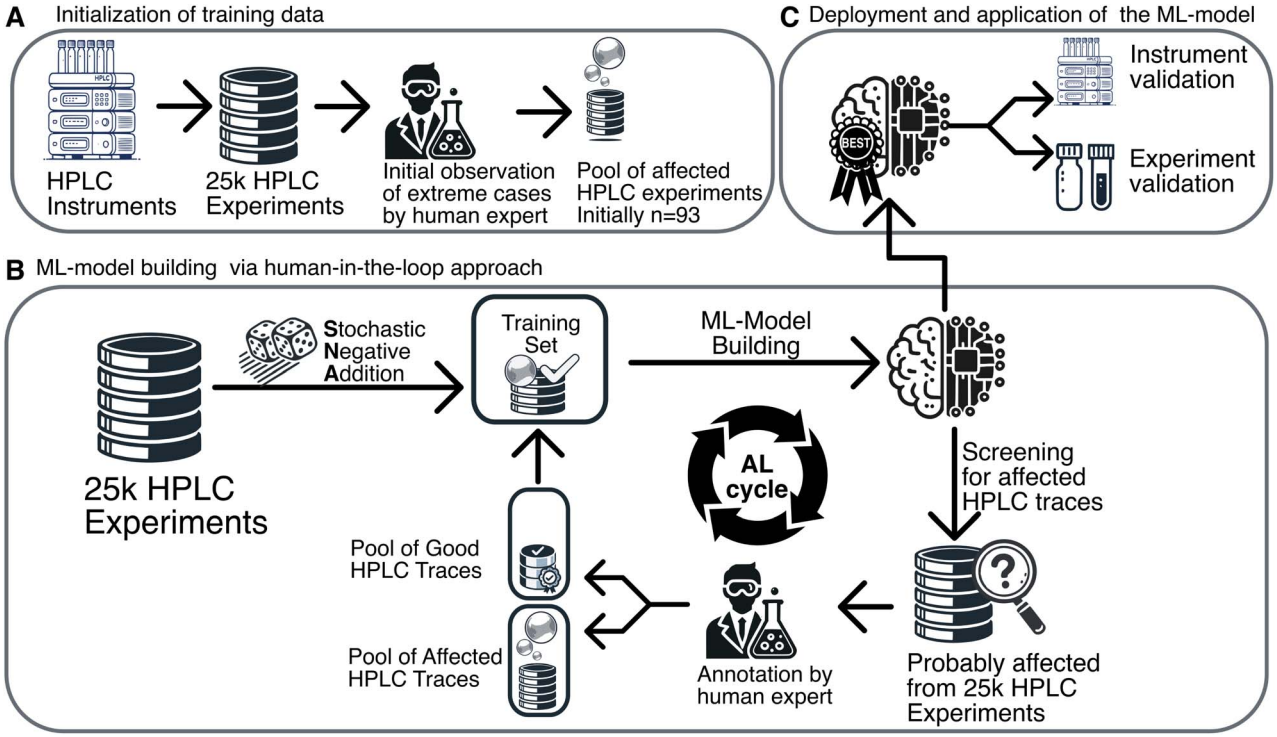

Fig. 2 The ML-based anomaly (HPLC air bubble contamination) detection workflow. The overall workflow is organized in three major steps: (A). Creation of the initial training data set; (B). ML-model building organized in iterative Active Learning Cycle with a human expert as an Oracle; (C). Deployment of the trained final model to Cloud Lab and its application.

model treats HPLC experiments affected by air bubble contamination as the positive class (class 1) and unaffected as the negative class (class 0). Focusing on the air bubbles, we analyzed HPLC pressure data—which exhibit a characteristic pattern when air is introduced into the HPLC tubing—and employed active learning combined with a human-in-the-loop approach to develop the model efficiently. The overall workflow (see Fig. 2) comprises three major steps: (i) Initialization of Training Data (Fig. 2A), (ii) ML Model Building *via* Human-in-the-Loop Approach (Fig. 2B), (iii) Deployment and Performance measurement of the final ML Model (Fig. 2C).

Once the ML model reached optimal performance, it was deployed in the Cloud Lab to autonomously screen HPLC experiments in real time. The purpose of the trained model is to screen and identify affected HPLC traces autonomously. Two prospective validation steps—one at the experiment level and one at the instrument level—were performed to ensure that the model's predictions align with real-world scenarios, thereby confirming its reliability and effectiveness.

2 Results

2.1 ML framework for automated anomaly detection in HPLC experiments

Traditional quality control for HPLC experiments relies on the expertise of human operators or on hardware enhancements – such as a parallel dual plunger system[21,22] or optical sensors. However, in high-throughput, fully automated systems like Cloud Labs, these methods are impractical: frequent manual checks are infeasible, and manufacturer-specific solutions compromise the universal compatibility required to operate a diverse set of HPLC systems.

A variety of instrument- or software-specific methods currently exist to detect leaks, empty buffer bottles, or pressure fluctuations during experimental HPLC runs (*e.g.*, Shimadzu Nexera-40/LabSolutions, Waters Alliance iS/Empower, Thermo Scientific Vanquish Core/Chromeleon, and Agilent InfinityLab 1290 III/OpenLab).[23–26] However, these methods often lack transparency regarding their underlying mechanisms and accuracy. More importantly, users generally cannot modify or improve commercial models to suit their specific needs. Here, we present an open-source anomaly detection approach that is adaptable, retrainable, and can ultimately be tailored to the user's requirements.

We designed our automatic workflow (see Fig. 2) in a data-driven manner rather than relying on rule-based or hardware-based approaches. This strategy makes the system adaptable and generalizable to other rare pitfalls that become observable as the database of HPLC experiments accumulates at scale. Our workflow began with the collection of approximately 25 000 HPLC experiments from a diverse set of chromatographic methods, instruments, and protocols (see Initial dataset for details, Tables S2 and S5), a dataset large enough to capture infrequent events like air bubble contamination. An initial subset of this data was reviewed by a human expert, who observed and annotated anomalous cases, resulting in an initial pool of 93 HPLC experiments affected by air bubble contamination. The initial pool of affected cases is relatively small due

to infrequency of occurrence (a conservative *a priori* estimation was ≈1%) coupled with infeasibility of explicit annotation of the whole dataset due to its size. Although air bubble contamination is a common issue in HPLC, the low observed frequency is expected in a well-maintained system. Such an imbalance can be challenging for an ML model and may introduce bias. To address this, we employed Stochastic Negative Addition[27] (SNA), which stochastically adds negative ("unaffected") examples to the training set to ensure balanced representation while minimizing further annotation effort, after preliminary analysis (see Classical ML for details, Fig. S1) we decided to target 1 : 10 class ratio (the ratio of samples with positive, *e.g.* affected by air bubbles, class to samples with negative, *e.g.* normal, class) for the initial dataset and maintain it for the rest of the modeling stages. SNA has been successfully applied as a balancing strategy in other data-driven domains.[27,28]

During the model-building phase (see Fig. 2B), we employed an Active Learning (AL) cycle combined with a human-in-the-loop approach to iteratively refine the model with expert input while minimizing overall annotation effort by focusing on the most informative cases. This phase comprised the following steps: (i) training set creation: a training set comprising both affected (identified by expert annotation) and unaffected (selected using SNA[27] balancing strategy) HPLC traces were assembled in each annotation round. (ii) ML model building: a ML model was built using the training set, then the model screened the dataset of 25k HPLC experiments to identify traces that were potentially affected (requiring further annotation) as well as those most likely unaffected (which were used for SNA later). (iii) Human expert annotation: the flagged traces were reviewed by a human expert who annotated each as affected or unaffected, further enriching the dataset and improving the model's accuracy. This iterative cycle continued until the ML model achieved satisfactory performance; in total, only three rounds (one initial and two AL) of annotation were conducted to sufficiently train our model.

For the air bubble contamination we focused on HPLC pressure traces. Since the pressure trace, by its nature, is a time series, we started our modeling from classical ML approaches for time series data. The classical featurization approach[29] (see Methods for details) performed well (Fig. S1); however, its resource demands—in terms of memory footprint and processing delays—made it unsuitable for on-the-fly deployment in the Cloud Lab. Therefore, we transitioned to an end-to-end Deep Learning approach utilizing a 1D convolutional neural network (CNN) coupled with automatic architecture and hyperparameter optimization.[30]

With complete pressure traces available during analysis, the 1D CNN minimized model size, memory usage, and response time, while preserving the capacity to generalize to other HPLC anomalies in future developments. This approach also avoids the need for labor-intensive, manual feature engineering required by rule-based methods.

The model perception (Fig. 3A), visualized as a UMAP projection (see Methods for details) of the latent representation from the 1D CNN model, reveals the learned feature space structure. There are two regions of very high artifact probability that were sampled through the three rounds of expert annotation (Fig. 3B). Unlike normal experimental cases (Fig. 3C), Fig. 3D–G illustrate varying levels of uncertainty in trace annotation. These traces fall between "clean" samples and those clearly containing air bubbles. As annotation progressed through multiple rounds, the focus shifted from simply identifying air bubbles to investigating potential causes of anomalous behavior, particularly for traces near the ML-model's decision boundary. Fig. 3G highlights traces exhibiting pressure-related anomalies likely caused by factors other than air bubbles.

These anomalies can be attributed to several technical issues in the HPLC system. Insufficiently tightened barrel-tubing connections often lead to pressure fluctuations as fluid escapes through minute gaps in the assembly. When HPLC buffers run dry, a characteristic pattern emerges where pressure gradually drops toward zero as the system attempts to pull nonexistent fluid through the lines. Pump malfunctions represent another common source of pressure anomalies, creating irregular patterns in the trace data that differ distinctly from the signature patterns of air bubbles but nonetheless require identification and remediation to ensure experimental validity.

Validating these hypotheses would require either generating controlled error states in the laboratory or further accumulation of historical data. Although the anomalous traces shown in Fig. 3G represent only a minor fraction of the total HPLC experiments—and are not yet a significant concern—the continuous aggregation of data in the Cloud Lab facilitates ongoing model retraining. This will enable future refinements to distinguish among various pressure-related artifacts.

For deployment in Cloud Lab, the ML model was serialized in ONNX format. This enabled seamless integration into the Emerald Cloud Lab backend by loading it directly into Wolfram Language to analyze HPLC data for bubble likelihood, expressed as a Class 1 probability. Immediately after HPLC data from the experiment are parsed and imported into the Cloud Lab database, the pressure traces undergo brief preprocessing to ensure compatibility with the model and to eliminate false positives due to early retention-time pressure instabilities. Each preprocessed pressure data set is then passed to the model to yield a predicted likelihood (between 0 and 1) that the corresponding experiment was contaminated by air bubbles, and the predictions are added to the experiment's metadata.

2.2 Prospective experiment validation

For prospective validation, we compiled a dataset of 967 HPLC traces (see Methods) that were fully annotated by a human expert. In this real-world evaluation (Fig. 4), the model achieved an accuracy of 0.96, an F1 score of 0.92, an AUC of 0.98, and an average precision of 0.91. These metrics confirm that our model effectively distinguishes between bubble-affected and unaffected traces and handles class imbalance, supporting its deployment in the Cloud Lab environment.

Interestingly, the frequency of air bubble-affected HPLC experiments was higher than expected. This observation prompted us to apply the ML model for instrument validation-

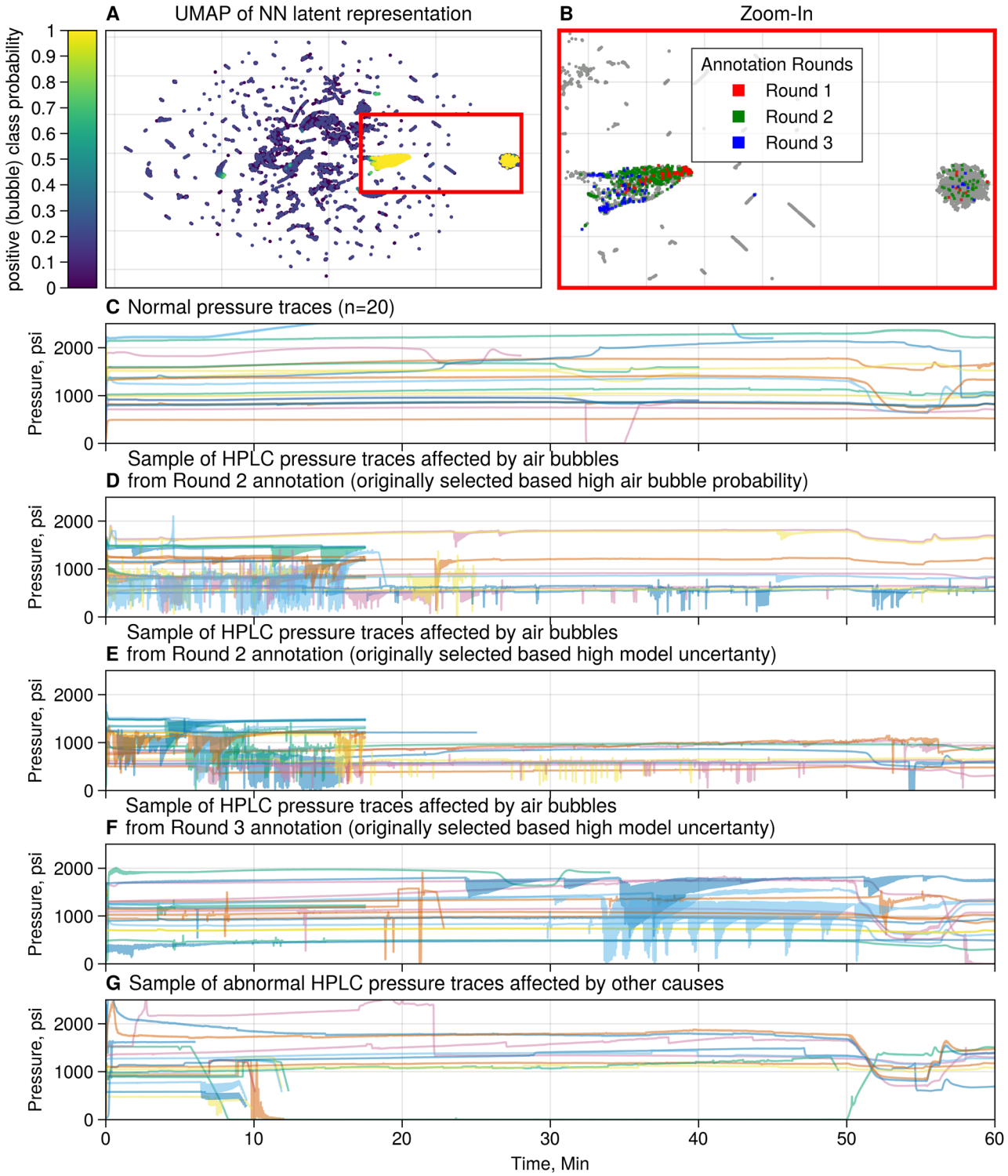

Fig. 3 Characterization of ML Model perception and representative traces from investigated dataset. (A) UMAP projection of the CNN latent space for the initial dataset (25k traces). Points are colored by the model's predicted probability that a trace contains air bubbles. The red rectangle marks the bubble-rich region presented at (B). (B). Zoom-in of the boxed area in (A), with points colored by the round in which they were annotated, illustrating how active learning progressively sampled this region. (C) Examples of normal pressure traces (n = 20); each experiment is shown in a different color. (D) Examples of pressure traces drawn from Round 2 annotations, selected for high predicted air bubble probability. (E) Examples of pressure traces drawn from Round 2 annotations, selected for high model uncertainty. (F) Examples of pressure traces drawn from Round 3 annotations, selected for high model uncertainty. (G) Examples of abnormal pressure traces attributed to causes other than air bubbles.

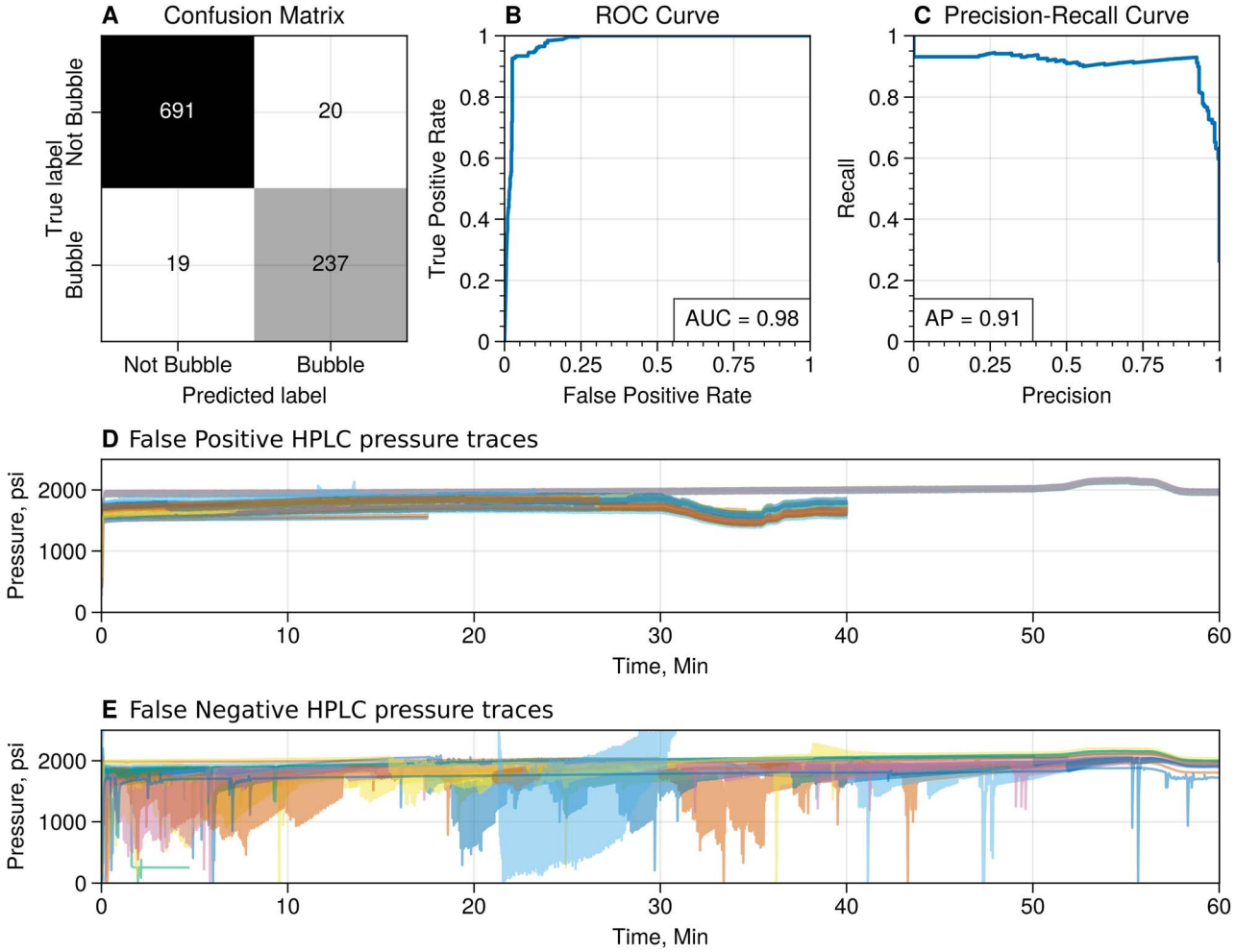

Fig. 4 Prospective validation of the deployed HPLC air bubble detection model. (A) Confusion matrix illustrating model performance in classifying pressure traces as containing bubbles or not. (B) Receiver operating characteristic (ROC) curve, with an area under the curve (AUC) of 0.98, demonstrating high classification performance. (C) Precision-recall (PR) curve, with an average precision (AP) of 0.91, indicating strong model reliability in detecting bubbles. (D) False positive HPLC pressure traces, where the model incorrectly identified bubble presence. (E) False negative HPLC pressure traces, showing cases where the model failed to detect bubbles. The model exhibits high accuracy and generalizability, with misclassifications predominantly occurring in ambiguous pressure fluctuations.

to assess whether the fraction of bubble-affected traces for a given instrument deviates from expected norms. Although routine qualification experiments (using known standards) ensure instrument reproducibility, their infrequent scheduling (weekly or monthly) may overlook subtle, stochastic shifts in experimental quality.

A qualification experiment is an in-depth control experiment that tests the performance and health of an instrument. In the Cloud Lab, every qualification generates an automated report that can be easily compared to previous reports to give users confidence in their experiments. Each automated report is assigned a pass/fail grade that is confirmed by a human expert. If an instrument is passing its latest qualification test, it is "qualified" to run experimental samples.

Most qualification tests are focused on testing for reproducible experimental outcomes. For HPLC the test targets autosampler, fraction collection, lineshape, *etc.* For the most part (>90% of the time), the air bubbles in the lines are a transient issue and do not cause any appreciable difference in the experimental result based on what was tested in the qualification runs.

2.3 Prospective instrument validation

We collected a prospective dataset comprising at least 100 HPLC experiments per instrument from eight systems equipped with UV-vis detectors (see Methods). Kernel density estimation of the air bubble class probabilities revealed a noticeable shift for instrument #8 (Fig. 5A), which led to the identification of malfunctioning check valves. These valves, critical for preventing backflow and maintaining pressure stability, were subsequently replaced on the HPLC pump module, restoring normal performance (Fig. 5B).

This approach appears to be more sensitive than the qualification experiments in detecting air bubble-associated issues (*e.g.* pump malfunctions). Incorporating this model into the instrument quality control pipeline will further enhance overall Cloud Lab performance. This enabled us to "flag" all affected

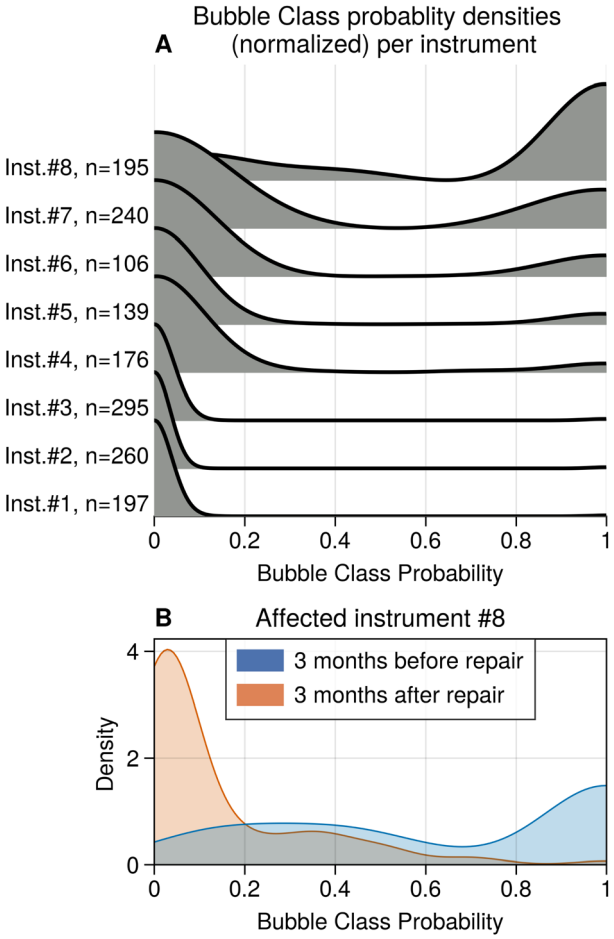

Fig. 5 Bubble Class probability densities (normalized) per instrument and pre/post-repair comparison for instrument #8. (A) Density distribution of Bubble Class probabilities across eight instruments, with the number of HPLC traces more than 100. (B) Density distributions for instrument #8 comparing data collected 3 months before (blue) and 3 months after (orange) a repair event. The post-repair distribution shows a shift in Bubble Class probability profiles after maintenance, with a reduction in high-probability occurrences.

HPLC experiments that had bubbles in the lines, not just those that had obvious noticeable negative impacts on the sample elution data. Overall, this will have a few positive impacts for the lab: (1) increase reproducibility of elution times and peak widths (2) improve troubleshooting turnaround time for HPLC.

Implementing the air bubble detection ML model reduces the learning curve for scientists new to HPLC by demystifying one of the major error modes. Researchers will no longer need months or years of experience analyzing various pressure traces alongside experimental outcomes to pinpoint failures. Instead, this model represents a step toward making complex experimentation accessible across disciplines and skill levels. For the more experienced users who typically review key experimental parameters (such as column age, standard and blank traces in a neat window, *etc.*) as part of their workflow during troubleshooting, the ML model predictions are used for displaying the average bubble likelihood for an entire batch run—as well as the minimum and maximum values—providing a rapid quality assessment.

3 Discussion

Recent advances in autonomous scientific discovery offer exciting opportunities but also present formidable challenges, specifically in managing the accumulation and propagation of errors in closed-loop systems, an issue that is often overlooked. One effective strategy to mitigate this risk is to integrate autonomous workflows with robust, on-the-fly quality control and validation of experimental execution.

In this study, we proposed a protocol-agnostic, instrument-agnostic, and, in principle, vendor-neutral framework for on-the-fly detection of common errors in High-Performance Liquid Chromatography (HPLC) experiments. Leveraging a Cloud Lab's rigorous management of all experimental data provides a foundation for adapting and generalizing our end-to-end, data-driven anomaly detection framework to address rarer types of errors as they accumulate over time.

The machine learning model developed in this study demonstrated strong performance in prospective validation across a diverse set of HPLC traces, achieving an accuracy of 0.96 and an F1 score of 0.92 in detecting HPLC traces affected by air bubble contamination, formulated as a binary classification problem. Furthermore, we provided a proof-of-principle demonstration of repurposing the ML model for validation of HPLC instruments based on systematic performance evaluation over a large set of experiments, which appeared to have higher sensitivity compared to individual control experiments.

Future development could enhance the feedback loop in Cloud Lab environments to operate at the level of individual experiments rather than post-batch analysis, enabling automatic retries for affected experiments. The successful application of our framework, using HPLC experiments as a case study, demonstrates its viability for both experiment and instrument validation, addressing key challenges in closed-loop experimental automation.

This study represents a significant advancement in the field of autonomous laboratory operations and has several far-reaching implications for scientific research. At its core, the work addresses a fundamental challenge in automated laboratories: the need for continuous quality control without human oversight. By developing a system that can detect experimental anomalies in real-time, we have eliminated a task that traditionally required experienced human operators.

The implications for democratizing science are particularly noteworthy. The system significantly reduces the learning curve for scientists new to HPLC by automating the detection of common error modes. This means researchers no longer need months or years of experience to identify certain types of experimental failures, making complex experimentation more accessible across disciplines and experience levels.

Our approach to maintenance and quality control represents a significant improvement over traditional methods. It demonstrated higher sensitivity in detecting certain equipment issues compared to conventional qualification tests, thereby enabling proactive maintenance through early identification of systematic problems before they cause major failures. Unlike periodic checks, our method provides continuous quality monitoring.

Looking toward the future, this work opens possibilities for automatic retrial of failed experiments in real-time and represents a crucial step toward fully closed-loop experimental automation. The framework could be expanded to detect other types of experimental anomalies as more data accumulates in the Cloud Lab. From a practical standpoint, this work helps prevent waste of valuable samples, reduces equipment downtime through predictive maintenance, and potentially lowers operational costs by preventing failed experiments.

The present work marks an important milestone in making automated laboratories more reliable and accessible, while potentially reducing costs and enhancing research quality across scientific disciplines. As laboratories increasingly move toward automation and cloud-based operations, models like this will become essential for maintaining rigorous standards of scientific research while democratizing access to advanced analytical instrumentation.

4 Methods

We formulated the anomaly detection problem as a binary classification task on the pressure traces. Experiments known to be affected by air bubbles were labeled as the positive class (class 1), while experiments known to be unaffected—along with additional samples selected *via* (SNA) from the initial dataset (randomly chosen based on the low expected frequency of air bubble anomalies or guided by model predictions, forming)—were labeled as the negative class (class 0).

The model development workflow (Fig. 2B) was organized into three iterative rounds of annotation. In Round 1 (the initial round) we obtained the initial dataset to start the Active Learning cycle. In Round 2, a classical ML modeling approach was employed to create an ensemble of models, allowing us to balance selecting candidates for annotation between cases that were likely positive (specifically "class1_prob_mean > 0.5 and class1_prob_std < 0.1" yielding 287 candidate cases) and uncertain cases (specifically "class1_prob_std ≥ 0.1 yielding 213 candidate cases). This process—after expert evaluation—resulted in a pool of 567 experiments affected by air bubbles. In Round 3 we utilized DeepLearning modeling (see Methods for details), which led—after expert evaluation—to a final accumulation of 700 experiments affected by air bubbles.

At this stage, the pool of traces with high uncertainty (0.1 < "class1_prob < 0.9") significantly decreased, compared to previous rounds, to 93, suggesting that no further rounds of annotation were required. The accumulated data were then used to train the final ML model for deployment—with an optimized architecture and hyperparameters (see S1). The model performance was measured on prospective experiment validation—to assess generalizability—and on prospective equipment validation.

4.1 Datasets

4.1.1 Initial dataset. The initial dataset—comprising HPLC experiments collected over an extended period—consisted of approximately 25 thousand experiments (25 423). The majority of the data fell into the following three categories: (1) semi-preparative size exclusion chromatography experiments separating small molecules away from target oligonucleotides; (2) preparative reverse phase ion pair chromatography experiments separating a desired oligonucleotide from a mixture of small molecules and undesired oligonucleotides; and (3) reverse phase chromatography experiments aimed at small molecule analysis mainly for the purpose of qualifying and ensuring the health of the HPLC instruments. (see Table S2 for detailed counts per chromatography type, instrument model and manufacturer) Among these, 93 experiments were annotated by a human expert as being affected by air bubble contamination.

4.1.2 Prospective experiment validation. To measure model performance in a real-world scenario, we constructed a prospective experiment validation dataset consisting of 967 HPLC experiments conducted after deploying the final model to the Cloud Lab. All experiments were annotated by human expert and compared with the deployed model's predictions (Fig. 4). Detailed performance breakdown across chromatography types and instrument models is provided in Table S4.

4.1.3 Prospective instrument validation. To validate model's ability to detect instruments for which air bubble affected HPLC records are overrepresented, we constructed a separate dataset for eight instruments with each of which performed at least 100 experiments. For the affected instrument #8, data were collected for the three months preceding and the three months following a repair, ensuring an adequate sample size (over 100 experiments for each time period).

4.2 Machine learning modeling

4.2.1 Classical ML. The initial dataset (see above) was processed as follows. First, pressure traces were extracted and treated as time-series data. Experiments with traces exceeding 10 000 time steps, shorter than 100 time steps, or longer than 75 minutes were discarded, yielding a final set of 25 036 experiments. Next, we featurized the remaining experiments using the tsfresh[29] default set of 783 features. Features that contained undefined values in any experiment were removed, reducing the feature set to 585. We then processed these features using Scikit-learn:[31] (1) each feature was scaled using a MinMaxScaler; (2) features with a variance of less than 0.01 were filtered out—resulting in 122 features; and (3) a pairwise Spearman correlation matrix was computed; for each pair of features with an absolute correlation greater than 0.9, only one was retained. This procedure resulted in a final processed dataset comprising 99 features suitable for classical machine learning modeling.

Notably, before applying classical ML modeling, we tried several non-ML, simple mathematical models like pressure oscillation or the derivative of pressure with respect to time, and signal processing approaches. All of them were deemed not suitable for the project because of their lack of transferability and extensibility.

Since the classification task is sensitive to class imbalance, and due to usage of SNA framework we evaluated three class ratio variants: 1 : 1, 1 : 10, 1 : 100. The positive class was represented by the 93 initially annotated as affected by air bubble contamination, while the negative class examples were

randomly sampled from the Initial dataset according to desired class ratio. The Random Forest algorithm was used for classification. The dataset was split using StratifiedKFold in 5 folds, hyperparameters ({'max_depth':[2,4,8,16,32,64,None], 'n_estimators':[10,25,50,100,250,500,1000], "max_features":['auto', 'sqrt', 'log 2']}) were optimized using GridSearch inner loop cross-validation with F1 score as objective function. The model with the best hyperparameters was fitted on the whole fold of the outer 5-fold cross-validation loop.

Based on model performance (Fig. S1) 1 : 10 class ratio was considered optimal (forming a training set of 1023 traces: 93 affected + 930 SNA-sampled unaffected) and used for all other modeling stages. For Round 1, an ensemble of 5 models obtained for the 1 : 10 class ratio was used.

4.2.2 Deep learning ML. We employed a 1D deep convolutional neural network for binary classification. Pressure traces were represented as 1D vectors, with first SKIP_FIRST_N values skipped; the remaining values were trimmed to lengths of 1000 time steps and left-padded with zeros. SKIP_FIRST_N was treated as a hyperparameter and optimized. The model was implemented using PyTorch Lightning. The CNN consists of a series of 1D convolutional layers interleaved with ReLU activations and Batch Normalization layers, followed by fully connected layers with ReLU activation and a dropout layer (dropout rate: 0.1) for aggregation.

The model was trained using binary cross-entropy loss (BCELoss) with the Adam optimizer. The initial learning rate was treated as a hyperparameter, and a learning rate scheduler (ReduceLROnPlateau: factor 0.5, patience 20) was employed. The model was trained for up to 500 epochs, with early stopping based on validation loss (patience: 150) and batch size of 100. Performance metrics—including F1 score, accuracy, precision, and recall—were tracked across the training, validation, and test splits.

The model's hyperparameters (and architecture) were fixed for Round 2 and later optimized for deployment using Optuna[30] by maximizing the validation F1 score with an optimization budget of 2000 trials.

For Round 3 annotation using the CNN, the dataset was constructed as follows. From the Round 2 and initial sets, 567 traces annotated as affected by air bubbles were combined with 5670 SNA samples randomly drawn from the initial dataset—preselected using an Upper Confidence Bound (UCB) threshold of < 0.05 (where UCB is defined as "class1_prob_mean" + "class1_prob_std" from the Round 1 RandomForest ensemble). The combined dataset—forming the dataset for Round 3—was then split into training, testing, and validation sets in an 80 : 10 : 10 ratio with class stratification.

For training the CNN intended for deployment to Cloud Lab, we used 700 traces annotated as affected by air bubbles (acquired by the end of Round 3), 261 traces annotated as normal, and sampled SNA examples from the initial dataset (to accumulate in total 7000 normal traces preserving the desired class ratio)—preselecting those with a Round 3 ML model predicted class 1 probability < 0.05—to construct the deployment training dataset (Table S1 and Fig. S2 for further details). This dataset was split into training, testing, and validation sets in an 80 : 10 : 10 ratio with class stratification. Finally, the trained model was converted to ONNX to ensure native compatibility with the Wolfram Mathematica-based backend of the Cloud Lab.

Author contributions

Conceptualization – F. G.; O. I., Methodology—F. G. (development/design of the machine learning models), software—F. G. (implementation of machine learning models), formal analysis—F. G. (data analysis), investigation—B. C. K.; R. Q.; A. X.; B. S.; B. F. (conducting experiments, collecting data, cloud lab integration), resources—B. C. K.; R. Q.; A. X.; B. S.; B. F. (providing experimental data, technical and computational infrastructure), writing – original draft—F. G.; O. I., writing – review & editing—F. G.; O. I.; B. C. K.; R. Q.; A. X.; B. S.; B. F. (input from all authors).

Conflicts of interest

B. C. K., R. Q., A. X., B. S., and B. F. are employees of Emerald Cloud Lab and Emerald Therapeutics. The other authors declare no competing interests.

Data availability

All scripts and pretrained models are available at **https://github.com/isayevlab/HPLC_anomaly**; archived version available by DOI at **https://doi.org/10.1184/R1/30389245.v1**.

Supplementary information is available. See DOI: **https://doi.org/10.1039/d5dd00253b**.

Acknowledgements

Authors would like to thank Carrie McDonough and Kevin Noonan for critical reading of the manuscript and stimulating discussions. O. I. is supported by the National Science Foundation (NSF) through the Center for Computer-Assisted Synthesis (C-CAS) CHE-2202693 award. This work used Expanse (SDSC) and Delta (NCSA) systems through allocation CHE-200122 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Notes and references

1 M. Baker, *Nature*, 2016, **533**, 452–454.
2 C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune and D. Ha, The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, *arXiv*, 2024, preprint, arXiv:2408.06292, DOI: **10.48550/arXiv.2504.08066**.
3 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, *Nature*, 2023, **624**, 570–578.
4 G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G. D. Akkoc, F. Strieth-Kalthoff,

M. Seifrid and A. Aspuru-Guzik, *Chem. Rev.*, 2024, **124**, 9633–9732.
5 F. Strieth-Kalthoff, H. Hao, V. Rathore, J. Derasp, T. Gaudin, N. H. Angello, M. Seifrid, E. Trushina, M. Guy, J. Liu, X. Tang, M. Mamada, W. Wang, T. Tsagaantsooj, C. Lavigne, R. Pollice, T. C. Wu, K. Hotta, L. Bodo, S. Li, M. Haddadnia, A. Wołos, R. Roszak, C. T. Ser, C. Bozal-Ginesta, R. J. Hickman, J. Vestfrid, A. Aguilar-Granda, E. L. Klimareva, R. C. Sigerson, W. Hou, D. Gahler, S. Lach, A. Warzybok, O. Borodin, S. Rohrbach, B. Sanchez-Lengeling, C. Adachi, B. A. Grzybowski, L. Cronin, J. E. Hein, M. D. Burke and A. Aspuru-Guzik, *Science*, 2024, **384**, eadk9227.
6 N. H. Angello, V. Rathore, W. Beker, A. Wołos, E. R. Jira, R. Roszak, T. C. Wu, C. M. Schroeder, A. Aspuru-Guzik, B. A. Grzybowski and M. D. Burke, *Science*, 2022, **378**, 399–405.
7 M. Seifrid, R. Pollice, A. Aguilar-Granda, Z. Morgan Chan, K. Hotta, C. T. Ser, J. Vestfrid, T. C. Wu and A. Aspuru-Guzik, *Acc. Chem. Res.*, 2022, **55**, 2454–2466.
8 M. Jones and R. L. Goodyear, *ACS Med. Chem. Lett.*, 2023, **14**, 916–919.
9 C. Ginsburg-Moraff, J. Grob, K. Chin, G. Eastman, S. Wildhaber, M. Bayliss, H. M. Mues, M. Palmieri, J. Poirier, M. Reck, A. Luneau, S. Rodde, J. Reilly, T. Wagner, C. E. Brocklehurst, R. Wyler, D. Dunstan and A. N. Marziale, *SLAS Technol.*, 2022, **27**, 350–360.
10 N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng and G. Ceder, *Nature*, 2023, **624**, 86–91.
11 J. M. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377–381.
12 M. Reis, F. Gusev, N. G. Taylor, S. H. Chung, M. D. Verber, Y. Z. Lee, O. Isayev and F. A. Leibfarth, *J. Am. Chem. Soc.*, 2021, **143**, 17677–17689.
13 M. Abolhasani and E. Kumacheva, *Nat. Synth.*, 2023, **2**, 483–492.
14 Y. Jiang, D. Salley, A. Sharma, G. Keenan, M. Mullin and L. Cronin, *Sci. Adv.*, 2022, **8**, eabo2626.
15 J. T. Rapp, B. J. Bremer and P. A. Romero, *Nat. Chem. Eng.*, 2024, **1**, 97–107.
16 D. Fourches, E. Muratov and A. Tropsha, *J. Chem. Inf. Model.*, 2010, **50**, 1189–1204.
17 M. Hart, K. Idanwekhai, V. M. Alves, A. J. M. Miller, J. L. Dempsey, J. F. Cahoon, C.-H. Chen, D. A. Winkler, E. N. Muratov and A. Tropsha, *Chem. Mater.*, 2024, **36**, 9046–9055.
18 W. D. Mellin, *Hammond Times*, 1957, p. 65.
19 A. A. Volk and M. Abolhasani, *Nat. Commun.*, 2024, **15**, 1378.
20 L. R. Snyder, J. J. Kirkland and J. W. Dolan, *Introduction to Modern Liquid Chromatography*, Wiley, 1st edn, 2009.
21 H. Terada, K. Koterasawa and T. Kihara, Increased Analysis Throughput by Overlapped Injection Using the SIL-40 Series Autosampler, *Shimadzu Technical Report C190-E235*, 2019.
22 S. Akita and K. Watanabe, New Analytical Intelligence Concept—Support for Automating Analytical Operations, *Shimadzu Technical Report C190-E245*, 2019.
23 Shimadzu Corporation, Nexera Series UHPLC: AI, IoT, Auto-Diagnostics and Recovery, *Product brochure C196-E096C*, 2021, **https://www.shim-pol.pl/files/115283891/broszura-lc-40.pdf**.
24 Waters Corporation, *Alliance iS HPLC System Instrument Software 1.2.0 — Release Notes*, 715009116, 2024, **https://help.waters.com/content/dam/waters/en/support/releasenotes/2024/715009116/715009116v00.pdf**.
25 Thermo Fisher Scientific, *Vanquish Core HPLC Systems — Simple to the Core*, Brochure BR-73271, 2023, **https://documents.thermofisher.com/TFS-Assets/CMD/brochures/br-73271-hplc-vanquish-core-simple-to-the-core-br73271-en.pdf**.
26 Agilent Technologies, *1290 Infinity III LC System — System Manual*, User manual G7104A, 2025, **https://www.agilent.com/cs/library/usermanuals/public/G7104ASystem.pdf**.
27 E. L. Cáceres, N. C. Mew and M. J. Keiser, *J. Chem. Inf. Model.*, 2020, **60**, 5957–5970.
28 M. Brocidiacono, P. Francoeur, R. Aggarwal, K. I. Popov, D. R. Koes and A. Tropsha, *J. Chem. Inf. Model.*, 2024, **64**, 2488–2495.
29 M. Christ, N. Braun, J. Neuffer and A. W. Kempa-Liehr, *Neurocomputing*, 2018, **307**, 72–77.
30 T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage AK, USA, 2019, pp. 2623–2631.
31 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.