Machine learning approaches for biomolecular, biophysical, and biomaterials research

Cite as: Biophysics Rev. **3**, 021306 (2022); doi: 10.1063/5.0082179
Submitted: 13 December 2021 · Accepted: 12 May 2022 ·
Published Online: 3 June 2022 · Publisher error corrected: 6 June 2022

View Online Export Citation CrossMark

Carolin A. Rickert[1,2] and Oliver Lieleg[1,2,a)]

AFFILIATIONS

[1]Department of Materials Engineering, TUM School of Engineering and Design, Technical University of Munich, Boltzmannstr. 15, 85748 Garching b. München, Germany
[2]Center for Functional Protein Assemblies (CPA), Technical University of Munich, Ernst-Otto-Fischer Straße 8, 85748 Garching b. München, Germany

[a)]Author to whom correspondence should be addressed: oliver.lieleg@tum.de

ABSTRACT

A fluent conversation with a virtual assistant, person-tailored news feeds, and deep-fake images created within seconds—all those things that have been unthinkable for a long time are now a part of our everyday lives. What these examples have in common is that they are realized by different means of machine learning (ML), a technology that has fundamentally changed many aspects of the modern world. The possibility to process enormous amount of data in multi-hierarchical, digital constructs has paved the way not only for creating intelligent systems but also for obtaining surprising new insight into many scientific problems. However, in the different areas of biosciences, which typically rely heavily on the collection of time-consuming experimental data, applying ML methods is a bit more challenging: Here, difficulties can arise from small datasets and the inherent, broad variability, and complexity associated with studying biological objects and phenomena. In this Review, we give an overview of commonly used ML algorithms (which are often referred to as "machines") and learning strategies as well as their applications in different bio-disciplines such as molecular biology, drug development, biophysics, and biomaterials science. We highlight how selected research questions from those fields were successfully translated into machine readable formats, discuss typical problems that can arise in this context, and provide an overview of how to resolve those encountered difficulties.

TABLE OF CONTENTS

I. INTRODUCTION

In many areas of medicine and materials science, analyzing complex datasets is a crucial task; those datasets, for instance, consist of images that can be used to identify pathologies or to quantify the progress of diseases[1–3] as well as for detecting defects on materials[4–6] and monitoring experimental[7–9] and production[10,11] processes. When performed manually, those tasks require time-consuming expert involvement but, nevertheless, may remain error-prone and biased. This is where computer-based decision processes can help. In the recent decade, machine learning (ML) approaches have gained vastly increased attention and have been successfully applied to different problems. Machine learning is a field of data science that encompasses a variety of algorithms that automatically learn from provided information and then draw conclusions. Such approaches aim at simplifying, extending, or replacing human decision and analysis processes. Examples include object detection[12,13] and monitoring,[14,15] identification of patterns or correlations between datasets,[16,17] as well as data classification,[18–20] regression,[21,22] or clustering[23,24] (Fig. 1).

A key task for which machine learning has turned out to be highly helpful is image analysis.[25–28] Here, image segmentation and object detection methods can be used to automatically identify and locate the presence of certain objects within an image or video.[29–31] By receiving example images as an input, the algorithms learn to find informative regions in the pictures and extract characteristic features such as edges or specific shapes from them.[32,33] At the moment, such approaches are extensively applied to face recognition or autonomous driving tasks; yet, this technique offers great potential in other areas as well where decisions are made based on visual impressions: The progression of glaucoma,[34–36] dementia,[37,38] or cancer[39–42] was successfully extracted from medical images, cell nuclei were detected in microscope images,[43,44] microtissue-contraction measurements were automatically analyzed in laboratory experiments,[45] and additive manufacturing processes of biomaterials were optimized.[46,47]

In addition to analyzing images, ML algorithms can also handle other data types such as numerical values or text. Instead of an image, the samples then comprise multiple input parameters (commonly referred to as features) and—optionally—an output label or value. In materials science, such data analyses can uncover links among the composition, structure, and characteristics of known materials and extrapolate this knowledge to propose potential new materials with predefined properties.[48–50] Here, the algorithms search for patterns and correlations in the dataset, from which conclusions can be drawn.[51] With such an approach, it was possible to explore therapeutics that target specific diseases[52–55] to study glycan functions,[56,57] to enhance single molecule sensing,[58] and to improve manufacturing processes such as 3D bioprinting[59] or microparticle production.[60]

By mapping such input data, e.g., experimental findings, onto output labels, predictive algorithms can be established. Depending on the type of possible outputs, one can distinguish between classification and regression attempts. Classification describes the prediction of discrete outputs, i.e., samples are assigned to specific classes. Examples are the categorization of surfaces with regard to their wetting behavior[61] or sorting the state of polymer conformations.[62] In contrast, regression algorithms predict properties that can be described by continuous values such as interaction affinities,[63–65] transcriptional

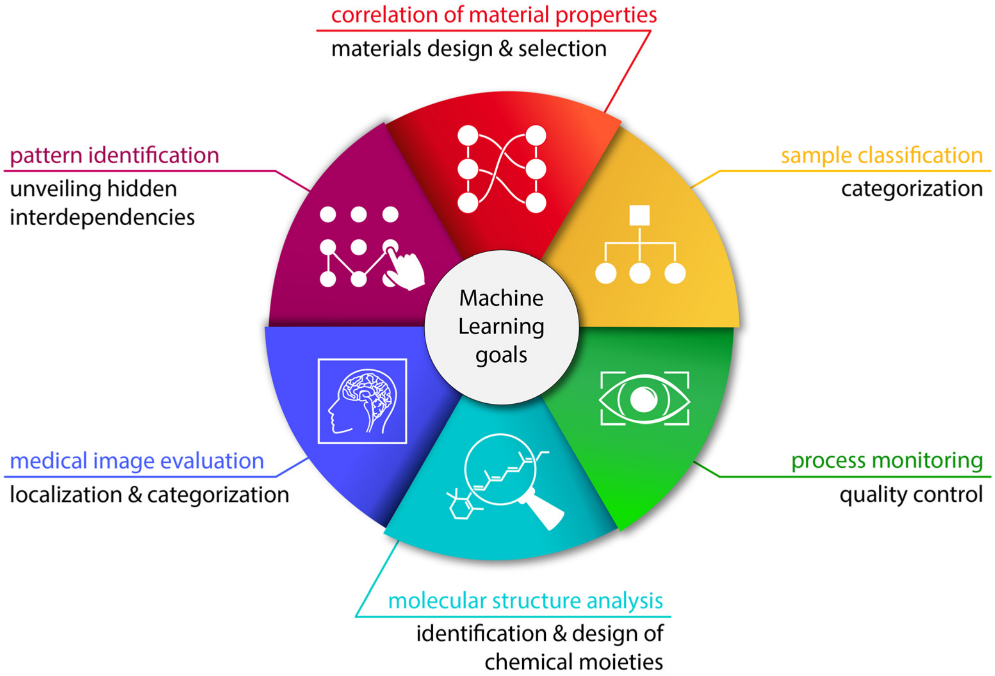
correlation of material properties
materials design & selection
pattern identification
unveiling hidden interdependencies
sample classification
categorization
Machine Learning goals
medical image evaluation
localization & categorization
process monitoring
quality control
molecular structure analysis
identification & design of chemical moieties

FIG. 1. Typical objectives of machine learning approaches. A ML-based analysis of data from the biosciences can have different goals. Typical examples include the correlation of material properties, the classification of samples, the identification of patterns, process monitoring, molecular structure analysis, and the evaluation of medical images. Correlating material properties can, for instance, be useful to predict the behavior or certain characteristics of materials to provide guidance for a target-oriented design or selection process. Sample classification finds broad applications in areas where samples need to be assigned to discrete categories, e.g., for the classification of disease patterns based on various biomarkers. ML-based process monitoring can be an essential part of quality control to automatically identify and react to defects or variations in the process flow. Analyzing molecular structures by means of ML allows us to scan large databases to identify or even to design chemical moieties with certain properties. Another growing area of application for ML is the automated evaluation of medical images to, e.g., localize and categorize organs or pathological manifestations in tissues. Finally, one versatile purpose of ML is to identify patterns in databases to unveil hidden dependencies between different characteristics and attributes.

activities of DNA motifs,[66] or material parameters describing mechanical responses.[67,68] These approaches are especially useful when mathematical equations based on physical models are still unknown.

II. PRINCIPLES, ADVANTAGES, AND LIMITATIONS OF DIFFERENT ML ALGORITHMS

Considering the large variety of available ML algorithms, selecting the most suitable one for a given problem is not always trivial: The best choice depends on the problem statement, the database, the desired output, interpretability, and many other factors. In Sec. II, we give an overview over common learning strategies, we highlight selected ML models (including random ensemble-based, probabilistic, linear, and deep learning methods), and we explain their working principles and characteristics. Although some models can make use of different learning strategies, in the following, each of them is assigned to the most commonly used one. Graphical representations of the algorithms discussed here are depicted in Fig. 2, and an overview of the advantages and disadvantages is given in Table I.

Overall, data fed into an algorithm can serve three different purposes: First, a "training set" is required to allow the algorithms to develop a model. Second, a "test set" is used for validation, and this set contains data the algorithms are only confronted with once they have established the model. Third, once validation was successful, so-called "query samples" are fed into the algorithm with the aim to get classified or to make predictions for. In all those datasets, input variables that quantify individual measurable characteristics of a data point are referred to as "features," outputs assigned to training or test samples are called "labels," and the output of the algorithm (be it continuous or discrete values) created for a query sample is called "prediction."

A. Supervised learning

In supervised learning, models are developed based on labeled data—similar to how parents teach their children to name objects. The algorithm needs to be provided a training dataset, containing a sufficiently large number of samples; each of them is represented by input data—i.e., information (descriptors) that is likely to characterize the desired output—and corresponding output labels. Such datasets could, for example, comprise histological images of cancerous tissue (input) labeled with the name of the affected organ (output),[114] or they could link the composition of a polymeric biomaterial (input) to its mechanical behavior (output).[115,116] With such information offered, the ML models aim at identifying relationships between the input and the output and can then perform classification or prediction tasks for new data they were not confronted with before.

1. k nearest neighbor (KNN) algorithms

The simple but powerful *k* nearest neighbor algorithm follows the assumption that similarity between samples is accompanied by proximity in the data space; in other words, similar samples are expected to come with similar inputs. Instead of developing a generalized model, predictions are made by comparing a query sample to the

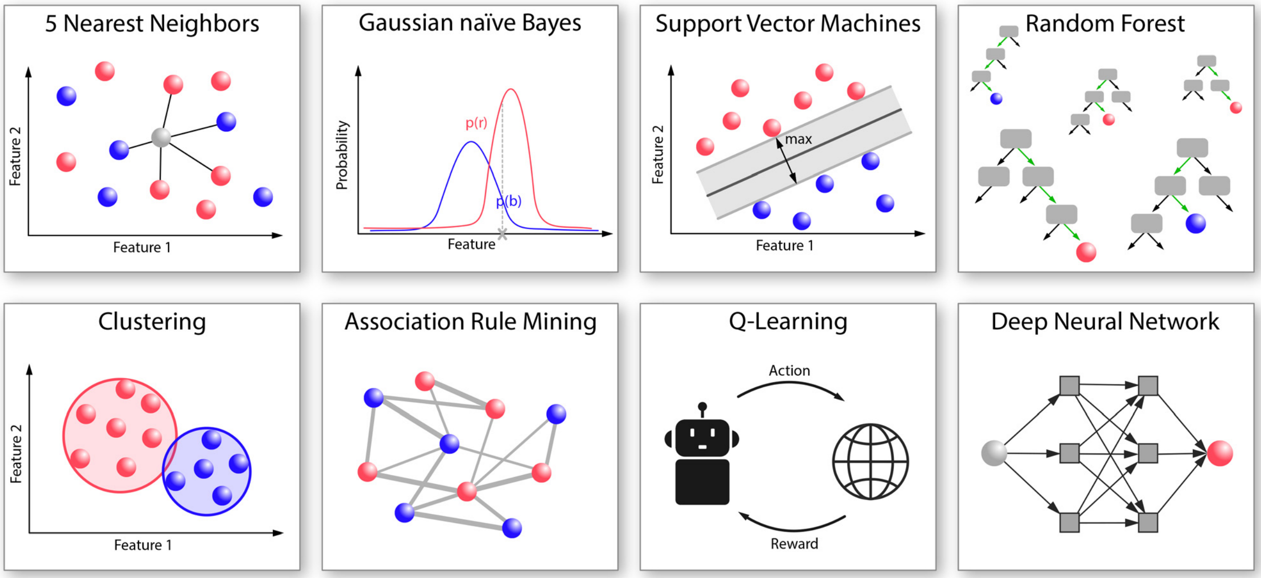

FIG. 2. Schematic representation of typical ML algorithms used for analyzing problems from the different fields of biosciences. The *k* nearest neighbor (KNN) algorithm classifies a query sample according to the *k* samples that are most similar to it, i.e., which have the lowest distance in an *n*-dimensional hyperspace (here, *n* corresponds to the number of analyzed features). The Gaussian Naïve Bayes algorithm determines conditional probabilities and classifies samples based on a "most probable" principle. Support vector machines define hyperplanes in the *n*-dimensional feature space to distinctly separate samples of different classes while maximizing the distance of all samples to this separating hyperplane. The Random Forest classifier combines many randomly generated, uncorrelated decision trees to perform predictions in a popular-vote-like manner. Clustering refers to algorithms that group unlabeled data based on their characteristics. Association rule mining describes the process of finding dependencies that govern correlations and associations between samples. Q-learning assesses the quality of each action available for a given state by rewarding a subset of desired outcomes. Deep neural networks mimic the structure of the human brain by combining activatable units in consecutive, interconnected layers that process information in various manners.

TABLE I. Overview of the advantages and disadvantages of the different ML algorithms discussed here.

K nearest neighbors[69–72]	
No training phase needed	High dimensionality leads to decreased accuracies
Intuitive and simple algorithm	Can become slow for big datasets
Easily adapts to new training data	Needs feature scaling
Only one hyperparameter to tune	Has problems with imbalanced datasets
	Missing values are problematic
Naïve Bayes[73–78]	
Very fast	The assumption of independent features that equally contribute to the output rarely holds true
Needs less training data than most other algorithms	Zero probability problem: If one feature of a sample exhibits a value of zero probability according to the trained model, the class will be assigned a probability of zero.
Works well with high-dimensional data	
Support vector machines[79–86]	
Kernel functions can be used to solve complex problems	Choosing an appropriate kernel can be difficult
Effective in high-dimensional spaces even for comparably small sample sizes	Training times can become long with large datasets
Memory efficient, as it uses a subset of training points for the decision function	Limited capability to handle noisy or strongly overlapping classes
Random forest[87–93]	
Robust to outliers, noise, and imbalanced datasets	Long training times for large datasets
Lower risk of overfitting	Little control over model formation
Runs efficiently with large datasets	Limited ability to extrapolate
Easy data preparation	
Can handle high dimensionalities	
Clustering[94–99]	
Can handle unlabeled data	It can be difficult to interpret the sorting decision
Algorithms of different complexity are available	Big datasets can lead to long running times
Can be used on very small and very large datasets	The criteria to stop clustering or the number of clusters need to be defined
Association rule mining[100–104]	
Offers an easy way to detect correlations in unsorted datasets	Does not guarantee statistical significance
Unveils relationships between elements	Requires nominal variables; continuous values need to be translated
Q-Learning[105–109]	
After sufficient training, it finds optimal actions	Can be computationally expensive since each state/action pair needs to be evaluated multiple times
Can solve problems without explicitly being told how to	Does not include risk assessments into the decision making
	Can have problems with high dimensionality
Deep neural networks[110–113]	
Highly flexible and suitable to approximate complex functions	Requires lots of training data
	Can be difficult to interpret (black box)
Once trained, the predictions are fast	Training can be computationally expensive
There are multiple different network architectures already available	Finding the best network architecture can be challenging

training data. Then, the k nearest neighbors, i.e., the most similar samples according to their feature values, are identified, and a prediction is made considering the labels of those data points in a popular-vote-like manner. The number of neighbors k can be varied to find a valid compromise between robustness toward outliers (which is achieved for high values of k) and distinctness (which is a typical result for low values of k).[117,118]

KNN algorithms can be used for multi-class problems,[119] and their accuracy can easily be improved by adding more data points to the training set. Providing more input data, however, typically comes at the cost of long computational runtimes.[69] Moreover, KNN algorithms have limitations when it comes to handling imbalanced datasets[70] (e.g., training data with a dominant class): For predictions to be reliable, a certain amount of data points from all classes is required to achieve a suitable (local) density in the data space. Also, KNN algorithms tend to struggle with large numbers of input features—a phenomenon, which is known as "curse of dimensionality."[71] Finally, as the input features are usually weighted equally when calculating the distance of a query sample to its nearest neighbors, it is important to ensure that the input features have the same scale[72] (which is why some preprocessing of the data might be required).

2. Naïve Bayes methods

Naïve Bayes approaches are probabilistic learning methods that are mostly used for classification tasks. Here, the training data are used to determine likelihood distributions (e.g., Gaussian, multinomial, Bernoulli, or categorical distributions[120,121]) of the feature values representing each class. Then, the probability that a query sample belongs to one of the classes is calculated based on the Naïve assumption that all features are independent and contribute equally to the output. The corresponding mathematical relationship is formulated in Bayes' theorem.[122] Although Naïve Bayes approaches typically rely on over-simplified assumptions, those algorithms can outperform even highly sophisticated methods.[123]

Compared to other algorithms, Naïve Bayes classifiers can be extremely fast[73] and require a small amount of training data only.[74] Owing to the independent likelihood estimation applied to each feature, those algorithms also perform well when tasked with high-dimensional problems[75] (i.e., those, where many input features are considered) and multi-class classifications[119]—and they can process both, categorical[124,125] and continuous input data.[126] However, the simplified assumptions made by Naïve Bayes classifiers do not always hold true when real-life problems are studied: Here, only rarely all features of a sample are truly independent;[76] similarly, it is not likely that all sample features contribute equally to the output[77] and all feature distributions meet the assumed profile. Furthermore, categorical inputs of the query sample that were not present in the training data will lead to an incorrect probability of zero, known as the "zero frequency problem."[78]

3. Support vector machines (SVMs)

Support vector machines (SVMs) define hyperplanes in the n-dimensional feature space, which then can be used to either distinctly separate the dataset into single-variety classes (i.e., for classification) or to approximate the training data (i.e., for regression). To allow for handling problems that would otherwise involve complex mathematical operations, kernel functions that transform input data into higher dimensionality can be integrated into those models.[127,128]

Since only a subset of training points is used for calculating the decision function, support vector methods can handle data spaces of high dimensionality[79,80] while remaining efficient regarding memory and runtime.[81] However, for large datasets, the training times can increase significantly.[82] Due to the large variety of kernel functions that can be selected and specified for creating the decision function,[83,84] the algorithms are very versatile and can even be applied to unstructured data. Still, support vector classifiers can have problems with handling very noisy data[129] or classes that strongly overlap.[85,86]

4. Decision trees and random forest (RF) algorithms

Decision trees are flow chart-like representations of hierarchical decision-making models that are created by analyzing a labeled training set. They consist of nodes (i.e., consecutive stages in which distinct decisions are made) and branches that connect these nodes. Starting with a root node, the training data are (based on individual input features) split in a stepwise manner by creating and answering simple true/false questions. A new (=query) sample can then be classified/predicted by running through the tree using the input values of this new sample and the previously established decision rules.

According to the principle of swarm intelligence, the accuracy of such an approach can be improved by combining an ensemble of non-correlating decision trees—a random forest.[130] Enforcing this mandatory variation among the trees is mainly achieved by applying two methods known as feature randomization (here, only a random subset of features is provided for splitting the data) and bootstrap aggregation (short: bagging, i.e., randomly eliminating samples of the training set and replacing them with duplicates of the remaining samples).[131]

Random forest algorithms can achieve very high accuracies even in high-dimensional data spaces.[87] These algorithms run efficiently for large datasets,[88] and they can handle variable input data types, including binary, categorical, and numerical features.[132] They are well suitable for unbalanced data,[89] robust toward non-linearity,[90] and outliers[91] and—when a sufficient number of independent decision trees is used—rather insensitive to overfitting. Moreover, the decision criteria chosen by the decision trees can be extracted and used to rank the importance of individual features for the categorization process.[133,134] However, the self-directed formation of the different trees strongly restricts options to influence random forest algorithms. Importantly, random forest models are not able to extrapolate correlations, and this limits them to making predictions within the created knowledge space.[92] Finally, even though running efficiently once the model has been established, training can be computationally costly[93] since many trees (usually between 100 and 1000) must be created to obtain a robust random forest.

B. Unsupervised learning

When it is not clear yet what the algorithm is supposed to find, or if labeled data are not available, unsupervised machine learning is more suitable. In such a data-driven approach, the algorithm is simply fed with unsorted input data and allowed to draw its own conclusions by either autonomously clustering the samples or by identifying trends, similarities, extreme points, or patterns in the data. With such a strategy, it was possible to quantify the morphological heterogeneity of cells based on a specified set of geometrical parameters[135] and to automatically control the quality of electro-spun nanofibers.[136]

1. Clustering

An important concept in the field of unsupervised learning is clustering; this approach can be used to identify patterns in a set of unlabeled data. Here, a dataset (containing input values only) is analyzed by sorting the samples into subgroups (clusters) by identifying similarities among them. A common subtype of this approach is k-means clustering. Here, the samples are assigned to k clusters in an exclusive manner by iteratively adjusting cluster centroids until the variety of samples within the formed clusters is minimized while the variety between the clusters is maximized. K-means clustering algorithms are simple and fast, which is why they can handle large datasets.[94] They can easily adapt to new samples or data, and their sorting result can be influenced by predefining the initial centroids.[95,96] Yet, identifying the correct number k of clusters to be formed can be far from trivial and might require preliminary analyses.[97,98] Also, as common for distance-based algorithms, high data dimensionalities can cause issues.[99] Finally, basic k-means algorithms encounter problems when the created clusters differ in terms of size or density; however, generalization methods can be applied to deal with this particular issue.[137]

In addition to the rather simple k-means clustering algorithms, there are also other clustering variants that are selected when more complex datasets need to be processed. Mean-shift clustering, for example, searches regions of high data density by sliding pre-defined analysis windows over the data until the windows containing the highest number of data points are identified. There are two main advantages of this algorithm variant: First, the number of final clusters does not need to be pre-defined; second, centroids in close proximity to each other are automatically merged. A very powerful extension of such mean-shift clustering is the DBSCAN method (density-based spatial clustering of applications with noise), which is capable of identifying clusters of any shape and size while detecting and ignoring outliers. In addition, methods that establish clusters of different hierarchies were shown to work efficiently as well.[138]

2. Association rule mining

Another popular example of an unsupervised ML method is association rule mining. This approach aims at unveiling correlations between variables in a set of unlabeled data. Such association rules can be interpreted as "if–then" statements, where certain variables (antecedents) are linked to correlating ones (consequents). To identify the most important rules, the dataset is first searched for such if-then patterns, which are then ranked using different significance measures. A major drawback of this approach is that calculating those metrics for all identified relations becomes computationally expensive rather soon. The so-called *a priori* algorithm provides a good solution to this problem: Here, item sets containing variables or subsets with low importance in one metric are quickly eliminated, and this drastically reduces the amount of data that need to be analyzed regarding the other measures. In addition, there is a broad variety of other approaches for association rule mining that allow for handling different datasets and problems of higher complexity.[100–102] Yet, in any case, a sufficiently high data density is essential for these algorithms to avoid random correlations from becoming too prominent.

C. Reinforcement learning

A third learning strategy is reinforcement learning—an action-focused training approach. Here, the machine chooses from different possible actions and is punished or rewarded depending on whether or not it made a "correct" choice. Typically, this is implemented by the algorithm trying to optimize a reward function: Here, positive values are assigned when the algorithm chooses the desired outcome, which presents an incentive for the machine to make this choice; consistently, assigning negative values to "wrong" choices serves as a punishment rendering undesired behavior less likely. With this reward/penalty strategy, a machine can, for example, learn to play a simple board game by repeatedly exploring possible actions in a trial-and-error like fashion and trying to maximize the cumulative reward that is granted upon victory. So far, in materials science, reinforcement learning has been applied to a lower extent than supervised or unsupervised learning strategies. Nevertheless, reinforcement-based training strategies were shown to be suitable for controlling the growth of microbial co-cultures in bioreactors[139] and for automatically designing RNA sequences with desired secondary structures.[140]

1. Q-learning

Q-learning is a simple but efficient method to teach an algorithm to automatically act and react in the context of playing a game or to perform certain workflows. By repeatedly (over thousands or even millions of trials) exploring all available actions during the training phase and iteratively assessing their quality based on the final received reward, the algorithm learns to identify the best available action for a given state.

A major advantage of Q-learning is that it does not require an actual model of the environment. The algorithm does not undergo any explicit external teaching step but learns on its own by autonomously exploring the possible options. This allows for gaining competence in areas that might otherwise remain unexplored by humans. Such wide-ranging exploration, however, can easily become computationally expensive. Another drawback is that—in its basic form—Q-learning is only useful for stationary environments; for non-stationary problems, new training is required to adapt the decision values. However, there are several modified versions of Q-learning, where these issues are dealt with.[105–107]

D. Deep learning

In addition to the learning strategies discussed so far, there are also "deep learning" approaches. Deep learning can be performed in a supervised, unsupervised, or reinforced manner and aims at mimicking the anatomical structure of biological neural networks and the decision-making process of the human brain. Therefore, multi-hierarchical structures of algorithms are established that can handle and analyze data at different levels of abstraction. This approach holds the potential to analyze even highly complex problems but comes at a prize: Owing to the autonomous, multi-stage data processing procedure, such algorithms act as a black-box. In addition to the provided input, only the generated results are accessible: It remains concealed how exactly the algorithm arrived at a particular decision, and this makes it difficult to rationalize the models suggested by deep learning. Nevertheless, deep learning models have demonstrated tremendous success across a plethora of research areas including biomaterials science; for instance, they precisely predicted the skin permeation behavior of drugs released from biopolymeric films,[141] supported the design of anti-fouling polymer coatings and materials,[142–145] size-tunable poly(lactic-co-glycolic acid) particles,[146] or nucleus-targeting polypeptides,[147] they successfully detected single molecule activity from patch-clamp electrophysiology trials,[148] and they could accurately model biopolymerization processes.[149,150]

1. Deep neural networks (DNNs)

Deep neural networks (DNNs) denote digital constructs that mimic the architecture and mode of operation of the human brain. Here, the key players are artificial neurons—small, digital units that can be triggered with a (typically) non-linear activation function. Those neurons are structured in subsequent, interconnected layers, and the individual computations made by each neuron are eventually combined into a final output. Each neuron transforms the received input variables and transmits the result to the next layer. Between each input and output layer, there can be a variable number of "hidden" layers comprising different numbers of neurons with distinct activation functions. A basic example of a DNN making use of forward-only data processing is the so-called multi-layer perceptron (MLP). MLPs are suitable for supervised learning problems (both, regression and classification tasks) and are basically able to model any non-linear function, which is why they are also referred to as "universal function approximators." Recurrent neural networks (RNNs) are extensions of such DNNs and aim at including more complex information into the decision-making process: Different from MLPs, RNNs combine information from preceding and subsequent layers with the goal of not only to analyze single elements but also to consider their context as well.

DNNs are especially suitable for large-scale datasets, for problems that are too complex for other ML algorithms, and when the problem space is not well understood. Their architecture can be flexibly adapted to other problems, applications, learning strategies, or data types. These networks are able to handle data of high dimensionality, can analyze problems at different levels of abstraction, and learn progressively over time. For DNNs to outperform other ML techniques, though, usually a very large amount of data is needed, and this comes with high computational costs. However, once the costly training phase is completed, making predictions on query samples can be very fast. For instance, a deep model that learned to segment and track cells from microscopy images (which involved large experimental and computational costs) was able, after training, to perform segmentation tasks in less than a second.[151] Owing to the high complexity of DNNs in combination with the low transparency of their decision-making process, choosing the right approach and interpreting the obtained results or models can be extremely challenging.

2. Convolutional neural networks (CNNs)

When aiming at processing images or videos, convolutional neural networks (CNNs) usually are the method of choice. When given an image as an input, CNNs use trainable weights to assign importance gradings to different aspects of an image or to objects within the image. The networks can then be used to analyze or classify images, or to identify trained objects within an image. For this purpose, CNNs mainly make use of three procedures: convolution, pooling, and flattening. For image convolution, filters are applied to each pixel. This can help the network to identify certain structures such as edges or peaks. Pooling can lower the computational cost by combining pixels from the same region into one, thus reducing the size of the image. After applying (multiple) convolution and pooling steps, the individual pixels of the resulting image matrix are fed into a standard neural network—a process, which is referred to as "flattening."

III. SELECTED EXAMPLES OF MACHINE LEARNING APPLICATIONS FROM DIFFERENT BIOSCIENCES

For years, ML approaches have been an integral part of many scientific areas and have been used to develop computer vision for autonomous systems,[152,153] to design synthetic materials,[154,155] or for human behavioral analysis.[156–159] Yet, their application in biophysics or biomaterials science has been less frequent. The scientific questions addressed in these bio-disciplines are characterized by a very high complexity that arises from biological variance and, thus, noisy, divergent data. Hence, it can be quite challenging to translate experimental results from those areas into a format that can be well interpreted by ML models and algorithms. However, once this major hurdle is taken, ML approaches can deliver highly valuable insight into bio-based data as well: Implementation of ML was successfully achieved in the fields of biofabrication,[160–165] biosensors and -markers,[166–174] pharmaceutical science,[175–185] pathophysiology,[186–198] biomacromolecule science,[199–210] gene analysis,[211–221] biomaterials,[222–231] and process optimization[232–240] (Fig. 3; for more details, see Table II). In this section, we discuss selected examples from those areas, and we highlight what type of data was used by the different ML algorithms to obtain predictions or classifications that—using classical data analysis approaches—would have either been way more time consuming to achieve or outright impossible.

A. Supervised learning approaches

When applying supervised learning strategies, the researchers still have a good level of control over how the algorithms are trained and what type of predictions they try to achieve. For instance, Tourlomousis *et al.*[241] used a supervised SVM algorithm to investigate the mechano-sensing response of cells to electrospun fibrous materials (Fig. 4). Therefore, they compared the morphologies of cells after they were cultivated on different substrate geometries. They correlated cell morphology parameters (e.g., cell area, ellipticity, or number of focal adhesions per cell) obtained from confocal microscopy images with architectural features of the substrate (e.g., fiber diameter, pore size, or degree of uniform fiber alignment). With this ML strategy, it was possible to investigate yet unexplored design spaces to yield specific designs qualified at the single-cell level. The authors demonstrated that certain geometrical characteristics of fiber-based materials can be mapped onto unique aspects of cell morphologies—and this is an important step toward a shape-driven pathway to controlling cellular phenotypes.

Other studies went beyond purely analyzing datasets and used the knowledge generated by ML algorithms to tailor materials for specific applications. For instance, Sujeeun *et al.*[242] utilized multiple supervised learning algorithms for the development of scaffolds for tissue regeneration; such scaffolds are typically used to provide structural support for cell attachment and to enable cell proliferation. Here, the main challenge was to browse through a plethora of available polymeric materials to identify the most suitable candidate that meets specified requirements regarding, e.g., biocompatibility, biodegradability, mechanical strength, porosity, and wound healing behavior. To do so, *in vitro* cell viability data (obtained from an MTT [3–(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide] assay) were

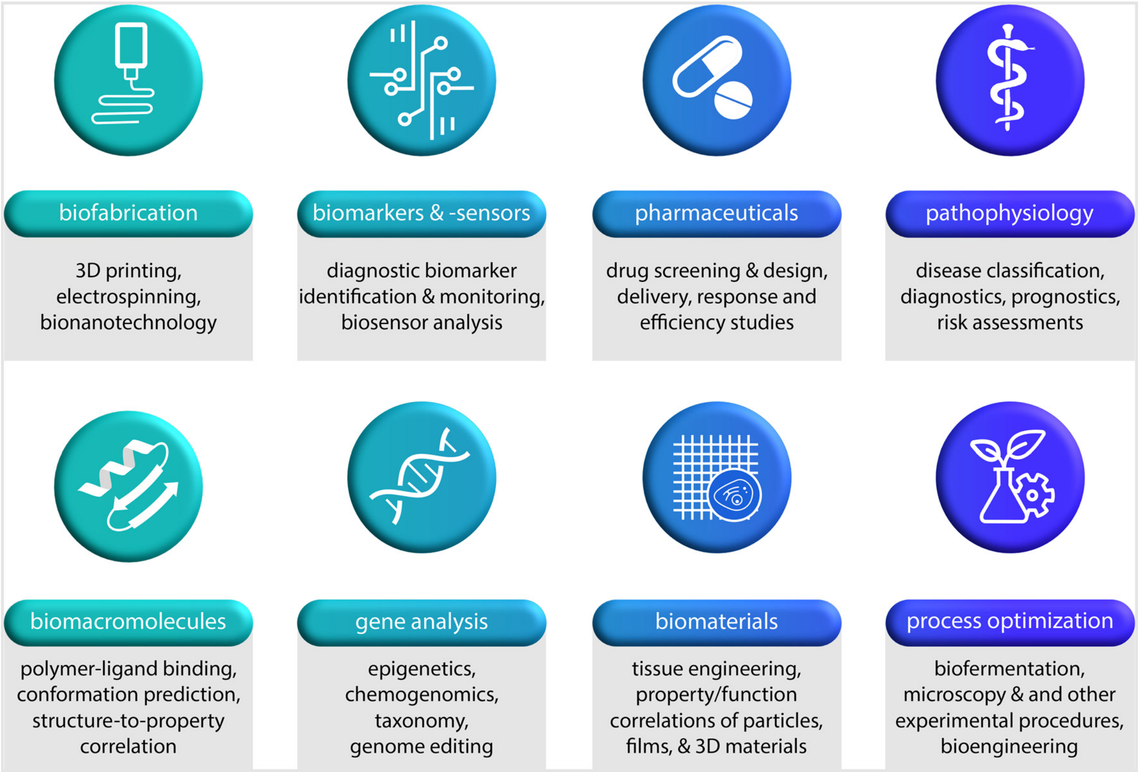
biofabrication
3D printing, electrospinning, bionanotechnology
biomarkers & -sensors
diagnostic biomarker identification & monitoring, biosensor analysis
pharmaceuticals
drug screening & design, delivery, response and efficiency studies
pathophysiology
disease classification, diagnostics, prognostics, risk assessments
biomacromolecules
polymer-ligand binding, conformation prediction, structure-to-property correlation
gene analysis
epigenetics, chemogenomics, taxonomy, genome editing
biomaterials
tissue engineering, property/function correlations of particles, films, & 3D materials
process optimization
biofermentation, microscopy & and other experimental procedures, bioengineering

FIG. 3. Research areas from the biosciences in which machine learning has already been successfully applied. ML approaches were successfully implemented in different fields dealing with biofabrication, biomarkers and sensors, pharmaceuticals, pathophysiology, biomacromolecules, gene analysis, biomaterials, or process optimization. Biofabrication includes various production methods, such as 3D printing or electrospinning; here, ML can be used for process and quality control or for the *a priori* definition of process parameters. In the context of biomarkers and biosensors, ML can support the identification and the monitoring of diagnostic molecules, and it can assist in the analysis of signals. Pharmaceutical sciences benefit from ML in drug screening and design applications as well as in extensive studies on drug delivery, response, and efficiency. In the context of pathophysiology, ML can help with the classification of diseases as well with diagnostics, prognostics, and the assessment of risk levels. Moreover, a ML-driven analysis of biomacromolecules can help us to investigate polymer-ligand binding, to predict molecule conformations, and to correlate molecular structures with their properties. As part of gene analysis, ML can be employed in the fields of epigenetics, chemogenomics, taxonomy, and genome editing. Biomaterials science and development profit strongly from an ML-driven correlation of properties and functions of different materials including particles, films, or three-dimensional bulk materials. As a final example, process optimization can be achieved by ML-based monitoring and an analysis of microscopy or other experimental procedures/bioengineering processes.

combined with physico-chemical properties (e.g., the dimensions of fibers and pores, Young's modulus, or water contact angles) of different scaffolds to model the material-cell interactions. The established correlations then served for reverse-engineering scaffolds with desired performance. Six basic supervised approaches, including KNN and SVM, were compared, and a random forest classifier achieved the highest accuracy. Moreover, this RF algorithm could provide deeper insight into the identified correlations and demonstrated that two selected material characteristics (the pore and fiber diameter) have the strongest influence on the material-cell interaction. Finally, by performing preliminary *in vivo* biocompatibility experiments, the authors were able to show that the determined correlations also hold true (at least to a certain extent) when the material is placed into a living organism. The authors mention, however, that an integration of more advanced techniques, such as reinforcement learning or transfer learning (see Chap. 4), should be considered to obtain a more generalized and robust model that is applicable to unknown scaffolds.

In addition to those applications related to tissue engineering, basic supervised ML algorithms were proven to be handy for various other tasks: RF models, for example, can support the design of self-assembling dipeptide hydrogels[243] and anti-biofouling surfaces[244] and can supervise 3D bioprinting.[245] With KNN and SVM algorithms, it is possible to differentiate healthy from apoptotic cells,[246] to detect pneumonia[247] or COVID-19[248,249] by extracting features from x-ray images, to diagnose Parkinson based on recordings of speech disorders,[250] and to classify white blood cells.[251] Finally, Naïve Bayes models can classify protein folding patterns,[252] identify post-transcriptional modifications in RNA sequences,[253] and support the detection of brain tumors.[254]

B. Unsupervised learning approaches

Different from supervised learning approaches, unsupervised algorithms process unlabeled data. For instance, Gamage *et al.*[255] employed a *k*-means clustering algorithm to group seismocardiographic signals (SCG) according to the patients' different respiratory states (Fig. 4). SCG is a noninvasive technique that monitors heart function by measuring cardiac-related vibrations on the chest surface.

TABLE II. Overview of studies from various research areas, in which ML was applied.

Question	Approach	Outcome	Study
Predicting biophysical interactions			
Affinity of protein-peptide interactions across multiple protein families	Hierarchical statistical model	Interaction affinities were successfully predicted based on the amino acid sequences and the inferred structured Hamiltonians (mathematical functions that map the state of a system to its energy). The model outperformed both, other computational methods[293–295] and high-throughput experimental assays developed for the same purpose Good performance in high-data and low-data domains	16
Protein-ligand binding	SVM, random forest, gradient boosting tree, and a CNN	Successful prediction of protein-ligand binding affinities based on molecular descriptors obtained from topological models Comparable to or even outperforming other state-of-the-art models[296–299] Powerful feature engineering	63
Compound-protein interactions	Combination of GNNs and CNNs (both supervised); networks were analyzed with neural attention mechanisms	Data-driven representations of compounds (as graphs) and proteins (as sequences of characters) were achieved that proved to be more robust than traditional chemical and biological feature vectors Competitive or even better performance compared to state-of-the-art models[300,301]	64
Wettability of a surface based on its topography	KNN, linear regression, Naïve Bayes, random forest, and a DNN	Successful mapping of surface topography parameters to the wetting behavior of the surfaces Feature elimination was performed to reduce dimensionality and to identify the most influential surface parameters, the choice of which otherwise relies on expert assessment The random forest outperformed the other models	61
Pathogen attachment to macromolecular coatings	Bayesian regularized artificial neural networks	Successful mapping of individual pathogen attachment to copolymers represented by a set of molecular descriptors Multiple-pathogen modeling was achieved	145
Functional interactions between human genes	Decision tree, logistic regression, Naïve Bayes, random forest	Phylogenetic profiling was performed, and the combination with ML considerably improved the prediction of functional interactions between genes The random forest outperformed the other models	217
Cytotoxicity of nanoparticles (NPs)	Association rule mining	Knowledge about the toxicity of inorganic, organic and carbon-based NPs was extracted from the literature NPs properties most relevant for their toxicity were identified with a focus on hidden relationships	257
Molecular analysis			
Identifying polymer states	DNN	Based on a simulated 3D polymer configuration represented by spatial coordinates, the model can identify different configurational patterns Phase transition points identified by the model compared well with those obtained from independent specific-heat calculations	62
Designing functional protein sequences	Generative model	The model was trained on evolutional protein sequence data and, by this, learned sequence constraints A diverse library of nanobody sequences was designed that significantly increases the efficiency of discovering stable, functional nanobodies compared to synthetic libraries	202

TABLE II. *(Continued.)*

Question	Approach	Outcome	Study
Predicting protein liquid–liquid phase separation	DNN	The ML classifier was trained based on a pre-analysis of datasets comprising proteins of different phase separation tendencies and learned the underlying principles of phase separation behavior with similar accuracy to classifiers using knowledge-based features	209
Analyzing the structural folding of proteins	Naïve Bayes, SVM, Bayesian generalized linear model	The classifiers accurately predicted mainfolds of proteins based on provided biophysical properties of the amino acids The Bayesian model outperformed the other two models	252
Investigating structures and functions of proteins	Unsupervised language processing (transformer neural networks)	Based on the amino acid character sequences of more than 250 $\times 10^6$ proteins as an input, knowledge of intrinsic biological properties was developed without supervision	259
Sensing of single molecules	CNN	A CNN was trained to classify translocation events of single molecules based on time-series signals obtained from nanopore sensors The network was able to automatically extract such information with higher accuracies than previously possible	58
		Disease classification	
Automated detection of glaucoma	Modified CNN (DenseNet), decision trees	Multiple different models were combined to automatically detect glaucoma based on medical images as well as demographic and systemic data The model shed light onto features that were previously not considered for diagnosis	36
Predicting the primary origin of cancer	CNN with an attention model	The model was trained based on labeled images of tumors of known primary origin The trained model first classified unknown tumors to be either metastatic or primary; then it predicted its site of origin with high accuracy	114
Detection of brain tumors	Random forest, SVM, decision trees	Based on geometric features extracted from MRI images, the different models were able to distinguish normal from abnormal brain images The SVM had the highest sensitivity for detecting brain tumors, whereas the RF had the highest accuracy	88
Assessing sepsis through biomarker host response	Naïve Bayes, decision trees	Multiple biomarker measures from plasma samples were used to distinguish septic from healthy cohorts with high accuracies Naïve Bayes and decision trees performed better than other classifiers—especially regarding the small data size	168
COVID-19 detection from x-ray images	Pretrained CNN	Transfer learning (based on a CNN trained on images of general objects) was employed to train a CNN to analyze chest x-ray images The model successfully distinguished between healthy patients and those suffering from pulmonary diseases; from ill patients, it could identify those with COVID-19 and marked regions of interest in the x-ray images	249
Classification of EEG signals in dementia	MLP, logistic regression, SVM	Different feature sets extracted from EEG signals obtained from neurological patients were analyzed and used to make highly accurate predictions of cognitive disorders The MLP outperformed the other models, and a combination of two different feature sets was shown to entail the most accurate results	196

TABLE II. (*Continued.*)

Question	Approach	Outcome	Study
		Biomaterials design	
Antifouling polymer brushes	DNN, SVR	A DNN was trained on a benchmark database to rationalize the antifouling properties of existing polymer brushes A functional group-based SVR was then used to design new antifouling polymer brushes that indeed showed excellent protein resistance properties	142
Abiotic nuclear-targeting mini-proteins	Directed, evolution-inspired deep learning	The ML model was provided with data from high-throughput experiments and was then capable of predicting activities of mini-proteins in cells and to decipher sequence-activity predictions for new designs The ML-designed mini-proteins were more effective than any previously known variant	147
Gas-separation polymer membranes	Regression	A rather small set of known polymer membranes (represented by binary fingerprints) and their experimental gas permeability data were used to train the model to predict the gas-separation behavior of a large dataset of polymers that have not been tested for these properties yet Tested membranes produced from the most promising candidates (based on the prediction) were shown to exhibit excellent gas-separation performance	21
Mechanically tough bio-nano-composites	Decision tree and random forest (both as regressors)	Using material compositions linked to the resulting fracture toughness obtained from experimental trials and finite elements analysis, the ML models successfully predicted composition/strength relationships which assist the design of new composites without time-consuming trial-and-error experimentation	68
Stabilized silver clusters	SVM	The algorithm learned how the sequence of 10 base pair DNA strands correlates to the wavelength of fluorescent light emitted from silver-DNA clusters With the motifs extracted from the analysis, the model was able to predict the fluorescence color of silver clusters with DNA sequences of variable length	166
3D-printable bioinks	Regression	Different bioink formulations were evaluated regarding their rheological properties and printability, and a general relationship between those properties was established	59
		Cell image analysis	
Extracting biological information from bright field images	Generative adversarial neural network	After being trained on a dataset comprising bright field and fluorescently labeled cell images, the model was able to virtually stain cellular compartments, which eliminates the need for actual (possibly toxic) staining Quantitative measures of cellular structures were then extracted from the virtually stained images	278
Identifying cell morphologies	Image segmentation, principal component analysis, k-means clustering	Cell contours were first identified by image segmentation. After aligning the cell shapes, a principal component analysis was conducted and the cell shape was reconstructed based on the determined eigen-vectors. Finally, different shape modes were identified by k-means clustering. The protocol is highly automated and very fast in quantifying the cell morphologies	135

TABLE II. (*Continued.*)

Question	Approach	Outcome	Study
Predicting osteogenic differentiation	SVM	Based on the cell morphology recorded after 1 day of incubation on nanofiber scaffolds, a pretrained classifier was able to successfully predict the osteogenic differentiation fate of cells	226
Detecting leukemia	CNN	Characteristic features of white blood cell leukemia were extracted from images and sorted regarding importance By applying statistics-based feature elimination, the model outperformed several CNN-only based models	302
Tracking cell migration	CNN	Stain-free, instance-aware segmentation of cells from phase contrast images was achieved with a CNN and provided unique identifiers for each cell Based on those identifiers, the same cell could be followed in a series of images taken at different times Highly accurate visualization and analysis of cell migration was achieved	237
		Pharmaceutical development	
Analyzing existing drugs regarding their suitability to target SARS-CoV-2	Natural language processing with self-attention mechanism	A pre-trained model was used to predict binding affinities between antiviral drugs (represented as strings) and amino acid sequences of the target proteins without providing explicit structural information on the binding epitope A list of antiviral drugs with good inhibitory potencies against SARS-CoV-2 related proteins was identified	55
Identifying self-aggregating drug formulations	Random forest	First, a RF model was used to identify self-aggregating drugs Then, another RF model precisely predicted the co-aggregation properties of different drugs and excipients and was able to find suitable excipients for a novel drug	177
Generation of anticancer molecules	Conditional generative model	A reinforcement learning-based model was trained to design anticancer molecules with specific drug sensitivity and toxicity properties to target individual transcriptomic profiles Such designed molecules exhibit (*in silico*) comparable physico-chemical properties as existing cancer drugs	271
Predicting cancer patient drug responses	Linear regression, ridge regression, support vector regression	Based on transcriptomic data obtained from 3D culture models, different biomarkers were identified that allow for accurate patient/drug response predictions	273
Identifying drug targets	Naïve Bayes	Multiple different data types were combined to train the model based on a dataset of known molecule/target correlations Novel drug binding targets were predicted	180
		Biofabrication	
Predicting the molecular weight of synthesized biomolecules	MLP, SVM	Biopolymers were synthesized via enzymatic polymerization, and various reaction parameters were tuned to alter the molecular weight of the product An SVM was shown to be highly suitable to predict the molecular weight despite the small training data size	150
Controlling the size of elastin-based particles	K-means clustering	A dataset comprising the properties of elastin-based particles and the corresponding fabrication parameters were analyzed by the clustering algorithm The influence of the fabrication parameters on the size of the created particles was revealed, and this information was used to fine-tune the fabrication process	60

TABLE II. (*Continued.*)

Question	Approach	Outcome	Study
Controlling microbial co-cultures in bioreactors	Q-learning	Process feedback *via* a trained reinforcement learning model successfully supported maintaining populations at pre-defined target levels The model was shown to be robust toward variations in the initial states and targets and outperformed standard control approaches	139
Identifying high-quality printing configurations	Random forest	With the printing conditions (resulting from the material composition) and the printing parameters as inputs, a classification model could distinguish between "high" and "low" quality prints, and a regression model returned a direct quality metric The random forest outperformed a simple linear model	245
Monitoring anomalies in 3D bioprinting	CNN, SVM	SVM models were trained to predict whether a specific defect is directly visible in the image of a printed object A CNN was trained to provide information about the applied printing pattern and the occurring printing anomalies The combined model accurately detected and recognized anomalies in various different printing patterns	274

Since the measured signals are typically a convolution of respiratory movements and heart contractions, a direct comparison of two different measurements is difficult. By subdividing the obtained signals and using the vibration amplitudes in those subsignals as input features to cluster the generated subsequences based on their similarity, the SCG data were automatically separated into classes of different lung volumes (high or low) or different flow directions (inhaling or exhaling process). Indeed, within those categories, a comparison of the vibration signals to assess cardiac health (and to detect anomalies) is feasible. Hence, an ML-supported analysis of SCG signals may eliminate the necessity of additional (simultaneous, but independent) respiratory measurements.

Another unsupervised clustering approach was reported by Helfrecht *et al.*[256] who aimed at identifying secondary and tertiary structures in proteins and rationalizing their formation. Here, the idea was not to use common structural descriptions of molecules that are based on predefined motifs such as intramolecular hydrogen bonds or distinct dihedral angle patterns (two strategies, which often rely on human intuition/approximations and only cover a predefined subset of molecular motifs) but to develop a more general approach that is readily applicable to various macromolecules. Therefore, the positions of all atoms in a given protein backbone were combined into an input vector whose complexity was reduced into 6–10 features based on a principle component analysis; then, a density-based algorithm was employed to cluster those reduced vectors: Regions in the feature space with high data density were defined as clusters that are separated from each other by low density areas. Even though several of the formed molecule clusters can belong to the same category of secondary structures (e.g., α-helices or β-strands), a similarly good over-all classification could be achieved with this ML-based approach as with traditional methods. Furthermore, the authors compared unsupervised and supervised methods: Their example highlighted that a supervised approach is suitable to adapt existing motif definitions or to test whether the chosen input data sufficiently represent the output. Unsupervised learning, in contrast, turned out to be better suitable for finding new patterns in the feature space.

In addition to clustering, which is certainly one of the most important techniques where unsupervised ML is applied for, unsupervised association rule mining was shown to be a useful tool to highlight hidden correlations between data such as between the material properties and production process of nanoparticles and their cytotoxicity.[257] Moreover, unsupervised learning methods were successfully employed for image processing or pattern analysis. For instance, the autonomous detection of characteristic features from abdominal computed tomography (CT) images enabled the reconstruction of CT images captured with low radiation doses.[258] Owing to this reduced radiation exposure, the concomitant risk of side effects for patients (such as developing new cancer) is minimized while sufficient image quality is maintained. Furthermore, unsupervised models were able to rationalize and predict selected functional properties (e.g., the biological activity[259] or thermostability[260]) of proteins based on their sequences only, they could unravel the structure of block copolymer micelles,[261] and they managed to successfully and automatically recognize the origin tissue of metastatic tumor cells.[262]

C. Reinforcement learning approaches

Reinforcement learning does not aim at identifying correlations or classifying samples according to given labels (which are typical goals for supervised and unsupervised learning strategies) but makes use of learning procedures to perform certain actions "correctly." An interesting example for a reinforcement based learning approach was presented by Jafari and Javidi.[263] Here, the researchers tried to obtain a complete prediction of the conformation of a polypeptide based on hydrophobic interactions only (Fig. 4). For this purpose, polypeptides

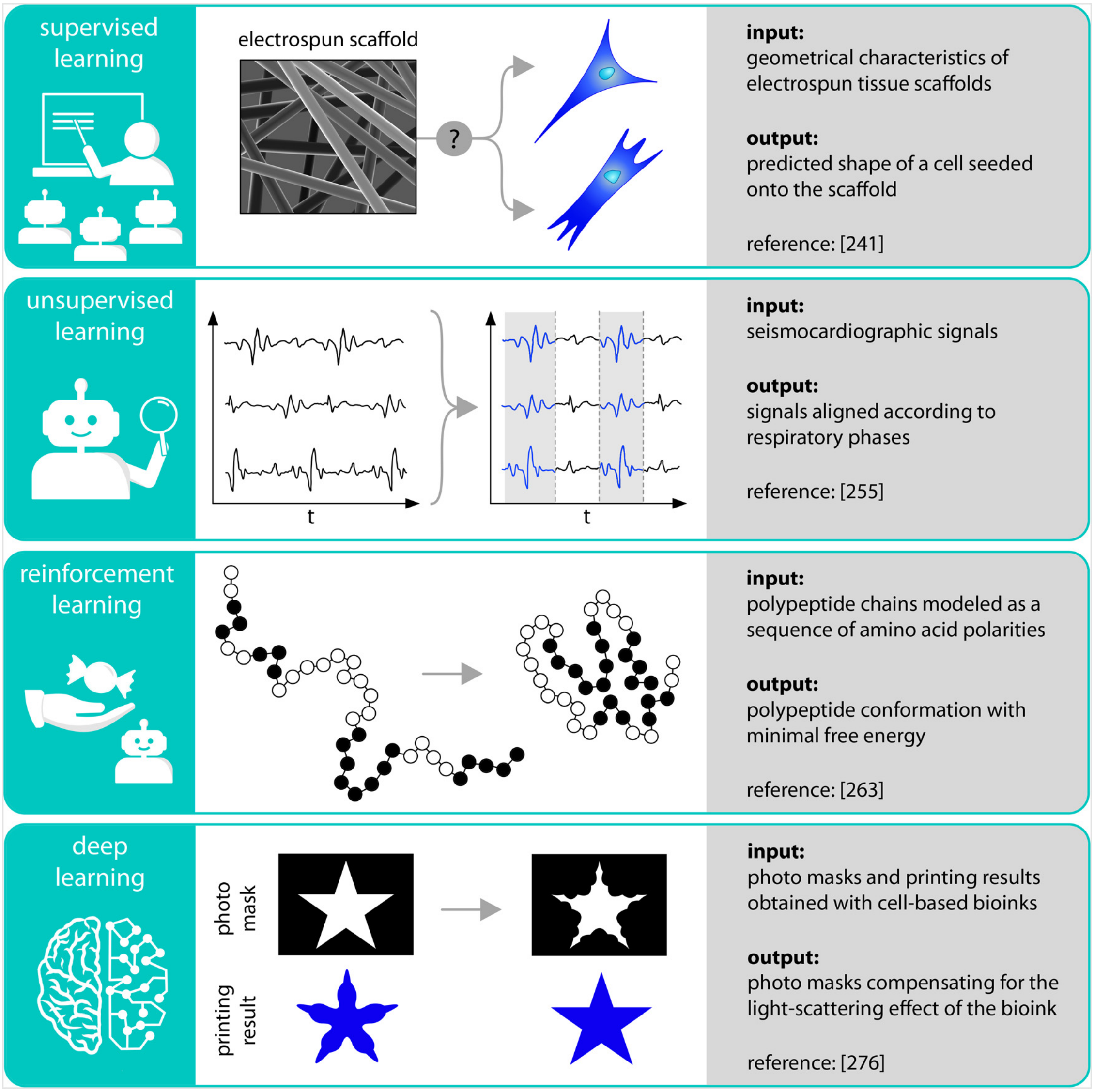
supervised learning
electrospun scaffold
?
input:
geometrical characteristics of electrospun tissue scaffolds
output:
predicted shape of a cell seeded onto the scaffold
reference: [241]
unsupervised learning
t
t
input:
seismocardiographic signals
output:
signals aligned according to respiratory phases
reference: [255]
reinforcement learning
input:
polypeptide chains modeled as a sequence of amino acid polarities
output:
polypeptide conformation with minimal free energy
reference: [263]
deep learning
photo mask
printing result
input:
photo masks and printing results obtained with cell-based bioinks
output:
photo masks compensating for the light-scattering effect of the bioink
reference: [276]

FIG. 4. Schematic representation of selected examples from the biosciences, where ML algorithms have been successfully applied. By using a supervised approach, the geometrical characteristics of electrospun scaffolds were successfully linked to the resulting shape of cells seeded onto the scaffold. An unsupervised ML algorithm could group seismocardiographic signals according to the respiratory phases during which they were acquired to allow for a more direct signal comparison. Reinforcement learning was employed to find energetically optimal conformations of polypeptides. Finally, deep learning was applied to generate photo masks that compensate for the light-scattering effects of cells present in the used bioink.

were modeled as a sequence of amino acid polarities: For instance, the sequence "HHHPP" would represent a polypeptide with three hydrophobic (H) amino acids followed by two polar (P) ones. Then, the possible conformational space of a polypeptide is given as a bidimensional Cartesian grid with two constraints: First, two consecutive amino acids must be vertical or horizontal neighbors in the grid; second, two amino acids cannot be superimposed. To find the ideal overall conformation, a Q-learning algorithm with a dedicated reward function was employed, which aimed at minimizing the free energy of the polypeptide. In this model, the only actions available to the algorithm are moving a given amino acid from its current position in the grid to a neighboring position. With this approach, conformations of minimal free energy were identified (and found to agree with classical calculations using complex models) without explicitly implementing biophysical knowledge; moreover, it was faster than other state-of-the-art approaches. Remarkably, the "long short-term memory" network (a subtype of recurrent neural networks) used in this study proved to be particularly capable of handling sequential data such as chains of amino acids.

Interestingly, reinforcement learning was also successfully used for target-oriented design tasks such as *de novo* drug development: Popova *et al.*[264] employed reinforcement learning to combine two independent supervised learning algorithms; here, the first one was capable of creating drug-like molecules, and the second one could predict certain properties of molecular structures. After individual, supervised training phases of both algorithms (in which either learned how to fulfill its particular task), they were jointly re-trained in a reinforcement approach to deliberately bias the creation of new molecules toward variants with desired properties: The first algorithm received a reward only if the properties predicted by the second matched the predefined goal. By adjusting this reward, the created molecule library was successfully tailored to contain drugs with specific physical properties, biological activity, or chemical substructures. Overall, this study impressively demonstrated how reinforcement learning can be used for generating property-optimized chemical libraries of novel compounds.

Overall, reinforcement learning is currently gaining an importance. It was recently used to control and optimize bioprocesses,[265] to adapt cold atmospheric plasma conditions to optimally eliminate cancer cells,[266] or to identify efficient surgical cardiac ablation strategies for atrial fibrillation.[267] Moreover, reinforcement learning was shown to be useful for controlling tumor growth,[268] to optimize cancer therapy,[269,270] and for the development and dosing of anti-cancer drugs.[271–273]

D. Deep learning approaches

Deep learning is a special subtype of machine learning, where all types of (supervised, unsupervised, or reinforcement) approaches are solved by algorithms that try to mimic the structure and function of the human brain. These algorithms are often difficult to interpret, but they come with the advantage of high variability and the potential to model even highly complex systems. A process-oriented application of deep learning that recently gained considerable importance addresses 3D bioprinting: Here, deep learning-based algorithms can be used for monitoring the printing procedure to determine optimal process parameters or for detecting anomalies in the printed products.[274,275] Moreover, an advanced deep learning approach was demonstrated by Guan *et al.*[276]; here, the researchers set out to compensate for cell-induced light scattering effects in light-based bioprinting—a common fabrication technology used for tissue engineering and regenerative medicine purposes (Fig. 4). To obtain the desired structures, a typical approach is to illuminate a reservoir containing the bioink while using a photo mask that only allows curing in predefined regions. However, the light-scattering effect brought about by cells embedded into the bioink impacts the photopolymerization process and entails a reduced printing resolution. To determine the correlation between the used photo mask and the resulting printing pattern, a convolutional neural network was employed: Pairs of graphical representations of the photo mask on the one hand and the printing result on the other hand were processed with several subsequent convolution and deconvolution steps to model the transformation of the former into the latter. With such a trained network, a photo mask was generated that was supposed to compensate the light-scattering effect of this particular bioink sample based on a desired printing output. Indeed, with this approach, a considerable improvement of the printing resolution was achieved; without the help provided by ML, a similar result would have required an extensive and costly trial-and-error style optimization for each individual structure.

Overall, deep learning techniques have been proven to be particularly useful for processing and analyzing images. This includes assessing the damage mechanics of bone tissue based on microCT images,[277] extracting quantitative properties of cells from bright-field images,[278] or compensating optical errors in microscopy images to obtain reliable images even under difficult conditions.[279] Skärberg *et al.*[280] employed a deep learning approach to analyze images of porous polymer films; here, the aim was to obtain a better understanding of how to tune those materials for controlled drug release. Therefore, they collected combined focused ion beam and scanning electron microscopy images of polymer films with different porosities and fed them into a convolutional neural network for segmentation. From the obtained dataset, 100 images (which corresponds to ∼0.4% of the total dataset) were manually segmented and used for training. To increase the dataset size, those images were subdivided, resulting in over 19×10^6 training samples. The trained CNN was then able to automatically identify pores in the images; thus, important information was retrieved that is needed for further sample analysis but that otherwise could only be gathered through expensive expert assessments. In fact, the results received with the CNN were comparable to manual segmentations and better than those previously obtained with a random forest classifier that was trained on scale-space features. Hence, extending the training set by augmenting data (for more information on this particular method, see Sec. IV) was an important step to achieve a robust ML model capable of competing with actual expert judgments.

The potential applications of deep learning approaches are virtually limitless, and many highly sophisticated neural network architectures have been developed and applied to different problem sets. For instance, generative models, such as generative adversarial neural networks, Gaussian mixture models, or hidden Markov models, are unsupervised approaches that can learn patterns from given input data; then, those models can generate new examples that could plausibly stem from the original dataset. Such algorithms were shown to be useful for the design and discovery of drugs,[281,282] for the development of complex materials with desired elasticity and porosity[283] or tissue engineering-related properties,[284] to create synthetic data (e.g., photo-realistic images[285] or biomedical signals[286]) for network training, and for analytical tasks such as identifying cell morphologies typical for cancer.[287] Whether for an automated evaluation of tumor spheroid behavior in 3D cultures[288] or for identifying cancer based on RNA data,[289] for predicting the *in vivo* fate of nanomaterials based on mass spectrometry,[290] to detect the presence of viral DNA sequences from metagenomic contigs,[291] or to autonomously detect sleep apnea events from electrocardiogram signals,[292] deep learning can be considered the ML equivalent of a Swiss-Army Knife as it can be a helpful tool in many fields of research.

IV. BIGGER IS BETTER BUT HARD TO GET—HOW TO HANDLE SMALL DATA

The performance of all ML algorithms critically depends on the amount of existing knowledge, i.e., the size of the database available for training. Whereas "Big Data" are a phrase commonly used in the context of machine learning, generating large volumes of data from experimental trials is often very challenging: The costs and time requirements associated with experimental studies are typically significant. When the training set is too small, commonly encountered problems include overfitting, biased predictions, or a phenomenon known as the "curse of dimensionality" (Fig. 5). Overfitting refers to algorithms that represent the training data in too much detail. Typically, this happens when a model depicts the variations (and, sometimes, even noise) in the training data to such an extent that it negatively impacts the performance of the model when confronted with new data.[303] Data bias denotes a type of prejudice or favoritism toward a certain class or a decision that is based on wrong assumptions, which are made based on (non-ideal) training data.[304] This can, for instance, occur when the sample set used does not sufficiently represent the whole problem, hence (possibly) neglecting concealed factors or if the model does not properly fit the training data.[305,306] Finally, a prominent issue of small datasets occurs with increasing dimensionality (i.e., with increasing numbers of features added): When the total amount of training data stays the same, the density of data points decreases with every dimension added to a multidimensional feature space, and low data density can lead to reduced accuracy. Thus, a frequently asked question is: How many data points

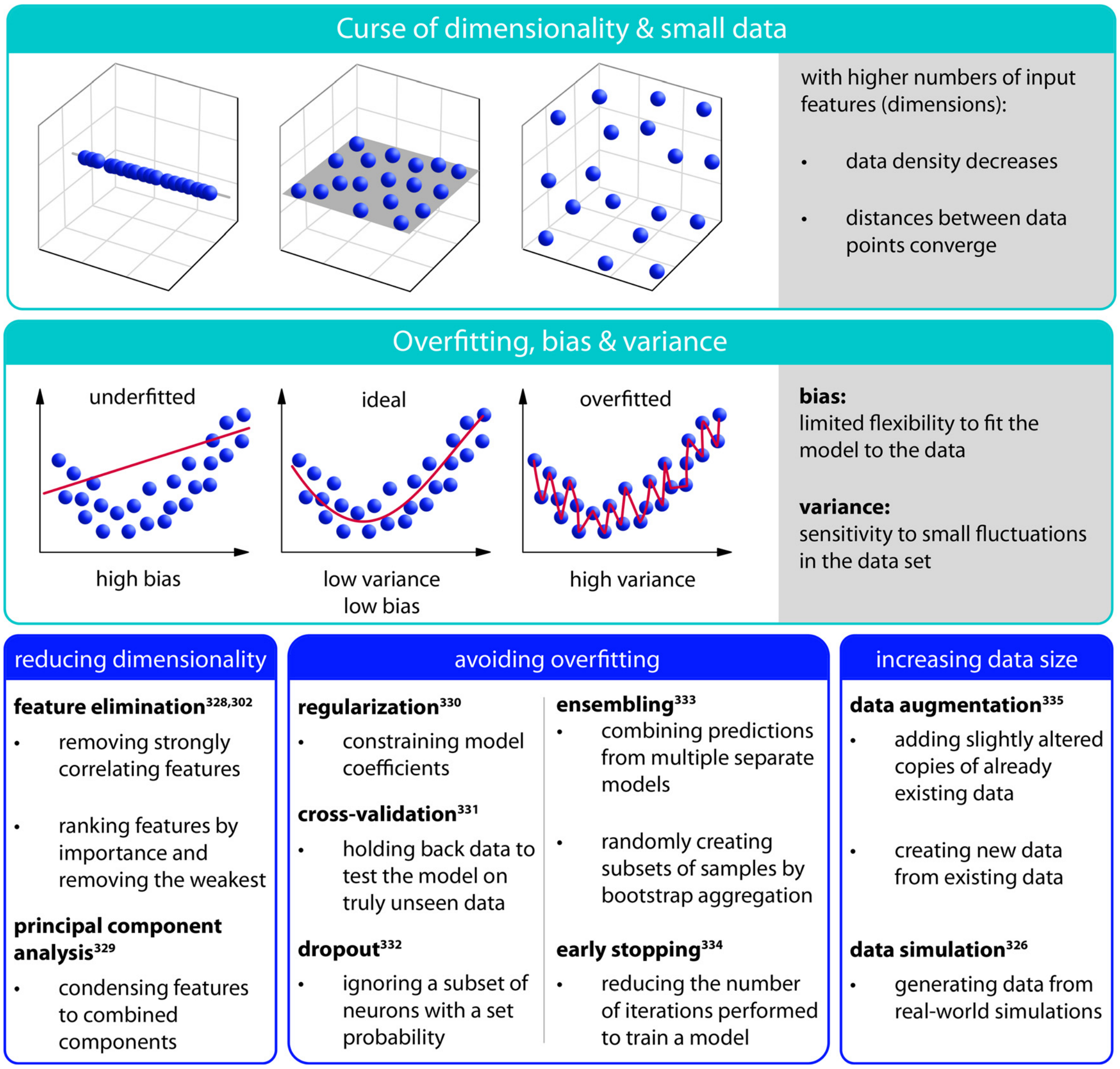
Curse of dimensionality & small data
with higher numbers of input features (dimensions):
• data density decreases
• distances between data points converge
Overfitting, bias & variance
underfitted
ideal
overfitted
high bias
low variance
low bias
high variance
bias:
limited flexibility to fit the model to the data
variance:
sensitivity to small fluctuations in the data set
reducing dimensionality
feature elimination328,302
• removing strongly correlating features
• ranking features by importance and removing the weakest
principal component analysis329
• condensing features to combined components
avoiding overfitting
regularization330
• constraining model coefficients
cross-validation331
• holding back data to test the model on truly unseen data
dropout332
• ignoring a subset of neurons with a set probability
ensembling333
• combining predictions from multiple separate models
• randomly creating subsets of samples by bootstrap aggregation
early stopping334
• reducing the number of iterations performed to train a model
increasing data size
data augmentation335
• adding slightly altered copies of already existing data
• creating new data from existing data
data simulation326
• generating data from real-world simulations

FIG. 5. Typical challenges that can arise when applying ML methods and available remedies to deal with them. High dimensionality, small datasets, overfitting, bias, and variance are common difficulties encountered when using ML. High dimensionality entails a decrease in the data density in the feature space and leads to an equalization of distances between data points. This becomes particularly problematic when datasets are too small to compensate for these effects. Overfitting refers to ML models that approximate the training data too well. Overfitted models show a high sensitivity to small fluctuations in the dataset—a phenomenon which is referred to as "high variance." In contrast, when the models are not able to sufficiently capture the relationship between input and output, the model is underfitting the training data. Such limited flexibility to fit the model to the data is called "bias." Those problems, however, can be tackled for the following strategies: Reducing the dimensionality can be achieved by performing a feature elimination[302,328] or by condensing the feature space via a principal component analysis.[329] Overfitting can be avoided or at least reduced by including regularization,[330] early stopping,[334] or dropouts[332] into the ML models, by using multiple independent predictors (ensembles),[333] or by validating the models using cross-validation.[331] Finally, the size of a dataset can be increased by simulating[326] or augmenting data.[335]

are actually necessary to establish robust models that provide reliable results? Answering this question is, however, not trivial as several factors need to be taken into account: the complexity of the problem, the chosen algorithm, the number and type of input features, and the noise level in the available data.

There are some established rules of thumb that can help researchers to navigate this issue: In a regression problem, the number of training samples should be ten times as high as the number of dimensions of the investigated problem; and at least 1000 images per class should be available for computer vision tasks.[307] However, under certain conditions, good prediction accuracies have also been reported for much smaller datasets. For instance, Shaikhina *et al.*[308] successfully established a deep neural network for predicting the compressive strength of human trabecular bones in severe osteoarthritic conditions, and they could achieve this by using data from 35 bone specimens only. Here, the versatile design of DNNs came in handy: The number of hidden layers as well as the number or neurons and their activation functions were iteratively adjusted until the predictive accuracy of the model reached a maximum. Similarly, basic (non-deep) ML models can be optimized with respect to both, the desired problem and the available dataset: Every ML model is characterized by a set of distinct parameters, which are typically referred to as hyperparameters. Examples for such hyperparameters are the number of neighbors considered in a KNN model, the allowed dimensions of the trees in a RF model, or the set amount of penalty for misclassified samples in an SVM; however, also more advanced parameters can be adjusted. With such optimized algorithms, even fewer than 65 samples were shown to be sufficient to train various algorithms including RF or SVM models.[309,310]

Another important realization in this context is that, even though each research topic is distinct, most questions asked are not entirely unique. Thus, machine learning models that were trained for a certain task can often be used as a starting point for similar problems (this is referred to as transfer learning).[311] Then, only few data points of the target problem are needed to transfer models generated from the source task to the target task—a procedure known as few-shot[312] or even one-shot learning.[313] With this approach, neural networks trained on large-scale image datasets of various macroscopic objects were successfully employed to classify electroencephalogram (EEG) signals obtained from patients diagnosed with delirium,[314] or to identify diseases on grape leaves.[315]

Of course, no algorithm can generate knowledge where no data exist—all models are based on the assumption that the training data cover a suitable and representative subset of the problem at hand. Inter- or extrapolation procedures can (to a certain extent) fill in local gaps, where data are missing, but the machines and models generated by them will only be as reliable as the data fed into them. Even though there might not be a pre-trained algorithm for every research problem, there is a huge amount of data documented in the literature or even stored in readily accessible repositories. From those sources, it is often possible to selectively extract a subset of data to complement one's own dataset, thus increasing the amount of training data. Intriguingly, the collection of such supplemental data is not limited to data already available in a numerical form; especially the extraction of data from texts has been quite successful recently:[316] for instance, unsupervised algorithms were—without having been provided with explicit chemical knowledge—able to understand the structure of the periodic table from text-based sources only, and they could recognize complex structure-property relationships of materials for specific applications, such as energy conversion,[317] nanomedicine,[318] or pharmaceutics,[319] even years before they were actually realized.[320,321]

Gathering data from various sources can, of course, involve considerable effort in terms of retrieving and formatting. Other—possibly less expensive—approaches to extend the training dataset (to improve model generalization and robustness) make use of augmented or synthetic data. Data augmentation refers to a strategy where slightly altered copies of existing data are added to the training set. In the case of images, for example, augmented data can be created by rotating, shifting, splitting, zooming, or flipping the original pixel matrix.[322] With these transformations, Liang *et al.*[323] used 48 microscopy images obtained from collagenous tissue to create >300 000 training images; with this augmented dataset, they then successfully trained a CNN to predict non-linear stress-strain responses of the tissue. Importantly, such an approach is not limited to images—also other data types can be augmented, e.g., by superimposing random noise[324] or by adding synthetically generated features; examples for the latter include crude estimations of the property-to-predict[325] or calculated characteristics derived from empirical models.[326] When training samples are created entirely from simulations, this is referred to as synthetic or *in silico* data. Indeed, by complementing experimental datasets with large amounts of such *in silico* data, Tulsyan *et al.*[327] were able to develop a reliable ML-based monitoring system for biopharmaceutical manufacturing processes—a task that was previously very difficult due to the lack of data.

V. CONCLUSION AND OUTLOOK

Ongoing challenges encountered in the context of ML include having to deal with insufficient data quality, data scarcity, under- or overfitting of the models on the training data, biased training sets, and high computational costs. Indeed, for a long time, the application of ML techniques for bio-related research questions has been severely restricted by the range of difficulties associated with such problems, i.e., small datasets, complex problem definitions, and biological variability. However, some of those issues can now successfully be dealt with: Once the research questions have been translated into computer-readable formats, various methods can be used to increase the data density and to optimize the models in a way that common problems, such as overfitting and bias, are reduced. Even though the training phase of such algorithms might be computationally and/or experimentally costly, once trained, the models can make predictions very quickly.[328–335]

The black-box character of most deep learning methods and the increasing complexity of advanced algorithms in combination with the lack of experienced users especially entails a completely new set of hurdles on the path to fully exploiting the potential of ML. ML nowadays includes a diverse spectrum of different algorithms that can be employed for a plethora of different purposes, and the continuous advancement and expansion of the ML portfolio open up an ever-increasing number of possible applications in all kinds of scientific areas. Generative adversarial neural networks, for example, have successfully been employed to mimic any type of data (including images, numerical, or binary data), which then can be used to either increase the training dataset and/or to generate results. As neural networks are automatically developed inspired by human evolution, evolutionary machine learning approaches can decrease the required expert knowledge needed for creating deep ML models. Attention mechanisms are very recent but promising strategies to improve deep model performances by putting a stronger focus on a few, more relevant aspects while paying less attention to the rest. Finally, by integrating statistical properties into variables, Bayesian neural networks are especially suitable for research problems dealing with sparse data. With these improved techniques available now, current ML models are well-equipped to explore the diverse range of structures, effects, and mechanisms of bio-related systems in more detail, and it is clear that we will encounter many more exciting results in the near future.

ACKNOWLEDGMENTS

The authors thank Jochen Mück for helpful discussions regarding terminology. This project was conducted in the framework of the innovation network "ARTEMIS" by the Technical University of Munich.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

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

REFERENCES

[1] R. Patriat, J. Niederer, J. Kaplan, S. A. Huffmaster, M. Petrucci, L. Eberly, N. Harel, and C. MacKinnon, "Morphological changes in the subthalamic nucleus of people with mild-to-moderate Parkinson's disease: A 7T MRI study," Sci. Rep. **10**(1), 8785 (2020).

[2] L. Wang, Y. Yan, L. Zhang, Y. Liu, R. Luo, and Y. Chang, "Substantia nigra neuromelanin magnetic resonance imaging in patients with different subtypes of Parkinson disease," J. Neural Transm. **128**(2), 171–179 (2021).

[3] S. Dorbala, S. Cuddy, and R. H. Falk, "How to image cardiac amyloidosis: A practical approach," Cardiovasc. Imaging **13**(6), 1368–1383 (2020).

[4] M. Abdelrahman, E. W. Reutzel, A. R. Nassar, and T. L. Starr, "Flaw detection in powder bed fusion using optical imaging," Addit. Manuf. **15**, 1–11 (2017).

[5] C. Gobert, E. W. Reutzel, J. Petrich, A. R. Nassar, and S. Phoha, "Application of supervised machine learning for defect detection during metallic powder bed fusion additive manufacturing using high resolution imaging," Addit. Manuf. **21**, 517–528 (2018).

[6] C. K. Groschner, C. Choi, and M. C. Scott, "Machine learning pipeline for segmentation and defect identification from high-resolution transmission electron microscopy data," Microsc. Microanal. **27**(3), 549–556 (2021).

[7] D. D. Rhoads, "Computer vision and artificial intelligence are emerging diagnostic tools for the clinical microbiologist," J. Clin. Microbiol. **58**(6), e00511–e00520 (2020).

[8] X. Zhu, A. Mohsin, W. Q. Zaman, Z. Liu, Z. Wang, Z. Yu, X. Tian, Y. Zhuang, M. Guo, and J. Chu, "Development of a novel noninvasive quantitative method to monitor Siraitia grosvenorii cell growth and browning degree using an integrated computer-aided vision technology and machine learning," Biotechnol. Bioeng. **118**(10), 4092–4104 (2021).

[9] C. Fei, X. Cao, D. Zang, C. Hu, C. Wu, E. Morris, J. Tao, T. Liu, and G. Lampropoulos, "Machine learning techniques for real-time UV-Vis spectral analysis to monitor dissolved nutrients in surface water," in *AI and Optical Data Sciences II* (International Society for Optics and Photonics, 2021), Vol. 11703, p. 117031D.

[10] D. J. Roach, A. Rohskopf, C. M. Hamel, W. D. Reinholtz, R. Bernstein, H. J. Qi, and A. W. Cook, "Utilizing computer vision and artificial intelligence algorithms to predict and design the mechanical compression response of direct ink write 3D printed foam replacement structures," Addit. Manuf. **41**, 101950 (2021).

[11] A. Martynenko, "Computer vision for real-time control in drying," Food Eng. Rev. **9**(2), 91–111 (2017).

[12] M. M. Alam and M. T. Islam, "Machine learning approach of automatic identification and counting of blood cells," Healthcare Technol. Lett. **6**(4), 103–108 (2019).

[13] C. Yamanishi, E. Parigoris, and S. Takayama, "Kinetic analysis of label-free microscale collagen gel contraction using machine learning-aided image analysis," Front. Bioeng. Biotechnol. **8**, 1–8 (2020).

[14] S. Park, J. W. Ahn, Y. Jo, H.-Y. Kang, H. J. Kim, Y. Cheon, J. W. Kim, Y. Park, S. Lee, and K. Park, "Label-free tomographic imaging of lipid droplets in foam cells for machine-learning-assisted therapeutic evaluation of targeted nanodrugs," ACS Nano **14**(2), 1856–1865 (2020).

[15] V. Spanoudaki, J. C. Doloff, W. Huang, S. R. Norcross, S. Farah, R. Langer, and D. G. Anderson, "Simultaneous spatiotemporal tracking and oxygen sensing of transient implants *in vivo* using hot-spot MRI and machine learning," Proc. Natl. Acad. Sci. **116**(11), 4861–4870 (2019).

[16] J. M. Cunningham, G. Koytiger, P. K. Sorger, and M. AlQuraishi, "Biophysical prediction of protein–peptide interactions and signaling networks using machine learning," Nat. Methods **17**(2), 175–183 (2020).

[17] P. Jones, F. Coupette, A. Härtel, and A. A. Lee, "Bayesian unsupervised learning reveals hidden structure in concentrated electrolytes," J. Chem. Phys. **154**(13), 134902 (2021).

[18] J. C. Clauser, J. Maas, J. Arens, T. Schmitz-Rode, U. Steinseifer, and B. Berkels, "Automation of hemocompatibility analysis using image segmentation and supervised classification," Eng. Appl. Artif. Intell. **97**, 104009 (2021).

[19] A. Chu, D. Nguyen, S. S. Talathi, A. C. Wilson, C. Ye, W. L. Smith, A. D. Kaplan, E. B. Duoss, J. K. Stolaroff, and B. Giera, "Automated detection and sorting of microencapsulation via machine learning," Lab Chip **19**(10), 1808–1817 (2019).

[20] R. M. Madiona, D. A. Winkler, B. W. Muir, and P. J. Pigram, "Optimal machine learning models for robust materials classification using ToF-SIMS data," Appl. Surf. Sci. **487**, 773–783 (2019).

[21] J. W. Barnett, C. R. Bilchak, Y. Wang, B. C. Benicewicz, L. A. Murdock, T. Bereau, and S. K. Kumar, "Designing exceptional gas-separation polymer membranes using machine learning," Sci. Adv. **6**(20), eaaz4301 (2020).

[22] N. Liang, B. Li, Z. Jia, C. Wang, P. Wu, T. Zheng, Y. Wang, F. Qiu, Y. Wu, and J. Su, "Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning," Nat. Biomed. Eng. **5**, 586–599 (2021).

[23] C. Campano, P. Lopez-Exposito, L. Gonzalez-Aguilera, Á. Blanco, and C. Negro, "In-depth characterization of the aggregation state of cellulose nanocrystals through analysis of transmission electron microscopy images," Carbohydr. Polym. **254**, 117271 (2021).

[24] F. S. Ruggeri, P. Flagmeier, J. R. Kumita, G. Meisl, D. Y. Chirgadze, M. N. Bongiovanni, T. P. Knowles, and C. M. Dobson, "The influence of pathogenic mutations in α-synuclein on biophysical and structural characteristics of amyloid fibrils," ACS Nano **14**(5), 5213–5222 (2020).

[25] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Med. Image Anal. **42**, 60–88 (2017).

[26] P. Yin, R. Yuan, Y. Cheng, and Q. Wu, "Deep guidance network for biomedical image segmentation," IEEE Access **8**, 116106–116116 (2020).

[27] R. LaLonde, Z. Xu, I. Irmakci, S. Jain, and U. Bagci, "Capsules for biomedical image segmentation," Med. Image Anal. **68**, 101889 (2021).

[28] A. Sekuboyina, M. E. Husseini, A. Bayat, M. Löffler, H. Liebl, H. Li, G. Tetteh, J. Kukačka, C. Payer, and D. Štern, "VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images," Med. Image Anal. **73**, 102166 (2021).

[29] S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, and M. Rudy, "Ilastik: Interactive machine learning for (bio) image analysis," Nat. Methods **16**(12), 1226–1232 (2019).

[30] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," J. Digital Imaging **32**(4), 582–596 (2019).

[31] H. Li, A. Menegaux, B. Schmitz-Koep, A. Neubauer, F. J. Bäuerlein, S. Shit, C. Sorg, B. Menze, and D. Hedderich, "Automated claustrum segmentation in human brain MRI using deep learning," Hum. Brain Mapp. **42**(18), 5862–5872 (2021).
[32] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," IEEE Access **7**, 128837–128868 (2019).
[33] Z.-Q. Zhao, P. Zheng, S-t. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE Trans. Neural Networks Learn. Syst. **30**(11), 3212–3232 (2019).
[34] X. Wang, H. Chen, A.-R. Ran, L. Luo, P. P. Chan, C. C. Tham, R. T. Chang, S. S. Mannil, C. Y. Cheung, and P.-A. Heng, "Towards multi-center glaucoma OCT image screening with semi-supervised joint structure and function multi-task learning," Med. Image Anal. **63**, 101695 (2020).
[35] G. An, K. Omodaka, K. Hashimoto, S. Tsuda, Y. Shiga, N. Takada, T. Kikawa, H. Yokota, M. Akiba, and T. Nakazawa, "Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images," J. Healthcare Eng. **2019**, 4061313.
[36] P. Mehta, C. A. Petersen, J. C. Wen, M. R. Banitt, P. P. Chen, K. D. Bojikian, C. Egan, S.-I. Lee, M. Balazinska, and A. Y. Lee, "Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images," Am. J. Ophthalmol. **231**, 154–169 (2021).
[37] M. Bruun, J. Koikkalainen, H. F. Rhodius-Meester, M. Baroni, L. Gjerum, M. van Gils, H. Soininen, A. M. Remes, P. Hartikainen, and G. Waldemar, "Detecting frontotemporal dementia syndromes using MRI biomarkers," NeuroImage: Clinical **22**, 101711 (2019).
[38] C. V. Dolph, M. Alam, Z. Shboul, M. D. Samad, and K. M. Iftekharuddin, "Deep learning of texture and structural features for multiclass Alzheimer's disease classification," in *2017 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2017), pp. 2259–2266.
[39] V. Kasivisvanathan, A. S. Rannikko, M. Borghi, V. Panebianco, L. A. Mynderse, M. H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, and R. G. Hindley, "MRI-targeted or standard biopsy for prostate-cancer diagnosis," New Engl. J. Med. **378**(19), 1767–1777 (2018).
[40] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (IEEE, 2018), pp. 1–4.
[41] S. L. Goldenberg, G. Nir, and S. E. Salcudean, "A new era: Artificial intelligence and machine learning in prostate cancer," Nat. Rev. Urol. **16**(7), 391–403 (2019).
[42] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review," Expert Syst. Appl. **167**, 114161 (2021).
[43] J. C. Caicedo, J. Roth, A. Goodman, T. Becker, K. W. Karhohs, M. Broisin, C. Molnar, C. McQuin, S. Singh, and F. J. Theis, "Evaluation of deep learning strategies for nucleus segmentation in fluorescence images," Cytom. Part A **95**(9), 952–965 (2019).
[44] F. Englbrecht, I. E. Ruider, and A. R. Bausch, "Automatic image annotation for fluorescent cell nuclei segmentation," PLoS One **16**(4), e0250093 (2021).
[45] A. M. Gracioso Martins, M. D. Wilkins, F. S. Ligler, M. A. Daniele, and D. O. Freytes, "Microphysiological system for high-throughput computer vision measurement of microtissue contraction," ACS Sens. **6**(3), 985–994 (2021).
[46] W. L. Ng, A. Chan, Y. S. Ong, and C. K. Chua, "Deep learning for fabrication and maturation of 3D bioprinted tissues and organs," Virtual Phys. Prototyping **15**(3), 340–358 (2020).
[47] A. J. Radcliffe and G. V. Reklaitis, "An application of computer vision for optimal sensor placement in drop printing," in *Computer Aided Chemical Engineering* (Elsevier, 2020), Vol. 48, pp. 457–462.
[48] C. T. Chen and G. X. Gu, "Effect of constituent materials on composite performance: Exploring design strategies via machine learning," Adv. Theory Simul. **2**(6), 1900056 (2019).
[49] J. R. Hattrick-Simpers, J. M. Gregoire, and A. G. Kusne, "Perspective: Composition–structure–property mapping in high-throughput experiments: Turning data into knowledge," APL Mater. **4**(5), 053211 (2016).
[50] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**(7715), 547–555 (2018).
[51] P. F. McMillan, *Machine Learning Reveals the Complexity of Dense Amorphous Silicon* (Nature Publishing Group, 2021).
[52] A. Zhavoronkov, Y. A. Ivanenkov, A. Aliper, M. S. Veselov, V. A. Aladinskiy, A. V. Aladinskaya, V. A. Terentiev, D. A. Polykovskiy, M. D. Kuznetsov, and A. Asadulaev, "Deep learning enables rapid identification of potent DDR1 kinase inhibitors," Nat. Biotechnol. **37**(9), 1038–1040 (2019).
[53] G. Yamanluirt, E. J. Berns, A. Xue, A. Lee, N. Bagheri, M. Mrksich, and C. A. Mirkin, "Exploration of the nanomedicine-design space with high-throughput screening and machine learning," in *Spherical Nucleic Acids* (Jenny Stanford Publishing, 2020), pp. 1687–1716.
[54] S. Rodriguez, C. Hug, P. Todorov, N. Moret, S. A. Boswell, K. Evans, G. Zhou, N. T. Johnson, B. T. Hyman, and P. K. Sorger, "Machine learning identifies candidates for drug repurposing in Alzheimer's disease," Nat. Commun. **12**(1), 1033 (2021).
[55] B. R. Beck, B. Shin, Y. Choi, S. Park, and K. Kang, "Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model," Comput. Struct. Biotechnol. J. **18**, 784–790 (2020).
[56] D. Bojar, R. K. Powers, D. M. Camacho, and J. J. Collins, "Deep-learning resources for studying glycan-mediated host-microbe interactions," Cell Host Microbe **29**(1), 132–144.e133 (2021).
[57] R. Burkholz, J. Quackenbush, and D. Bojar, "Using graph convolutional neural networks to learn a representation for glycans," Cell Rep. **35**(11), 109251 (2021).
[58] K. Misiunas, N. Ermann, and U. F. Keyser, "QuipuNet: Convolutional neural network for single-molecule nanopore sensing," Nano Lett. **18**(6), 4040–4045 (2018).
[59] J. Lee, S. J. Oh, S. H. An, W.-D. Kim, and S.-H. Kim, "Machine learning-based design strategy for 3D printable bioink: Elastic modulus and yield stress determine printability," Biofabrication **12**(3), 035018 (2020).
[60] J. S. Cobb, A. Engel, M. A. Seale, and A. V. Janorkar, "Machine learning to determine optimal conditions for controlling the size of elastin-based particles," Sci. Rep. **11**(1), 6343 (2021).
[61] C. A. Rickert, E. N. Hayta, D. M. Selle, I. Kouroudis, M. Harth, A. Gagliardi, and O. Lieleg, "Machine learning approach to analyze the surface properties of biological materials," ACS Biomater. Sci. Eng. **7**(9), 4614–4625 (2021).
[62] Q. Wei, R. G. Melko, and J. Z. Chen, "Identifying polymer states by machine learning," Phys. Rev. E **95**(3), 032504 (2017).
[63] Z. Meng and K. Xia, "Persistent spectral–based machine learning (PerSpect ML) for protein-ligand binding affinity prediction," Sci. Adv. **7**(19), eabc5329 (2021).
[64] M. Tsubaki, K. Tomii, and J. Sese, "Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences," Bioinformatics **35**(2), 309–318 (2019).
[65] W. Pronobis, A. Tkatchenko, and K.-R. Müller, "Many-body descriptors for predicting molecular properties with machine learning: Analysis of pairwise and three-body interactions in molecules," J. Chem. Theory Comput. **14**(6), 2991–3003 (2018).
[66] C. Y. Huang, C. J. Cassidy, C. Medrano, and J. T. Kadonaga, "Identification of the human DPR core promoter element using machine learning," Nature **585**(7825), 459–463 (2020).
[67] T. Wang, M. Shao, R. Guo, F. Tao, G. Zhang, H. Snoussi, and X. Tang, "Surrogate model via artificial intelligence method for accelerating screening materials and performance prediction," Adv. Funct. Mater. **31**(8), 2006245 (2021).
[68] V. Daghigh, T. E. Lacy, Jr., H. Daghigh, G. Gu, K. T. Baghaei, M. F. Horstemeyer, and C. U. Pittman, Jr., "Machine learning predictions on fracture toughness of multiscale bio-nano-composites," J. Reinf. Plast. Compos. **39**(15–16), 587–598 (2020).
[69] J. Maillo, S. Ramírez, I. Triguero, and F. Herrera, "kNN-IS: An iterative Spark-based design of the k-nearest neighbors classifier for big data," Knowl.-Based Syst. **117**, 3–15 (2017).
[70] S. S. Mullick, S. Datta, and S. Das, "Adaptive learning-based *k*-nearest neighbor classifiers with resilience to class imbalance," IEEE Trans. Neural Networks Learn. Syst. **29**(11), 5713–5725 (2018).
[71] A. D. Bhat, H. R. Acharya, and H. Srikanth, "A novel solution to the curse of dimensionality in using KNNs for image classification," in *2019 2nd*

International Conference on Intelligent Autonomous Systems (ICoIAS) (IEEE, 2019), pp. 32–36.

[72] A. Pandey and A. Jain, "Comparative analysis of KNN algorithm using various normalization techniques," Int. J. Comput. Network Inf. Secur. **9**(11), 36 (2017).

[73] A. Singh and R. Lakshmiganthan, "Impact of different data types on classifier performance of random forest, Naive Bayes, and k-nearest neighbors algorithms," (IJACSA) International Journal of Advanced Computer Science and Applications **8**(12), 1–10 (2017).

[74] O. Abdelwahab, M. Bahgat, C. J. Lowrance, and A. Elmaghraby, "Effect of training set size on SVM and Naive Bayes for Twitter sentiment analysis," in *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)* (IEEE, 2015), pp. 46–51.

[75] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, "Analysis of dimensionality reduction techniques on big data," IEEE Access **8**, 54776–54788 (2020).

[76] Ö. F. Arar and K. Ayan, "A feature dependent Naive Bayes approach and its application to the software defect prediction problem," Appl. Soft Comput. **59**, 197–209 (2017).

[77] J. Yang, Z. Ye, X. Zhang, W. Liu, and H. Jin, "Attribute weighted Naive Bayes for remote sensing image classification based on cuckoo search algorithm," in *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)* (IEEE, 2017), pp. 169–174.

[78] M. Kikuchi, K. Kawakami, K. Watanabe, M. Yoshida, and K. Umemura, "Unified likelihood ratio estimation for high-to zero-frequency *N*-grams," IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **E104.A**(8), 1059–1074 (2021).

[79] Y. Bai, Z. Sun, B. Zeng, J. Long, L. Li, J. V. de Oliveira, and C. Li, "A comparison of dimension reduction techniques for support vector machine modeling of multi-parameter manufacturing quality prediction," J. Intell. Manuf. **30**(5), 2245–2256 (2019).

[80] W. U. Adiwijaya, E. Lisnawati, A. Aditsania, and D. S. Kusumo, "Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification," J. Comput. Sci. **14**(11), 1521–1530 (2018).

[81] S. Hossain, R. M. Mou, M. M. Hasan, S. Chakraborty, and M. A. Razzak, "Recognition and detection of tea leaf's diseases using support vector machine," in *2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)* (IEEE, 2018), pp. 150–154.

[82] M. Zareapoor, P. Shamsolmoali, D. K. Jain, H. Wang, and J. Yang, "Kernelized support vector machine with deep learning: An efficient approach for extreme multiclass dataset," Pattern Recognit. Lett. **115**, 4–13 (2018).

[83] B. Feizizadeh, M. S. Roodposhti, T. Blaschke, and J. Aryal, "Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping," Arabian J. Geosci. **10**(5), 122 (2017).

[84] M. Achirul Nanda, K. Boro Seminar, D. Nandika, and A. Maddu, "A comparison study of kernel functions in the support vector machine and its application for termite detection," Information **9**(1), 5 (2018).

[85] H. K. Lee and S. B. Kim, "An overlap-sensitive margin classifier for imbalanced and overlapping data," Expert Syst. Appl. **98**, 72–83 (2018).

[86] H. Wang, Y. Shao, S. Zhou, C. Zhang, and N. Xiu, "Support vector machine classifier via $L_{0/1}$ soft-margin loss," IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[87] K. C. Dewi, H. Murfi, and S. Abdullah, "Analysis accuracy of random forest model for Big Data—A case study of claim severity prediction in car insurance," in *2019 5th International Conference on Science in Information Technology (ICSITech)* (IEEE, 2019), pp. 60–65.

[88] M. Thayumanavan and A. Ramasamy, "An efficient approach for brain tumor detection and segmentation in MR brain images using random forest classifier," Concurrent Eng. **29**(3), 266–274 (2021).

[89] X. Zhou, P. Lu, Z. Zheng, D. Tolliver, and A. Keramati, "Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree," Reliab. Eng. Syst. Saf. **200**, 106931 (2020).

[90] Z. Pu, Z. Li, R. Ke, X. Hua, and Y. Wang, "Evaluating the nonlinear correlation between vertical curve features and crash frequency on highways using random forests," J. Transp. Eng., Part A: Syst. **146**(10), 04020115 (2020).

[91] F. B. de Santana, W. B. Neto, and R. J. Poppi, "Random forest as one-class classifier and infrared spectroscopy for food adulteration detection," Food Chem. **293**, 323–332 (2019).

[92] T. Zhu, "Analysis on the applicability of the random forest," in *Journal of Physics: Conference Series* (IOP Publishing, 2020), Vol. 1607, p. 012123.

[93] Y. Y. Aung and M. M. Min, "An analysis of random forest algorithm based network intrusion detection system," in *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)* (IEEE, 2017), pp. 127–132.

[94] M. Capó, A. Pérez, and J. A. Lozano, "An efficient approximation to the K-means clustering for massive data," Knowl.-Based Syst. **117**, 56–69 (2017).

[95] M. Motwani, N. Arora, and A. Gupta, "A study on initial centroids selection for partitional clustering algorithms," in *Software Engineering* (Springer, 2019), pp. 211–220.

[96] N. Nidheesh, K. A. Nazeer, and P. Ameer, "An enhanced deterministic K-means clustering algorithm for cancer subtype prediction from gene expression data," Comput. Biol. Med. **91**, 213–221 (2017).

[97] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP Conference Series: Materials Science and Engineering* (IOP Publishing, 2018), Vol. 336, p. 012017.

[98] C. Yuan and H. Yang, "Research on K-value selection method of K-means clustering algorithm," J **2**(2), 226–235 (2019).

[99] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," Appl. Intell. **48**(12), 4743–4759 (2018).

[100] S. Rathee and A. Kashyap, "Adaptive-miner: An efficient distributed association rule mining algorithm on Spark," J. Big Data **5**(1), 1–17 (2018).

[101] M. Abdel-Basset, M. Mohamed, F. Smarandache, and V. Chang, "Neutrosophic association rule mining algorithm for big data analysis," Symmetry **10**(4), 106 (2018).

[102] F. Chiclana, R. Kumar, M. Mittal, M. Khari, J. M. Chatterjee, and S. W. Baik, "ARM–AMO: An efficient association rule mining algorithm based on animal migration optimization," Knowl.-Based Syst. **154**, 68–80 (2018).

[103] M. Kaushik, R. Sharma, S. A. Peious, M. Shahin, S. B. Yahia, and D. Draheim, "A systematic assessment of numerical association rule mining methods," SN Comput. Sci. **2**(5), 1–13 (2021).

[104] P. Yazgana and A. O. Kusakci, "A literature survey on association rule mining algorithms," Southeast Eur. J. Soft Comput. **5**(1), 5–14 (2016).

[105] S. J. Majeed and M. Hutter, "On Q-learning convergence for non-Markov decision processes," in *International Joint Conference on Artificial Intelligence* (AAAI Press, 2018), pp. 2546–2552.

[106] S. Padakandla, K. Prabuchandran, and S. Bhatnagar, "Reinforcement learning algorithm for non-stationary environments," Appl. Intell. **50**(11), 3590–3606 (2020).

[107] H. Malik and A. Almutairi, "Modified fuzzy-Q-learning (MFQL)-based mechanical fault diagnosis for direct-drive wind turbines using electrical signals," IEEE Access **9**, 52569–52579 (2021).

[108] E. S. Low, P. Ong, and K. C. Cheah, "Solving the optimal path planning of a mobile robot using improved Q-learning," Rob. Auton. Syst. **115**, 143–161 (2019).

[109] L. Yang and M. Wang, "Sample-optimal parametric q-learning using linearly additive features," in *International Conference on Machine Learning* (PMLR, 2019), pp. 6995–7004.

[110] R. M. Cichy and D. Kaiser, "Deep neural networks as scientific models," Trends Cognit. Sci. **23**(4), 305–317 (2019).

[111] R. Miikkulainen, J. Liang, E. Meyerson, A. Rawal, D. Fink, O. Francon, B. Raju, H. Shahrzad, A. Navruzyan, and N. Duffy, "Evolving deep neural networks," in *Artificial Intelligence in the Age of Neural Networks and Brain Computing* (Elsevier, 2019), pp. 293–312.

[112] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," Front. Plant Sci. **10**, 621 (2019).

[113] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," Proc. IEEE **109**(3), 247–278 (2021).

[114] M. Y. Lu, T. Y. Chen, D. F. Williamson, M. Zhao, M. Shady, J. Lipkova, and F. Mahmood, "AI-based pathology predicts origins for cancers of unknown primary," Nature **594**(7861), 106–110 (2021).

[115] G. X. Gu, C.-T. Chen, and M. J. Buehler, "De novo composite design based on machine learning algorithm," Extreme Mech. Lett. **18**, 19–28 (2018).
[116] S. Ghouli, M. R. Ayatollahi, B. Bahrami, and J. Jamali, "*In-situ* optical approach to predict mixed mode fracture in a polymeric biomaterial," Theor. Appl. Fract. Mech. **115**, 103211 (2021).
[117] P. Thanh Noi and M. Kappas, "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery," Sensors **18**(1), 18 (2018).
[118] W. Gao, B.-B. Yang, and Z.-H. Zhou, "On the resistance of nearest neighbor to random noisy labels," e-print arXiv:1607.07526 (2016).
[119] L. Mandal and N. D. Jana, "A comparative study of Naive Bayes and k-NN algorithm for multi-class drug molecule classification," in *2019 IEEE 16th India Council International Conference (INDICON)* (IEEE, 2019), pp. 1–4.
[120] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between multinomial and Bernoulli Naïve Bayes for text classification," in *2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (IEEE, 2019), pp. 593–596.
[121] R. Panigrahi and L. Kumar, "Application of Naïve Bayes classifiers for refactoring prediction at the method level," in *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)* (IEEE, 2020), pp. 1–6.
[122] D. Berrar, "Bayes' theorem and Naive Bayes classifier," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (Elsevier Science Publisher, Amsterdam, 2018), pp. 403–412.
[123] J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data* (O'Reilly Media, Inc., 2016).
[124] Y. Li, Q. Pu, S. Li, H. Zhang, X. Wang, H. Yao, and L. Zhao, "Machine learning methods for research highlight prediction in biomedical effects of nanomaterial application," Pattern Recognit. Lett. **117**, 111–118 (2019).
[125] Z. Zheng, Y. Cai, Y. Yang, and Y. Li, "Sparse weighted Naive Bayes classifier for efficient classification of categorical data," in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)* (IEEE, 2018), pp. 691–696.
[126] Z. K. Abass, T. M. Hasan, and A. K. Abdullah, "Brain computer interface enhancement based on stones blind source separation and Naive Bayes classifier," in *International Conference on New Trends in Information and Communications Technology Applications* (Springer, 2020), pp. 17–28.
[127] L. C. Padierna, M. Carpio, A. Rojas-Dominguez, H. Puga, and H. Fraire, "A novel formulation of orthogonal polynomial kernel functions for SVM classifiers: The Gegenbauer family," Pattern Recognit. **84**, 211–225 (2018).
[128] H. Hong, B. Pradhan, D. T. Bui, C. Xu, A. M. Youssef, and W. Chen, "Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: A case study at Suichuan area (China)," Geomatics Nat. Hazards Risk **8**(2), 544–569 (2017).
[129] X. Shen, L. Niu, Z. Qi, and Y. Tian, "Support vector machine classifier with truncated pinball loss," Pattern Recognit. **68**, 199–210 (2017).
[130] L. Breiman, "Random forests," Mach. Learn. **45**(1), 5–32 (2001).
[131] T.-H. Lee, A. Ullah, and R. Wang, "Bootstrap aggregating and random forest," in *Macroeconomic Forecasting in the Era of Big Data* (Springer, 2020), pp. 389–429.
[132] K. Kirasich, T. Smith, and B. Sadler, "Random forest vs logistic regression: Binary classification for heterogeneous datasets," SMU Data Sci. Rev. **1**(3), 9 (2018).
[133] F. Fabris, A. Doherty, D. Palmer, J. P. De Magalhães, and A. A. Freitas, "A new approach for interpreting random forest models and its application to the biology of ageing," Bioinformatics **34**(14), 2449–2456 (2018).
[134] A. V. Gonçalves, I. J. C. Schneider, F. V. Amaral, L. P. Garcia, and G. M. de Araújo, "Feature importance investigation for estimating COVID-19 infection by random forest algorithm," in *International Conference on Data and Information in Online* (Springer, 2021), pp. 272–285.
[135] J. M. Phillip, K.-S. Han, W.-C. Chen, D. Wirtz, and P.-H. Wu, "A robust unsupervised machine-learning method to quantify the morphological heterogeneity of cells and nuclei," Nat. Protoc. **16**(2), 754–774 (2021).
[136] C. Ieracitano, A. Paviglianiti, M. Campolo, A. Hussain, E. Pasero, and F. C. Morabito, "A novel automatic classification system based on hybrid unsupervised and supervised machine learning for electrospun nanofibers," IEEE/CAA J. Autom. Sin. **8**(1), 64–76 (2020).
[137] L. Bai, J. Liang, and Y. Guo, "An ensemble clusterer of multiple fuzzy *k*-means clusterings to recognize arbitrarily shaped clusters," IEEE Trans. Fuzzy Syst. **26**(6), 3524–3533 (2018).
[138] N. T. Gupta, K. D. Adams, A. W. Briggs, S. C. Timberlake, F. Vigneault, and S. H. Kleinstein, "Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data," J. Immunol. **198**(6), 2489–2499 (2017).
[139] N. J. Treloar, A. J. Fedorec, B. Ingalls, and C. P. Barnes, "Deep reinforcement learning for the control of microbial co-cultures in bioreactors," PLoS Comput. Biol. **16**(4), e1007783 (2020).
[140] P. Eastman, J. Shi, B. Ramsundar, and V. S. Pande, "Solving the RNA design problem with reinforcement learning," PLoS Comput. Biol. **14**(6), e1006176 (2018).
[141] H. Salma, Y. M. Melha, L. Sonia, H. Hamza, and N. Salim, "Efficient prediction of *in vitro* piroxicam release and diffusion from topical films based on biopolymers using deep learning models and generative adversarial networks," J. Pharm. Sciences **110**(6), 2531–2543 (2021).
[142] Y. Liu, D. Zhang, Y. Tang, Y. Zhang, X. Gong, S. Xie, and J. Zheng, "Machine learning-enabled repurposing and design of antifouling polymer brushes," Chem. Eng. J. **420**, 129872 (2021).
[143] T. C. Le, M. Penna, D. A. Winkler, and I. Yarovsky, "Quantitative design rules for protein-resistant surface coatings using machine learning," Sci. Rep. **9**(1), 265 (2019).
[144] M. Echezarreta-López and M. Landin, "Using machine learning for improving knowledge on antibacterial effect of bioactive glass," Int. J. Pharm. **453**(2), 641–647 (2013).
[145] P. Mikulskis, A. Hook, A. A. Dundas, D. Irvine, O. Sanni, D. Anderson, R. Langer, M. R. Alexander, P. Williams, and D. A. Winkler, "Prediction of broad-spectrum pathogen attachment to coating materials for biomedical devices," ACS Appl. Mater. Interfaces **10**(1), 139–149 (2018).
[146] S. A. Damiati, D. Rossi, H. N. Joensson, and S. Damiati, "Artificial intelligence application for rapid fabrication of size-tunable PLGA microparticles in microfluidics," Sci. Rep. **10**(1), 19517 (2020).
[147] C. K. Schissel, S. Mohapatra, J. M. Wolfe, C. M. Fadzen, K. Bellovoda, C.-L. Wu, J. A. Wood, A. B. Malmberg, A. Loas, and R. Gómez-Bombarelli, "Deep learning to design nuclear-targeting abiotic miniproteins," Nat. Chem. **13**(10), 992–1000 (2021).
[148] N. Celik, F. O'Brien, S. Brennan, R. D. Rainbow, C. Dart, Y. Zheng, F. Coenen, and R. Barrett-Jolley, "Deep-channel uses deep neural networks to detect single-molecule events from patch-clamp data," Commun. Biol. **3**(1), 3 (2020).
[149] Y. J. Wong, S. K. Arumugasamy, and J. Jewaratnam, "Performance comparison of feedforward neural network training algorithms in modeling for synthesis of polycaprolactone via biopolymerization," Clean Technol. Environ. Policy **20**(9), 1971–1986 (2018).
[150] S. K. Arumugasamy, Z. Chen, L. D. Van Khoa, and H. Pakalapati, "Comparison between artificial neural networks and support vector machine modeling for polycaprolactone synthesis via enzyme catalyzed polymerization," Process Integr. Optim. Sustainability **5**(3), 599–607 (2021).
[151] J.-B. Lugagne, H. Lin, and M. J. Dunlop, "DeLTA: Automated cell segmentation, tracking, and lineage reconstruction using deep learning," PLoS Comput. Biol. **16**(4), e1007673 (2020).
[152] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao, "Adversarial sensor attack on lidar-based perception in autonomous driving," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2019), pp. 2267–2281.
[153] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," Found. Trends® Comput. Graph. Vision **12**(1–3), 1–308 (2020).
[154] J. Schmidt, M. R. Marques, S. Botti, and M. A. Marques, "Recent advances and applications of machine learning in solid-state materials science," npj Comput. Mater. **5**(1), 83 (2019).
[155] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, "Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning," Phys. Rev. B **99**(6), 064114 (2019).
[156] M. Kashif, A. Hussain, A. Munir, A. B. Siddiqui, A. Abbasi, M. Aakif, A. J. Malik, F. E. Alazemi, and O.-Y. Song, "A machine learning approach for

expression detection in healthcare monitoring systems," Comput. Mater. Continua **67**(2), 2123–2139 (2021).

[157]X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," Comput. Hum. Behav. **98**, 166–173 (2019).

[158]Z. Lv, L. Qiao, and A. K. Singh, "Advanced machine learning on cognitive computing for human behavior analysis," in *IEEE Transactions on Computational Social Systems* (IEEE, 2020), pp. 1194–1202.

[159]J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, and T. L. Griffiths, "Using large-scale experiments and machine learning to discover theories of human decision-making," Science **372**(6547), 1209–1214 (2021).

[160]J. M. Bone, C. M. Childs, A. Menon, B. Poczos, A. W. Feinberg, P. R. LeDuc, and N. R. Washburn, "Hierarchical machine learning for high-fidelity 3D printed biopolymers," ACS Biomater. Sci. Eng. **6**(12), 7021–7031 (2020).

[161]Z. Zhu, D. W. H. Ng, H. S. Park, and M. C. McAlpine, "3D-printed multifunctional materials enabled by artificial-intelligence-assisted fabrication technologies," Nat. Rev. Mater. **6**(1), 27–47 (2021).

[162]Y. Liu, F. Han, F. Li, Y. Zhao, M. Chen, Z. Xu, X. Zheng, H. Hu, J. Yao, and T. Guo, "Inkjet-printed unclonable quantum dot fluorescent anti-counterfeiting labels with artificial intelligence authentication," Nat. Commun. **10**(1), 2409 (2019).

[163]J. D. Toscano, Z. Li, L. J. Segura, and H. Sun, "A machine learning approach to model the electrospinning process of biocompatible materials," in *International Manufacturing Science and Engineering Conference* (American Society of Mechanical Engineers, 2020), Vol. 84263, p V002T006A031.

[164]A. B. Ramzi, S. N. Baharum, H. Bunawan, and N. S. Scrutton, "Streamlining natural products biomanufacturing with omics and machine learning driven microbial engineering," Front. Bioeng. Biotechnol. **8**, 608918 (2020).

[165]T. Oyetunde, F. S. Bao, J.-W. Chen, H. G. Martin, and Y. J. Tang, "Leveraging knowledge engineering and machine learning for microbial bio-manufacturing," Biotechnol. Adv. **36**(4), 1308–1315 (2018).

[166]S. M. Copp, S. M. Swasey, A. Gorovits, P. Bogdanov, and E. G. Gwinn, "General approach for machine learning-aided design of DNA-stabilized silver clusters," Chem. Mater. **32**(1), 430–437 (2019).

[167]E. Becht, D. Tolstrup, C.-A. Dutertre, P. A. Morawski, D. J. Campbell, F. Ginhoux, E. W. Newell, R. Gottardo, and M. B. Headley, "High-throughput single-cell quantification of hundreds of proteins using conventional flow cytometry and machine learning," Sci. Adv. **7**(39), eabg0505 (2021).

[168]A. U. Sardesai, A. S. Tanak, S. Krishnan, D. A. Striegel, K. L. Schully, D. V. Clark, S. Muthukumar, and S. Prasad, "An approach to rapidly assess sepsis through multi-biomarker host response using machine learning algorithm," Sci. Rep. **11**(1), 16905 (2021).

[169]R. F. Rojas, X. Huang, and K.-L. Ou, "A machine learning approach for the identification of a biomarker of human pain using fNIRS," Sci. Rep. **9**(1), 5645 (2019).

[170]O. F. Odish, K. Johnsen, P. van Someren, R. A. Roos, and J. G. van Dijk, "EEG may serve as a biomarker in Huntington's disease using machine learning automatic classification," Sci. Rep. **8**(1), 16090 (2018).

[171]S. Baik, J. Lee, E. J. Jeon, B-y. Park, D. W. Kim, J. H. Song, H. J. Lee, S. Y. Han, S.-W. Cho, and C. Pang, "Diving beetle–like miniaturized plungers with reversible, rapid biofluid capturing for machine learning–based care of skin disease," Sci. Adv. **7**(25), eabf5695 (2021).

[172]H. M. Robison, C. A. Chapman, H. Zhou, C. L. Erskine, E. Theel, T. Peikert, C. S. Lindestam Arlehamn, A. Sette, C. Bushell, and M. Welge, "Risk assessment of latent tuberculosis infection through a multiplexed cytokine biosensor assay and machine learning feature selection," Sci. Rep. **11**(1), 20544 (2021).

[173]E. M. Green, R. van Mourik, C. Wolfus, S. B. Heitner, O. Dur, and M. J. Semigran, "Machine learning detection of obstructive hypertrophic cardiomyopathy using a wearable biosensor," npj Digital Med. **2**(1), 57 (2019).

[174]X. Mi, B. Zou, F. Zou, and J. Hu, "Permutation-based identification of important biomarkers for complex diseases via machine learning models," Nat. Commun. **12**(1), 3008 (2021).

[175]R. Kumar, N. Le, Z. Tan, M. E. Brown, S. Jiang, and T. M. Reineke, "Efficient polymer-mediated delivery of gene-editing ribonucleoprotein payloads through combinatorial design, parallelized experimentation, and machine learning," ACS Nano **14**(12), 17626–17639 (2020).

[176]J. Tréguier, L. Bugnicourt, G. Gay, M. Diallo, S. T. Islam, A. Toro, L. David, O. Théodoly, G. Sudre, and T. Mignot, "Chitosan films for microfluidic studies of single bacteria and perspectives for antibiotic susceptibility testing," mBio **10**(4), e01375-19 (2019).

[177]D. Reker, Y. Rybakova, A. R. Kirtane, R. Cao, J. W. Yang, N. Navamajiti, A. Gardner, R. M. Zhang, T. Esfandiary, and J. L'Heureux, "Computationally guided high-throughput design of self-assembling drug nanoparticles," Nat. Nanotechnol. **16**(6), 725–733 (2021).

[178]S. Golriz Khatami, S. Mubeen, V. S. Bharadhwaj, A. T. Kodamullil, M. Hofmann-Apitius, and D. Domingo-Fernández, "Using predictive machine learning models for drug response simulation by calibrating patient-specific pathway signatures," npj Syst. Biol. Appl. **7**(1), 40 (2021).

[179]I. Piazza, N. Beaton, R. Bruderer, T. Knobloch, C. Barbisan, L. Chandat, A. Sudau, I. Siepe, O. Rinner, and N. de Souza, "A machine learning-based chemoproteomic approach to identify drug targets and binding sites in complex proteomes," Nat. Commun. **11**(1), 4200 (2020).

[180]N. S. Madhukar, P. K. Khade, L. Huang, K. Gayvert, G. Galletti, M. Stogniew, J. E. Allen, P. Giannakakou, and O. Elemento, "A Bayesian machine learning approach for drug target identification using diverse data types," Nat. Commun. **10**(1), 5221 (2019).

[181]H. Kobayashi, C. Lei, Y. Wu, A. Mao, Y. Jiang, B. Guo, Y. Ozeki, and K. Goda, "Label-free detection of cellular drug responses by high-throughput bright-field imaging and machine learning," Sci. Rep. **7**(1), 12454 (2017).

[182]M. Sarmadi, A. M. Behrens, K. J. McHugh, H. T. Contreras, Z. L. Tochka, X. Lu, R. Langer, and A. Jaklenec, "Modeling, design, and machine learning-based framework for optimal injectability of microparticle-based drug formulations," Sci. Adv. **6**(28), eabb6594 (2020).

[183]C. V. Theodoris, P. Zhou, L. Liu, Y. Zhang, T. Nishino, Y. Huang, A. Kostina, S. S. Ranade, C. A. Gifford, V. Uspenskiy, and A. Malashicheva, "Network-based screen in iPSC-derived cells reveals therapeutic candidate for heart valve disease," Science **371**(6530), eabd0724 (2021).

[184]A. Morris, W. McCorkindale, N. Drayman, J. D. Chodera, S. Tay, N. London, and C. M. Consortium, "Discovery of SARS-CoV-2 main protease inhibitors using a synthesis-directed de novo design model," Chem. Commun. **57**, 5909–5912 (2021).

[185]J. Chodera, A. A. Lee, N. London, and F. von Delft, "Crowdsourcing drug discovery for pandemics," Nat. Chem. **12**(7), 581–581 (2020).

[186]D. E. Wood, J. R. White, A. Georgiadis, B. Van Emburgh, S. Parpart-Li, J. Mitchell, V. Anagnostou, N. Niknafs, R. Karchin, and E. Papp, "A machine learning approach for somatic mutation discovery," Sci. Transl. Med. **10**(457), eaar7939 (2018).

[187]A. Huda, A. Castaño, A. Niyogi, J. Schumacher, M. Stewart, M. Bruno, M. Hu, F. S. Ahmad, R. C. Deo, and S. J. Shah, "A machine learning model for identifying patients at risk for wild-type transthyretin amyloid cardiomyopathy," Nat. Commun. **12**(1), 2725 (2021).

[188]S. H. Kim, E.-T. Jeon, S. Yu, O. Kyungmi, C. K. Kim, T.-J. Song, Y.-J. Kim, S. H. Heo, K.-Y. Park, and J.-M. Kim, "Interpretable machine learning for early neurological deterioration prediction in atrial fibrillation-related stroke," Sci. Rep. **11**, 20610 (2021).

[189]C. Ricciardi, K. J. Edmunds, M. Recenti, S. Sigurdsson, V. Gudnason, U. Carraro, and P. Gargiulo, "Assessing cardiovascular risks from a mid-thigh CT image: A tree-based machine learning approach using radiodensitometric distributions," Sci. Rep. **10**(1), 2863 (2020).

[190]P. Jurmeister, M. Bockmayr, P. Seegerer, T. Bockmayr, D. Treue, G. Montavon, C. Vollbrecht, A. Arnold, D. Teichmann, and K. Bressem, "Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases," Sci. Transl. Med. **11**(509), eaaw8513 (2019).

[191]S. Chen, L. Jiang, F. Gao, E. Zhang, T. Wang, N. Zhang, X. Wang, and J. Zheng, "Machine learning-based pathomics signature could act as a novel prognostic marker for patients with clear cell renal cell carcinoma," Br. J. Cancer **126**, 771–777 (2021).

[192]S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, G. H. Chang, A. S. Joshi, B. Dwyer, and S. Zhu, "Development and validation of an interpretable deep learning framework for Alzheimer's disease classification," Brain **143**(6), 1920–1933 (2020).

[193] J. Yu, Y. Zhou, Q. Yang, X. Liu, L. Huang, P. Yu, and S. Chu, "Machine learning models for screening carotid atherosclerosis in asymptomatic adults," Sci. Rep. **11**(1), 22236 (2021).

[194] Y.-C. Chiu, S. Zheng, L.-J. Wang, B. S. Iskra, M. K. Rao, P. J. Houghton, Y. Huang, and Y. Chen, "Predicting and characterizing a cancer dependency map of tumors with deep learning," Sci. Adv. **7**(34), eabh1275 (2021).

[195] L. A. Gemein, R. T. Schirrmeister, P. Chrabąszcz, D. Wilson, J. Boedecker, A. Schulze-Bonhage, F. Hutter, and T. Ball, "Machine-learning-based diagnostics of EEG pathology," NeuroImage **220**, 117021 (2020).

[196] C. Ieracitano, N. Mammone, A. Hussain, and F. C. Morabito, "A novel multi-modal machine learning based approach for automatic classification of EEG recordings in dementia," Neural Networks **123**, 176–190 (2020).

[197] J.-Y. An, K. Lin, L. Zhu, D. M. Werling, S. Dong, H. Brand, H. Z. Wang, X. Zhao, G. B. Schwartz, and R. L. Collins, "Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder," Science **362**(6420), eaat6576 (2018).

[198] D. B. Brückner, N. Arlt, A. Fink, P. Ronceray, J. O. Rädler, and C. P. Broedersz, "Learning the dynamics of cell–cell interactions in confined cell migration," Proc. Natl. Acad. Sci. **118**(7), e2016602118 (2021).

[199] O. Méndez-Lucio, M. Ahmad, E. A. del Rio-Chanona, and J. K. Wegner, "A geometric deep learning approach to predict binding conformations of bioactive molecules," Nat. Mach. Intell. **3**, 1033–1039 (2021).

[200] M. A. Webb, N. E. Jackson, P. S. Gil, and J. J. de Pablo, "Targeted sequence design within the coarse-grained polymer genome," Sci. Adv. **6**(43), eabc6216 (2020).

[201] Y. Uesawa, "Quantitative structure–activity relationship analysis using deep learning based on a novel molecular image input technique," Bioorg. Med. Chem. Lett. **28**(20), 3400–3403 (2018).

[202] J.-E. Shin, A. J. Riesselman, A. W. Kollasch, C. McMahon, E. Simon, C. Sander, A. Manglik, A. C. Kruse, and D. S. Marks, "Protein design and variant prediction using autoregressive generative models," Nat. Commun. **12**(1), 2403 (2021).

[203] Y. Matsunaga and Y. Sugita, "Linking time-series of single-molecule experiments with molecular dynamics simulations by machine learning," Elife **7**, e32668 (2018).

[204] J. Wang, A. Ferguson, and J. W. Team, "Machine learning of protein folding funnels from experimentally measurable observables," in APS March Meeting Abstracts, 2018.

[205] X. Chen, B. Yang, and Z. Lin, "A random forest learning assisted "divide and conquer" approach for peptide conformation search," Sci. Rep. **8**(1), 8796 (2018).

[206] E. Moman, M. A. Grishina, and V. A. Potemkin, "Nonparametric chemical descriptors for the calculation of ligand-biopolymer affinities with machine-learning scoring functions," J. Comput.-Aided Mol. Des. **33**(11), 943–953 (2019).

[207] E. Moebel, A. Martinez-Sanchez, L. Lamm, R. Righetto, W. Wietrzynski, S. Albert, D. Lariviere, E. Fourmentin, S. Pfeffer, and J. Ortiz, "Deep learning improves macromolecule identification in 3D cellular cryo-electron tomograms," Nat. Methods **18**, 1386–1394 (2021).

[208] S. Bandyopadhyay and J. Mondal, "A deep autoencoder framework for discovery of metastable ensembles in biomacromolecules," e-print arXiv:2106.00724 (2021).

[209] K. L. Saar, A. S. Morgunov, R. Qi, W. E. Arter, G. Krainer, and T. P. Knowles, "Learning the molecular grammar of protein condensates from sequence determinants and embeddings," Proc. Natl. Acad. Sci. **118**(15), e2019053118 (2021).

[210] Q. Yang, A. Bassyouni, C. R. Butler, X. Hou, S. Jenkinson, and D. A. Price, "Ligand biological activity predicted by cleaning positive and negative chemical correlations," Proc. Natl. Acad. Sci. **116**(9), 3373–3378 (2019).

[211] S. Sharifi, A. Pakdel, M. Ebrahimi, J. M. Reecy, S. Fazeli Farsani, and E. Ebrahimie, "Integration of machine learning and meta-analysis identifies the transcriptomic bio-signature of mastitis disease in cattle," PLoS one **13**(2), e0191227 (2018).

[212] C. Fu, X. Zhang, A. O. Veri, K. R. Iyer, E. Lash, A. Xue, H. Yan, N. M. Revie, C. Wong, and Z.-Y. Lin, "Leveraging machine learning essentiality predictions and chemogenomic interactions to identify antifungal targets," Nat. Commun. **12**(1), 6497 (2021).

[213] B. Jiang, Q. Mu, F. Qiu, X. Li, W. Xu, J. Yu, W. Fu, Y. Cao, and J. Wang, "Machine learning of genomic features in organotropic metastases stratifies progression risk of primary tumors," Nat. Commun. **12**(1), 6692 (2021).

[214] W. Kim, T. H. Kim, S. J. Oh, H. J. Kim, J. H. Kim, H.-A. Kim, J.-Y. Jung, I. A. Choi, and K. E. Lee, "Association of TLR 9 gene polymorphisms with remission in patients with rheumatoid arthritis receiving TNF-α inhibitors and development of machine learning models," Sci. Rep. **11**(1), 20169 (2021).

[215] Y. Huang, X. Sun, H. Jiang, S. Yu, C. Robins, M. J. Armstrong, R. Li, Z. Mei, X. Shi, and E. S. Gerasimov, "A machine learning approach to brain epigenetic analysis reveals kinases associated with Alzheimer's disease," Nat. Commun. **12**(1), 4472 (2021).

[216] M. A. Scott, A. R. Woolums, C. E. Swiderski, A. D. Perkins, and B. Nanduri, "Genes and regulatory mechanisms associated with experimentally-induced bovine respiratory disease identified using supervised machine learning methodology," Sci. Rep. **11**(1), 22916 (2021).

[217] D. Stupp, E. Sharon, I. Bloch, M. Zitnik, O. Zuk, and Y. Tabach, "Co-evolution based machine-learning for predicting functional interactions between human genes," Nat. Commun. **12**(1), 6454 (2021).

[218] A. B. Gussow, A. E. Park, A. L. Borges, S. A. Shmakov, K. S. Makarova, Y. I. Wolf, J. Bondy-Denomy, and E. V. Koonin, "Machine-learning approach expands the repertoire of anti-CRISPR protein families," Nat. Commun. **11**(1), 3784 (2020).

[219] A. A. Trofimov, A. A. Pawlicki, N. Borodinov, S. Mandal, T. J. Mathews, M. Hildebrand, M. A. Ziatdinov, K. A. Hausladen, P. K. Urbanowicz, and C. A. Steed, "Deep data analytics for genetic engineering of diatoms linking genotype to phenotype via machine learning," npj Comput. Mater. **5**(1), 67 (2019).

[220] C.-Y. Cheng, Y. Li, K. Varala, J. Bubert, J. Huang, G. J. Kim, J. Halim, J. Arp, H.-J. S. Shih, and G. Levinson, "Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships," Nat. Commun. **12**(1), 5627 (2021).

[221] J. X. Zhang, B. Yordanov, A. Gaunt, M. X. Wang, P. Dai, Y.-J. Chen, K. Zhang, J. Z. Fang, N. Dalchau, and J. Li, "A deep learning model for predicting next-generation sequencing depth from DNA sequence," Nat. Commun. **12**(1), 4387 (2021).

[222] Y. Leng, V. Tac, S. Calve, and A. B. Tepole, "Predicting the mechanical properties of biopolymer gels using neural networks trained on discrete fiber network data," Comput. Methods Appl. Mech. Eng. **387**, 114160 (2021).

[223] E. Entekhabi, M. H. Nazarpak, M. Sedighi, and A. Kazemzadeh, "Predicting degradation rate of genipin cross-linked gelatin scaffolds with machine learning," Mater. Sci. Eng.: C **107**, 110362 (2020).

[224] M. Özkan, M. Borghei, A. Karakoç, O. J. Rojas, and J. Paltakari, "Films based on crosslinked TEMPO-oxidized cellulose and predictive analysis via machine learning," Sci. Rep. **8**(1), 4748 (2018).

[225] M. Röding, C. Fager, A. Olsson, C. von Corswant, E. Olsson, and N. Lorén, "Three-dimensional reconstruction of porous polymer films from FIB-SEM nanotomography data using random forests," J. Microsc. **281**(1), 76–86 (2021).

[226] D. Chen, J. P. Dunkers, W. Losert, and S. Sarkar, "Early time-point cell morphology classifiers successfully predict human bone marrow stromal cell differentiation modulated by fiber density in nanofiber scaffolds," Biomaterials **274**, 120812 (2021).

[227] Y. Robles-Bykbaev, S. Naya, S. Díaz-Prado, D. Calle-López, V. Robles-Bykbaev, L. Garzón, C. Sanjurjo-Rodríguez, and J. Tarrío-Saavedra, "An artificial-vision-and statistical-learning-based method for studying the biodegradation of type I collagen scaffolds in bone regeneration systems," PeerJ **7**, e7233 (2019).

[228] Z. Liu, Y. Shi, H. Chen, T. Qin, X. Zhou, J. Huo, H. Dong, X. Yang, X. Zhu, and X. Chen, "Machine learning on properties of multiscale multisource hydroxyapatite nanoparticles datasets with different morphologies and sizes," npj Comput. Mater. **7**(1), 142 (2021).

[229] Y. Cao, M. Karimi, E. Kamrani, P. Nourani, A. M. Manesh, H. Momenieskandari, and A. E. Anqi, "Machine learning methods help accurate estimation of the hydrogen solubility in biomaterials," Int. J. Hydrogen Energy **47**(6), 3611–3624 (2022).

[230] V. Daghigh, T. E. Lacy, Jr., H. Daghigh, G. Gu, K. T. Baghaei, M. F. Horstemeyer, and C. U. Pittman, Jr., "Heat deflection temperatures of bio-

nano-composites using experiments and machine learning predictions," Mater. Today Commun. **22**, 100789 (2020).

[231] B. G. Mitterwallner, C. Schreiber, J. O. Daldrop, J. O. Rädler, and R. R. Netz, "Non-Markovian data-driven modeling of single-cell motility," Phys. Rev. E **101**(3), 032408 (2020).

[232] J. Zhang, S. D. Petersen, T. Radivojevic, A. Ramirez, A. Pérez-Manríquez, E. Abeliuk, B. J. Sánchez, Z. Costello, Y. Chen, and M. J. Fero, "Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism," Nat. Commun. **11**(1), 4880 (2020).

[233] C. Li, Y. Wang, S. Sha, H. Yin, H. Zhang, Y. Wang, B. Zhao, and F. Song, "Analysis of the tendency for the electronic conductivity to change during alcoholic fermentation," Sci. Rep. **9**(1), 5512 (2019).

[234] E. Hlangwani, W. Doorsamy, J. A. Adebiyi, L. I. Fajimi, and O. A. Adebo, "A modeling method for the development of a bioprocess to optimally produce umqombothi (a South African traditional beer)," Sci. Rep. **11**(1), 20626 (2021).

[235] A. Durand, T. Wiesner, M.-A. Gardner, L.-É. Robitaille, A. Bilodeau, C. Gagné, P. De Koninck, and F. Lavoie-Cardinal, "A machine learning approach for online automated optimization of super-resolution optical microscopy," Nat. Commun. **9**(1), 5247 (2018).

[236] Y. V. Kistenev, D. Vrazhnov, V. Nikolaev, E. Sandykova, and N. Krivova, "Analysis of collagen spatial structure using multiphoton microscopy and machine learning methods," Biochemistry **84**(1), S108–S123 (2019).

[237] H.-F. Tsai, J. Gajda, T. F. Sloan, A. Rares, and A. Q. Shen, "Usiigaci: Instance-aware cell tracking in stain-free phase contrast microscopy enabled by machine learning," SoftwareX **9**, 230–237 (2019).

[238] T. I. Anderson, B. Vega, and A. R. Kovscek, "Multimodal imaging and machine learning to enhance microscope images of shale," Comput. Geosci. **145**, 104593 (2020).

[239] Y. He, W. Xu, Y. Zhi, R. Tyagi, Z. Hu, and G. Cao, "Rapid bacteria identification using structured illumination microscopy and machine learning," J. Innovative Opt. Health Sci. **11**(1), 1850007 (2018).

[240] S. Mazurenko, Z. Prokop, and J. Damborsky, "Machine learning in enzyme engineering," ACS Catal. **10**(2), 1210–1223 (2019).

[241] F. Tourlomousis, C. Jia, T. Karydis, A. Mershin, H. Wang, D. M. Kalyon, and R. C. Chang, "Machine learning metrology of cell confinement in melt electrowritten three-dimensional biomaterial substrates," Microsyst. Nanoeng. **5**(1), 15 (2019).

[242] L. Y. Sujeeun, N. Goonoo, H. Ramphul, I. Chummun, F. Gimié, S. Baichoo, and A. Bhaw-Luximon, "Correlating in vitro performance with physicochemical characteristics of nanofibrous scaffolds for skin tissue engineering using supervised machine learning algorithms," R. Soc. Open Sci. **7**(12), 201293 (2020).

[243] F. Li, J. Han, T. Cao, W. Lam, B. Fan, W. Tang, S. Chen, K. L. Fok, and L. Li, "Design of self-assembly dipeptide hydrogels and machine learning via their chemical features," Proc. Natl. Acad. Sci. **116**(23), 11259–11264 (2019).

[244] Y. Liu, D. Zhang, Y. Tang, Y. Zhang, Y. Chang, and J. Zheng, "Machine learning-enabled design and prediction of protein resistance on self-assembled monolayers and beyond," ACS Appl. Mater. Interfaces **13**(9), 11306–11319 (2021).

[245] A. Conev, E. E. Litsa, M. R. Perez, M. Diba, A. G. Mikos, and L. E. Kavraki, "Machine learning-guided three-dimensional printing of tissue engineering scaffolds," Tissue Eng. Part A **26**(23–24), 1359–1368 (2020).

[246] Y. Li, C. M. Nowak, U. Pham, K. Nguyen, and L. Bleris, "Cell morphology-based machine learning models for human cell state classification," npj Syst. Biol. Appl. **7**(1), 23 (2021).

[247] I. S. Masad, A. Alqudah, A. M. Alqudah, and S. Almashaqbeh, "A hybrid deep learning approach towards building an intelligent system for pneumonia detection in chest X-ray images," Int. J. Electr. Comput. Eng. **11**(6), 5530–5540 (2021).

[248] T. Tuncer, S. Dogan, and F. Ozyurt, "An automated residual exemplar local binary pattern and iterative ReliefF based COVID-19 detection method using chest X-ray image," Chemom. Intell. Lab. Syst. **203**, 104054 (2020).

[249] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays," Comput. Methods Programs Biomed. **196**, 105608 (2020).

[250] S. Lahmiri, D. A. Dawson, and A. Shmuel, "Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures," Biomed. Eng. Lett. **8**(1), 29–39 (2018).

[251] A. Malkawi, R. Al-Assi, T. Salameh, H. Alquran, and A. M. Alqudah, "White blood cells classification using convolutional neural network hybrid system," in *2020 IEEE 5th Middle East and Africa Conference on Biomedical Engineering (MECBME)* (IEEE, 2020), pp. 1–5.

[252] S. Rajendran and A. Jothi, "Sequentially distant but structurally similar proteins exhibit fold specific patterns based on their biophysical properties," Comput. Biol. Chem. **75**, 143–153 (2018).

[253] L. Dou, X. Li, H. Ding, L. Xu, and H. Xiang, "iRNA-m5C_NB: A novel predictor to identify RNA 5-methylcytosine sites based on the Naive Bayes classifier," IEEE Access **8**, 84906–84917 (2020).

[254] H. T. Zaw, N. Maneerat, and K. Y. Win, "Brain tumor detection based on Naïve Bayes Classification," in *2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST)* (IEEE, 2019), pp. 1–4.

[255] P. T. Gamage, M. K. Azad, A. Taebi, R. H. Sandler, and H. A. Mansy, "Clustering seismocardiographic events using unsupervised machine learning," in *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (IEEE, 2018), pp. 1–5.

[256] B. A. Helfrecht, P. Gasparotto, F. Giberti, and M. Ceriotti, "Atomic motif recognition in (bio) polymers: Benchmarks from the protein data bank," Front. Mol. Biosci. **6**, 24 (2019).

[257] G. Gul, R. Yildirim, and N. Ileri-Ercan, "Cytotoxicity analysis of nanoparticles by association rule mining," Environ. Sci.: Nano **8**(4), 937–949 (2021).

[258] S. Kuanar, V. Athitsos, D. Mahapatra, K. Rao, Z. Akhtar, and D. Dasgupta, "Low dose abdominal CT image reconstruction: An unsupervised learning based approach," in *2019 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2019), pp. 1351–1355.

[259] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, and J. Ma, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," Proc. Natl. Acad. Sci. **118**(15), e2016239118 (2021).

[260] K. Yoshida, S. Kawai, M. Fujitani, S. Koikeda, R. Kato, and T. Ema, "Enhancement of protein thermostability by three consecutive mutations using loop-walking method and machine learning," Sci. Rep. **11**(1), 11883 (2021).

[261] R. M. Ziolek, P. Smith, D. L. Pink, C. A. Dreiss, and C. D. Lorenz, "Unsupervised learning unravels the structure of four-arm and linear block copolymer micelles," Macromolecules **54**(8), 3755–3768 (2021).

[262] G. G. Bushnell, T. P. Hardas, R. M. Hartfield, Y. Zhang, R. S. Oakes, S. Ronquist, H. Chen, I. Rajapakse, M. S. Wicha, and J. S. Jeruss, "Biomaterial scaffolds recruit an aggressive population of metastatic tumor cells *in vivo*," Cancer Res. **79**(8), 2042–2053 (2019).

[263] R. Jafari and M. M. Javidi, "Solving the protein folding problem in hydrophobic-polar model using deep reinforcement learning," SN Appl. Sci. **2**(2), 259 (2020).

[264] M. Popova, O. Isayev, and A. Tropsha, "Deep reinforcement learning for *de novo* drug design," Sci. Adv. **4**(7), eaap7885 (2018).

[265] P. Petsagkourakis, I. O. Sandoval, E. Bradford, D. Zhang, and E. A. del Rio-Chanona, "Reinforcement learning for batch bioprocess optimization," Comput. Chem. Eng. **133**, 106649 (2020).

[266] Z. Hou, T. Lee, and M. Keidar, "Reinforcement learning with safe exploration for adaptive plasma cancer treatment," IEEE Trans. Radiat. Plasma Med. Sci. **6**(4), 482–492 (2022).

[267] H. Seno, M. Yamazaki, N. Shibata, I. Sakuma, and N. Tomii, "*In-silico* deep reinforcement learning for effective cardiac ablation strategy," J. Med. Biol. Eng. **41**, 935–965 (2021).

[268] P. Yazdjerdi, N. Meskin, M. Al-Naemi, A.-E. Al Moustafa, and L. Kovács, "Reinforcement learning-based control of tumor growth under anti-angiogenic therapy," Comput. Methods Programs Biomed. **173**, 15–26 (2019).

[269] H. H. Tseng, Y. Luo, S. Cui, J. T. Chien, R. K. Ten Haken, and I. E. Naqa, "Deep reinforcement learning for automated radiation adaptation in lung cancer," Med. Phys. **44**(12), 6690–6705 (2017).

[270] R. Padmanabhan, N. Meskin, and W. M. Haddad, "Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment," Math. Biosci. **293**, 11–20 (2017).

[271] J. Born, M. Manica, A. Oskooei, J. Cadow, G. Markert, and M. R. Martínez, "PaccMannRL: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning," Iscience **24**(4), 102269 (2021).

[272] J.-N. Eckardt, K. Wendt, M. Bornhäuser, and J. M. Middeke, "Reinforcement learning for precision oncology," Cancers **13**(18), 4624 (2021).

[273] J. Kong, H. Lee, D. Kim, S. K. Han, D. Ha, K. Shin, and S. Kim, "Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients," Nat. Commun. **11**(1), 5485 (2020).

[274] Z. Jin, Z. Zhang, X. Shao, and G. X. Gu, "Monitoring anomalies in 3D bioprinting with deep neural networks," ACS Biomater. Sci. Eng. (2021).

[275] C. Yu and J. Jiang, "A perspective on using machine learning in 3D bioprinting," Int. J. Bioprint. **6**(1), 253 (2020).

[276] J. Guan, S. You, Y. Xiang, J. Schimelman, J. Alido, X. Ma, M. Tang, and S. Chen, "Compensating the cell-induced light scattering effect in light-based bioprinting using deep learning," Biofabrication **14**(1), 015011 (2021).

[277] S. C. Shen, M. P. Fernández, G. Tozzi, and M. J. Buehler, "Deep learning approach to assess damage mechanics of bone tissue," J. Mech. Behav. Biomed. Mater. **123**, 104761 (2021).

[278] S. Helgadottir, B. Midtvedt, J. Pineda, A. Sabirsh, C. B. Adiels, S. Romeo, D. Midtvedt, and G. Volpe, "Extracting quantitative biological information from bright-field cell images using deep learning," Biophys. Rev. **2**(3), 031401 (2021).

[279] L. Xin, W. Xiao, R. Cao, X. Wu, P. Ferraro, and F. Pan, "Automatic compensation of phase aberration in digital holographic microscopy with deep neural networks for monitoring the morphological response of bone cells under fluid shear stress," in *Optical Methods for Inspection, Characterization, and Imaging of Biomaterials V* (International Society for Optics and Photonics, 2021), Vol. 11786, p. 117860O.

[280] F. Skärberg, C. Fager, F. Mendoza-Lara, M. Josefson, E. Olsson, N. Lorén, and M. Röding, "Convolutional neural networks for segmentation of FIB-SEM nanotomography data from porous polymer films for controlled drug release," J. Microsc. **283**(1), 51–63 (2021).

[281] E. Lin, C.-H. Lin, and H.-Y. Lane, "Relevant applications of generative adversarial networks in drug design and discovery: Molecular *de novo* design, dimensionality reduction, and *de novo* peptide and protein design," Molecules **25**(14), 3250 (2020).

[282] Y. Bian and X.-Q. Xie, "Generative chemistry: Drug discovery with deep learning generative models," J. Mol. Model. **27**(3), 1–18 (2021).

[283] Y. Mao, Q. He, and X. Zhao, "Designing complex architectured materials with generative adversarial networks," Sci. Adv. **6**(17), eaaz4169 (2020).

[284] H. Zhang, L. Yang, C. Li, B. Wu, and W. Wang, "Scaffoldgan: Synthesis of scaffold materials based on generative adversarial networks," Comput.-Aided Des. **138**, 103041 (2021).

[285] F. Calimeri, A. Marzullo, C. Stamile, and G. Terracina, "Biomedical data augmentation using generative adversarial neural networks," in *International Conference on Artificial Neural Networks* (Springer, 2017), pp. 626–634.

[286] D. Hazra and Y.-C. Byun, "SynSigGAN: Generative adversarial networks for synthetic biomedical signal generation," Biology **9**(12), 441 (2020).

[287] S. Aida, J. Okugawa, S. Fujisaka, T. Kasai, H. Kameda, and T. Sugiyama, "Deep learning of cancer stem cell morphology using conditional generative adversarial networks," Biomolecules **10**(6), 931 (2020).

[288] Z. Chen, N. Ma, X. Sun, Q. Li, Y. Zeng, F. Chen, S. Sun, J. Xu, J. Zhang, and H. Ye, "Automated evaluation of tumor spheroid behavior in 3D culture using deep learning-based recognition," Biomaterials **272**, 120770 (2021).

[289] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data," Comput. Methods Programs Biomed. **166**, 99–105 (2018).

[290] J. Lazarovits, S. Sindhwani, A. J. Tavares, Y. Zhang, F. Song, J. Audet, J. R. Krieger, A. M. Syed, B. Stordy, and W. C. Chan, "Supervised learning and mass spectrometry predicts the *in vivo* fate of nanomaterials," ACS Nano **13**(7), 8023–8034 (2019).

[291] A. Tampuu, Z. Bzhalava, J. Dillner, and R. Vicente, "ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples," PLoS One **14**(9), e0222271 (2019).

[292] U. Erdenebayar, Y. J. Kim, J.-U. Park, E. Y. Joo, and K.-J. Lee, "Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram," Comput. Methods Programs Biomed. **180**, 105001 (2019).

[293] K. Kundu, M. Mann, F. Costa, and R. Backofen, "MoDPepInt: An interactive web server for prediction of modular domain–peptide interactions," Bioinformatics **30**(18), 2668–2669 (2014).

[294] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht, "Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*," Nucl. Acids Res. **10**(9), 2997–3011 (1982).

[295] M. L. Miller, L. J. Jensen, F. Diella, C. Jørgensen, M. Tinti, L. Li, M. Hsiung, S. A. Parker, J. Bordeaux, and T. Sicheritz-Ponten, "Linear motif atlas for phosphorylation-dependent signaling," Sci. Signal. **1**(35), ra2 (2008).

[296] M. Wójcikowski, M. Kukiełka, M. M. Stepniewska-Dziubinska, and P. Siedlecki, "Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions," Bioinformatics **35**(8), 1334–1341 (2019).

[297] J. Jiménez, M. Skalic, G. Martinez-Rosell, and G. De Fabritiis, "K DEEP: Protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks," J. Chem. Inf. Model. **58**(2), 287–296 (2018).

[298] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, "Development and evaluation of a deep learning model for protein–ligand binding affinity prediction," Bioinformatics **34**(21), 3666–3674 (2018).

[299] F. Boyles, C. M. Deane, and G. M. Morris, "Learning from the ligand: Using ligand-based features to improve binding affinity prediction," Bioinformatics **36**(3), 758–764 (2020).

[300] I. Wallach, M. Dzamba, and A. Heifets, "AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery," e-print arXiv:1510.02855 (2015).

[301] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, and D. R. Koes, "Protein-ligand scoring with convolutional neural networks," J. Chem. Inf. Model. **57**(4), 942–957 (2017).

[302] A. T. Sahlol, P. Kollmannsberger, and A. A. Ewees, "Efficient classification of white blood cell leukemia with improved swarm optimization of deep features," Sci. Rep. **10**(1), 2536 (2020).

[303] X. Ying, "An overview of overfitting and its solutions," in *Journal of Physics: Conference Series* (IOP Publishing, 2019), Vol. 1168, p. 022022.

[304] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Comput. Surv. **54**(6), 1–35 (2021).

[305] K. N. Vokinger, S. Feuerriegel, and A. S. Kesselheim, "Mitigating bias in machine learning for medicine," Commun. Med. **1**(1), 25 (2021).

[306] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science **366**(6464), 447–453 (2019).

[307] M. van Smeden, K. G. Moons, J. A. de Groot, G. S. Collins, D. G. Altman, M. J. Eijkemans, and J. B. Reitsma, "Sample size for binary logistic prediction models: Beyond events per variable criteria," Stat. Methods Med. Res. **28**(8), 2455–2474 (2019).

[308] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Machine learning for predictive modelling based on small data in biomedical engineering," IFAC-PapersOnLine **48**(20), 469–474 (2015).

[309] O. Rahmati, N. Tahmasebipour, A. Haghizadeh, H. R. Pourghasemi, and B. Feizizadeh, "Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion," Geomorphology **298**, 118–137 (2017).

[310] G. L. Perry and M. E. Dickson, "Using machine learning to predict geomorphic disturbance: The effects of sample size, sample prevalence, and sampling strategy," J. Geophys. Res.: Earth Surf. **123**(11), 2954–2970, https://doi.org/10.1029/2018JF004640 (2018).

[311] A. W. Rogers, F. Vega-Ramon, J. Yan, E. A. del Río-Chanona, K. Jing, and D. Zhang, "A transfer learning approach for predictive modelling of bioprocesses using small data," Biotechnol. Bioeng. **119**(2), 411–422 (2021).

[312] G.-J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," in *IEEE Transactions on Pattern Analysis and Machine Intelligence* (IEEE, 2020), pp. 2168–2187.

[313] H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande, "Low data drug discovery with one-shot learning," ACS Cent. Sci. **3**(4), 283–293 (2017).

[314] N. Bahador and J. Kortelainen, "Deep learning-based classification of multichannel bio-signals using directedness transfer learning," Biomed. Signal Process. Control **72**, 103300 (2022).

[315] K. Aravind, P. Raja, R. Aniirudh, K. Mukesh, R. Ashiwin, and G. Vikas, "Grape crop disease classification using transfer learning approach," in *International Conference on ISMAC in Computational Vision and Bio-Engineering* (Springer, 2018), pp. 1623–1633.

[316] O. Hakimi, M. Krallinger, and M.-P. Ginebra, "Time to kick-start text mining for biomaterials," Nat. Rev. Mater. **5**(8), 553–556 (2020).

[317] C. J. Court, A. Jain, and J. M. Cole, "Inverse design of materials that exhibit the magnetocaloric effect by text-mining of the scientific literature and generative deep learning," Chem. Mater. **33**(18), 7217–7231 (2021).

[318] J. Ye, B. Xu, B. Fan, J. Zhang, F. Yuan, Y. Chen, Z. Sun, X. Yan, Y. Song, and S. Song, "Discovery of selenocysteine as a potential nanomedicine promotes cartilage regeneration with enhanced immune response by text mining and biomedical databases," Front. Pharmacol. **11**, 1138 (2020).

[319] J. Rincón-López, Y. C. Almanza-Arjona, A. P. Riascos, and Y. Rojas-Aguirre, "When cyclodextrins met data science: Unveiling their pharmaceutical applications through network science and text-mining," Pharmaceutics **13**(8), 1297 (2021).

[320] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, "Unsupervised word embeddings capture latent knowledge from materials science literature," Nature **571**(7763), 95–98 (2019).

[321] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, "Materials synthesis insights from scientific literature via text extraction and machine learning," Chem. Mater. **29**(21), 9436–9444 (2017).

[322] S. Mongkhonthanaphon and Y. Limpiyakorn, "A deep neural network for pixel-wise classification of titanium microstructure," Int. J. Mach. Learn. Comput. **10**(1), 128–133 (2020).

[323] L. Liang, M. Liu, and W. Sun, "A deep learning approach to estimate chemically-treated collagenous tissue nonlinear anisotropic stress-strain responses from microscopy images," Acta Biomater. **63**, 227–235 (2017).

[324] J. Korpela, H. Suzuki, S. Matsumoto, Y. Mizutani, M. Samejima, T. Maekawa, J. Nakai, and K. Yoda, "Machine learning enables improved runtime and precision for bio-loggers on seabirds," Commun. Biol. **3**(1), 633 (2020).

[325] Y. Zhang and C. Ling, "A strategy to apply machine learning to small datasets in materials science," npj Comput. Mater. **4**(1), 25 (2018).

[326] T. Oyetunde, D. Liu, H. G. Martin, and Y. J. Tang, "Machine learning framework for assessment of microbial factory performance," PLoS One **14**(1), e0210558 (2019).

[327] A. Tulsyan, C. Garvin, and C. Ündey, "Advances in industrial biopharmaceutical batch process monitoring: Machine-learning methods for small data problems," Biotechnol. Bioeng. **115**(8), 1915–1924 (2018).

[328] M. Nair, I. Bica, S. M. Best, and R. E. Cameron, "Feature importance in multi-dimensional tissue-engineering datasets: Random forest assisted optimization of experimental variables for collagen scaffolds," Appl. Phys. Rev. **8**(4), 041403 (2021).

[329] W. Yang, Y. Si, D. Wang, and B. Guo, "Automatic recognition of arrhythmia based on principal component analysis network and linear support vector machine," Comput. Biol. Med. **101**, 22–32 (2018).

[330] Y. Tian and Y. Zhang, "A comprehensive survey on regularization strategies in machine learning," Inf. Fusion **80**, 146–166 (2021).

[331] Z. Xiong, Y. Cui, Z. Liu, Y. Zhao, M. Hu, and J. Hu, "Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation," Comput. Mater. Sci. **171**, 109203 (2020).

[332] Y.-D. Zhang, C. Pan, J. Sun, and C. Tang, "Multiple sclerosis identification by convolutional neural network with dropout and parametric ReLU," J. Comput. Sci. **28**, 1–10 (2018).

[333] J. C. Hyun, E. S. Kavvas, J. M. Monk, and B. O. Palsson, "Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens," PLoS Comput. Biol. **16**(3), e1007608 (2020).

[334] M. Lippeveld, C. Knill, E. Ladlow, A. Fuller, L. J. Michaelis, Y. Saeys, A. Filby, and D. Peralta, "Classification of human white blood cells using machine learning for stain-free imaging flow cytometry," Cytom. Part A **97**(3), 308–319 (2020).

[335] M. Loey, G. Manogaran, and N. E. M. Khalifa, "A deep transfer learning model with classical data augmentation and CGAN to detect COVID-19 from chest CT radiography digital images," Neural Comput. Appl. **2020**, 1–13.