Accelerating the Discovery of Abiotic Vesicles with AI-Guided Automated Experimentation

Christelle Ekosso,[†,¶] Hao Liu,[‡,¶] Avery Glagovich,[†] Dustin Nguyen,[‡] Sarah Maurer,[*,†] and Joshua Schrier[*,‡]

[†]*Department of Chemistry and Biochemistry, Central Connecticut State University, 1615 Stanley Street, New Britain, Connecticut, 06050, USA*

[‡]*Department of Chemistry and Biochemistry, Fordham University, 441 East Fordham Road, The Bronx, New York, 10458, USA*

[¶]*Contributed equally to this work*

E-mail: smaurer@ccsu.edu; jschrier@fordham.edu

Abstract

The first protocells are speculated to have arisen from the self-assembly of simple abiotic carboxylic acids, alcohols, and other amphiphiles into vesicles. To study the complex process of vesicle formation, we combined laboratory automation with AI-guided experimentation to accelerate the discovery of specific compositions and underlying principles governing vesicle formation. Using a low-cost commercial liquid handling robot, we automated experimental procedures, enabling high-throughput testing of various reaction conditions for mixtures of seven (7) amphiphiles. Multi-template Multi-scale Template Matching (MMTM) was used to automate confocal microscopy image analysis, enabling us to quantify vesicle formation without tedious manual counting. The results were used to create a Gaussian process surrogate model, and then active learning was used to iteratively direct the laboratory experiments to

reduce model uncertainty. Mixtures containing primarily trimetyl decylammonium and decylsulfate in equal amounts formed vesicles at sub-millimolar critical vesicle concentrations, and that more than 20% glycerol monodecanoate prevented vesicles from forming even at high total amphiphile concentrations.

Keywords

Gaussian process classifier, Template matching, Microscopy, Amphiphiles, Vesicles, Origins of life

1 Introduction

Compartmentalization is a crucial step in the formation of life.[1] Many studies over the past decades have highlighted the role of non-biological amphiphilic compounds in forming vesicles as compartments for the origins of life.[2–5] This self-assembly of simple amphiphiles into membrane structures (protocells) could have compartmentalized early biological processes on prebiotic Earth and allowed for the selection of more fit individuals.[6] However, it is unlikely that only a single amphiphile type would be present in such an environment, as prebiotic synthesis produces mixtures of compounds.[7] Therefore, we assume that the first cells would have self-assembled from a mixture of available compounds in the environment. Furthermore, this process would need to occur at low amphiphile concentrations that would have been present on early Earth. Despite progress on understanding the prebiotic synthesis of diverse amphiphiles,[8,9] studies of protocell self-assembly have generally considered relatively simple (1-3 component) mixtures.

A combination of automation and machine learning could enable the study of otherwise intractable mixtures, like the diverse compositions and conditions for protocell formation. Recent review articles describe this combination as a "self-driving laboratory"[10–13] and "autonomous experimentation system".[14–16] Automation assists by reducing the time and labor

needed to prepare, dehydrate, and rehydrate, and analyze samples. Thus, in addition to taking advantage of automated liquid handling, computer vision methods can be used to automate the detection of vesicles in microscopy images.[17–20] But brute force automation alone is insufficient. A complete combinatorial exploration of N amphiphiles at only 10 different concentrations would require 10^N experiments, which would be impossible to perform for reasonably complex mixtures. Therefore, AI-guided techniques, such as active learning,[21] are essential to accelerate the process. Active learning has been applied to a wide variety of chemical problems from drug discovery[22] to material synthesis,[23] to nanoscience.[24] The closest related work are studies by Lapkin and co-workers which used active learning methods to optimize the viscosity of 5-,[25] 7-,[26] and 18-component[27] surfactant/polymer/thickener formulations. Cronin and co-workers used a combination of automation and machine learning to study the behavior of 2- and 4-component fatty acid and alcohol oil droplets and their interaction with dissolved surfactants as a model for protocell behavior.[28–30]

In this paper, we demonstrate the use of automated sample preparation and confocal microscopy characterization to find seven-component mixtures of abiotic amphiphiles capable of forming vesicles at low concentrations. Active learning enables efficient updates to the underlying machine learning model, allowing us to generate experimentally testable predictions and extract insights about the role of different amphiphile species in vesicle formation in water.

2 Methods

The overall experimental cycle is depicted schematically in Figure 1.

Sample preparation. We explored mixed compositions of seven (7) different 10-carbon amphiphiles (Chart 1). These amphiphiles were chosen as they represent a chemically interesting set of headgroups, both polar and charged, and are short enough to be considered prebiotically relevant. These amphiphiles are also relatively inexpensive and available from

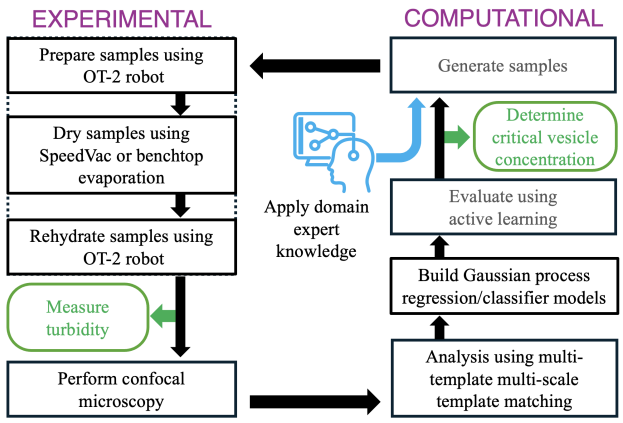


Figure 1: Schematic representation of the workflow process

commercial sources.

Chart 1: Chemical structures of chosen seven 10-carbon amphiphiles: (1) decanoic acid, (2) decanoate, (3) decylamine, (4) decyl trimethyl ammonium bromide, (5) sodium decyl sulfate, (6) decanol, (7) glycerol monodecanoate (GMD).

Stock solutions were prepared as 50 mM of decanoic acid (Thermo Fisher Scientific, Agawam, MA, USA), decylamine (Thermo Fisher Scientific, Agawam, MA USA), decyltrimethyl ammonium bromide (DTAB; TCI, Portland, OR, USA), were prepared in 70% ethanol. Decanoate was prepared from decanoic acid in 0.1 M NaOH. Using a serial dilution, 10 mM and 2 mM stock solutions were prepared from the 50 mM and 10 mM solution respectively. The same process was followed to prepare 50 mM, 10 mM, and 2 mM of decyl sodium sulfate (Chem-Impex Int'l Inc, Wood Dale, IL, USA) in deionized water. Decanol (Thermo Fisher

Scientific, Agawam, MA, USA) was brought up in 70% ethanol at concentrations of 15 mM and 3 mM. Glycerol monodecanoate (TCI, Portland, OR, USA) was dissolved in 70% ethanol and brought to concentrations of 10 mM and 2 mM.

Samples were prepared by using an Opentrons OT-2 liquid handling robot (Brooklyn, NY, USA)[31] to dispense stock solutions into 96-well SBS microplates. We verified that the use of 70% ethanol solutions was reliable in the liquid handler as shown in Figure S1. The dispensing process used both "touch tip" and "blow out" instructions to ensure all the liquid was pushed out from the tip for each operation. Experiments were typically performed in batches of 48 compositions, repeated in quadruplicate over two SBS microplates, unless noted otherwise. After the dispensing process, samples were dried at ambient temperature by leaving the well plates on the bench (approximately 12 h) or using a vacuum concentrator SpeedVac SPD120 at 35°C (approximately 2 h 30 min). Samples were then rehydrated using the OT-2 to dispense 200 µL of deionized water, and mixing by repeat pipetting 20 times. After all samples were rehydrated, the SBS plate was placed on a shaker for at least 12 h to obtain well-mixed homogeneous solutions. Python codes used for all OT-2 processes are available at `https://github.com/mfbliposome/opentrons`

Turbidity measurements. Turbidity measurements were performed using a UV/vis plate reader Multiskan SkyHigh Microplate Spectrophotometer (ThermoScientific, USA). UV/vis data were collected from 350 nm to 700 nm (in 10 nm intervals). Turbidity was evaluated at 450 nm.

Critical Vesicle Concentration. A serial dilution of an amphiphile mixture was prepared in well plates. To each 100 µL sample, 1 µL of 4 mM Merocyanine 540 was added. Visible absorbance was collected from 400-600 nm. The baseline was normalized at 600 nm to 0, and then the intensity ratio of the 570 to 540 nm peak was calculated. The CVC was then determined as the first set of three points that had a slope above the baseline value.

Confocal microscopy. Using the OT-2, 10 µL of Nile red was added to 100 µL of each sample, then 25 µL of each sample was transferred from 96 well-plate to 384 well-plate fitted with a coverslip base (Cellvis, USA) for automated confocal microscopy. Samples were imaged using a Nikon AXR confocal microscope with a 60X/1.40 Plan Apo Lambda D oil immersion lens and 561 nm excitation. A custom routine for automated 384 well-plate imaging was programmed using NIS Elements' JOBS function. Each well was imaged from 4 random fields, and exported as a TIFF file. The raw microscopy data is available at https://doi.org/10.5281/zenodo.12522610

Image analysis. Depending on the image quality (determined by manual inspection), images were first preprocessed by background subtraction using a Gaussian filter, contrast enhancement using CLAHE (Contrast Limited Adaptive Histogram Equalization), and noise suppression using a Gaussian blur. Vesicle detection and counting were performed using template matching, a computer vision technique that finds the location of a small template within a larger image by sliding a template image over the target image to find high-similarity regions. More specifically, a multi-template multi-scale template matching (MMTM) algorithm was used based on previously developed for extracellular vesicle counting,[18] as implemented in the cv2 module from the OpenCV library.[32] Source code is available at https://github.com/mfbliposome/ActiveLearning_Campaign/

Surrogate model and active learning. Active learning is a machine learning approach that selects the most informative data points for labeling in an iterative setting.[21] In this work, we used active learning to efficiently design experiments, achieving desired performance with fewer iterations. This technique is particularly valuable in scenarios where experiments are labor- and time-intensive. This work used two strategies—Bayesian optimization (BO) with a Gaussian Process Regression (GPR) model and uncertainty sampling with a Gaussian Process Classifier (GPC) model. The input to both models was a seven-dimensional vector, where each dimension represents the concentration of the seven amphiphile species in units

of millimolar. BO was performed using a Matern kernel to identify amphiphile compositions most likely to form vesicles, focusing on minimizing total concentration. The probability of improvement (PI) and expected improvement (EI) acquisition functions were used to select subsequent experiments. GPC was employed to solve binary classification problems, with entropy-based uncertainty used to select the next experiments. GPR and GPC models were implemented by using the botorch[33] and scikit-learn[34] package. Program code is available at `https://github.com/mfbliposome/ActiveLearning_Campaign/`

3 Results and discussion

Reliability of automated confocal microscopy image analysis. Figure 2 shows example confocal microscopy and their analysis with the MMTM image processing algorithm. The MMTM algorithm reliably detected vesicles across a wide range of concentrations, from abundant to sparse. It also distinguished vesicles from non-vesicular structures such as crystalline formations and oil droplets, reducing the likelihood of false positives. For example, the algorithm correctly excludes oil droplets (solid round shapes in Figure 2e) and crystalline shapes (long strips in Figure 2c, f). A complete analysis of the 768 images in a batch of experiments requires only 84 minutes of CPU time on a laptop computer and is trivially parallelized. This is much faster than human experts. Although the present study only required counting vesicles, the algorithm can also calculate vesicle sizes and automate preprocessing and post-processing tasks, including the recording of individual image analyses and compiling summary results for entire batches. This efficiency accelerates the pace of research and enhances reproducibility by minimizing human error.

Active learning directed experimentation. Finding low total amphiphile concentration compositions that reliably form vesicles informs our understanding these systems as precursors to life in dilute pools on the early Earth. We used a human-in-the-loop active learning approach, in which human experts set the concentration bounds to explore and the

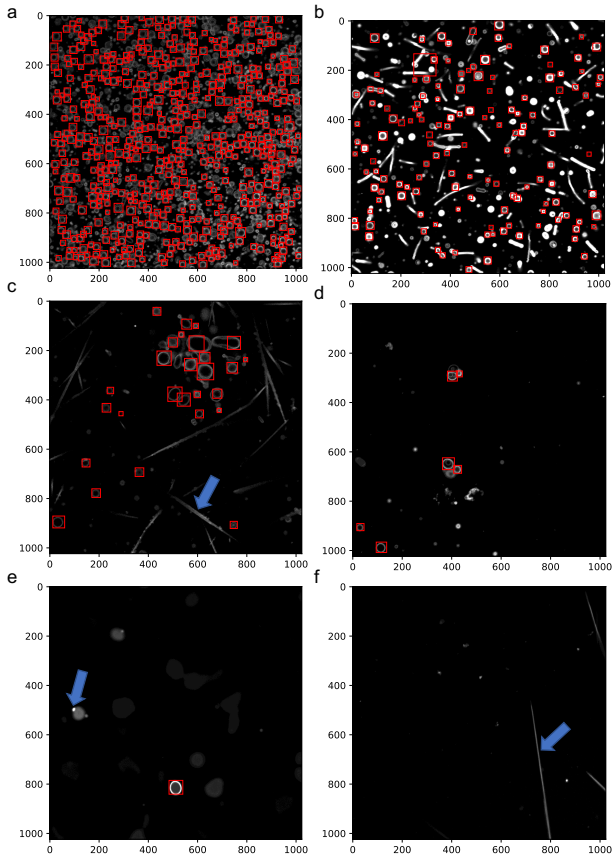


Figure 2: Analysis of 6 randomly selected confocal microscopy images. Each image was 102 µm x 102 µm. The detected vesicles are located inside red bounding boxes. Blue arrows show oil droplets and crystals that are correctly not identified as vesicles by the MMTM algorithm.

active learning (AL) algorithm selected the compositions within those bounds that would be most useful for improving the model quality. Each AL round was performed in batches of 48 compositions, each replicated in quadruplicate. The outcomes and total concentrations are shown in Figure 3a. The initial round (AL 0, green) was selected by uniform random sampling at concentrations of the 7 amphiphiles in the hypercuboid between the origin and the upper bounds of [5., 5., 5., 5., 5., 1.5, 1.0] mmol for each species. The smaller upper bounds for decanol and glycerol monodecanoate were selected because of their lower solubility in ethanol. The distribution of total concentrations is a consequence of the central limit theorem, as that the sum of uniformly distributed random numbers converges to a normal distribution. Surprisingly, all but one of these randomly selected compositions results in the

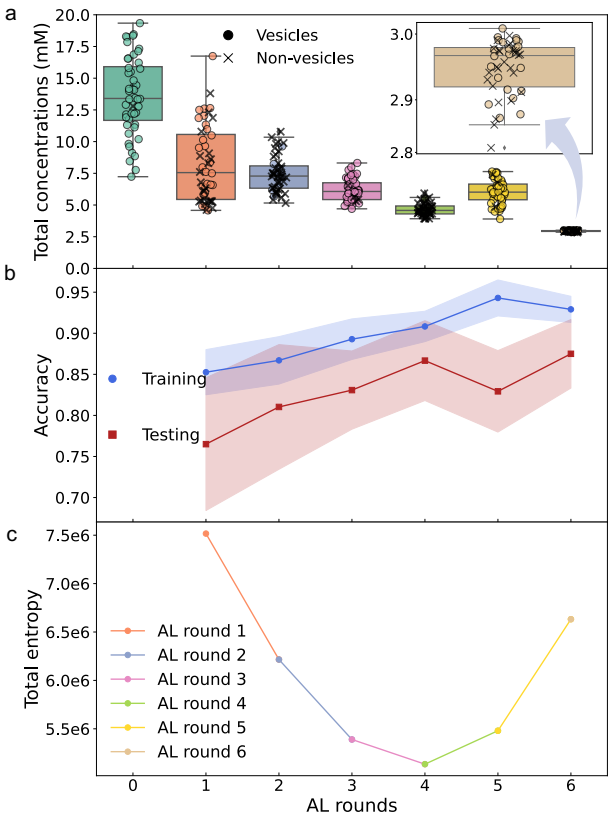


Figure 3: Active learning results and model performance. (a) Total concentration of amphiphiles evolution across the active learning rounds. Box and whisker plots indicate the distribution of total concentrations for samples in each AL round. The box represents the interquartile range (IQR), which captures the middle 50% of the data. The top and bottom edges of the box correspond to the third quartile (75th percentile, Q3) and first quartile (25th percentile, Q1) respectively. The whiskers represent the data points within 1.5 times the IQR from Q1 and Q3, which capture the variability outside the IQR. Outliers are values that fall outside the whiskers' range. (b) Model performance (classification accuracy) along with active learning rounds on the training (blue) and testing (red) datasets. (c) Model uncertainty along active learning rounds, quantified by the total entropy of all samples in the input space.

formation of vesicles, with concentrations as low as 7.2 mM.

Using the data from AL 0, a BO/GPR model was used to select the AL 1 experiments. The AL 1 results (orange) show that 54% of samples had vesicles, and the total concentration was decreased since the objective function in BO is designed to lower the total concentration of amphiphiles (see Equation S7). However, backtesting on the collected AL 0 and AL 1 data at this stage revealed that the GPR model tended to overfit and had poor generalizability to unseen data, because of a discontinuity in the penalty function for samples without vesicles (see discussion in Supporting Information). Similar backtesting with a GPC model did not have these problems, so the GPC model was used for all the subsequent AL rounds. This transforms the question from an optimization into a binary classification problem. Samples with more than zero vesicles were treated as positive samples, and others were treated as negative samples. The most uncertain samples (i.e., samples near the classification boundary) were selected for the subsequent round, starting with AL 2. First, 10 million samples were randomly selected within the input space bounds. Based on predicted probabilities from the GPC model, the most uncertain 48 samples were chosen as the next experimental plan. The uncertainty was evaluated by the entropy of each data point, which is defined in Equation S12 and Equation S13

About 13% of samples in AL 2 (blue) had vesicles, the average total concentration in AL 2 was about 7.5 mM. To focus attention on lower concentrations in AL 3, the input concentration maxima were reduced by half, and still, 92% of AL 3 samples (pink) had vesicles! The total concentration bounds were limited to less than 1 mM for every amphiphile in in AL 4 , but none of these samples contained vesicles (green). Hence, in AL 5 (yellow), a decision was made to revert the input space bounds to the same as in AL 3. Although the GPC model had been updated based on the new data, the results were similar to AL 3, and 94% samples had vesicles. This suggested that finding vesicles was relatively easy in this range of concentrations.

To test the capability of the GPC model to extrapolate to lower concentrations, we used

a GPC model trained on all data collected from AL 0 to AL 5 to predict the outcomes of experiments with total concentrations less than 3 mM, 2 mM, and 1 mM (See Figure S7). The possibilities of finding vesicles in the region where the total concentration was less than 3 mM were most promising. Therefore, in AL 6, half of the samples (24 samples) that had the highest probability and the other half that had the lowest probability of forming vesicles were chosen. The composition of those samples can be seen in Figure S2. Of the 24 samples predicted to have the highest probability of vesicle formation, only 1 did not have vesicles present (4% false positive). For the 24 samples having the lowest probabilities of vesicle formation, only 2 samples were observed to have vesicles (8% false negative rate). Thus, the experiment results are consistent with the model prediction, which demonstrates the ability of the GPC model to extrapolate to low concentration conditions.

CVC measurements were performed on the four mixtures from AL 6 with the largest number of vesicles to better characterize the lower limits of formation. The critical point for vesicle formation was around 1 mM with the lowest value tested at 0.85 mM. These values are exceedingly low for single-chain amphiphile membranes. For example, the most well characterized prebiotic vesicles, formed from a mixture of decanoic acid/decanoate have the CVC between 10 and 40 mM.[35] While these membranes were still orders of magnitude greater than that of stable phospholipid membranes (e.g., dipalmitoylphosphatidylcholine or DPPC is estimated to have CVCs in the sub picomolar range),[36] this work demonstrates how active learning can find much lower CVCs for prebiotic amphiphile mixtures than previously known.

How does the model evolve during this process? Having demonstrated the effectiveness of uncertainty-driven active learning to find compositions that form vesicles under low total amphiphile concentration, a natural question is *how* the model evolves during this process. In practice, monitoring this type of information is useful for guiding the decision about when to stop the exploration stage (AL 0 - AL 5) and begin the exploitation stage (AL 6)

in the process. Figure 3b shows the GPC model accuracy for AL 1 to AL 6, as AL 0 was not directly comparable using the GPR model. The data pool was randomly split to 80% for training and 20% for testing and repeated 10 times. Overall, the training and testing accuracy increased as the AL round progressed, indicating good predictive capabilities and generalizability of the model.

Figure 3c shows the model uncertainty for each AL round, as quantified by the total information entropy for all samples in the initial input space. Predictions made by the GPC model established after AL 6 were performed on 10^7 randomly selected samples in the initial input space, then the entropy for each sample was calculated according to Equation S13. For AL 1 to AL 4, the model uncertainty decreased monotonically, indicating that the model predictions become more confident with each round. In AL 5, the data in AL 3 were revisited, thus the model became more uncertain about existing data, and the total entropy increased. In AL 6, more exploitation of the low concentration regions was investigated, thus, the model gained new knowledge about this area, and the total entropy was also increased.

Another way to visualize the model's evolution is by visualizing the decision boundary between vesicle forming and non-forming compositions. To allow for visualization, the original 7-dimensional compositional data were reduced to 2-dimensions using principal component analysis (PCA). These first two components accounted for only about 50% of the variance, demonstrating the complexities of understanding the input space. The top row of Figure 4 shows how the classifier decision boundary evolved as the AL round progressed. In AL 2, the decision boundary roughly split data into two parts in 2D input space, and PC1 separated as non-vesicles having negative values and vesicles having positive values. Then, AL 4 sampled increased concentrations of glycerol monodecanoate, changing the decision boundary so that negative values in both PC1 and 2 represented a low probability of forming vesicles. In AL 6, the uncertain area became larger because new data explored an unknown concentration/composition region for the model, and is consistent with the total entropy calculation in Figure 3c. The increase in uncertain/grey areas, indicates that as the total

amphiphile concentration ranges changed in AL 2 and 4 at 5-10 mM to <3 mM in AL 6, the 2D representation did not capture the fitness landscape in a meaningful way.

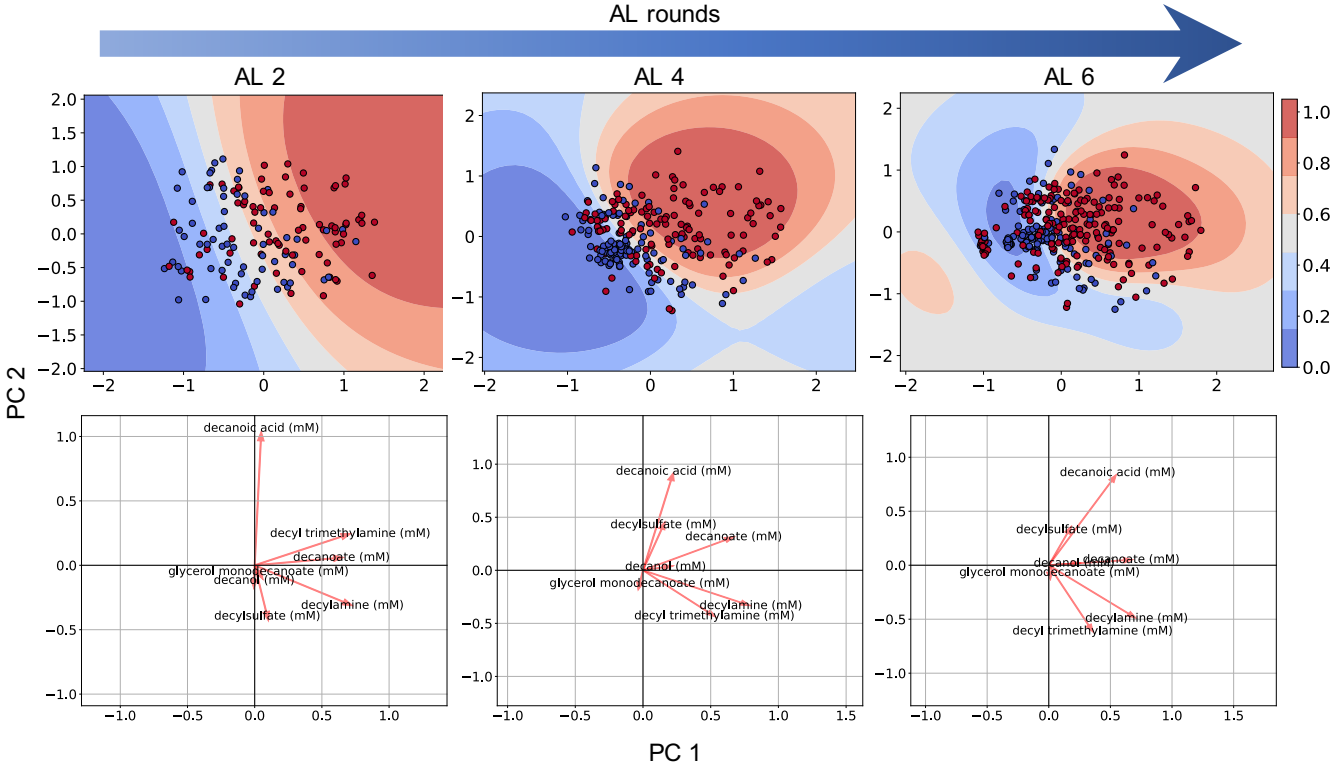


Figure 4: Updates of the classifier decision boundary(first row) and PCA loading plot (second row) during active learning from AL round 2 (left) to AL round 6 (right). Red and blue points represent experimental observations that form and do not form vesicles, respectively. Contours indicate the probability of finding vesicles estimated by the classifier (red is high and blue is low). The second row shows the evolution of PCA loading plots. Each plot shows the contribution of 7 amphiphile concentrations to the first two principal components (PCs). Vector lengths indicate the strength of feature relationships with the PCs, with larger vectors representing higher feature correlation.

Model interpretability and chemical insight. Chemical insights about how different amphiphiles contributed to vesicle formation could also be extracted from the model. The bottom row of Figure 4 shows the PCA loading plot, which explains how different features contributed to each PC (refer to Figure S6 for all rounds). For AL 2, with "high" total amphiphile concentrations, three charged amphiphiles, decanoate, decyl trimethylamine, and decylamine, contributed most strongly to the high probability vesicle region (positive

PC1 values). As the AL iteration progressed, decylsulfate inverted and also became more important to the vesicle forming region as total concentrations of amphiphiles decreased.

Interestingly, glycerol monodecanoate became negatively related to PC1 in AL 4, which is consistent with Figure 3 (a) and Figure S2 where those samples that have a larger percentage of glycerol monodecanoate did not to form vesicles. This is in contrast to previous studies that used glycerol monodecanoate to *stabilize* membranes when mixed with decanoic acid/decanoate in compositional amounts from 10 to 50%.[37,38] Glycerol monodecanoate is the glycerol ester of decanoic acid, and thus has been suggested as an "evolutionary step" in membrane composition between the simplest amphiphiles and phosphoglycerolipids.[39] However, our results show that an abundance of this molecule, at about 10-20% of the total amphiphile concentration, reduced the probability of finding vesicles. This result highlights the complexities of predicting vesicle-forming compositions for the origins of life and suggests that glycerol monodecanoate may be less prebiotically relevant than previously thought.

Among the low total concentration region (< 3 mM), those samples that had high probabilities of forming vesicles were from 88-96% total decyl trimethylamine and decyl sulfate compared to those samples having low probabilities of finding vesicles (Figure S2, top half samples). The ratio of decyl trimethylamine to decylsulfate ranged from about 1:1 to 1:4 in vesicle-forming samples. Moreover, finding vesicles in the region (< 3 mM) is very difficult. As shown in the bottom of Figure 5, only 813 out of 10^7 samples were found to have 70-80% probabilities of finding vesicles. This low percentage can also seen in the histogram shown in Figure 6. The majority of samples concentrated on the probability of 0.1, which means they may not form vesicles according to the model-predicted probabilities. Interestingly, as a quaternary amine, decyl trimethylamine has the same functional group that is found in phosphatidylcholine head groups in modern lipids. Likewise, the decyl sulfate anion is an analogue for a phosphate group, giving this composition type a similar charge structure to lipid bilayers in extant cells.

These results highlight the ability of "messy" mixtures of amphiphiles to generate abun-

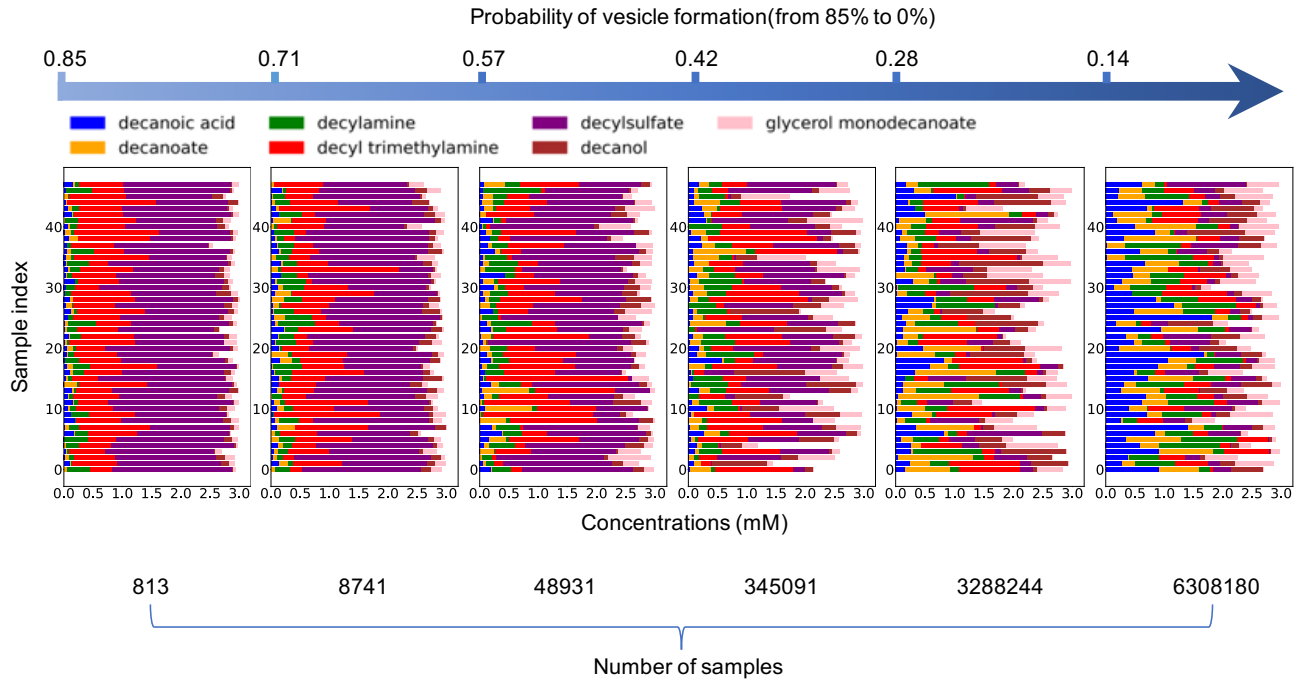


Figure 5: The model prediction of amphiphile compositions for the input space in AL round 6. The total 10 million samples were divided into 6 regions according to vesicle formation probabilities as indicated by the numbers in the top arrow. The figure displays amphiphile concentrations for each region, with absolute concentrations shown and varying total concentrations per sample. Sample counts per region are provided at the bottom.

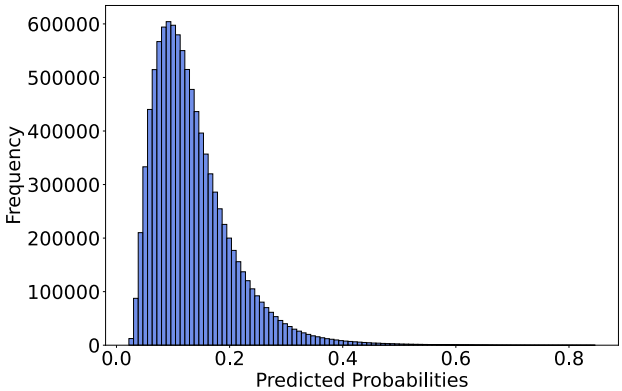


Figure 6: Model predicting probabilities of finding vesicles for amphiphiles compositions in input space in AL round 6.

dant self-assembled structures in water from non-biological lipids and are the first time such a large chemical space has been explored for this purpose. Previous reports have looked at up to three-amphiphile mixtures, generally carboxylate, alcohol, and glycerol. The mixture of decanoate with either decanol or glycerol monoester reduces the critical vesicle concentration to the 1-2 millimolar range.[38,40] Our results are produced in pure water, with some NaBr(aq) generated from the counter ions on our charged amphiphiles, sodium decanoate, sodium decylsulfate and trimethy decylamine bromide. Most of the previous literature did little to control the ionic strength of the solution but likely had between 50 and 150 mM NaCl, which increases the aggregation of decanoate containing vesicles.[35]

Previous reports of vesicles generated from single-chain amphiphiles have been primarily composed of a mixture of a polar headgroup (-OH containing) and a charged head group (either NH_3^+ or COO^-, sometimes PO_4^-).[41–43] Interestingly, our "best" membranes at low concentrations were composed of a mixture of negative sulfate headgroups and positive quaternary amine head groups. It is possible that ion pairing between headgroups contributes to the stability of our novel aggregates. In biological membranes, lipids are generally zwitterionic, with a phosphophoester-linked amine as a majority of head groups in both *E. coli* and human plasma membranes,[44] suggesting the doubly charged amphiphiles are a preferred evolutionary outcome.

Is turbidity a good measure of vesicle formation? Turbidity—quantified as optical absorption at 450 nm—has previously been used to determine the critical vesicle concentration of amphiphiles.[42,45] This would be especially useful for rapid initial screening, as acquiring UV/vis spectra is much faster and easier to process than confocal microscopy. However, we observed that the turbidity was very low (< 0.05 AU at 450 nm) in many of our vesicle-containing samples, and therefore the typical single-wavelength measurement was of limited value. During the experimental process, we generated 5644 samples for which we had both UV/vis spectra and confocal microscopy quantification, which could be used as

inputs for machine learning analysis. We trained three different classifier models to predict whether there was at least one vesicle observed in the (ground truth) confocal microscopy image. Figure 7 shows the tradeoff between the optimal tradeoff between false positive and true positive rates for the different methods; a perfect method would have an area under the curve (AUC) of one, and higher values are better. The baseline model (green curve) implements the typical strategy[42,45] of defining a threshold value for the absorbance at 450 nm to indicate the presence of vesicles; each point along the curve is a different choice of the threshold value, illustrating the sensitivity/selectivity tradeoff. Better performance is obtained by training a random forest classifier on the 450nm values (blue) curve, but the best performance is obtained by using the entire UV/vis spectral data as input to the random forest model (red). We suspect this is because it accounts for non-linearities in the input. (Similar results are obtained with other classifier models, such as support vector machines.) This high accuracy of the RF method (as measured against the ground truth of confocal microscopy) and its much faster and simpler operation mean that it would be well suited for serving as an initial, high-throughput screening technique for multi-fidelity optimization methods,[46] which could be explored in future studies. More generally, this also demonstrates an additional benefit of automated experimentation—the acquisition a systematic, machine-readable dataset of results which can be data-mined after the fact for new insights.[14,47]

4 Conclusion

Automation and active learning can improve our ability to understand complex mixtures for astrobiology, specifically membrane formation. These results demonstrate that submillimolar concentrations of amphiphiles are sufficient to form primitive membranes and that an abundance of compositions form membranes when total amphiphile concentrations are between 1 and 10 mM. A computer vision technique—MMTM has been established to auto-

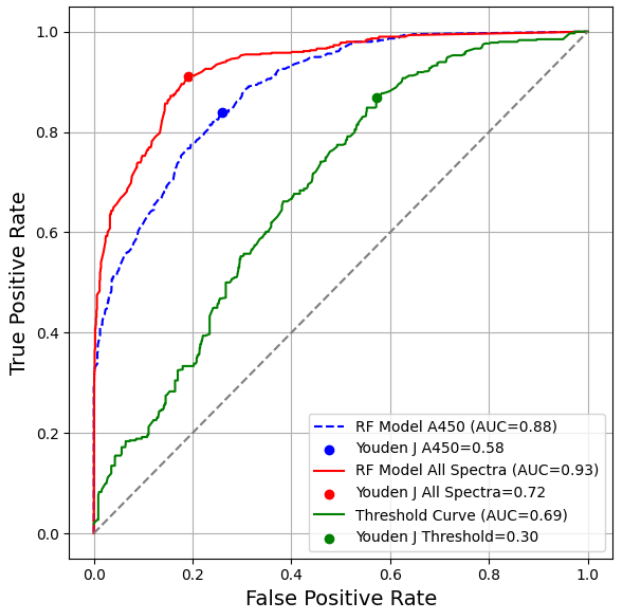


Figure 7: Receiver-Operator Characteristic Curves for predicting the presence of vesicles from UV/vis absorption spectra. Markers indicate the Youdon's-J optimal[48] values for each method.

mate confocal microscopy image analysis, which expedites the research process compared to manual methods. Active learning incorporating human expert domain knowledge was employed to optimize experimental design, offering an efficient approach to explore the chemical landscape without extensive prior knowledge or ground truth. The Gaussian process model built during the active learning process not only quantifies uncertainty but also provides chemical insights. Our model revealed that vesicle formation at lower concentrations is more likely with higher percentages of decyl trimethylamine and decyl sulfate and lower glycerol monodecanoate. While this work focused on a limited set of seven amphiphiles, the approach is versatile and can be extended to more diverse chemical systems. Thus, considering more diverse amphiphiles or constraining relative amphiphile abundances by knowledge of abiotic processes/thermochemistry[7–9] could be future steps. In addition to including other types of fatty acids and alcohols, other prebiotic molecules could be incorporated. A promising example would be dipeptides which increase the growth rate of decanoic acid vesicles.[49] Expanding the scope to include a broader range of environmental conditions, such as varying salinity or

pH, could further refine our understanding of primitive membrane formation. Finally, our current model does not include the chemical structure of the amphiphiles, and so methods that can incorporate the structural aspects of the amphiphiles and their mixtures[27,50] may improve transferability and extrapolation.

5 Data and code availability

The data in this work are openly available on Zenodo at `https://doi.org/10.5281/zenodo.12522610`, The code in this work are openly available on Github at `https://github.com/mfbliposome/ActiveLearning_Campaign/tree/main`, with a persistent archival copy deposited in Zenodo at DOI: [to be deposited after article passes peer review]

Acknowledgement

This work was supported by the National Science Foundation (PHY-2226511). We also thank the MERCURY Consortium for providing computational resources, supported by National Science Foundation OAC-2320718.

References

(1) Rasmussen, S.; Constantinescu, A.; Svaneborg, C. Generating minimal living systems from non-living materials and increasing their evolutionary abilities. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2016**, *371*, 20150440.

(2) Jordan, S. F.; Rammu, H.; Zheludev, I. N.; Hartley, A. M.; Maréchal, A.; Lane, N. Promotion of protocell self-assembly from mixed amphiphiles at the origin of life. *Nature Ecology & Evolution* **2019**, *3*, 1705–1714.

(3) Maurer, S. E.; Tølbøl Sørensen, K.; Iqbal, Z.; Nicholas, J.; Quirion, K.; Gioia, M.;

Monnard, P.-A.; Hanczyc, M. M. Vesicle self-assembly of monoalkyl amphiphiles under the effects of high ionic strength, extreme pH, and high temperature environments. *Langmuir* **2018**, *34*, 15560–15568.

(4) Fiore, M.; Madanamoothoo, W.; Berlioz-Barbier, A.; Maniti, O.; Girard-Egrot, A.; Buchet, R.; Strazewski, P. Giant vesicles from rehydrated crude mixtures containing unexpected mixtures of amphiphiles formed under plausibly prebiotic conditions. *Organic & Biomolecular Chemistry* **2017**, *15*, 4231–4240.

(5) Joshi, M. P.; Sawant, A. A.; Rajamani, S. Spontaneous emergence of membrane-forming protoamphiphiles from a lipid–amino acid mixture under wet–dry cycles. *Chemical science* **2021**, *12*, 2970–2978.

(6) Rasmussen, S.; Chen, L.; Deamer, D.; Krakauer, D. C.; Packard, N. H.; Stadler, P. F.; Bedau, M. A. Transitions from nonliving to living matter. *Science* **2004**, *303*, 963–965.

(7) Mißbach, H.; Schmidt, B. C.; Duda, J.-P.; Lünsdorf, N. K.; Goetz, W.; Thiel, V. Assessing the diversity of lipids formed via Fischer-Tropsch-type reactions. *Organic Geochemistry* **2018**, *119*, 110–121.

(8) Cohen, Z. R.; Todd, Z. R.; Wogan, N.; Black, R. A.; Keller, S. L.; Catling, D. C. Plausible Sources of Membrane-Forming Fatty Acids on the Early Earth: A Review of the Literature and an Estimation of Amounts. *ACS Earth and Space Chemistry* **2023**, *7*, 11–27.

(9) Zhu, P.; Wang, C.; Lang, J.; He, D.; Jin, F. Prebiotic Synthesis of Microdroplets from Formate over a Bimetallic Cobalt–Nickel Nanomotif. *Journal of the American Chemical Society* **2024**, *146*, 25005–25015, PMID: 39219062.

(10) Tom, G. et al. Self-Driving Laboratories for Chemistry and Materials Science. *Chemical Reviews* **2024**, *124*, 9633–9732, PMID: 39137296.

(11) Lo, S.; Baird, S. G.; Schrier, J.; Blaiszik, B.; Carson, N.; Foster, I.; Aguilar-Granda, A.; Kalinin, S. V.; Maruyama, B.; Politi, M.; Tran, H.; Sparks, T. D.; Aspuru-Guzik, A. Review of low-cost self-driving laboratories in chemistry and materials science: the "frugal twin" concept. *Digital Discovery* **2024**, *3*, 842–868.

(12) Hysmith, H.; Foadian, E.; Padhy, S. P.; Kalinin, S. V.; Moore, R. G.; Ovchinnikova, O. S.; Ahmadi, M. The future of self-driving laboratories: from human in the loop interactive AI to gamification. *Digital Discovery* **2024**, *3*, 621–636.

(13) Bayley, O.; Savino, E.; Slattery, A.; Noël, T. Autonomous chemistry: Navigating self-driving labs in chemical and material sciences. *Matter* **2024**, *7*, 2382–2398.

(14) Stach, E. et al. Autonomous experimentation systems for materials development: A community perspective. *Matter* **2021**, *4*, 2702–2726.

(15) Xie, Y.; Sattari, K.; Zhang, C.; Lin, J. Toward autonomous laboratories: Convergence of artificial intelligence and experimental automation. *Progress in Materials Science* **2023**, *132*, 101043.

(16) Noack, M.; Ushizima, D. In *Methods and Applications of Autonomous Experimentation*, 1st ed.; Noack, M., Ushizima, D., Eds.; Chapman and Hall/CRC, 2023.

(17) Gómez de Mariscal, E.; Maška, M.; Kotrbová, A.; Pospichalova, V.; Matula, P.; Muñoz-Barrutia, A. Deep-Learning-Based Segmentation of Small Extracellular Vesicles in Transmission Electron Microscopy Images. *Scientific Reports* **2019**, *9*, 1–10.

(18) Van Buren, L.; Koenderink, G. H.; Martinez-Torres, C. DisGUVery: a versatile open-source software for high-throughput image analysis of Giant Unilamellar Vesicles. *ACS Synthetic Biology* **2022**, *12*, 120–135.

(19) Lee, I.-H.; Passaro, S.; Ozturk, S.; Ureña, J.; Wang, W. Intelligent fluorescence image

analysis of giant unilamellar vesicles using convolutional neural network. *BMC Bioinformatics* **2022**, *23*, 48.

(20) Caliari, A.; Hanczyc, M. M.; Imai, M.; Xu, J.; Yomo, T. Quantification of giant unilamellar vesicle fusion products by high-throughput image analysis. *International Journal of Molecular Sciences* **2023**, *24*, 8241.

(21) Settles, B. *Active Learning Literature Survey*; University of Wisconsin–Madison, 2009; Technical Report 1648.

(22) Reker, D.; Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today* **2015**, *20*, 458–465.

(23) Li, Z.; Nega, P. W.; Nellikkal, M. A. N.; Dun, C.; Zeller, M.; Urban, J. J.; Saidi, W. A.; Schrier, J.; Norquist, A. J.; Chan, E. M. Dimensional Control over Metal Halide Perovskite Crystallization Guided by Active Learning. *Chemistry of Materials* **2022**, *34*, 756–767.

(24) Kim, M. A.; Ai, Q.; Norquist, A. J.; Schrier, J.; Chan, E. M. Active Learning of Ligands That Enhance Perovskite Nanocrystal Luminescence. *ACS Nano* **2024**, *18*, 14514–14522, PMID: 38776469.

(25) Cao, L.; Russo, D.; Felton, K.; Salley, D.; Sharma, A.; Keenan, G.; Mauer, W.; Gao, H.; Cronin, L.; Lapkin, A. A. Optimization of Formulations Using Robotic Experiments Driven by Machine Learning DoE. *Cell Reports Physical Science* **2021**, *2*, 100295.

(26) Cao, L.; Russo, D.; Matthews, E.; Lapkin, A.; Woods, D. Computer-aided design of formulated products: A bridge design of experiments for ingredient selection. *Computers Chemical Engineering* **2023**, *169*, 108083.

(27) Chitre, A.; Woods, D.; Lapkin, A. Machine Learning-Guided Space-filling Designs for

High Throughput Liquid Formulations Development. *ChemRxiv* **2024**, DOI: https://doi.org/10.26434/chemrxiv-2024-18rgd.

(28) Doran, D.; Rodriguez-Garcia, M.; Turk-MacLeod, R.; Cooper, G. J. T.; Cronin, L. A recursive microfluidic platform to explore the emergence of chemical evolution. *Beilstein Journal of Organic Chemistry* **2017**, *13*, 1702–1709.

(29) Points, L. J.; Taylor, J. W.; Grizou, J.; Donkers, K.; Cronin, L. Artificial intelligence exploration of unstable protocells leads to predictable properties and discovery of collective behavior. *Proceedings of the National Academy of Sciences* **2018**, *115*, 885–890.

(30) Grizou, J.; Points, L. J.; Sharma, A.; Cronin, L. A curious formulation robot enables the discovery of a novel protocell behavior. *Science Advances* **2020**, *6*, eaay4237.

(31) Opentrons Opentrons OT-2 Pipetting Robot. https://opentrons.com/ot-2, Accessed: 2024-07-04.

(32) Bradski, G. The OpenCV Library. https://opencv.org/, 2000; Accessed: 2024-10-02.

(33) Balandat, M.; Karrer, B.; Jiang, D.; Daulton, S.; Letham, B.; Wilson, A. G.; Bakshy, E. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Advances in neural information processing systems* **2020**, *33*, 21524–21538.

(34) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(35) Maurer, S. E.; Nguyen, G. Prebiotic Vesicle Formation and the Necessity of Salts. *Origins of Life and Evolution of Biospheres* **2015**, *46*, 215–222.

(36) Buboltz, J. T.; Feigenson, G. W. Phospholipid Solubility Determined by Equilibrium Distribution between Surface and Bulk Phases. *Langmuir* **2005**, *21*, 6296–6301.

(37) Monnard, P.-A.; Apel, C. L.; Kanavarioti, A.; Deamer, D. W. Influence of Ionic Inorganic Solutes on Self-Assembly and Polymerization Processes Related to Early Forms of Life: Implications for a Prebiotic Aqueous Medium. *Astrobiology* **2002**, *2*, 139–152.

(38) Maurer, S.; Deamer, D.; Boncella, J.; Monnard, P.-A. Chemical Evolution of Amphiphiles: Glycerol Monoacyl Derivatives Stabilize Plausible Prebiotic Membranes. *Astrobiology* **2009**, *9*, 979–987.

(39) Pulletikurti, S.; Veena, K. S.; Yadav, M.; Deniz, A. A.; Krishnamurthy, R. Experimentally modeling the emergence of prebiotically plausible phospholipid vesicles. *Chem* **2024**, *10*, 1839–1867.

(40) Jordan, S. F.; Nee, E.; Lane, N. Isoprenoids enhance the stability of fatty acid membranes at the emergence of life potentially leading to an early lipid divide. *Interface Focus* **2019**, *9*, 20190067.

(41) Albertsen, A.; Duffy, C.; Sutherland, J.; Monnard, P.-A. Self-Assembly of Phosphate Amphiphiles in Mixtures of Prebiotically Plausible Surfactants. *Astrobiology* **2014**, *14*, 462–472.

(42) Apel, C. L.; Deamer, D. W.; Mautner, M. N. Self-assembled vesicles of monocarboxylic acids and alcohols: conditions for stability and for the encapsulation of biopolymers. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2002**, *1559*, 1–9.

(43) Toparlak, O. D.; Mansy, S. S. Progress in synthesizing protocells. *Experimental Biology and Medicine* **2018**, *244*, 304–313.

(44) Voet, D.; Voet, J. *Biochemistry*; Wiley, 2010.

(45) Namani, T.; Walde, P. From Decanoate Micelles to Decanoic Acid/Dodecylbenzenesulfonate Vesicles. *Langmuir* **2005**, *21*, 6210–6219.

(46) Sabanza-Gil, V.; Barbano, R.; Gutiérrez, D. P.; Luterbacher, J. S.; Hernández-Lobato, J. M.; Schwaller, P.; Roch, L. Best Practices for Multi-Fidelity Bayesian Optimization in Materials and Molecular Research. 2024; https://arxiv.org/abs/2410.00544.

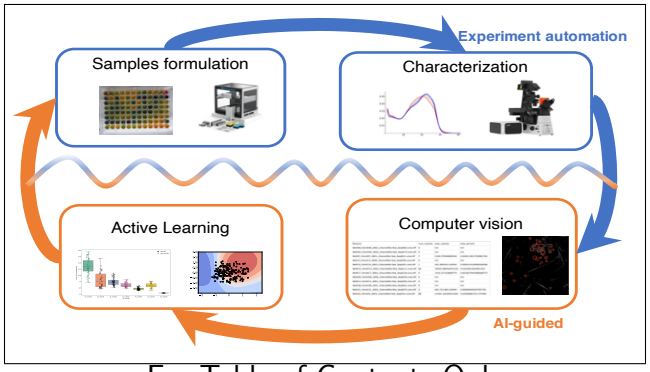
(47) Yano, J.; Gaffney, K. J.; Gregoire, J.; Hung, L.; Ourmazd, A.; Schrier, J.; Sethian, J. A.; Toma, F. M. The case for data science in experimental chemistry: examples and recommendations. *Nature Reviews Chemistry* **2022**, *6*, 357–370.

(48) Ruopp, M. D.; Perkins, N. J.; Whitcomb, B. W.; Schisterman, E. F. Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection. *Biometrical Journal* **2008**, *50*, 419–430.

(49) Todd, Z. R.; Cohen, Z. R.; Catling, D. C.; Keller, S. L.; Black, R. A. Growth of Prebiotically Plausible Fatty Acid Vesicles Proceeds in the Presence of Prebiotic Amino Acids, Dipeptides, Sugars, and Nucleic Acid Components. *Langmuir* **2022**, *38*, 15106–15112.

(50) Qin, S.; Jin, T.; Van Lehn, R. C.; Zavala, V. M. Predicting Critical Micelle Concentrations for Surfactants Using Graph Convolutional Neural Networks. *The Journal of Physical Chemistry B* **2021**, *125*, 10610–10620, PMID: 34498887.

TOC Graphic



For Table of Contents Only