Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2023, **2**, 298

Machine learning for soft and liquid molecular materials

Tetiana Orlova, [a] Anastasiia Piven, [a] Darina Darmoroz, [a] Timur Aliev, [a] Tamer Mahmoud Tamer Abdel Razik, [a] Anton Boitsev, [b] Natalia Grafeeva [b] and Ekaterina Skorb *[a]

This review discusses three types of soft matter and liquid molecular materials, namely hydrogels, liquid crystals and gas bubbles in liquids, which are explored with an emergent machine learning approach. We summarize specific examples of the use of machine learning technique to study the structure and properties of soft matter at the molecular, microscopic and macroscopic levels. The approaches of artificial intelligence have greatly improved the prediction of material properties, stimulated the progress in modeling methodologies capable of revealing physical phenomena, and opened up new perspectives in the design and use of soft material devices. For this reason we also provide guidance on machine learning methods and recommendations on best practices for data understanding.

Received 27th November 2022
Accepted 15th February 2023

DOI: 10.1039/d2dd00132b

rsc.li/digitaldiscovery

1 Machine learning methods: general introduction

Although ML technology appeared in the middle of the previous century, real opportunities for the practical application of the developed algorithms appeared only when the computing power of personal computers changed significantly. This is due to the fact that most of the algorithms developed require a considerable amount of computing resources. Today, ML algorithms are widely used for the tasks of recognizing the plots of drawings, faces, the tonality of texts, annotating texts, checking grammar, spelling, and many other tasks as well. Therefore, spheres of human activity where the road to ML approaches is open are developing at an unprecedented pace and many researchers in various branches of science question themselves how to apply ML methods in their own "sandbox".

The field of soft and liquid molecular materials can also be such a playground (Fig. 1). In this article, we consider the main categories of ML problems and algorithms and their applicability in the field. Among the tasks to be solved by ML methods, several categories are particularly in demand. Let us briefly introduce them to you.

The purpose of dimensionality reduction[1] is as follows. A large number of data features are reduced to fewer for future data visualization or application of other ML algorithms in real time. The anomaly detection[2] is also a very important field. Atypical objects (which are found in any industry) are separated from the usual (standard) ones based on various statistical characteristics of the entire sample or other ML algorithms. The regression[3] is the task of predicting the possible characteristics of new objects based on a sample of existing objects with a set of features, while the classification[4] is the task of assigning objects to predefined groups based on a set of features (there can be 2 or more classes). Finally, the clustering[5] is the task of dividing the analyzed objects into an unknown number of groups based on a set of features, highlighting the structure in the data.

The tasks mentioned above are solved by various methods, but perhaps the most well-known in this field are the following.

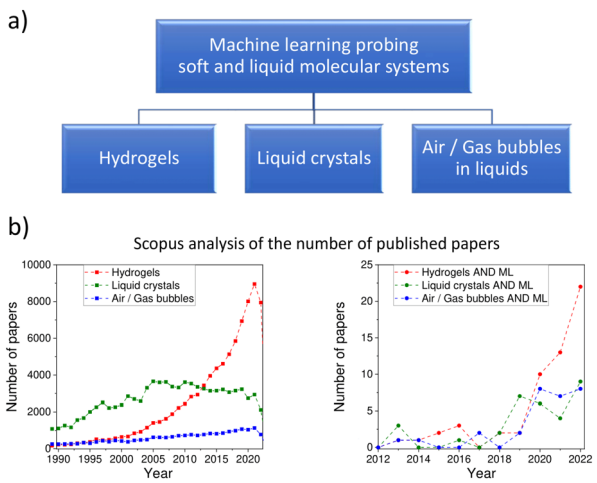

Fig. 1 (a) Soft and liquid molecular systems considered in this review. (b) Scopus analysis of the number of published papers, when the search query is associated with the title, keywords and abstract, and the results are limited to articles and conference papers.

[a]*Infochemistry Scientific Center, ITMO University, 9 Lomonosova Street, Saint-Petersburg 191002, Russia. E-mail: skorb@itmo.ru; Tel: +7 999 210 39 77*
[b]*The Higher School of Digital Culture, ITMO University, 9 Lomonosova Street, Saint-Petersburg 191002, Russia. E-mail: dc@itmo.ru; Tel: +7 921 997 97 91*

The Principal component method[6] (PCA) is basically an orthogonal transformation that allows to translate observations of interrelated variables into a set of principal components or linearly uncorrelated values. PCA is used to provide visualization of the source data, as well as to minimize them and facilitate the learning process itself.

The Nearest neighbor method[7] (k-NN) is a very popular classification method, sometimes used in regression problems. This is one of the most intuitive approaches to classification. The essence of the method is as follows: new objects are added to objects already divided into classes according to the principle of "the nearest neighbor". Hence, the new object falls into the class that is closer to it by attributes. The distance (proximity) between neighbors is determined by various metrics.

Decision trees[8] are a method that implements decision–making based on the use of a tree graph. Such a tree is formed from the minimum possible number of questions with an unambiguous answer (either "yes" or "no") based on existing labeled data. After entering the sequence of answers, the user comes to the right choice. As a rule, the method is used in classification tasks.

The Support Vector Machine[9] (SVM) is an algorithm that is actively used to solve classification problems. The main idea of the algorithm is iterative partitioning of N-dimensional objects by hyperplane of dimension $(N - 1)$.

Clustering algorithms[5] deal with the distribution of analyzed objects into clusters, in which similar elements should appear. Various algorithms based on probability, density, dimension reduction, *etc.* are used for clustering.

Neural networks[10] are an apparatus based on a mathematical model of the interaction of neurons in the human brain. To apply the model, the network is pre-trained on the basis of existing "marked up" data. Currently, the neural networks approach is one of the most popular branches of machine learning. Unlike classical ML algorithms, their advantage is that complex and important work on the formation of features before applying the algorithm (feature extraction) is alienated from the researcher and left to the algorithm.

In the next section of the article, we focus on the machine learning methods already applied in liquid molecular materials tasks and to what extent this application is justified.

2 Machine learning for hydrogels

2.1 Introduction in hydrogels

Due to their unique features like swelling and deswelling and stimuli-responsiveness, polymer hydrogels have attracted much attention as outstanding – soft and wet – materials (*e.g.*, temperature, pH, ionic strength, or chemical reactions).[11] On the other hand, traditional single-network hydrogels are too soft and fragile, with low fracture energies to maintain their strong resistance to crack propagation.[12] Hydrogel as a material can be classified according to different criteria. The preparation method is one of the earliest hydrogels that can be divided according to enveloped polymer sources into hydrogel, semi-synthetic, and synthetic hydrogels. Most natural-based hydrogels are biodegradable, whereas the synthetic-based ones can be nondegradable and have a long life (or be undegradable). Hydrogels can be classified as anionic, non-anionic, or neutral according to the ionic charges on the polymer and fragments. Classify hydrogels using homopolymers, copolymers, or interpenetrated polymers. The amphoteric change of hydrogel charges with different environmental conditions could give the hydrogel intelligent properties.[13]

Recently, there has been increased interest in hydrogel because of its unique properties that enable it to spread in several applications. Additionally, numerous books and articles on hydrogel materials have been published over the past few decades and offer more comprehensive and varied perspectives. There has been a fast growth of novel hydrogel materials and correlated research between 2000 and 2021, which has not been reviewed frequently. Fig. 1b shows a recent number of publications, including hydrogel and machine learning.

2.2 Models for functional hydrogel

Regardless of the materials class or performance criteria, hydrogel science research aims to build process–structure–property–performance correlations.[14] These relationships between a material's manufacturing process, its micro-or nanoscale structure, and its properties and performance in a given application frequently involve complex cause and effect interactions, requiring a sizeable parametric space and a variety of synthetic characterization and theoretical techniques. Materials scientists have already identified three interacting paradigms to generate those types of relationships: theory, experiment, and simulations. While tests directly measure and quantify material performance, the theory provides a mathematical framework for understanding and predicting correlations. The most recent of the three paradigms, computational simulations, uses theory to do granular hydrogel simulations of experiments to provide more insight into difficult or impossible phenomena to detect experimentally. If an experiment doesn't support a theoretical relationship, it isn't complete, and simulation approaches are often better at describing what happened in the experiment.[14]

The significantly increased ability of materials scientists to collect, share, and analyze large volumes of data in recent decades has led to what many are calling the fourth paradigm of materials science, also known as data-driven materials discovery or materials informatics (MI).[15,16]

MI is a discipline in which correlation relationships can be suggested or validated by examining massive data sets of materials using statistical techniques, many of which use machine learning. An informatics strategy uses these enormous data sets to alter more standard correlation relationship development methods. For example, a more traditional technique might utilize a theoretical model based on our understanding of chemistry and physics to predict a material's attributes based on its structure and composition. Experimental measurements of the structures and expected properties can then be used to validate or alter the theoretical model. Instead, an informatics method develops a model using data from the inputs (structure) and the measured responses (properties).

This model can then predict responses to similar or only slightly different inputs, and cross-validation can test and improve its ability to do this.

A wide range of technological applications are enabled by functional hydrogels, which are made up of crosslinked 3D macromolecules and molecular building blocks that self-organize into complex structures as a result of their adjustable connections. Inverse approaches allow designers to navigate their intrinsically high-dimensional design areas in order to generate materials with specific qualities. While a number of physically driven inverse techniques have been effectively applied in some circumstances, their application to directing experimental materials discovery has been confined to a few proof-of-concept investigations thus far.[17] We highlight recent improvements in inverse methods for hydrogel design that address two issues: (1) methodological limits that prevent such approaches from satisfying design requirements and (2) computing challenges that limit the pore size and network structure that may be addressed. Methods to identify order parameters that characterize complicated structural motifs, as well as approaches to effectively compute macroscopic features from the underlying structure, have proven to be particularly effective. We also talk about promising ways to improve the accuracy and computational efficiency of models that are relevant to experiments, such as finding materials that work in more than one thermodynamic state, making protocols for externally directed assembly that are easy to use in experiments, and coming up with other ways to improve the accuracy and computational efficiency of models that are relevant to experiments.

Drug delivery systems,[18] wound dressing membranes,[19] cell culture,[20] tissue engineering[21] and other critical applications have all benefited from tailored hydrogels. The use-inspired behaviors of these materials are caused by the physicochemical properties of their constituent components and their internal spatial organization (*i.e.*, structure). Synthetic polymers, polysaccharides, and proteins can serve as powerful material building blocks for hydrogel formation because their mutual interactions, which help determine the system's favored equilibrium state, can be systematically varied through, for example, their size, shape, charge, composition/sequence, and surface functionalization.

This opens up many design possibilities. It is a big job to determine which building components may consistently self-assemble a material with a given structure or desired macroscopic features. Innovative approaches to the search for new self-assembling materials are routinely used. In such approaches, an initial set of material building blocks is synthesized, and strategies are developed to make self-assembly easier in an experiment or a computer simulation. The structure and qualities of the final material are next investigated. These processes are repeated (typically numerous times) to hunt for materials with superior attributes using other building blocks or protocols. Instead of framing this process as an inverse problem, framing it as a methodology and amenable to meeting stated design objectives can be advantageous. For example, a figure of merit (FOM) can be set up based on the desired structure or macroscopic property, and constrained optimization methods can move through the multidimensional design space and determine which building blocks, interactions, or protocols are best for making material.

Kalasin and his coworkers developed The Remote Artificial Intelligence-Assisted Epidermal Wearable Sensing for Environmental Heat-Stress Sweat Creatinine Monitoring based on Satellite-Based Sensor. His group develop coated nylon with poly(vinyl alcohol) (PVA)-Cu^{2+}-poly(3,4-ethylenedioxythiophene) polystyrene sulfonate (PEDOT:PSS) and cuprous oxide nanoparticles. The system was equipped with heart rate monitoring and a satellite communication device to locate wearers, and incorporates machine learning to predict the levels of environmental heat stress. Electrochemical impedance spectroscopy (EIS) was used to investigate different charge-transfer resistances of PVA and PEDOT:PSS with cuprous and cuprite ions induced by single-chain and ionic cross-linking.[22]

To improve the performance of ultrasound-mediated chemical sensing using titanium dioxide (TiO_2) nanoparticles-embedded hydrogel, Islam and his colleagues used machine learning. A new wireless pH sensing technology has been developed using ultrasound transmission through titanium dioxide (TiO_2) nanoparticle-embedded hydrogels. An ultrasonic transmitter/receiver pair and a TiO_2-embedded hydrogel implanted under the skin make up the sensing system (or any other body area that requires pH monitoring). Filling hydrogel with TiO_2 nanoparticles improves ultrasonic wave backscattering, obviating the need for complex readout devices like ultrasound imaging. The physical behavior of ultrasonic waves changes as they flow through the hydrogel, depending on the thickness of the hydrogel. The ultrasonic receiver, located on the other side of the body, captures the changed ultrasonic wave caused by the hydrogel. Ultrasound activities were used to investigate the volumetric transition of hydrogel wirelessly. The delivered ultrasonic signals are gathered with feature extraction for machine learning applications in mind. This approach uses ultrasonic waves to measure pH information with low reflection and noise effects. Despite having many scopes, no machine learning-enabled direct pH measurement systems have been published previously. The measurement errors of today's state-of-the-art pH sensors (with analyte) are within 0.1 to 0.01 pH. So, this research aims to find ways to use machine learning to make the error rate even lower.[23]

For 3D printed bioink, Lee and his colleagues use a machine learning-based design technique. They used a model system using naturally produced biomaterials to build a machine learning-based method for designing a 3D-printable bioink. First, they showed that, compared to native collagen, atelocollagen (AC) had better physical qualities for printing (NC). NC gel had highly crosslinked and temperature-responsive irreversible behavior, resulting in brittleness and high yield stress. In contrast, AC gel had weakly elastic and temperature-responsive reversible activity, generating a soft cream-like structure with low yield stress. Then, using machine learning, they identified a universal relationship between the mechanical parameters of ink and printability: a high elastic modulus increases shape fidelity, and extrusion is possible below the

threshold yield stress. Based on this relationship, they were able to use multiple regression analysis to make a lot of different formulations of bio-inks made from natural materials with great shape fidelity. Finally, using a framework of high form fidelity bioink, a 3D model of a cell-laden hydrogel was created, revealing that the cells are incredibly viable and proliferative in the 3D structures.[24]

Liu and his colleagues created a system based on photography and machine learning, which they used to hydrogel pressure distribution sensors. The team proposed and built a hydrogel pressure distribution sensor that can monitor pressure distribution across the entire hydrogel component. This is performed *via* a technique called electrical impedance tomography (EIT), which involves putting electrodes just around the hydrogel (EIT) (Fig. 2). Meanwhile, PAAm/PAA-Fe^{3+} double-network hydrogels were developed as hydrogel pressure-sensitive substrates, and mechanical and electrical tests confirmed their suitability as sensitive elements for hydrogel pressure distribution sensors. A machine learning method based on the hydrogel pressure distribution sensor was also used to create a pressure distribution reconstruction model. Finally, the hydrogel pressure distribution sensor was used to test the viability of the EIT strategy-based hydrogel pressure distribution sensor by applying forces of known position and magnitude. The real sensor data was then reassembled and compared to the applied force.[25]

Li and his co-authors use chemical characteristics to examine the design of self-assembly dipeptide hydrogels and machine learning (Fig. 3). They built a peptide-like chemical library for screening chemicals that can produce hydrogels based on a Ugi four-component reaction. A rheometer and transmission electron microscopy (TEM) evaluated selected hydrogels, which were then grown with an adherent cell line. The machine learning method was designed to recognize these chemical properties and forecast whether a chemical structure may form a hydrogel at neutral pH without any divalent or trivalent metal ions. In addition, the molecular structure and gelation property connection was summarized.[26]

The researchers examined several polysaccharide hydrogels to find functional properties that could predict antibiotic accumulation in Gram-negative bacteria. A model composed of starch hydrogel was evaluated in the Gram-negative model bacterium *E. coli* and shown to be exceptionally capable of discriminating high from low-accumulating antibiotics. The rapid penetration of porin-dependent antibiotics in the starch gel matches Gram-negative specific porin-mediated absorption. However, findings on nalidixic acid permeation and structure-permeability connections suggest that this approach can also discover high-accumulating medicines with good porin-independent outer membrane penetration. Whether manually pipetted or printed, model preparation is simple, reproducible, cost-effective, and risk-free. Membrane-permeation tests can be

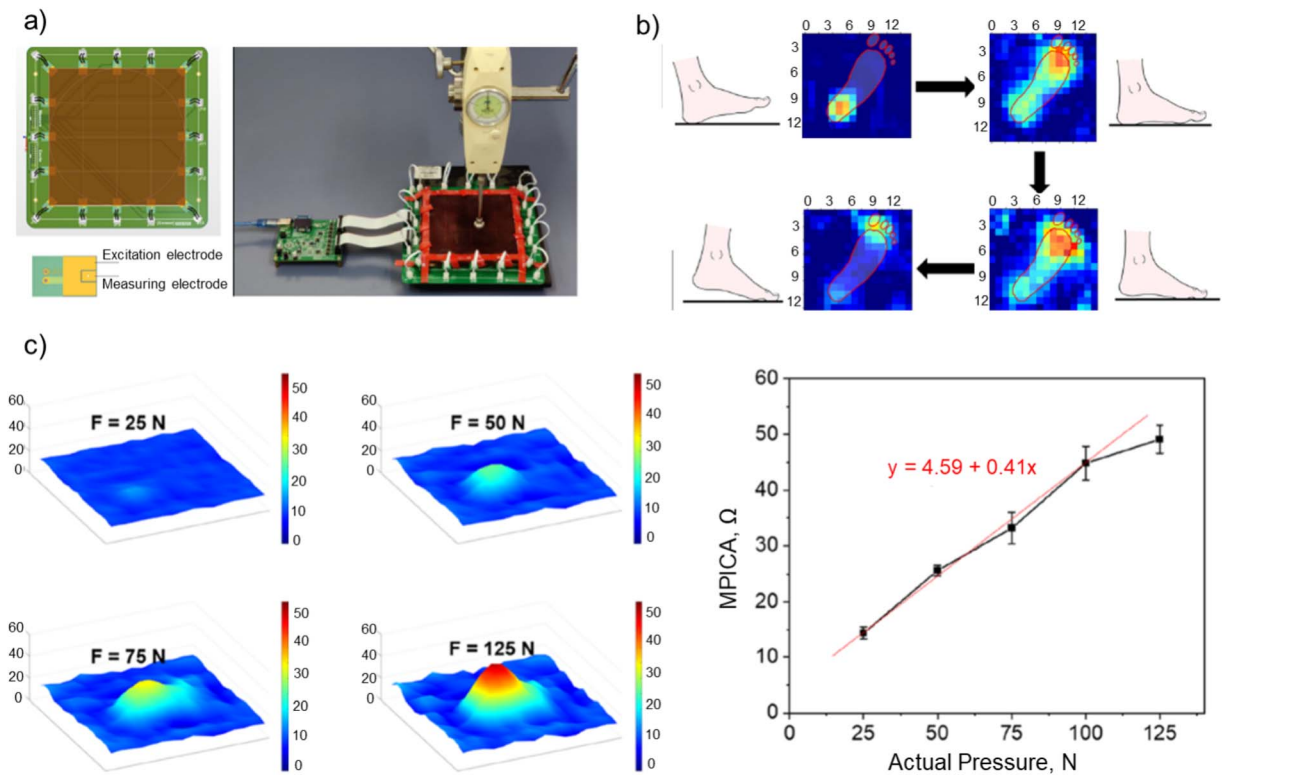

Fig. 2 (a) Hydrogel pressure distribution sensor consisting of two parts, an electrode array board where the hydrogel can be placed, and a driver board, tested with a force gauge. (b) Predicted maps of plantar pressure distribution during the simulated walking process, starting with heel under pressure, then with whole foot under pressure, next with main force on the forefoot, and finally with big toe under pressure while the foot is lifted. (c) The maximum predicted impedance value of the compressed area (MPICA) shows a linear correlation with the actual pressure applied to the sensor. All figures are adapted with permission from ref. 25.

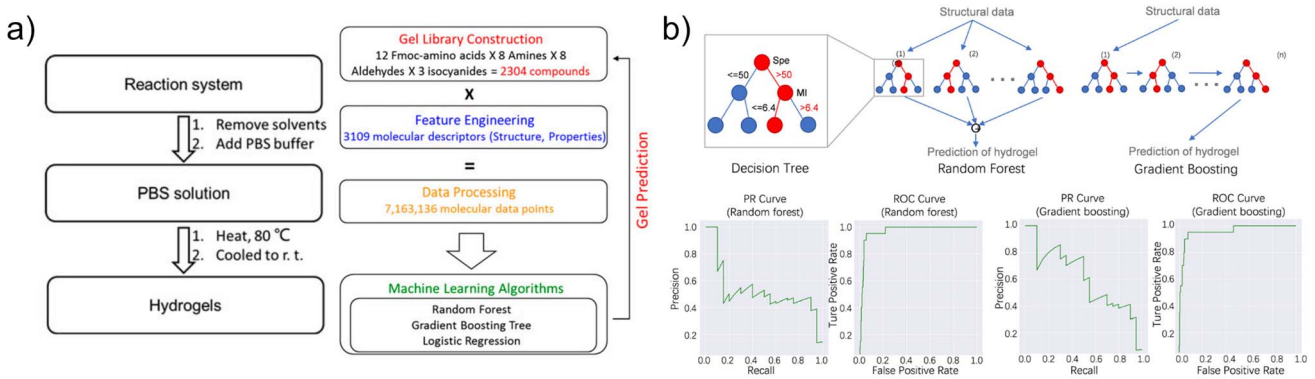

Fig. 3 (a) A typical method of peptide-based hydrogel preparation and the question of hydrogel formation as a binary classification problem. (b) The precision–recall (PR) and receiver operating characteristic (ROC) curves obtained for the random forest and gradient boosting models of hydrogel formation. All figures reprinted with permission from ref. 26.

done automatically and give accurate results in as little as 10 minutes. This makes them great for high-throughput screening of compounds with different physical and chemical properties.

Modern machine learning approaches to *in vitro* data offer evidence of the influence of previously revealed molecular characteristics in bacteria. Based on *in vitro* permeation data, it was determined that a small set of seven features was all that was needed to build a reliable machine learning model that predicted well. In testing the ability of different medicines to pass through cells in a lab, the first evidence of bacterial accumulation of aminoglycosides and sulfonamides, which are essential antibiotics for treating Gram-negative infections, was found. By optimizing the composition of the alginate formulation or using biological hydrogels, the experiment could be changed to study the permeability of biofilms, exopolysaccharides, or mucus on a larger scale.[27]

2.3 Machine learning approaches

Energy sources that integrate with a person have certain requirements: extensibility, softness and flexibility. The triboelectric nanogenerator (TENG) developed by Fan *et al.*[28] based on catechol-chitosan-diatom hydrogel allows these requirements to be taken into account. Based on the data obtained from the TENG based on the tremor of the hands of a sick and a healthy person, the KNN and SVM models were trained, of which the linear SVM showed an accuracy of 100%. Hydrogel-based TENG data show that hydrogels and artificial intelligence can be combined into smart electronics, which over time will help track the condition of athletes, as well as patients with certain diseases in real time[29] (Fig. 4a).

Hydrogels are being used in conjunction with machine learning for biological stem cell research. A group of scientists from Rutgers University created a system to analyze the high content of the true three-dimensional organization of SC-35 cells with machine learning approaches to classify the resulting cell states when cells are cultured in three-dimensional scaffolds. This system makes it possible to study cells without destroying them, since PCR, flow cytometry and immunoassays are time consuming and require cells to be extracted from the system (differentiation, apoptosis, transformation). Decision tree models for three types of cells were trained in the WEKA software package.[30] Each of the models showed an accuracy above 70% with an average error of no more than 0.4 (ref. 31) (Fig. 4b).

In an article by a research group from the ITMO University's ISC Infochemistry Scientific Center, hydrogels were used to collect a database using cyclic voltammetry. The obtained data was used to create a random forest model in the WEKA software package.[30] Four different hydrogels were used with different combinations of encephalitis antibody and antigen. The accuracy of the model was 93%[32] (Fig. 4c).

Hierarchical machine learning with small datasets has made it possible to create a pipeline that allows to analyze and predict the parameters required for the 3D printing method of soft hydrogel molds – freeform reversible embedding of suspended hydrogel (FRESH).[33] During the study, 48 seals were analyzed. They were analyzed based on a statistical comparison of CAD files and, obtained from them, real samples. LASSO regression with the parametrization of the middle layer to the upper one showed good values of the coefficient of determination $R^2 = 0.643$ [34] (Fig. 5a).

Hydrogels are used in supramolecular chemistry to study and analyze the process of release of supramolecular assemblies. The support vector machine was used to predict the binding energies of supramolecular assemblies with various molecules. Supervised learning was applied to create classification and regression models. In the training dataset, the data obtained by the DFT method were used, such as: geometry from optimized orientation, eigenvalues condensed to atoms, all electrons, condensed to atoms, geometry accompanying electronic data, electrostatic properties, electric field gradient, gradient eigenvalues. Also in the dataset, environmental data was used, such as: temperature, buffer and salt concentration and pH. In the course of the study, it was possible to train models that showed their effectiveness by correctly predicting the binding energy of cucurbit[7]uril with promising drugs against low-grade gliomas in children: the RAF type II inhibitor TAK-580 (ref. 35) and the MEK inhibitor selumetinib.[36] The

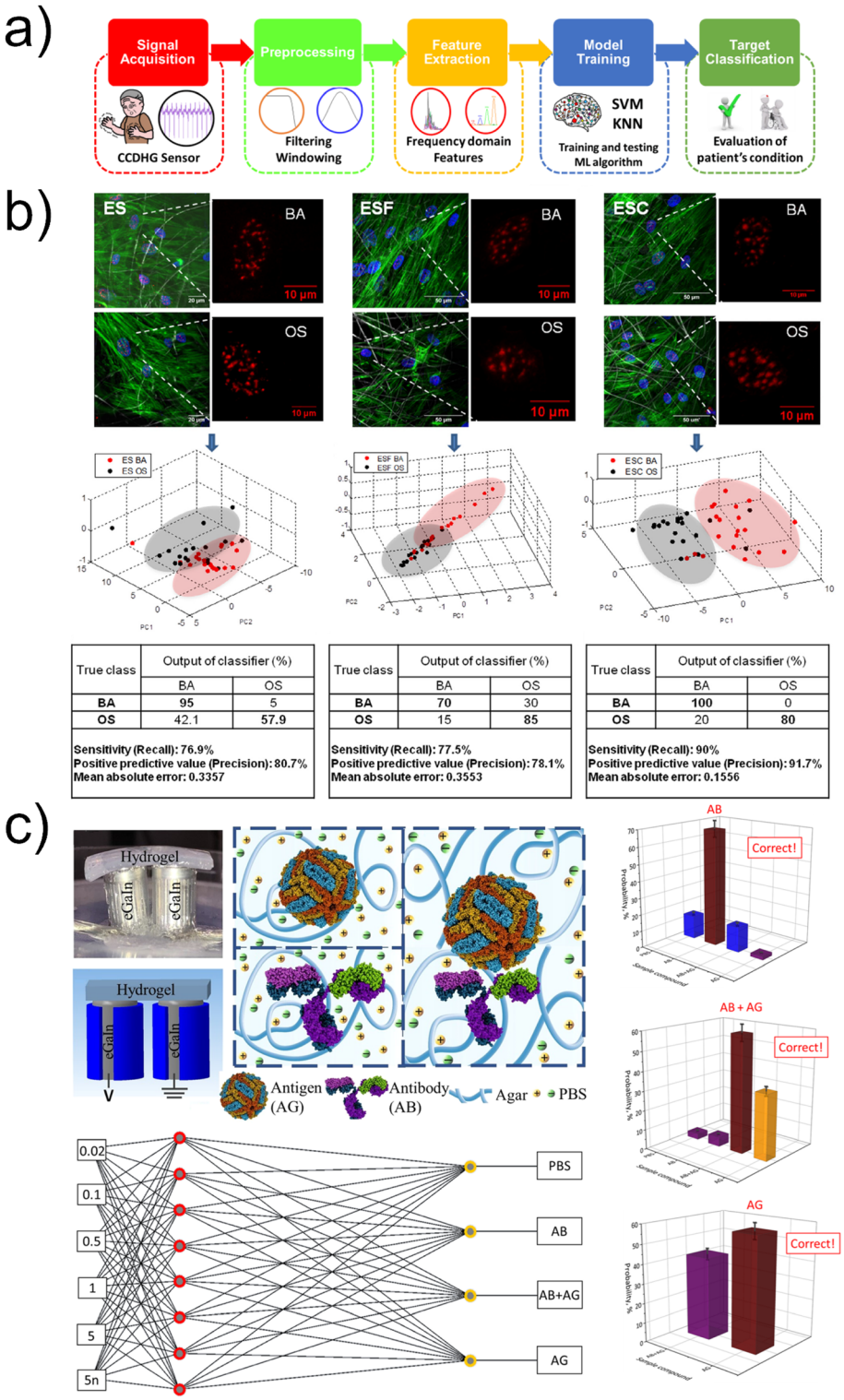

Fig. 4 (a) Catechol-chitosan-diatom hydrogel as triboelectric nanogenerators for collecting energy from human movements to monitor the health of a person with Parkinson's disease using machine learning methods. Reprinted with permission from ref. 29. (b) Predictive cell-state classification model based on computed quantitative 3D nuclear metrics for splicing factor SC-35. Reprinted with permission from ref. 31. (c) Using the random forest algorithm to determine the presence of encephalitis antigen in an electrochemical system. Reprinted with permission from ref. 32.

models predicted that RAF would have a high binding energy and MEK would have a low binding energy, which was confirmed experimentally. RAF and MEK were incorporated into a hydrogel based on cucurbit[7]uril with the same release kinetics of both guest structures for the needs local drug delivery[37] (Fig. 5b).

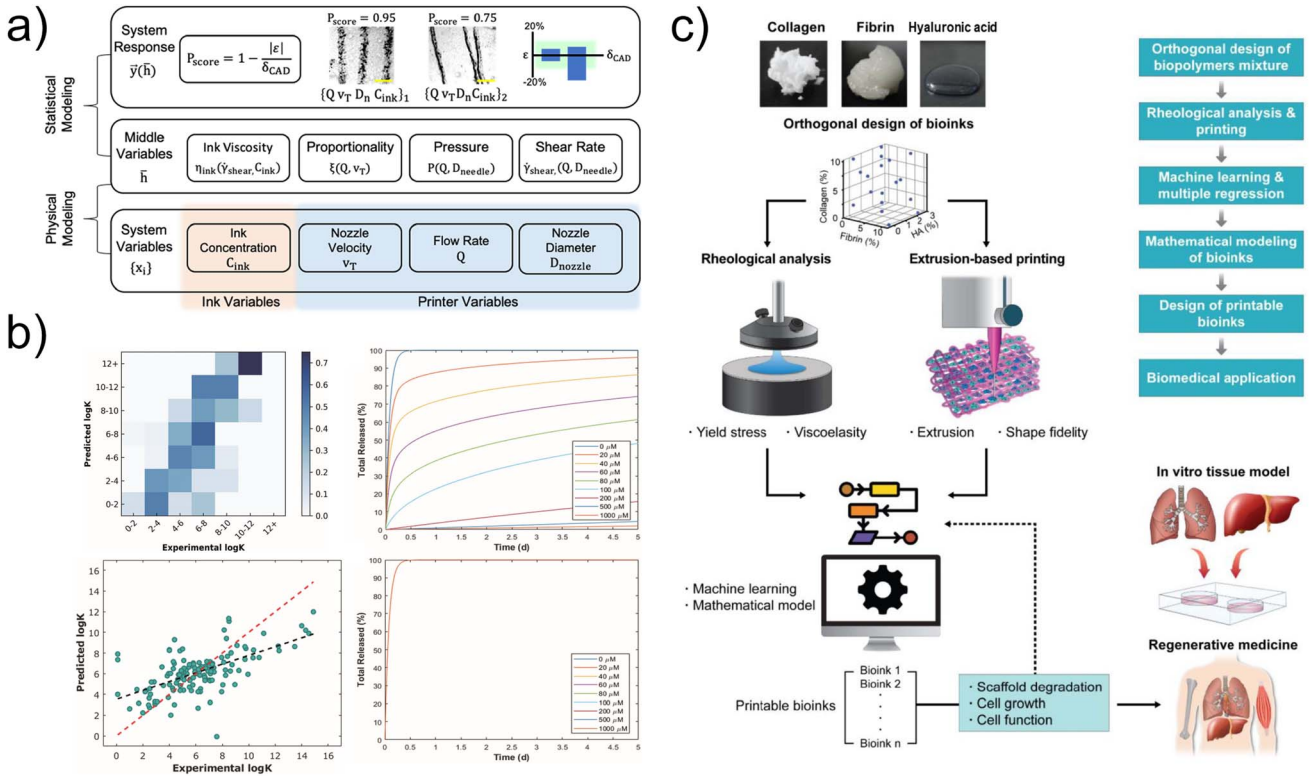

Fig. 5 (a) Hierarchical machine learning to predict the best parameters for 3D printing of elements from alginate hydrogel. Reprinted with permission from ref. 34. (b) Use of support vector machine for binding energy prediction in supramolecular chemistry with hydrogel depot release. Reprinted from ref. 37 with permission from the Royal Society of Chemistry. (c) The use of hydrogels based on collagen, fibrils and hyaluronic acid for the development of 3D biofilms using machine learning methods. Reprinted with permission from ref. 38.

Hydrogels based on collagen, fibril and hyaluronic acid are considered for 3D printable bioinks. The correlation between rheological properties and printability was analyzed using the relative least general generalization algorithm. The input parameters were the concentrations of three initial components of the hydrogel. Two outputs were obtained. The first output is the rheological parameters, and the second output were printing results such as shape fidelity. Combining it with multiple regression analysis, Lee with co-authors showed operating window maps to determine printability of natural hydrogels[38] (Fig. 5c).

Research is also using machine learning to automate the hydrogel manufacturing process. Hydrogels in which bacteria were introduced were used as an experiment. Differential dynamic microscopy was used to determine the gelation rate. For the selection of parameters, Bayesian methods of machine learning were used. A machine learning pipeline was obtained, which independently selected parameters for gelation, and after the experiment, estimated the speed and selected more efficient parameters.[39]

3 Machine learning for liquid crystals

3.1 Prediction of liquid crystal phases based on molecular-level description

For all known liquid crystal (LC) phases, thermotropic, lyotropic and metallotropic, the mesophase formation depends on the molecular architecture of a substance, which includes chemical structure, topology, polarity and polarizability.[40] The phase transition to the LC state is driven by macroscopic physical parameters such as temperature, concentration, pressure, and the ratio of organic and inorganic components for metallotropic LCs.[40] Despite the general understanding about the chemical structures typical of mesogenic compounds, it remains difficult to predict the formation of an LC state for a particular chemical substance. Thus, early attempts to use a machine learning approach for liquid crystals were focused on predicting the mesogenic properties of mono-component materials.

Various classification algorithms, such as multi-linear regression analysis and neural networks, were tested aiming to predict liquid-crystalline property from molecular structures of individual chemical compounds encoded by different sets of numerical descriptors along with the clearing temperatures.[41–44] Despite the large number of compounds used for to train neural networks by Kränz with co-authors,[41] the clearing temperatures were predicted by the best with a standard deviation of 13°, which is a rather high error, especially in the case of LC materials with a clearing temperature close to room temperature. Improvements in neural network model descriptors have reduced the RMS error to just a few degrees[42] along with successful encoding of clearing temperature trends into homologous trends.[43] A comparison of three classes of machine learning algorithms such as eager learners, lazy learners, and

neural networks has shown that no heuristic is better than another on the set of all possible problems.[44] All classification algorithms represented good methods for predicting the liquid-crystalline property, but their efficiency depended on a specific problem.

The quantitative structure–property relationship (QSPR) methodology, combined with a specific type of feed-forward artificial neural network, has been applied to predict the liquid crystallinity and phase transition temperature of bent-core molecules.[48] (Fig. 6a) The authors turned to nonlinear QSPR models and for the first time used a group method of data handling type neural network, testing several machine learning models with different sets of molecular structure descriptors. The developed neural network models demonstrate an improvement in determining the clearing temperatures compared to the results obtained in a previous study.[49] using the multivariate adaptive regression splines technique. Key structural features that affect the transition temperature of five-ring bent-core aromatic compounds have also been identified.

Recently, a general methodology was presented by Chen and coworkers[50] to identify appropriate machine learning algorithms and molecular descriptors for predicting a wide variety of liquid crystal behavior of organic compounds based on QSPR. Almost a dozen machine learning algorithms were compared using a dataset from the LiqCryst 5.2 database[51] with 3786 entries, of which 2780 compounds exhibit liquid crystal behavior. The most accurate for predicting the liquid-crystalline behavior was the random forest algorithm, while the molecular descriptor took into account the mesogen and wings of the chemical structure. Other advantages of the classifier included no pre-processing, quick training, simplicity and versatility for different descriptor inputs. This extensive study can serve as the basis for constructing a multipurpose QSPR model for each type of mesogenic molecule, including the prediction of desired LC properties for unknown or not yet synthesized compounds.

After a number of the above-mentioned studies predicting the properties of monodisperse molecular systems, a similar attempt was made for polydisperse systems.[45] (Fig. 6b) A

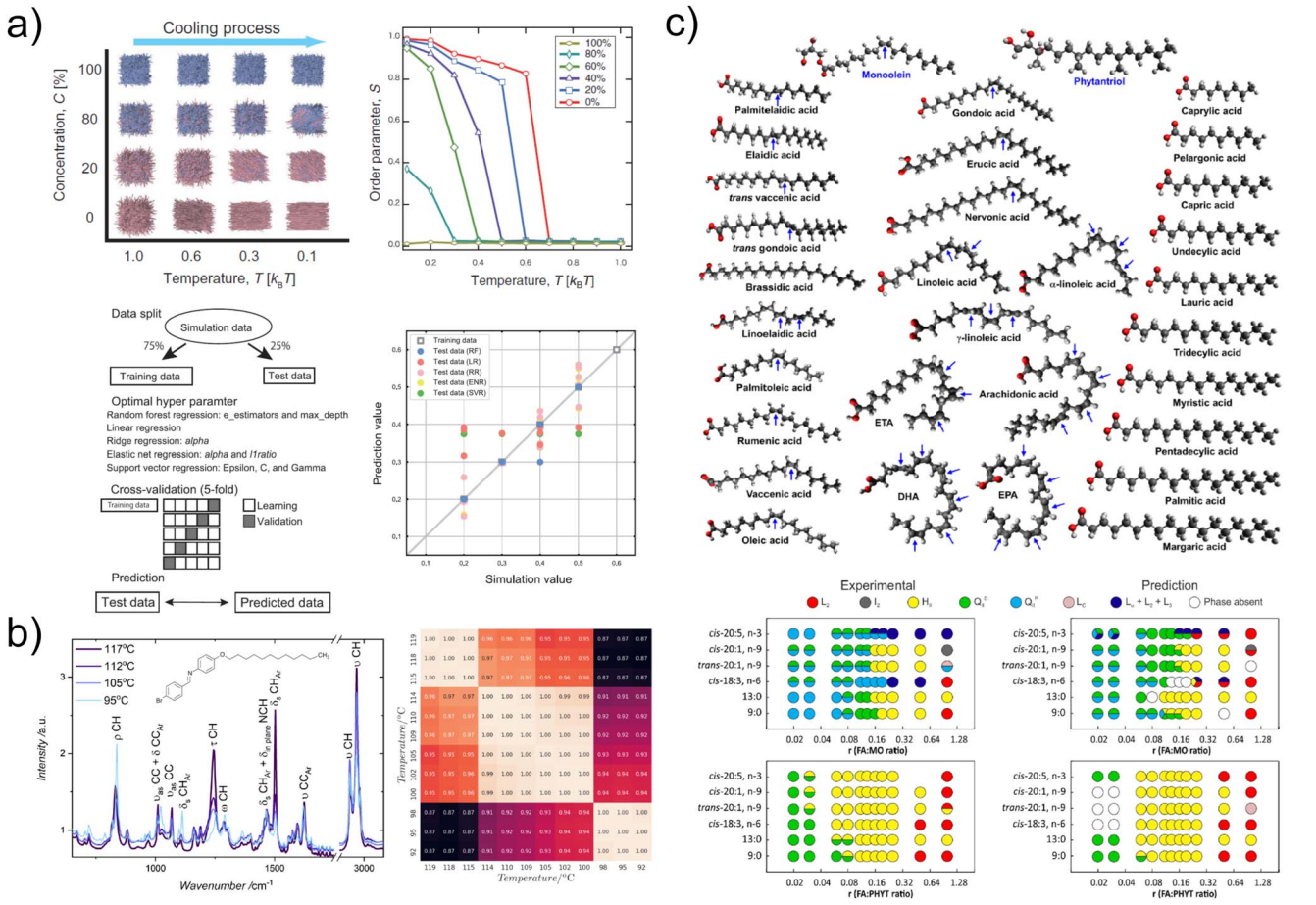

Fig. 6 (a) Phase transition temperature prediction for binary LC systems using machine learning methods (random forest, linear regression, ridge regression, elastic network regression, and support vector regression). Reprinted with permission from ref. 45. (b) Analysis of phase transitions in liquid crystal 12BBAA (4-bromobenzylidene-4′-dodecyloxyaniline) using the k-means cluster analysis of infrared spectra. Reprinted with permission from ref. 46. (c) Chemical structure of the two lipids and the twenty unsaturated and ten saturated fatty acids studied, where blue arrows indicate the location of the double bond in the hydrocarbon chains. Bottom part shows experimentally revealed (left) and predicted (right) phase diagrams for monoolein (top) and phytantriol (bottom) under the effect of different unsaturated fatty acids for samples. The mesophase type is indicated as the inverse bicontinuous diamond cubic (Q_{II}^{D}), inverse bicontinuous primitive cubic (Q_{II}^{P}), inverse hexagonal (H_{II}), micro-emulsion (L_2), micellar cubic (I_2), lamellar crystal (L_c), and a combination of L_α+L_2+L_3 phases. Reprinted with permission from ref. 47.

dissipative particle dynamics simulation method, developed specifically for soft matter and complex liquids,[52] was used to analyse self-assemble structures and phase transitions in the binary LC system. The order parameter and the phase transition temperature were predicted using several different machine learning algorithms, among which the random forest method showed the highest predictive ability. This study is an important step forward since many of technologically significant liquid crystals are multi-component mixtures, for instance, the widely popular thermotropic nematic LC E7 (Merck).[53]

Lipid-based lyotropic LCs are of considerable interest as potential delivery systems for drugs and *in vivo* imaging contrast agents. Therefore, Le and Tran extended machine learning to predict the complex phase behavior of monoolein and phytantriol based lyotropic LC nanomaterials.[54] Robust models were developed for seven different mesophases considering the effects of two types of lipids, 20 unsaturated fatty acids, 10 saturated fatty acids, a range of fatty acid/lipid ratios, and temperature. The phase behavior prediction was obtained with high accuracy using a Bayesian regularized artificial neural network. In addition, the developed models were able to interpolate data for the same fatty acids at temperatures that have not yet been tested, as well as extrapolate data for new lipid nanomaterials, thus elucidating rules that will be useful for the future development of advanced lipid systems for therapeutic delivery. It is worth noting that earlier Le with co-authors reported on the same Bayesian regularized neural network capable of predicting with high accuracy the complex non-stationary behaviour of amphiphilic nanostructured mesophases over time and under the influence of various crystallization screens.[47]

More than 50 new LC phases have been discovered during extensive studies of bent-shaped molecules over the last 20 years.[55] Among them, one of the most fascinating is the twist-bend nematic phase, when the director follows an oblique helicoid at a constant oblique angle to the helix axis. Recently,[56] it was demonstrated that a machine learning protocol can describe the helical trajectories of hard curved spherocylinders[57] that form the twist-bend phase in dynamic Monte Carlo simulations. The pitch and radius of the trajectories of diffusing hard particles are determined by the pitch and conical angle of twist-bend nematic phase, thereby relating the structural and dynamic properties of this complexly ordered LC. Such studies are important not only for fundamental science, but also for the industry of LC-based devices, for instance, optoelectronic elements with switching times that are determined by the diffusion rate.

The machine learning-based analysis of molecular descriptors or molecular dynamics data is not the solely way to predict complex phase behaviour. For example, the cluster analysis method was applied to the temperature-dependent infrared spectra of a mesogenic chemical compound 12BBAA (4-bromobenzylidene-40-dodecyloxyaniline)[46] (Fig. 6c). Changes in the FT-IR spectra are associated with the alkyloxy chain melting phenomena. Thus, phase transitions from an isotropic liquid to smectic A, crystalline smectic B, and a crystalline phase were successfully predicted from spectroscopy data on the characteristics of vibrational bands.

3.2 Prediction of liquid crystal characteristics from macroscopic data

Optical microscopy was the first early method to demonstrate that LCs represent a new state of matter.[58] Observations of the optical textures of liquid crystals can provide a lot of information about the macroscopic structure of LC phases.[59] There are many experimental methods for studying the structure and physical properties of LCs, but often routine inspection by polarized optical microscopy is sufficient to draw conclusions about the ordering and symmetry of the LC phase. Obviously, it is a tempting idea to predict the parameters of a LC phase only from textures images.

The group led by Ribeiro showed that machine learning methods are able to capture many fundamental characteristics of liquid crystals in an equilibrium state directly from optical images of LC textures.[60,62,63] Convolutional neural networks (CNNs) and k-nearest neighbors algorithm trained on simulated optical images of nematics and cholesterics have successfully predicted the LC phase (nematic or isotropic), the order parameter, and the pitch of the cholesteric helix.[60,62] (Fig. 7a) Furthermore, these algorithms demonstrated ability to predict the temperature of the nematic phase, the phase transition and its order, and the chiral dopant concentrations from experimentally obtained optical microscopy images(Fig. 7b). Newly,[63] it has been shown that ordinary networks of only 24 nodes encode enough optical information that, when combined with a simple machine learning method, is enough to identify and classify mesophase transitions with high accuracy, determine concentrations of chiral molecular dopants, and predict sample temperature.

Another useful technique is to apply machine learning technique for estimating the macroscopic parameters of physical models describing complex LC states. This has been demonstrated for active nematic hydrodynamics.[61,64] Neural networks predicted multiple hydrodynamic parameters using only movies of the director field (Fig. 7c) and forecasted the chaotic dynamics of these systems including point defect nucleation, splitting and annihilation.

An approach of analyzing spatiotemporal variations in the parameters of nonequilibrium systems, combined with machine learning methods, could explain the nucleation and behavior of dynamic supramolecular patterns in chiral nematics.[65] As the authors suggested, the continuous rotation of chiral patterns in a photoactive LC under constant illumination with a focused ligh beam is sustained by the reaction-diffusion process of embedded light-driven chiral molecular motors coupled to the long-range director field due by molecular diffusion. Despite this hypothesis and the analysis of the rotation period on the size of chiral pattern, the rotation equation depending on the physicochemical parameters of the system has not been explicitly obtained. Machine learning methods could reveal hidden connections between the macroscopic organization of the director field and molecular

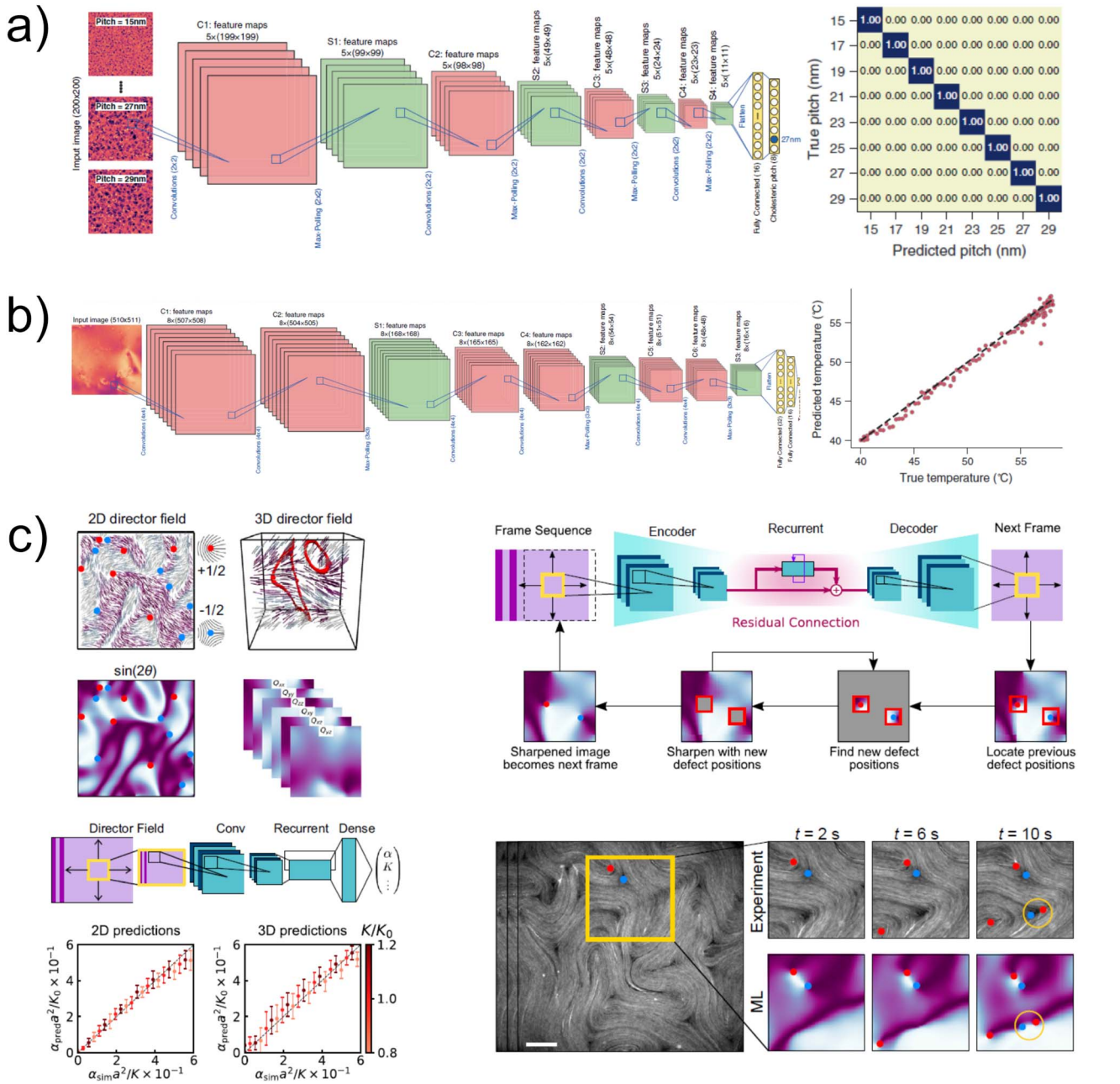

Fig. 7 (a) Prediction of the pitch length from numerically simulated optical images of the cholesteric LC phase with convolutional neural networks. Reprinted with permission from ref. 60. (b) Prediction of the sample temperature from experimental micro-photographs of E7 nematic LC textures. The network architecture is modified by including additional convolutional layers before each max-polling layer for high prediction accuracy. Reprinted with permission from ref. 60. (c) Extracting hydrodynamic parameters from a libraty of simulated LC director fields using a supervised neural network, and demonstrating the capabilities of a neural network as surrogate model of time evolution. Reprinted with permission from ref. 61.

photoisomerization along with diffusion, but this task remains to be explored.

3.3 Reading indicators of liquid crystal sensors

One of the basic features of liquid crystals is their extreme sensitivity to external physical and chemical stimuli, ranging from temperature and mechanical deformations to ultraviolet radiation and chemical agents.[66] External stimuli affect the orientational LC structure, which leads to the transformation and amplification of a physical, chemical, or biological event into an easily detectable optical signal due to the LC optical anisotropy.[67] The two main implementations of LC sensors involve measuring the light transmission intensity by a LC-based sample placed between a pair of polarizers, or the position of the selective reflection band in the case of chiral LCs with a cholesteric pitch comparable to the wavelengths of visible light.

Since all LCs are highly sensitive materials, the problem of analyzing sensor readings becomes especially important for chemical and biosensors, when precise agent detection at minimum concentration is crucial. Machine learning has great potential for filtering a useful optical signal caused by a change in the LC orientation structure under the action of a measured stimulus from random fluctuations in the director field or the influence of side physicochemical factors.[68] In other words, it is necessary to accurately and quickly solve the problem of pattern recognition and classification, for which machine learning was originally developed.[69]

In response to the global COVID-19 pandemic, Xu with co-workers developed a SARS-CoV-2 point-of-care detection kit based on the response of LC films to femtomolar concentrations of single-stranded ribonucleic acid (ssRNA).[70] For this purpose, the LC layer was decorated with a cationic surfactant and a complementary probe of 15-mer single-stranded deoxyribonucleic acid (ssDNA). The minimum concentration of SARS-CoV-2 RNA led to the ordering transition, causing changes in the optical response. Then micrographs were captured by a support vector machine for statistical classification into two categories, positive or negative. Furthermore, in order to obtain a reliable reading of test results for non-expert users, a smartphone application has been developed based on SVM(Fig. 8a).

Roque with co-authors applied CNNs and support vector machines to recognize more than 10 different volatile organic compounds (VOCs) using LC sensors.[71,72] (Fig. 8b) In various multicompartment gel films containing nematic or smectic droplets, variations in optical textures were registered due to orientational transitions of LC molecules in the presence of VOC vapors. Functional and structural differences in the selected VOCs were small. Depending on the LC material, VOC,

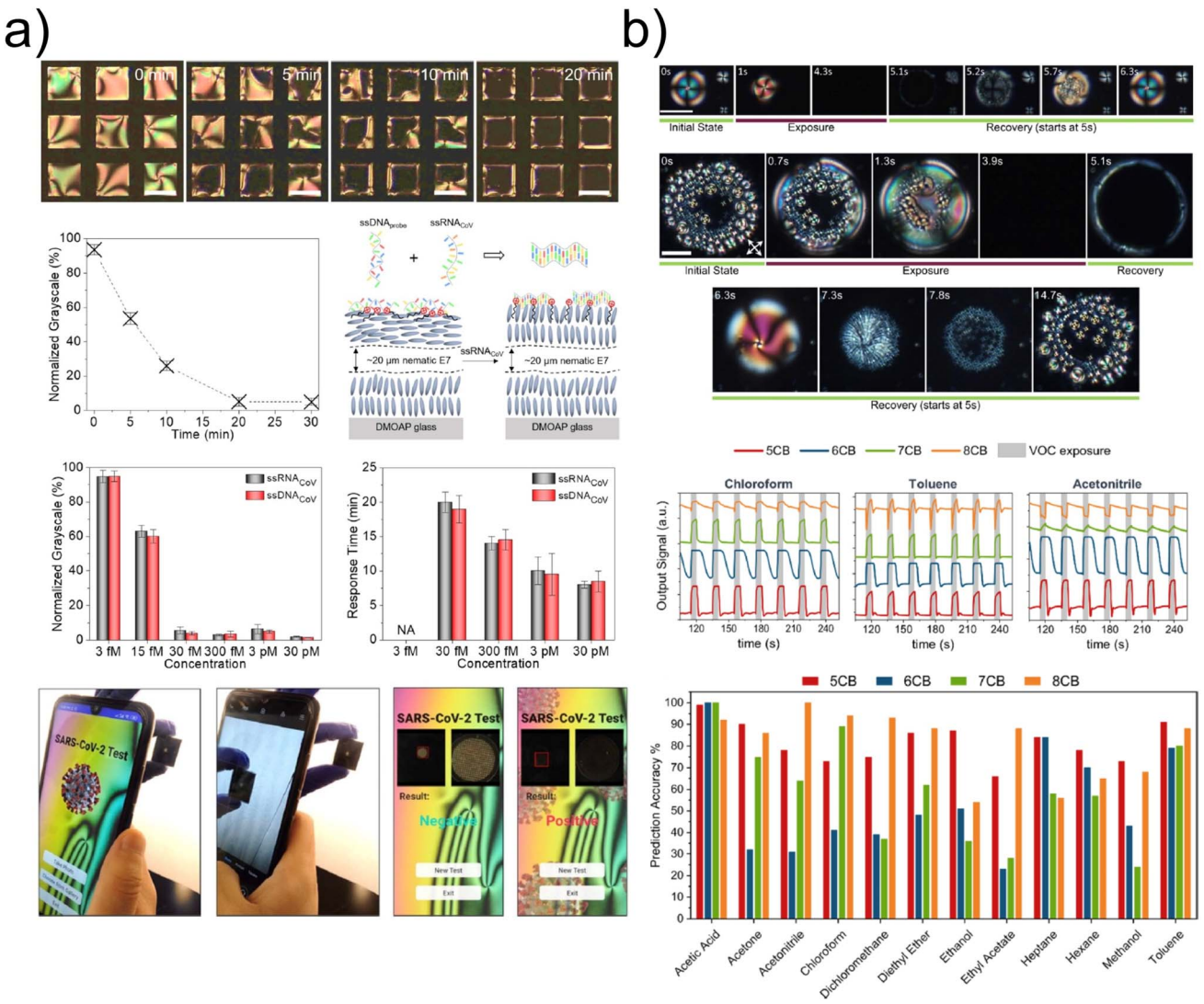

Fig. 8 (a) SARS-CoV-2 point-of-care detector kit for smartphone app with a reliable readout of the test result by machine learning based algorithm when adsorption of SARS-CoV-2 RNA at the water-LC interface results in an optical response of the LC film. Reprinted with permission from ref. 70. (b) Textural changes of nematic LC-based hybrid gels and corresponding signals during VOC vapor exposure and subsequent recovery along with a collective accuracy prediction plot for 12 VOCs. Reprinted with permission from ref. 72.

and droplet diameter, the prediction accuracy ranged from moderate 50–60% to near 100%. The change in optical texture also depended on the concentration of VOCs, for example, the mean absolute error in determining the concentration of acetone was below 0.25%.

The research group led by Abbott and Zavala has made significant contributions to the field of machine learning methods for LC sensors.[73–75] They developed LC-based sensors that exhibited optical responses to N_2-water (30% relative humidity) and N_2-DMMP (10 ppm) gaseous environments.[73,74] AlexNet, the CNN used in their first study,[73] showed a classified accuracy of 99% based on grayscale micrographs. However, such a high level of accuracy required the use of a large number of features, so the more compact VGG16-CNN using color micrographs was tested and demonstrated a perfect classification accuracy while the number of features was reduced to less than 100.[74]

The most original study by the group led by Abbott and Zavala is aimed at identifying bacterial agents and quantifying the concentration of bacterial endotoxins.[75] Instead of taking optical microphotographs, the authors for the first time used flow cytometry with measuring the intensity of side-scattered and forward-scattered light by LC droplets. The ratio of these intensities depended on the droplet orientational structure, the specific changes in which the authors attributed with the molecular structure of the lipid A domain of bacterial endotoxins. Endonet-CNN predicted endotoxin sources directly by the ratio of scattered light intensities and estimated endotoxin concentrations over a wide range of eight orders of magnitude from 1 μg mL^{-1} to 0.01 pg mL^{-1}.

3.4 Quality assessment of liquid crystal displays

The flat panel display is perhaps the most well-known LC-based device. We deal with LC displays (LCDs) in our daily life in laptops, smart watches, mobile phones, instrumental panels, and so on. A typical LC display has a multi-layer arrangement that consists of a uniformly oriented thin layer of LC molecules placed between glass substrates with ITO electrodes, which in turn are sandwiched between a pair of 90-degree crossed polarizers. Depending on whether the display is reflective or transmissive, a reflector or backlight is mounted behind the second polarizer. The electric field applied to the electrodes leads to the reorientation of the LC molecules and change in the intensity of the transmitted polarized light, thus switching the display pixels between the ON and OFF states.[76]

Image formation is affected by each layer of the LCD. ITO electrode defects result in permanent bright dots or black pixels on the screen. The brightness of the backlight affects the sharpness and contrast of the image. Therefore, it is important for manufacturers to evaluate the quality of LCDs directly at the factories in automatic mode without inspection by human eyes and at high speed. Clearly, machine learning could serve as a highly efficient method for solving this task.

The early attempts to apply various ML algorithms to assess the quality of LCDs were made more than 10 years ago aiming to detect and classify defects of TFT-LCD arrays.[77–80] Both assessments are crucial for LCD manufacturing, as identifying the type of defect allows the necessary corrective actions to be taken for its elimination and prevent future failures. In an early study,[77] different types of defects were determined with an accuracy of about 86% to almost 93% using various ML algorithms such as a support vector machine-based classifier and a backpropagation neural network. The data of the analyzed samples were obtained from a real manufacture process. Later, it was shown that modified support vector data descriptions[78,79] provide a high defect detection rate of 90–96% and capable of defect detection on an LCD image within 60 ms.[78] A method that combines K-means clustering with a backpropagation neural network machine learning algorithm outperforms others in the specific task of predicting the heights of thin film transistor-liquid crystal display photo-spacers.[80]

ML methods, primarily neural networks, have also been tested to control the local backlight dimming, which is important for less loss of image detail, higher contrast ratio, and low power consumption of LCDs.[81–83] A comparative study of local backlight dimming prediction accuracy applied to the subjective evaluation of video quality, impacted by ambient light exposure and peak white (maximum display brightness), revealed that Elastic Net algorithm performed best compared to partial least squares regression and support vector regression.[81] The CNN-based algorithm made it possible to control the backlight intensity along with reducing the loss of detail while achieving a high contrast ratio by taking into account the diffusion property of light and leakage property of liquid crystal.[82] However, this ML method required statistical information of pixel values in each local block. The local dimming algorithm based on the U-net convolutional network enabled the compensation of pixel data transferred to a panel directly from an input image, without any information about dimming levels of the backlight unit sub-blocks.[83]

Machine learning can be successfully used to predict the quality of a whole product through each process data.[84] Usually, the overall assessment of product quality in industry is carried out by a sampling method that is not comprehensive and has no timeliness. The random support vector machine method, which combines the support vector machine and random forest, showed the mean square error 0.6 percent lower than that of random forest, despite the fact that the random forest is considered the best traditional machine learning algorithm. A comprehensive three-stage data science framework has also been developed, consisting of variable selection, metrology prediction, and process control.[85] At each stage, different ML methods were used to identify the key factors, for instance, the decision tree, stepwise regression, and random forest. The proposed data science framework, applied to an empirical study of a leading TFT-LCD manufacturer, allowed to determine the variables affecting yield, predict the photo spacer thickness with higher performance than the company's method, and proposed the process control in the color filter manufacturing process. All this helps to reduce the cost for process monitoring and quality in TFT-LCD manufacturing.

Specific tasks, such as dead pixels and mura detection during manufacturing process were also adequately addressed

by ML.[86,87] Typically, such defects are detected by an operator, which is not always performed reliable and increases the cost of production. The support vector machine algorithm proved to be the most effective in detecting dead pixels with an accuracy of about 92% compared to random forest.[86] In contrast, mura (a variation in local brightness with no distinct contour on a uniform LCD surface) was predicted using the random forest algorithm with a detection rate above 99% and a processing time of 27 ms per image, which is a competitive result for industrial systems.[87]

4 Machine learning for bubbles

4.1 Cavitation bubbles' analysis

The cavitation process that occurs in liquid flows is an important characteristic of the system. The constant movement of bubbles complicates their analysis, for example, the identification of bubbles, determining the shape, size and quantity. Machine learning methods are successfully used to simplify bubble analysis.

Parameters such as bubble size and shape statistics generally determine bubbly flow. Using the analysis of cavitation bubbles and machine learning methods, it is possible to predict the structure of the flow. A similar method is demonstrated in a number of articles. Using graph files as inputs, Gao *et al.* classified samples by flow structure *via* convolutional neural networks.[88] Two convolutional neural networks were used to classify 6 flow structures and void fraction (Fig. 9a). The accuracy was more than 92%, the root mean square error was 0.0038 for the flow structure forecast.

Not only graph files can be used as machine learning inputs. Bubble images are a common inputs to cavitation bubbles analysis. Even images of micro-size bubbles can be processed successfully. CNNs can be also used for microbubble analysis. Qaddoori *et al.* performed a similar approach to determine the size of microbubbles using pre-processed images.[89] The images were preliminary converted into HSV format and then processed with CNN (Fig. 9b). According to this approach, the microbubble size was determined both by the number of pixels and luminosity intensity. The method showed 100% accuracy in microbubble size determination. Bubble' image analysis *via* CNNs also enables identification of the flow patterns.[90,91]

High-speed cameras are widely used to capture cavitation bubbles, but the resulting images are often of a low quality. Machine learning can also be used to improve defects of input images. Poletaev *et al.* used deep learning, specifically CNN, to build a model capable of detecting and tracing overlapping, blurry, and non-spherical cavitation bubbles, and demonstrated its further analysis.[92] Various models and training parameters were performed, and the proposed approach was found to be effective, as the accuracy was about 97% on the test dataset. CNNs showed high efficiency in a wide range of studies with defective images, as well as deep neural networks.[93,94] CNNs were also applied to reconstruct the bubble pattern and its further identification.[95] He *et al.* further developed the approach and used pulse-coupled neural networks (PCNN) for bubble segmentation.[96] The proposed scheme turned to be effective also for image enhancement (Fig. 9c). The segmentation accuracy was about 90%.

Deep learning and convolutional neural networks can be successfully replaced with classical machine learning methods

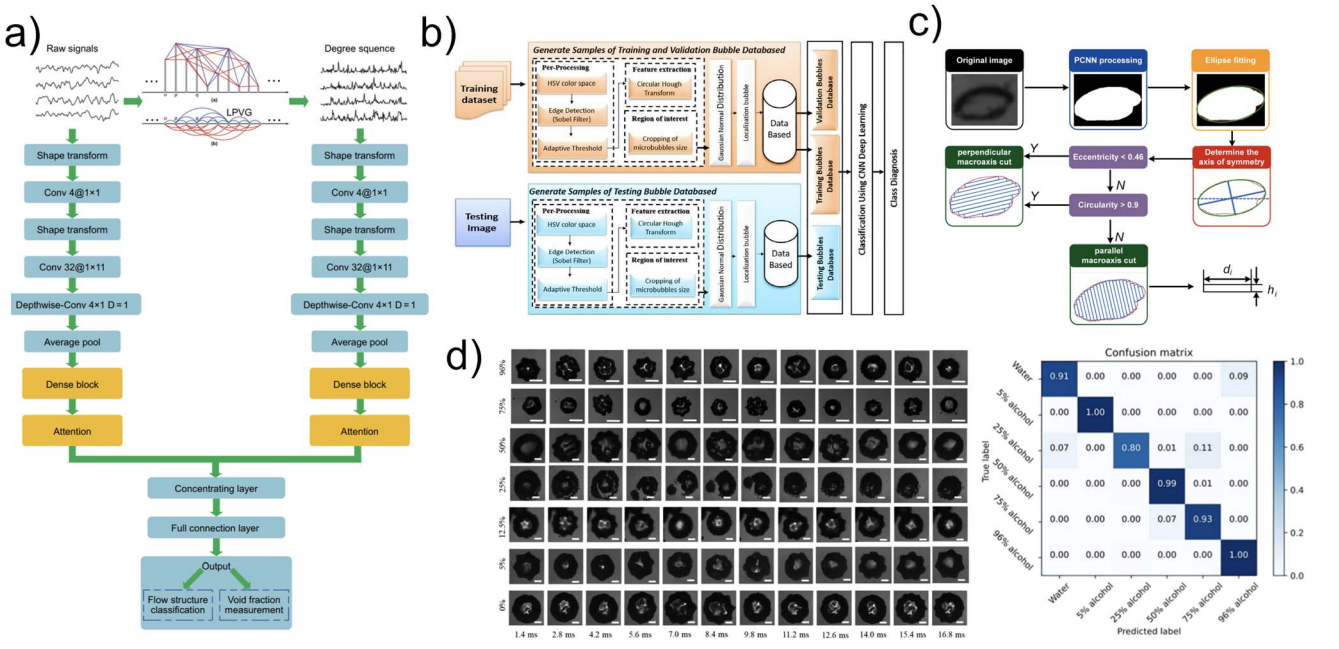

Fig. 9 (a) Convolutional neural network architecture with two inputs for stream structure classification and void fraction measurement. Reprinted with permission from ref. 88. (b) Bubble analysis based on image pre-processing and convolution neural network. Reprinted with permission from ref. 89. (c) Image processing algorithm based on PCNN. Reprinted with permission from ref. 96. (d) Classification of alcohol concentration based on cavitation bubble images and CNN. Reprinted with permission from ref. 103.

such as decision trees, support vector machines, linear regression, *etc.* Srivastava *et al.* demonstrated an approach to determine the size of bubbles depending on environment change. Cavitation bubble data was extracted from images using ImageJ scripts. Various regression models were used based om multilayer perceptron, decision tree, support vector machine. Performed models showed the same high efficiency as artificial neural network (ANN).[97] Classic ML methods were successfully applied to determine the size of bubbles in real column reactors.[98,99]

The discussed approaches require a relatively large dataset of input images. To solve the small dataset problem, Fu and Liu used generative adversarial networks to create a BubGAN model.[100] The performed model can generate image files that can later be used to create other models for cavitation bubble analysis. The algorithm showed high efficiency and image detection. The root mean square error (RMSE) was about 2–3%. This method is effective when it is necessary to create the intended applications of cavitation bubbles in the deficit of the liquid in which they were formed. This approach is proposed for both statistical machine learning methods, and is also applied to identify bubbles with high efficiency.[29]

Generative adversarial network (GAN) was also used for dataset enhancement and bubble image synthesis, while models for bubble segmentation were developed.[101] The combination of image synthesis and the U-Net model provided more accurate bubble segmentation in comparison with previous models.

Cavitation bubbles can interact with each other, which complicates the analysis of bubbles flow analysis and image defects. Bubbles' interaction is hardly described by 2D images. Shao *et al.* performed a 3D reconstruction *via* CNN with U-net architecture.[102] The result was a model that can be used to create a 3D graph with the distribution of bubbles.

Cavitation bubbles can transform not only *via* interactions with each other, but also by bubble growth, decomposition, or collapse. Bubble's shape transformation depends on the liquid content and can be used for quantitative content analysis of a liquid probe. CNN coupled with the transfer learning method allowed to classify samples by fluid content using a series of bubble images in dynamics (Fig. 9d).[103] Pretrained CNNs with frozen layers can be successfully used for efficient model training without requiring a large dataset.

4.2 Modelling of processes in bubbly flows

Machine learning can be combined with classic approaches to model bubbly flows and simulate complex liquid systems. Several complex liquid systems require the integration of machine learning methods into the previously performed simulation approach. The combination of precise numeric methods and neural networks is then turned into an advanced simulation model. Mosavi *et al.* performed this approach to create a model for predicting macroscopic parameters such as gas velocity in a multiphase reactor.[104] The combination of adaptive-network-based fuzzy inference system (ANFIS) and computational fluid dynamics (CFD) provided $R = 0.99$ and a significant reduction in simulation time. The same approach was successfully performed for a column reactor with three input parameters for ANFIS (Fig. 10a). ANFIS is widely used for hydrodynamics simulation in a wide range of real reactors.[105–109]

The bubble size prediction discussed earlier can also be combined with physics, mathematics, and machine learning to

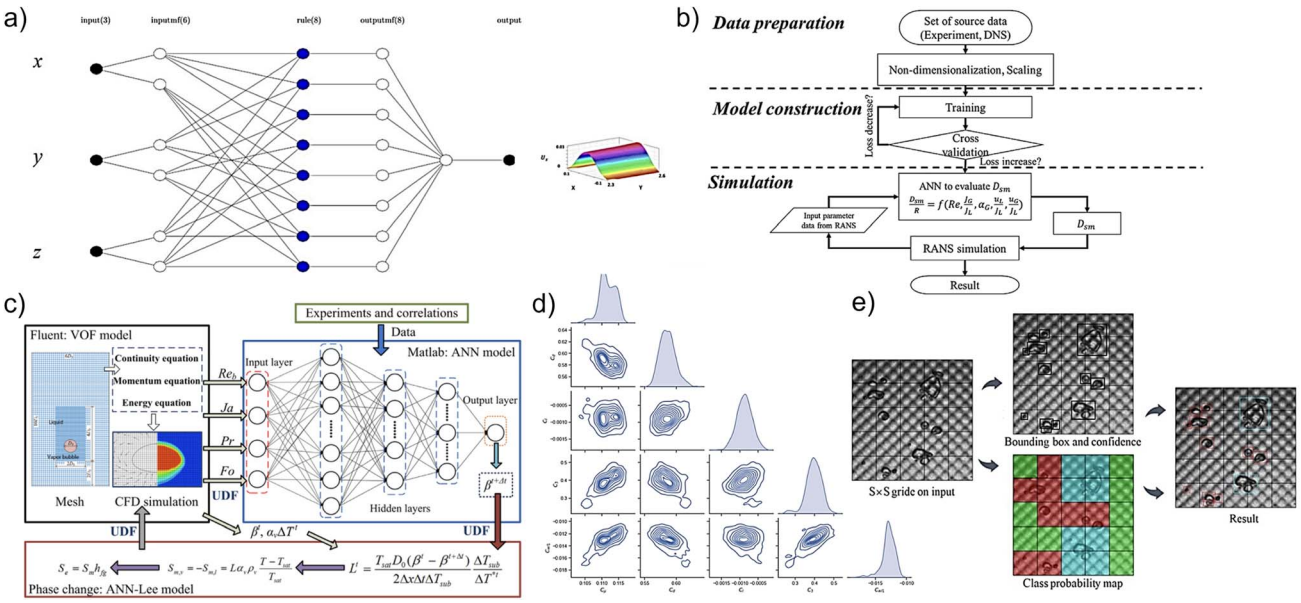

Fig. 10 (a) ANFIS structure scheme. Reprinted with permission from ref. 105. (b) Bubble size prediction based on a multi-layer artificial neural network. Reprinted with permission from ref. 110. (c) Workflow of applying ANN-Lee model for simulation of bubble condensation. Reprinted with permission from ref. 118. (d) Marginal and pairwise joint distribution of constitutive relation parameters. Reprinted with permission from ref. 119. (e) Using the YOLO net to predict the anomalous shape of cavitation bubbles. Reprinted with permission from ref. 120.

get a simulation model. Jung *et al.* performed an efficient model that can accurately predict the size of cavitation bubbles on test data and in real systems.[110] The approach was to create a pipeline with a multilayer perceptron (MLP) model capable of simulating the size of cavitation bubbles for various parameters (Fig. 10b). A low error of no more than 5% was obtained.

Similar pipelines were applied to identify the flow regime.[111] Two-phase flow data was measured and extracted using Ultrasound Doppler Velocimetry (UDV). The obtained data was analyzed using three classification algorithms such as decision tree, K nearest neighbor, and support vector machine. All three algorithms showed classification accuracy above 90% in real time. For comparison, CNN and LSTM were used to identify the flow regime with an accuracy of about 94%. The effectiveness of the proposed approach was also showed by another research groups.[112] Classic machine learning methods are also relevant for predicting the characteristics of the flow regime, the behavior of a single bubble in a flow, and its deformation.[113–117]

Integration of machine learning methods increases the efficiency of simulation models that are hardly performed for bubbles' transformation processes. Then, the bubble condensation simulation, which was previously performed using only the Lee model, was successfully implemented using artificial neural networks.[118] The back propagation artificial neural network was used to obtain an empirical coefficient for the Lee model (Fig. 10c). The database consisted of four numerical inputs such as Prandtl and Reynolds numbers. The predicted empirical coefficient was in a good agreement with literature data, as well as modeling *via* the coupled ANN-Lee model.

A similar advanced approach with the integration of machine learning to clarify inputs and outputs was applied to flow simulation. Liu *et al.* used machine learning techniques to reduce the computational costs for multiphase computational fluid dynamics (MCFD) simulation of bubbly flow.[119] Principal component analysis was performed to decrease the subspace of each MCFD simulation output, and then a feedforward neural network was used for surrogate modelling to predict MCFD results. A Gaussian process was used to obtain the model form uncertainty and parameters distribution (Fig. 10d).

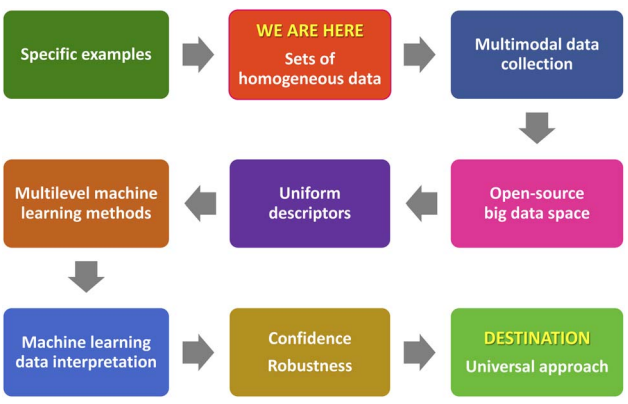

Fig. 11 Envisioned development of a machine learning approach for further deep analysis of complex soft and liquid material systems.

Combining machine learning methods with only numerical methods is also an efficient approach, especially for real systems such as reactor with complex flows. Wang *et al.* discussed the applicability of CNNs combined with improved three-frame difference (ITFD) method for recognizing and tracking bubbles in a plate heat exchanger (PHE).[120] The bubble size accuracy determination was 94% (Fig. 10e).

ML can also be applied for dynamic processes modelling without any support with classical simulation methods. This approach was performed to predict the migration of cavitation bubbles.[121] The research demonstrated that instead of traditional experimental and numerical methods, a two-branch model of a deep neural network with the embedding of a Kelvin impulse can be used. The demonstrated method showed high efficiency.

5 Conclusions

The first attempts to apply ML methods for understanding the soft matter structure and properties have been made since the 90 s, but the rapid growth of research in this area has occurred in the last few years. This happened due to the active development of machine learning and artificial intelligence approaches for many branches of science and is reflected both in the number of original research articles (Fig. 1) and in the appearance of the first reviews.[122–124]

The correct and appropriate use of ML techniques for soft and liquid molecular materials allows discovering their new properties at all scales, from molecular to macroscopic, based on the analysis of experimental or simulated homogeneous data sets, and demonstrate the promise of soft materials devices for complex technological applications. Our review provides many examples supporting this conclusion.

However, several remarks should be made to all the analyzed studies (Fig. 11). Generally, even recent studies do not consider multimodal data obtained by various methods. Also, most of the research papers do not offer open access to either the data or the source code of the used machine learning model. The importance of this practice has already been emphasized in chemistry[125] and should be recognized by researchers in the field of soft materials. An open data policy will enable in-depth studies using various big data analysis algorithms when considering existing scientific problems in soft matter from different points of view. Consideration of complex data sets involves the introduction of universal descriptors that could link together the various properties of the analyzed system at several levels of the hierarchy. This will most likely require the creation of new, integrated approaches for multivariate analysis of the systems under study, that will be based on several ML methods. Validation of the created complex ML approaches, interpretation of the obtained ML data, their reliability and validity remain critical issues. The implementation of these steps in the development of machine learning technique will ensure the formation of a universal approach to the analysis and prediction of the behavior of systems and materials with a complex architecture. It will also pave the path to new interdisciplinary areas of research and bring scientific results that

are difficult to imagine at the current level of applying ML and artificial intelligence methods for soft and liquid molecular materials.

Data availability

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Author contributions

T. O.: investigation; validation; writing – review editing; A. P.: investigation; writing – original draft; D. D.: investigation; writing – original draft; T. A.: data curation; investigation; writing – original draft; T. M. T. A. R.: investigation; writing – original draft; A. B.: investigation; writing – original draft; N. G.: writing – original draft; E. S.: conceptualization; supervision; validation.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The RSF grant no 21-13-00403 supports research related to hydrogel and machine learning approach. The RSF grant no 22-13-00185 supports research related to liquid crystals and machine learning approach. The goszadanie no. FSER-2021-0013 supports research related to bubble dynamics and machine learning approach. We thank the ITMO Fellowship and Professorship Program for the infrastructural support.

References

1 L. van der Maaten, E. Postma and J. van den Herik, *J. Mach. Learn. Res.*, 2009, **10**, 66–71.
2 V. Chandola, A. Banerjee and V. Kumar, *ACM Comput. Surv.*, 2009, **41**, 1–58.
3 D. A. Freedman, *Statistical Models: Theory and Practice*, Cambridge University Press, 1st edn, 2009.
4 R. A. Fisher, *Ann. Eugen.*, 1936, **7**, 179–188.
5 B. Everitt, *Cluster analysis*, Wiley, Chichester, West Sussex, U.K, 1st edn, 2011.
6 I. T. Jolliffe and J. Cadima, *Philos. Trans. R. Soc., A*, 2016, **374**, 20150202.
7 K. Beyer, *Database Theory—ICDT'99*, 1999, 217–235.
8 J. R. Quinlan, *Int. J. Man-Mach. Stud.*, 1987, **27**, 221–234.
9 C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273–297.
10 J. J. Hopfield, *Proc. Natl. Acad. Sci. U. S. A.*, 1982, **79**, 2554–2558.
11 J. Shang and P. Theato, *Soft Matter*, 2018, **14**, 8401–8407.
12 H. Lei, L. Dong, Y. Li, J. Zhang, H. Chen, J. Wu, Y. Zhang, Q. Fan, B. Xue, M. Qin, B. Chen, Y. Cao and W. Wang, *Nat. Commun.*, 2020, **11**, 4032.
13 F. Ullah, M. B. H. Othman, F. Javed, Z. Ahmad and H. M. Akil, *Mater. Sci. Eng., C*, 2015, **57**, 414–433.
14 J. S. Peerless, N. J. B. Milliken, T. J. Oweida, M. D. Manning and Y. G. Yingling, *Adv. Theory Simul.*, 2019, **2**, 1800129.
15 Y. Tsou, J. Khoneisser, P. Huang and X. Xu, *Bioact. Mater.*, 2016, **1**, 39–55.
16 G. Jose, K. Shalumon and J. Chen, *Curr. Med. Chem.*, 2020, **27**, 2734–2776.
17 Z. Sherman, M. Howard, B. Lindquist, R. Jadrich and T. Truskett, *J. Chem. Phys.*, 2020, **152**, 140902.
18 J. Li and D. Mooney, *Nat. Rev. Mater.*, 2016, **1**, 16071.
19 S. Tavakoli and A. S. Klar, *Biomolecules*, 2020, **10**, 1169.
20 S. Caliari and J. Burdick, *Nat. Methods*, 2016, **13**, 405–414.
21 S. Mantha, S. Pillai, P. Khayambashi, A. Upadhyay, Y. Zhang, O. Tao, H. M. Pham and S. D. Tran, *Materials*, 2019, **12**, 3323.
22 S. Kalasin, P. Sangnuang and W. Surareungchai, *ACS Biomater. Sci. Eng.*, 2020, **7**, 322–334.
23 S. Islam, M. Park, R. Campbell and A. Kim, *2020 IEEE Signal Processing in Medicine and Biology Symposium*, SPMB, 2020.
24 J. Lee, S. J. Oh, S. H. An, W.-D. Kim and S.-H. Kim, *Biofabrication*, 2020, **12**, 035018.
25 Z. Liu, T. Zhang, M. Yang, W. Gao, S. Wu, K. Wang, F. Dong, J. Dang, D. Zhou and J. Zhang, *ACS Appl. Electron. Mater.*, 2021, **3**, 3599–3609.
26 F. Lia, J. Hana, T. Caob, W. Lam, B. Fan, W. Tang, S. Chen, K. L. Fok and L. Li, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 11259–11264.
27 R. Richter, M. Kamal, M. A. García-Rivera, J. Kaspar, M. Junk, W. A. M. Elgaher, S. K. Srikakulam, A. Gress, A. Beckmann, A. Grißmer, C. Meier and M. Vielhaber, *Mater. Today Bio*, 2020, **8**, 100084.
28 F.-R. Fan, L. Lin, G. Zhu, W. Wu, Z.-Q. Tian and Z. L. Wang, *International Photonics and Optoelectronics Meetings*, POEM, 2013, p. NSa3A.17.
29 J.-N. Kim, J. Lee, H. Lee and I.-K. Oh, *Nano Energy*, 2021, **82**, 105705.
30 R. McQueen, D. Neal, R. DeWar, S. Garner and C. Nevill-Manning, *Proc. Canadian Machine Learning Workshop*, 1994, pp. 1–9.
31 A. Dhaliwal, M. Brenner, P. Wolujewicz, Z. Zhang, Y. Mao, M. Batish, J. Kohn and P. V. Moghe, *Acta Biomater.*, 2016, **45**, 98–109.
32 A. S. Ivanov, K. G. Nikolaev, A. A. Stekolshchikova, W. T. Tesfatsion, S. O. Yurchenko, K. S. Novoselov, D. V. Andreeva, M. Y. Rubtsova, M. F. Vorovitch, A. A. Ishmukhametov, A. M. Egorov and E. V. Skorb, *ACS Appl. Bio Mater.*, 2020, **3**, 7352–7356.
33 T. Hinton, Q. Jallerat, R. Palchesko, J. Park, M. Grodzicki, H. J. Shue, M. Ramadan, A. Hudson and A. Feinberg, *Sci. Adv.*, 2015, **1**, e1500758.
34 J. M. Bone, C. M. Childs, A. Menon, B. Póczos, A. W. Feinberg, P. R. LeDuc and N. R. Washburn, *ACS Biomater. Sci. Eng.*, 2020, **6**, 7021–7031.
35 K. D. Wright, M. A. Zimmerman, E. Fine, T. Aspri, M. W. Kieran and S. Chi, *Neurooncology*, 2018, **20**, i110.
36 S. Jackson, E. H. Baker, A. M. Gross, P. Whitcomb, A. Baldwin, J. Derdak, C. Tibery, J. Desanto, A. Carbonell,

K. Yohay, G. O'Sullivan, A. P. Chen, B. C. Widemann and E. Dombi, *Neuro-Oncol. Adv.*, 2020, **2**, 1–9.
37 A. Tabet, T. Gebhart, G. Wu, C. Readman, M. Pierson Smela, V. K. Rana, C. Baker, H. Bulstrode, P. Anikeeva, D. H. Rowitch and O. A. Scherman, *Phys. Chem. Chem. Phys.*, 2020, **22**, 14976–14982.
38 J. Lee, S. J. Oh, S. H. An, W.-D. Kim and S.-H. Kim, *Biofabrication*, 2020, **12**, 035018.
39 R. L. Martineau, A. V. Bayles, C.-S. Hung, K. G. Reyes, M. E. Helgeson and M. K. Gupta, *Adv. Biol.*, 2022, **6**, 2101070.
40 P. J. Collings and J. W. Goodby, *Introduction to Liquid Crystals*, CRC Press, 1st edn, 2019.
41 H. Kränz, V. Vill and B. Meyer, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 1173–1177.
42 S. R. Johnson and P. C. Jurs, *Chem. Mater.*, 1999, **11**, 1007–1023.
43 J. Xu, L. Wang, H. Zhang, C. Yi and W. Xu, *Mol. Simul.*, 2010, **36**, 26–34.
44 F. Leon, C. Lisa and S. Curteanu, *Mol. Cryst. Liq. Cryst.*, 2010, **518**, 129–148.
45 T. Inokuchi, R. Okamoto and N. Arai, *Liq. Cryst.*, 2020, **47**, 438–448.
46 N. Osiecka-Drewniak, M. A. Czarnecki and Z. Galewski, *J. Mol. Liq.*, 2021, **341**, 117233.
47 T. C. Le and N. Tran, *ACS Appl. Nano Mater.*, 2019, **2**, 1637–1647.
48 D. Antanasijević, J. Antanasijević, V. Pocajt and G. Uščumlić, *RSC Adv.*, 2016, **6**, 99676–99684.
49 J. Antanasijević, V. Pocajt, D. Antanasijević, N. Trišović and K. Fodor-Csorba, *Liq. Cryst.*, 2016, **43**, 1028–1037.
50 C.-H. Chen, K. Tanaka and K. Funatsu, *Mol. Inf.*, 2018, **38**, 1800095.
51 V. Vill, *LiqCryst 5.2 Advanced – Database of Liquid Crystals*, LCI Publisher, Hamburg, 2013.
52 U. D. Schiller, T. Krüger and O. Henrich, *Soft Matter*, 2018, **14**, 9–26.
53 S. Pestov and V. Vill, *Liquid Crystal, Springer Handbook of Condensed Matter and Materials Data*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 941–977.
54 T. C. Le, C. E. Conn, F. R. Burden and D. A. Winkler, *Cryst. Growth Des.*, 2013, **13**, 1267–1276.
55 A. Jákli, O. D. Lavrentovich and J. V. Selinger, *Rev. Mod. Phys.*, 2018, **90**, 045004.
56 M. Chiappini, A. Patti and M. Dijkstra, *Phys. Rev. E*, 2020, **102**, 040601.
57 A. Patti and A. Cuetos, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2012, **86**, 011403.
58 F. Reinitzer, *Monatsh. Chem.*, 1888, **9**, 421–441.
59 I. Dierking, *Textures of Liquid Crystals*, John Wiley Sons, Ltd, 2003.
60 H. Y. Sigaki, E. K. Lenzi, R. S. Zola, M. Perc and H. V. Ribeiro, *Sci. Rep.*, 2020, **10**, 1–10.
61 J. Colen, M. Han, R. Zhang, S. A. Redford, L. M. Lemma, L. Morgan, P. V. Ruijgrok, R. Adkins, Z. Bryant, Z. Dogic, M. L. Gardel, J. J. de Pablo and V. Vitelli, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2016708118.
62 H. Y. D. Sigaki, R. F. de Souza, R. T. de Souza, R. S. Zola and H. V. Ribeiro, *Phys. Rev. E*, 2019, **99**, 013311.
63 A. A. Pessa, R. S. Zola, M. Perc and H. V. Ribeiro, *Chaos, Solitons Fractals*, 2022, **154**, 111607.
64 Z. Zhou, C. Joshi, R. Liu, M. M. Norton, L. Lemma, Z. Dogic, M. F. Hagan, S. Fraden and P. Hong, *Soft Matter*, 2021, **17**, 738–747.
65 T. Orlova, F. Lancia, C. Loussert, S. Iamsaard, N. Katsonis and E. Brasselet, *Nat. Nanotechnol.*, 2018, **13**, 304–308.
66 D. J. B. Albert Schenning and G. P. Crawford, *Liquid Crystal Sensors*, CRC Press, 2017.
67 S. A. Oladepo, *Molecules*, 2022, **27**, 1453.
68 K. Nayani, Y. Yang, H. Yu, P. Jani, M. Mavrikakis and N. Abbott, *Liq. Cryst. Today*, 2020, **29**, 24–35.
69 N. J. Nilsson, *Learning machines*, New York, 1965.
70 Y. Xu, A. M. Rather, S. Song, J.-C. Fang, R. L. Dupont, U. I. Kara, Y. Chang, J. A. Paulson, R. Qin, X. Bao and X. Wang, *Cell Rep. Phys. Sci.*, 2020, **1**, 100276.
71 J. Frazão, S. I. C. J. Palma, H. M. A. Costa, C. Alves, A. C. A. Roque and M. Silveira, *Sensors*, 2021, **21**, 2854.
72 E. Ramou, S. I. C. J. Palma and A. C. A. Roque, *ACS Appl. Mater. Interfaces*, 2022, **14**, 6261–6273.
73 Y. Cao, H. Yu, N. L. Abbott and V. M. Zavala, *ACS Sens.*, 2018, **3**, 2237–2245.
74 A. D. Smith, N. Abbott and V. M. Zavala, *J. Phys. Chem. C*, 2020, **124**, 15152–15161.
75 S. Jiang, J. Noh, C. Park, A. D. Smith, N. L. Abbott and V. M. Zavala, *Analyst*, 2021, **146**, 1224–1233.
76 R. H. Chen, *Liquid Crystal Displays: Fundamental Physics and Technology*, John Wiley Sons, Ltd, 2011.
77 S. B. Kang, J. H. Lee, K. Y. Song and H. J. Pahk, *2009 IEEE International Symposium on Industrial Electronics*, 2009, pp. 2175–2177.
78 Y.-H. Liu and Y.-J. Chen, *Int. J. Mol. Sci.*, 2011, **12**, 5762–5781.
79 W. Huang and H. Lu, *Int. J. Image Graph.*, 2013, **13**, 1350011.
80 D.-C. Li, C.-C. Chang, C.-W. Liu and W.-C. Chen, *J. Intell. Manuf.*, 2013, **24**, 225–233.
81 C. Mantel, J. Søgaard, S. Bech, J. Korhonen, J. M. Pedersen and S. Forchhammer, *IEEE Trans. Image Process.*, 2016, **25**, 3751–3761.
82 J. Jo, J. W. Soh, J. S. Park and N. I. Cho, *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, APSIPA ASC, 2020, pp. 1067–1074.
83 S.-J. Song, Y. I. Kim, J. Bae and H. Nam, *Opt. Express*, 2019, **27**, 15907–15917.
84 T. Zhang, Y. Feng and B. Hao, *2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology*, IAICT, 2019, pp. 25–30.
85 C.-Y. Lee and T.-L. Tsai, *Robot. Comput.-Integr. Manuf.*, 2019, **55**, 76–87.
86 D. A. Ferreira, D. A. Amoedo, L. R. Costa, M. D. Valadão, A. S. Souza, K. Y. Ouchi, A. M. Pereira, G. M. Torres, M. O. Silva, C. F. Cruz, A. P. Silva, R. J. Belem, A. S. Jesus, A. S. Costa, L. G. Evangelista, O. R. Silva, T. B. Bezerra, W. S. Júnior and C. B. Carvalho, *2020 IEEE International*

Conference on Consumer Electronics – Taiwan, ICCE, Taiwan, 2020, pp. 1–2.
87 G. M. Torres, A. S. Souza, D. A. O. Ferreira, L. C. S. G. Júnior, K. Y. Ouchi, M. D. M. Valadão, M. O. Silva, V. L. G. Cavalcante, E. V. C. U. Mattos, A. M. C. Pereira, C. F. S. Cruz, A. P. Silva, R. J. S. Belem, A. S. Costa, L. G. C. Evangelista, W. C. C. Junior, R. G. Paula, T. B. Bezerra, W. S. S. Júnior and C. B. Carvalho, *2021 IEEE International Conference on Consumer Electronics*, ICCE, 2021, pp. 1–4.
88 Z.-K. Gao, M. Liu, W. Dang and Q. Cai, *Pet. Sci.*, 2021, **18**, 259–268.
89 A. S. Qaddoori, J. H. Saud and F. Hamad, *Materials Today: Proceedings*, 2021, available online 24 August 2021.
90 E. T. Brantson, M. Abdulkadir, P. H. Akwensi, H. Osei, T. F. Appiah, K. R. Assie and S. Samuel, *J. Nat. Gas Sci. Eng.*, 2022, **99**, 104406.
91 Y. Ju, L. Wu, M. Li, Q. Xiao and H. Wang, *Measurement*, 2022, **192**, 110861.
92 I. Poletaev, M. P. Tokarev and K. S. Pervunin, *Int. J. Multiphase Flow*, 2020, **126**, 103194.
93 I. E. Poletaev, K. S. Pervunin and M. P. Tokarev, *J. Phys.: Conf. Ser.*, 2016, **754**, 072002.
94 G. Montes-Atenas, F. Seguel, A. Valencia, S. M. Bhatti, M. S. Khan, I. Soto and N. B. Yoma, *Int. Commun. Heat Mass Transfer*, 2016, **76**, 197–201.
95 R. F. Cerqueira and E. E. Paladino, *Chem. Eng. Sci.*, 2021, **230**, 116163.
96 Y. He, C. Hu, H. Li, B. Jiang, X. Hu, K. Wang and D. Tang, *Chem. Eng. J.*, 2022, **429**, 132138.
97 A. Srivastava, R. Wang, S. K. Dinda and K. Chattopadhyay, *Mach. Learn. Appl.*, 2021, **6**, 100180.
98 C. Theßeling, M. Grünewald and P. Biessey, *Chem. Eng. Res. Des.*, 2020, **163**, 47–57.
99 P. Biessey, H. Bayer, C. Theßeling, E. Hilbrands and M. Grünewald, *Chem. Ing. Tech.*, 2021, **93**, 1968–1975.
100 Y. Fu and Y. Liu, *Chem. Eng. Sci.*, 2019, **204**, 35–47.
101 J. Li, S. Shao and J. Hong, *Meas. Sci. Technol.*, 2020, **32**, 015406.
102 S. Shao, K. Mallery, S. S. Kumar and J. Hong, *Opt. Express*, 2020, **28**, 2987–2999.
103 I. Korolev, T. A. Aliev, T. Orlova, S. A. Ulasevich, M. Nosonovsky and E. V. Skorb, *J. Phys. Chem. B*, 2022, **126**, 3161–3169.
104 A. Mosavi, S. Shamshirband, E. Salwana, K. wing Chau and J. H. M. Tah, *Eng. Appl. Comput. Fluid Mech.*, 2019, **13**, 482–492.
105 M. Babanezhad, A. Marjani and S. Shirazian, *Sci. Rep.*, 2020, **10**, 21502.
106 M. Babanezhad, A. T. Nakhjiri, M. Rezakazemi and S. Shirazian, *ACS Omega*, 2020, **5**, 20558–20566.
107 Q. Nguyen, I. Behroyan, M. Rezakazemi and S. Shirazian, *Arabian J. Sci. Eng.*, 2020, **45**, 7487–7498.
108 M. Babanezhad, A. T. Nakhjiri, M. Rezakazemi, A. Marjani and S. Shirazian, *Sci. Rep.*, 2020, **10**, 17793.
109 R. Pelalak, A. Nakhjiri, A. Marjani, M. Rezakazemi and S. Shirazian, *Sci. Rep.*, 2021, **11**, 1891.
110 H. Jung, S. Yoon, Y. Kim, J. H. Lee, H. Park, D. Kim, J. Kim and S. Kang, *Chem. Eng. Sci.*, 2020, **213**, 115357.
111 Y. Zhang, A. Azman, K.-W. Xu, C. Kang and H.-B. Kim, *Exp. Fluids*, 2021, **161**, 212.
112 O. N. Manjrekar and M. P. Dudukovic, *Chem. Eng. Sci.: X*, 2019, **2**, 100023.
113 G. Mask, X. Wu and K. Ling, *J. Pet. Sci. Eng.*, 2019, **183**, 106370.
114 B. Deng, R. J. Chin, Y. Tang, C. Jiang and S. H. Lai, *Appl. Sci.*, 2019, **9**, 3198.
115 F. D. Nunno, F. A. Pereira, G. de Marinis, F. D. Felice, R. Gargano, M. Miozzi and F. Granata, *Appl. Sci.*, 2020, **10**, 3879.
116 H. Chen, Y. Zeng and Y. Li, *Acta Mech. Sin.*, 2021, **37**, 35–46.
117 X. Wang, Z. Ning, M. Lv and C. Sun, *Results Phys.*, 2021, **25**, 104226.
118 J. Tang, H. Liu, M. Du, W. Yang and L. Sun, *Int. J. Heat Mass Transfer*, 2021, **178**, 121620.
119 Y. Liu, D. Wang, X. Sun, Y. Liu, N. Dinh and R. Hu, *Reliab. Eng. Syst. Saf.*, 2021, **212**, 107636.
120 Q. Wang, X. Li, C. Xu, T. Yan and Y. Li, *Int. J. Multiphase Flow*, 2021, **138**, 103593.
121 X. Ma, C. Wang, B. Huang and G. Wang, *Phys. Fluids*, 2019, **31**, 102003.
122 E. Bedolla, L. C. Padierna and R. Castañeda-Priego, *J. Phys.: Condens. Matter*, 2021, **33**, 053001.
123 P. S. Clegg, *Soft Matter*, 2021, **17**, 3991.
124 N. Jackson, M. Webb and J. de Pablo, *Curr. Opin. Chem. Eng.*, 2019, **23**, 106–114.
125 N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain and A. Walsh, *Nat. Chem.*, 2021, **13**, 505–508.