Acta Pharmaceutica Sinica B 2026;16(7):4024−4050

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

REVIEW

Machine learning reshapes the paradigm of nanomedicine research

Ziye Wei[a,b,†], Shijie Zhuo[a,b,†], Yixin Zhang[b,†], Lianlian Wu[b,c], Xiang Gao[d,*], Song He[b,*], Xiaochen Bo[b,*], Wenhu Zhou[a,e,*]

[a]*Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410013, China*
[b]*Academy of Military Medical Sciences, Beijing 100850, China*
[c]*Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin 300072, China*
[d]*State Key Laboratory of Toxicology and Medical Countermeasures, Beijing Institute of Pharmacology and Toxicology, Beijing 100850, China*
[e]*Hunan Key Laboratory of the Research and Development of Novel Pharmaceutical Preparations, School of Pharmaceutical Science, Changsha Medical University, Changsha 410219, China*

Received 12 December 2024; received in revised form 15 April 2025; accepted 30 April 2025

KEY WORDS

Nanomedicine;
Machine learning;
Deep learning;
Artificial intelligence;
Nanoinformatics;
Data science;
Drug delivery systems;
Pharmaceutics

Abstract Nanodrug delivery systems (NDDS) have demonstrated outstanding performance in drug delivery due to their efficient delivery capacity, targeting ability, and biocompatibility. However, the development of nanomedicines still heavily relies on the expertise of formulation scientists and extensive trial-and-error experiments. Despite the abundance of data in nanoscience, traditional biological research often struggles to effectively process, analyze, and utilize these datasets, limiting nanomedicine studies to a "one-to-one" approach. Against this backdrop, the rapid growth of artificial intelligence (AI) and machine learning (ML) offers a new paradigm for nanomedicine research. Unlike traditional statistical analyses and mathematical models, AI and ML provide deeper insights into big data, enhancing the efficiency of nanomedicine development while steering the field toward more intelligent and more precise research approaches. This review focuses on milestone studies that use ML to reshape nanomedicine research from a pharmaceutics perspective, highlighting how data-driven ML models can guide new directions in nanomedicine development.

This article is part of special issue entitled Machine Learning in Drug Discovery.

*Corresponding authors.
E-mail addresses: gaoxiang609@126.com (Xiang Gao), hes1224@163.com (Song He), boxiaoc@163.com (Xiaochen Bo), zhouwenhuyaoji@163.com (Wenhu Zhou).
[†]These authors made equal contributions to this work.

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2025.05.014

1. Introduction

With the continuous advancement of nanotechnology research, nanomaterials (NMs) have been widely applicable as a frontier material in many fields such as electronic sensing[1,2], chemical catalysis[3,4], medical diagnosis and therapy[5−7]. For instance, NMs possess unique value in catalyst and reactor design due to their increased surface active sites. The unique light absorption and luminescence properties also make NMs widely used in the field of optoelectronics and biomarkers. Furthermore, the development of nanomedicine offers broad opportunities for the precise design of biomedical materials with desirable properties of human needs. From the first FDA-approved nanodrug, Doxil (PEGylated liposomal doxorubicin, PLD)[8], to the approval of the world's first nanobody drug, ozoralizumab[9,10], NMs have completely reshaped the *in vivo* delivery paradigm of BCS Class III and IV drugs. NMs can deliver small molecules, proteins, nucleic acids or other active pharmaceutical ingredients (APIs) more safely and efficiently. By artificially regulating the carrier type, particle size, surface charge, surface modification, etc., NMs can be endowed with unique targeting, selectivity, and biocompatibility. The shortcomings of traditional preparations in biological distribution, safety and drug release performance were improved[11].

However, developing nanomedicines face a series of challenges[12]. The formulation composition and preparation process of nanomedicines significantly influence their pharmacodynamic and pharmacokinetic properties. However, Formulation design and optimization are complex and iterative processes, involving the selection of appropriate carrier materials, determination of material ratios, and establishment of optimal process parameters, with each of these variables can have a significant impact on the biological properties of nanomedicines. For example, nanoparticles(NPs) in the 100−200 nm range have been shown to accumulate in tumor tissues through enhanced permeability and retention (EPR) effects, while inappropriate particle sizes are more likely to result in poor delivery efficiency[13,14]; The shape of NMs also influences their flow and adhesion properties within blood vessels. disk-shaped NMs are more likely to roll and oscillate in the vascular system, increasing their extravasation in the vascular system[14,15]. However, traditional experimental methods involve complex combinations of characteristics, often relying on experience and trial and error, in addition, researchers have limited understanding of the quantitative/qualitative relationship between the properties of NMs and their biological outcomes, and they lack of effective tools to quickly analyze the structure−activity relationship of NMs. These challenges make it difficult to determine the preparation direction of NMs at the laboratory stage, restricting their clinical application prospects.

Machine learning (ML) offers potential solutions to the challenges mentioned above as an advanced data analysis tool[16−18]. As a key branch of artificial intelligence (AI), ML models can learn underlying patterns in data and independently draw conclusions. Leveraging existing experimental data on NMs, ML can identify key factors influencing NMs performance endpoints, including material properties, pharmacokinetics, and pharmacodynamics, and make accurate predictions[19−21]. By extracting the most influential features from high-dimensional datasets, it guides the optimization of nanomedicine formulations and processes, advancing their applications in drug delivery and clinical diagnosis. This data-driven approach is increasingly becoming an essential tool in nanomedicine development, driving rapid progress in the field.

In this review, we first provide a comprehensive overview of recent advances in ML-driven nanomedicine (including formulation preparation, pharmacokinetics, pharmacodynamics, nanotoxicology, and applications of nanomedicine, etc.), followed by a detailed summary of the available nanomedicine "reference tools" for interested pharmaceutical scientists and AI experts, which are key foundations to support ML-driven nanomedicine research and development. Although the application of ML in this field is still in its infancy, this data-driven strategy has reshaped the research paradigm of nanomedicine, as shown in Fig. 1.

2. Recent advances in ML-driven nanomedicine

Data-driven AI has permeated various industries, changing traditional ways of working. With its powerful data processing capabilities and ML/deep learning (DL) algorithms, AI can extract valuable patterns from large datasets, playing a crucial role in fields such as healthcare, finance, manufacturing, and transportation. AI, originally defined by Kurzweil et al.[21], is a branch of computer science aimed at developing algorithms that simulate human behavior, including thinking, perception, learning, and reasoning. ML, a subfield of AI, enables computers to learn from data without explicit programming, efficiently capturing nonlinear relationships in complex, high-dimensional datasets. This makes ML particularly suitable for cases where the relationship between experimental variables and target outcomes is unknown.

Various ML methods exist, including k-nearest neighbors (KNN), decision trees (DT), random forests (RF), LightGBM, and neural networks (NN). Although they differ in principle, all aim to establish mathematical relationships between input features (independent variables, such as formulation composition and experimental conditions) and target variables (dependent variables, such as nanoparticle size, shape, and encapsulation efficiency). By identifying patterns in feature space, these models enable the prediction of unknown properties.

KNN is a simple algorithm that makes predictions based on the K most similar (nearest neighbor) data points in the training set. For example, when predicting the reaction yield of a specific formulation, KNN uses results from the most similar prior formulations[22]. It is particularly suitable for datasets with a small sample size and moderate dimensionality. However, when computational complexity increases, it may suffer from the curse of dimensionality.

DT works like a flowchart, asking questions step by step based on data features (such as temperature or reagent concentration) until it reaches a final prediction (such as the characterization of NMs, or drug loading efficiency)[23]. DT has fast training speed, is

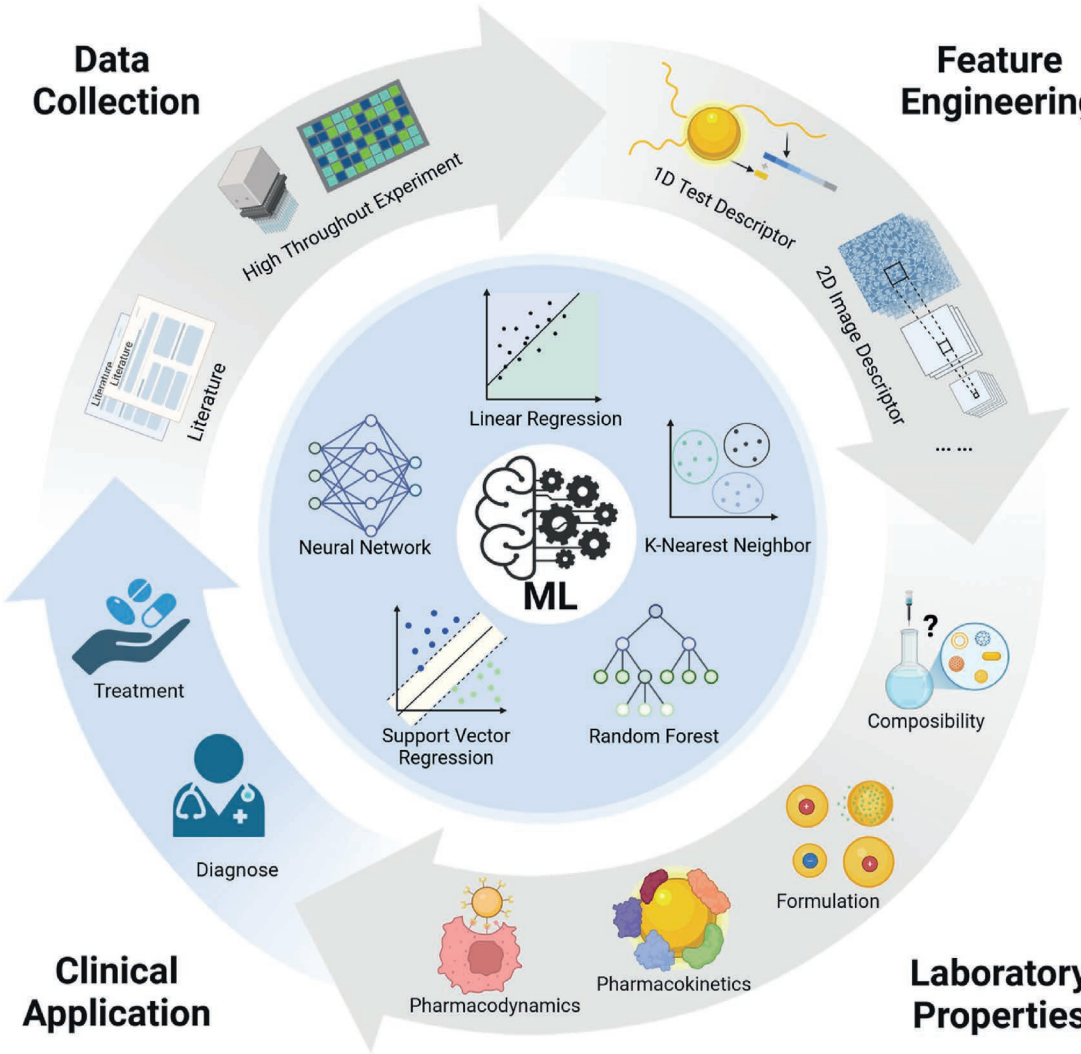

Figure 1 ML empowers nanomedicine development. Data from literature and experiments is processed into ML-recognizable formats to predict preparation, pharmacokinetics, pharmacology and toxicology, supporting diagnosis and therapy with NMs. Prior data can be reused for future studies in an iterative cycle (Created with bioRendor.com).

suitable for small datasets, and is one of the most interpretable ML models due to its traceable decision path, intuitive tree structure, and the ability to calculate feature importance (such as information gain, Gini index, or variance reduction), it can be used to reveal the main factors that affect the structure and function of nanomedicines. However, a single DT is easy to overfit and is sensitive to small data fluctuations, it has lower generalization ability. As a result, ensemble methods like RF[24], which improves stability by using multiple DT, and LightGBM, which speeds up computation and optimizes performance through gradient-boosted decision trees (GBDT), were developed to further enhance prediction accuracy and generalization[25,26]. These methods are also more suitable for large-scale datasets, adapt to high-dimensional feature data, and can be used for high-throughput screening (HTSc) of NMs and large-scale system optimization.

NN, inspired by the human nervous system, process and predict data by simulating neural information transmission[27]. A single neuron receives input signals (*e.g.*, x_1: reagent concentration, x_2: reaction time), processes them through a weighted sum ($w_1x_1 + w_2x_2 + b$, where w_1, w_2, and b are model parameters), applies an activation function, and generates an output, mimicking the firing mechanism of biological neurons. Neurons processing the same inputs form a layer, and stacking multiple layers allows the output of one neuron to serve as the input for others, ultimately constructing a NN.

Specifically, the development of ML models can generally be divided into five stages. 1. Data collection and mining: this involves constructing datasets through manual experiments, literature review, text mining, or utilizing existing data, followed by preprocessing steps like removing duplicates, eliminating outliers, imputing missing values, and performing unit and numerical conversions to ensure data quality[28]. 2. Feature engineering: computers can only recognize numerical data, so information must be converted into numerical data, referred to as descriptors, for use in ML models[29]. Subsequently, by removing highly correlated features and those with low variance to reduce data redundancy, a process known as "feature engineering". 3. Downstream task selection: determine whether the model addresses a regression or classification task. A regression task predicts continuous values, such as the particle size of NPs, while a classification task predicts

discrete outcomes, such as whether the particle size is smaller than 200 nm, with a "yes" or "no" result. The task type should be chosen based on practical application needs and modeling complexity. 4. ML algorithm selection: the algorithm is the framework for data analysis in ML models, determining the underlying logic for predictions. Simple algorithms, such as linear regression or DT, have low computational costs but often yield poor performance. Complex algorithms, like deep neural networks (DNNs), offer strong fitting capabilities but come with high computational costs[30,31]. The final choice of algorithm is typically a result of balancing model performance, computational resources, and time costs. 5. Model performance evaluation: this involves quantifying the model's predictive ability, generalization capability, and applicability using various metrics and methods. For regression models, accuracy can be assessed using metrics such as R^2, MSE, and RMSE[32,33], while for classification models, accuracy is evaluated using metrics like ACC, F1-score, AUC, and PRAUC[34,35].

DL is a subset of ML, developed from NN, and represents a highly complex class of algorithms. It primarily includes DNNs, convolutional neural networks (CNNs)[36,37], and graph neural networks (GNNs)[38,39].

DNNs are multi-layer neural networks that process input data through several layers of nodes, also known as "neurons". The output of each layer is used as the input to the next, allowing the network to progressively extract features from the data. The key to DNNs is their "depth", meaning the presence of several hidden layers (HL), which enables the network to capture more complex features. During training, DNNs adjust weights using the back-propagation algorithm to minimize prediction errors. The performance of DNNs depends on the amount of data, with performance typically improving as the data volume increases.

CNNs are the core technology of computer vision (CV), enabling computers to recognize and analyze images similar to humans. CNNs extract local features from images through convolution operations. They primarily consist of three core layers: convolutional layers (Conv layer), pooling layers, and fully connected layers (FC layer). The Conv layer applies convolution kernels (or filters) to the input image to extract local information from the feature map. The pooling layer reduces the dimensionality of the feature map, decreasing computational load while retaining key features. The FC layer combines the extracted features and outputs the final prediction. CNNs excel at automatically learning and extracting spatial hierarchies in images, making them particularly effective in image classification, object detection, and image segmentation tasks. The image analysis methods discussed in this review, such as U-net, DeepScreen, and Nano-ISML, are all based on CNNs.

GNNs are DL models specifically designed to handle graph-structured data. They pass information through the connections (edges) between nodes, allowing each node's representation to incorporate information from its neighboring nodes. GNNs typically update node feature representations iteratively to capture both local and global dependencies within the graph structure. GNNs have become a mainstream DL approach in the small molecule field. However, since nanomedicines are difficult to convert into standard graph structures, the application of GNNs in nanomedicine is still limited.

With the gradual integration of materials science and artificial intelligence, ML has made significant progress in the field of materials science, including atomic simulations, material imaging, and spectral analysis[40,41]. The ongoing integration of materials science and medicine has further promoted the application of ML in nanomedicine, leading to the interdisciplinary field of "nanoinformatics". As shown in Fig. 2, over the past decade, ML has been applied throughout the entire nanomedicine research process, including synthesis, characterization, *in vivo* processes, pharmacodynamics, and pharmacokinetics.

Fig. 3 shows how ML is reshaping the research paradigm of nanomedicine compared to traditional experimental methods. Data-driven ML can make intelligent prediction and decision support for various Formulation tasks through multi-parameter optimization algorithms. In terms of formulation exploration and optimization, NMs are often characterized in terms of formulation composition information, physical and chemical properties, process parameters, etc. The characterization dimensions of NMs include composition information (such as drug/excipient ratio), physicochemical properties (such as particle size, zeta potential) and process parameters (such as stirring rate and temperature). When it comes to pharmacokinetic/pharmacodynamic assessments, the characterization system is further extended to multi-scale data for cell lines (*e.g.*, uptake efficiency), organs (*e.g.*, targeted distribution), and animal levels (*e.g.*, blood concentrations, tumor inhibition rates). Once a suitable representation is obtained, the data is fed into the ML model for training and optimization. Therefore, in the face of the same nanomedicine task to be solved, ML only needs systematic data acquisition and algorithm iteration, which can avoid repetitive experiments in traditional trial and error methods, significantly shorten the research and development cycle and reduce labor costs. In addition, because ML is built on big data, unlike traditional methods, it can quickly explore a broader chemical space, helping to discover new materials, molecular structures or functional compounds that are difficult to consider in traditional research. Therefore, ML is expected to reshape the research paradigm of nanomedicine through the above process.

In this chapter, we review the application of ML in nanomedicine development including formulation design, pharmacokinetics, nanotoxicology, and application scenarios, aims to promote the convergence of materials science, nanomedicine, and artificial intelligence, and summarizes a paradigm for leveraging computational approaches to advance nanomedicine and facilitate its clinical translation.

2.1. *Formulation exploration and optimization*

The development and clinical application of most nanomedicines rely on exploring vast chemical design spaces and deducing structure−activity relationships governing delivery performance[42,43], combinatorial nanomedicine involves exploring an extremely broad potential combination space and requires a large amount of human, material and financial resources. During this process, ML enables formulation compatibility and predicts the synthetic feasibility of formulations, by incorporating preparation process data it can further directly predict key physicochemical properties, such as size, PDI, zeta potential, and encapsulation efficiency, thereby accelerating formulation optimization[44]. Moreover, ML significantly lowers the barriers to structural analysis of nanomedicines, accelerating the in-depth exploration of their structure−activity relationships. In this section, we discuss how ML explores vast chemical combinatorial spaces to identify synthesizable, well-characterized, and structurally rational nanomedicines.

2.1.1. *Composibility*

Self-assembled NMs (*e.g.*, ADDC, pure drug NPs, peptide-based NPs) have gained widespread attention due to the simple

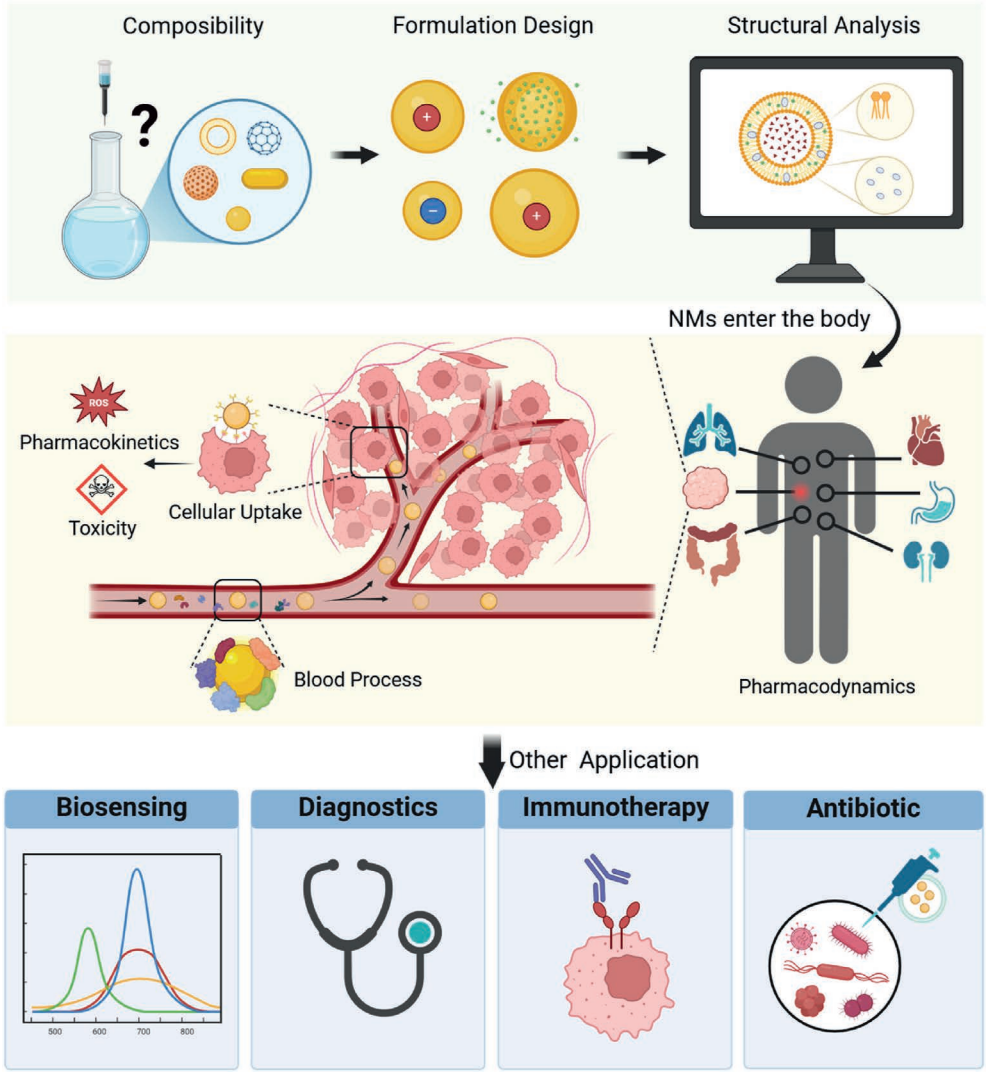

Figure 2 The full process of nanomedicine development and application (Created with bioRendor.com).

preparation, high drug-loading capacity, and potential for multi-drug combination applications[45–48]. To accelerate the development of self-assembled NMs, Shamay et al.[49] designed the Quantitative Nano-Structure Activity Relationships (QnSAR) model to identify the key role of small molecule descriptors in NPs assembly, using computational methods to reveal the excipient-driven self-assembly mechanism in an unprecedented manner. In the feature importance analysis, the descriptor SpMAX4_Bh demonstrated a Pearson correlation coefficient of 0.98, which guided the use of sulfonated organic dyes, such as indocyanine green, for the delivery of hydrophobic drugs. Inspired by the predictability of self-assembly between single drugs and near-infrared dyes, researchers attempted to extend their study to explore the self-assembly of drug combinations with dye stabilizers, which not only expanded the potential nanomaterial space but also introduced biological synergy based on chemical synergy, enabling the prediction of tumor-targeting self-assembled nanomedicine based on meta-synergy[50]. Meanwhile, Reker et al.[51] focused on the chemical aspects of drug self-assembly prediction. Using a large in-house dataset, they applied a RF model to predict drug self-assembly behavior. Based on this model, they developed two novel nanomedicines with pharmacological activity, demonstrating the practical application of AI-driven drug formulation.

2.1.2. *Formulation and process optimization*

The physicochemical properties of nanomedicines, such as size, shape, and surface chemistry, determine their stability, pharmacokinetics, and therapeutic effects. Therefore, the formulation and process of nanomedicines must be optimized based on specific requirements to meet application standards. This involves extensive exploration of materials, monitoring and adjusting process parameters, and repeated characterization, which is often time-consuming and labor-intensive[52–56]. ML can mine underlying patterns within accumulated formulation and process experimental data, predicting the nanomedicine properties and providing optimization suggestions, thereby significantly improving research efficiency, minimizing trial iterations, and accelerating the translation of nanomedicines from the laboratory to clinical[57].

He et al.[58] collected 910 size data points and 341 PDI data points of drug nanocrystals from different preparation methods, establishing eight ML methods for rapid screening of drug nanocrystal formulations. Among them, LightGBM showed the best performance and identified milling time, circulation index,

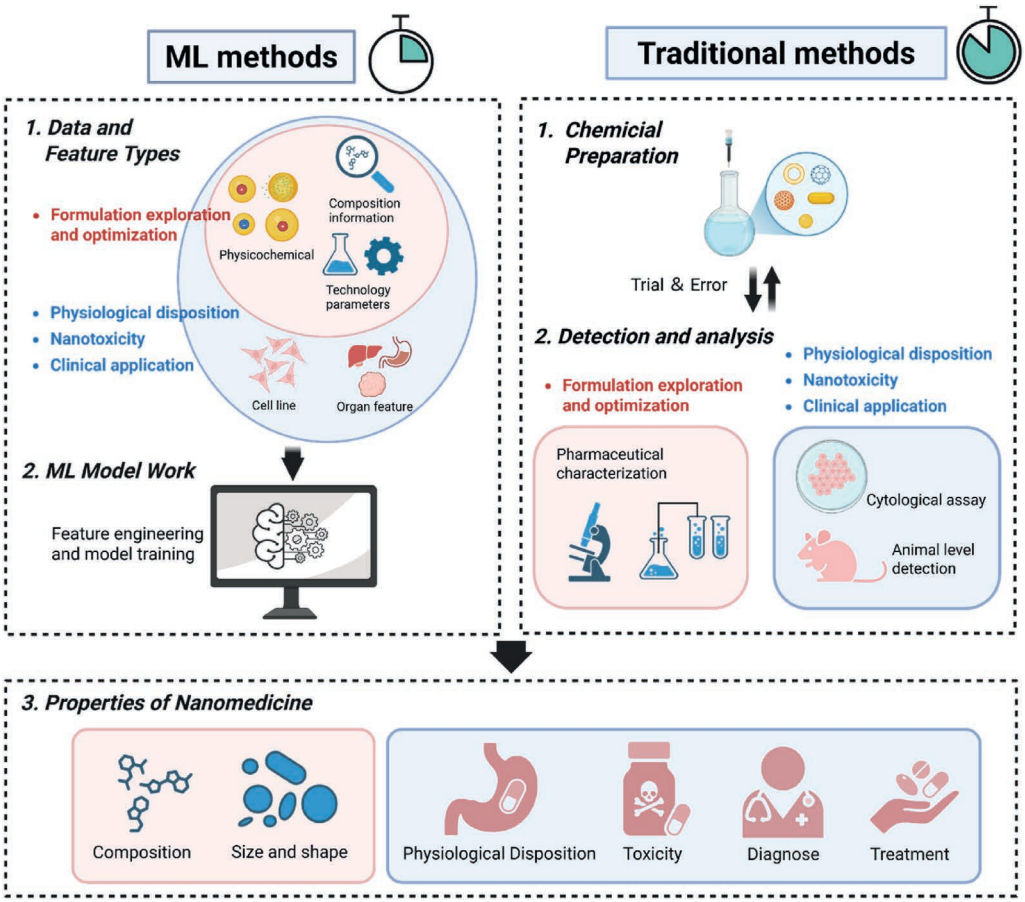

Figure 3 Workflow of ML in the development of nano-formulations (Created with bioRendor.com).

and concentration of stabilizer respectively as key factors in the preparation of nanocrystals by BWM, HPH, and ASP methods. In addition to ML, DL methods such as artificial neural networks (ANN)[59] and graph convolutional networks (GCN)[60] have also been applied for the formulation optimization of lipid nanoparticles (LNPs), polymer NMs, etc., demonstrating high accuracy and strong generalization ability. Stiepel et al.[61] combined DL with mathematical modeling to establish a neural network model for predicting the effective diffusion coefficient of drugs in polymer NMs based on their physicochemical properties and proposed a new diffusion-erosion model, ultimately achieving optimal prediction of drug release from acetylated chitosan NPs. In addition to directly optimizing the entire formulation, small-molecule modeling strategies can be used to refine specific components. Geometric deep learning (GDL) is particularly effective for optimizing small-molecule components[62]. Pretrained models based on message-passing networks can generate implicit stereogeometric information[63], enabling the analysis of molecular conformations. This approach successfully identified the bifunctional ligand Ilexgenin A from a natural product library containing over 300,000 compounds. Ilexgenin A exhibited both liposomal membrane modulation and tumor-targeting properties. Liposomes formulated with this ligand demonstrated excellent tumor-targeting ability and strong antitumor efficacy, highlighting significant potential for clinical translation.

Microfluidic devices enable precise manipulation and monitoring of reaction liquids within micron-sized channels, offering high reproducibility, cost-effectiveness, short synthesis time, and ease of automation, making them powerful tools for NMs production and testing, with the potential to enhance their controlled synthesis and expedite clinical translation[64,65]. Integrating microfluidic devices with ML could further accelerate the development and translation of NPs. As Francesco et al.[66,67] developed a DL model to predict the particle size and PDI of liposomes prepared by microfluidic chips based on preparation parameters, achieving rapid screening of microfluidic formulations. The predictive performance of ML is sensitive to the data quality, whereas high-throughput synthesis and screening based on microfluidic devices can rapidly generate large volumes of high-quality data, thereby addressing the limitations of ML and further advancing the development of NPs.

2.1.3. Structural analysis

Due to the complex preparation methods, the intrinsic structures of NMs in mixed systems are highly varied. These varied structures are closely related to the unique properties of NMs, such as catalytic activity and light response. Therefore, structural analysis of some NMs, such as bimetallic NPs[68], DNA NPs[69,70], and copper nanoclusters[71], is crucial. Traditional structural analysis methods rely on large-scale instruments such as X-ray diffraction or electron microscopy, which are costly and difficult to interpret. Density functional theory (DFT)-based computational simulations are also effective[72], but struggle with precise analysis of complex systems. ML offers a new approach for the structural analysis of NMs. Pink et al.[73] used NNs to analyze the conformations and distribution states of lipids and surfactants in LNPs, revealing that

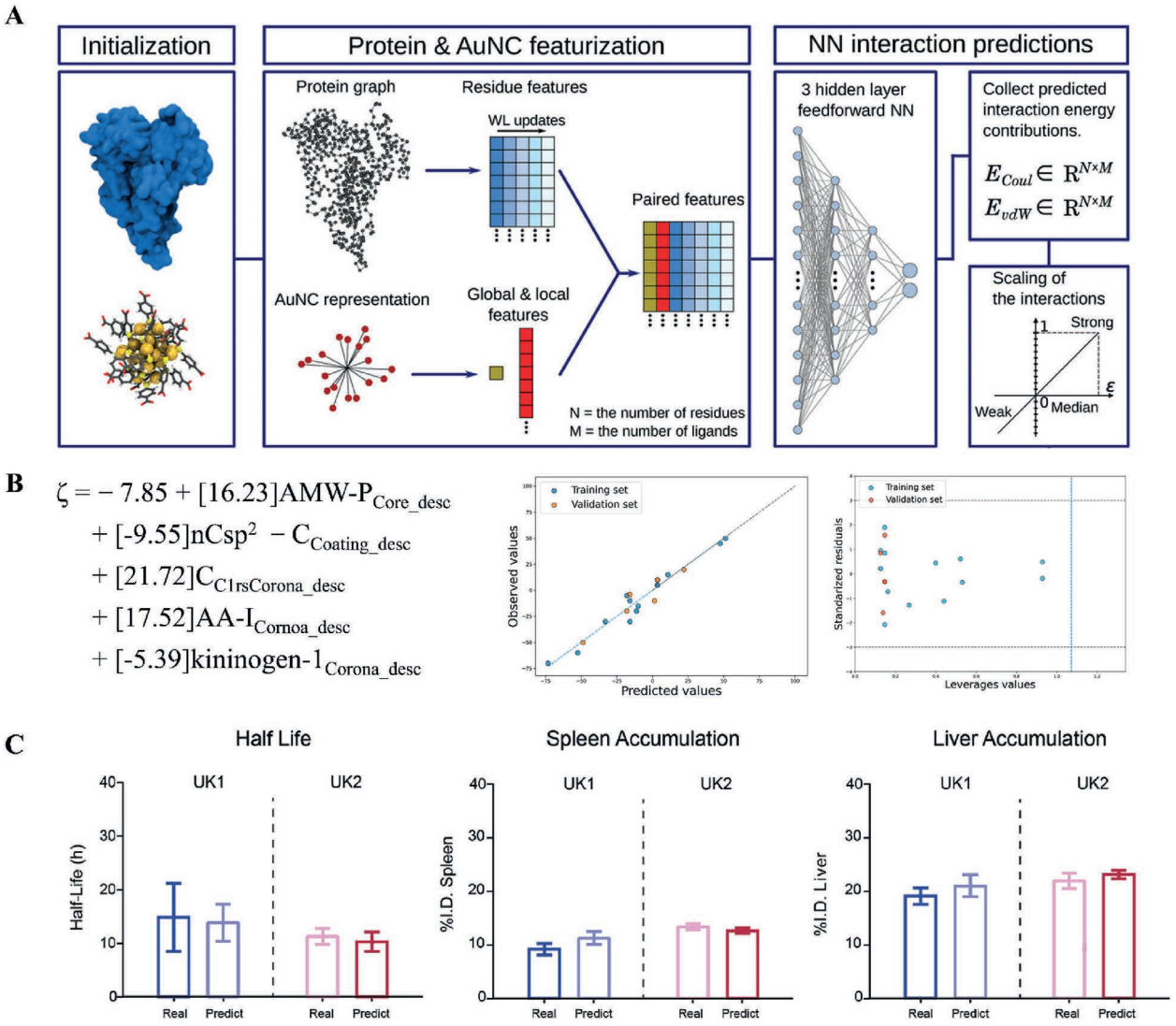

Figure 4 ML in nanomedicine−protein interaction. (A) The initialization, protein & AuNC featurization and NN interaction predictions of GraphBNC framework. Reprinted under the terms of the CC-BY 4.0 Creative Commons Attribution license from Ref. 76. Copyright © 2024 The Authors. Published by Wiley. (B) The nano-QSPR model that describe the relationship between the zeta potential (ζ) and protein and the predictive performance and Williams diagrams of the nano-QSPR model. Reprinted under the terms of the CC-BY 4.0 Creative Commons Attribution license from Ref. 79. Copyright © 2023 The Authors. Published by American Chemical Society. (C) Prediction accuracy of half life, spleen accumulation and liver accumulation of NPs by neural network algorithm. Reprinted with the permission from Ref. 80. Copyright © 2019 American Chemical Society.

even slight changes in molecular structure can significantly alter the morphology and structure of LNPs. Wang et al.[71] successfully applied convolutional neural network(CNN) to predict hydride sites in copper nanoclusters, demonstrating the potential of DL in analyzing NPs structures. Due to the severe imbalance between cost and benefit, the structural analysis of NPs has not received widespread attention in basic research. ML-based methods can significantly reduce analytical costs and thus increase the proportion of structural analysis in basic research, resulting in high-quality structural data for the development of better ML methods.

2.2. *Physiological disposition*

After entering the human body, the heterologous NMs will have special interaction with the molecules, cells, tissues and organs in the body. These interactions are strongly influenced by the characteristics and biological heterogeneity of the NMs themselves, and can reshape their physical and chemical properties, and finally make them show high heterogeneity and difficult to predict *in vivo* fate. In this section, we approach the topic from two perspectives. Firstly, we summarize the attempts of ML in analyzing the distribution mechanism of NMs. Although these studies cannot directly provide the accumulation outcome of NMs at the organ level, they provide unique insights for exploring the internal mechanism of the special distribution of NMs. Subsequently, we review studies that predict the *in vivo* distribution outcomes of NMs based on ML from a macroscopic perspective.

2.2.1. *Distribution mechanism of nanomedicine*

The interaction between NMs and proteins and cells is the basis of their interaction with organisms. Meanwhile, the hemodynamic behavior of NMs is the foundation for their circulation and

distribution *in vivo*. Only by understanding and analyzing the interactions between NMs and proteins and cells, and mastering the hemodynamics of NMs can we better analyze the interaction between NMs and organs or human bodies, which is crucial for the development of safer and more efficient NMs. At present, the mechanism analysis of the distribution fate of NMs *in vivo* by ML can be divided into three aspects: the analysis of nanomedicine−protein interaction, the analysis of nanomedicine−cell interaction and the hemodynamics of nanomedicine. These studies can not directly provide the accumulation outcome of NMs at the organ level, but provide unique insights for exploring the internal mechanism of the special distribution.

2.2.1.1. Nanomedicine−protein interaction. Upon entering biological tissues or fluids, nanomedicines spontaneously adsorb proteins from the surrounding biological environment, forming a single or multiple layers of protein corona. The composition, quantity, and conformation of the proteins in the corona are key factors that regulate the biocompatibility and pharmacokinetics of nanomedicines, ultimately influencing their *in vivo* and *in vitro* fate. Reliable characterization of the protein corona is a critical step in the development of safe and effective diagnostic and therapeutic nanomedicine products[74]. However, the interplay between the physicochemical properties of nanomedicines and the composition of the protein corona remains poorly understood, and predicting the protein corona composition without relying on experimental methods remains a significant challenge. ML offers a promising tool to overcome this obstacle.

ML is able to analyze the interaction of NMs with proteins. Saldinger et al.[75] developed NeCLAS, a DL framework that integrates atomic-level physicochemical properties and molecular surface characteristics to segment proteins and NMs into coarse-grained binding sites, employing a DNNs to predict binding affinities across all combinatorial site pairs and enabling domain-agnostic prediction of NPs-protein interaction hotspots. As shown in Fig. 4A, Pihlajamäki et al.[76] constructed GraphBNC, which employs graph convolutional neural networks (GCNs) to extract combinatorial features from graph representations of protein residues and NPs structures, and uses neural network to predict the binding affinity energies of NPs and protein. ML is able to analyze the interaction of NMs with proteins. Saldinger et al.[75] developed NeCLAS, which integrates coarse-grained low-dimensional representations of proteins and NMs with local environmental features, such as atomic-level physicochemical properties and molecular surface characteristics. This approach enables a generalizable prediction of nanomedicine−protein interaction. directly used the graph structure representation of proteins and established GraphBNC based on the graph convolutional neural network to successfully predict the binding sites of metal nanoclusters in proteins.

ML can also directly analyze the law of protein crown formation. Wheeler et al.[77] employed ML to analyze the relationship between the physicochemical properties of proteins and the formation patterns of the protein corona on silver nanoparticles (AgNPs). They collected enrichment data for 3012 proteins in the protein corona and developed a RF model based on 14 physicochemical properties of each protein, including isoelectric point and hydrophilicity, to predict whether specific proteins could adsorb onto AgNPs surfaces. This is an initial attempt to apply ML in protein corona analysis. Ban et al.[78] established a RF model based on 652 protein corona data across various NPs, linking the physicochemical properties of NPs to the composition of the protein corona and identifying the rules governing protein corona formation in SiO_2 NPs, gold nanoparticles (GNPs) and others. While the model could not accurately predict absolute RPA values for individual proteins, it captured protein binding patterns within the corona, offering a potential platform for predicting or designing NPs with specific protein corona fingerprints before administration.

In addition, ML can predict the *in vivo* fate of NPs based on high-dimensional protein corona omics data. Sengottiyan et al.[79] employed partial least squares (PLS) regression to construct a multivariate linear equation describing the relationship between the core, coating, and corona composition of NPs and their zeta potential in biological media (Fig. 4B), and found that complement C1r subcomponent, Apo A-I, and kininogen-1 in the protein corona were crucial for predicting zeta potential. Lazarovits et al.[80] trained a neural network using proteomic data of the protein corona on GNPs surfaces to predict the plasma clearance and hepatic and splenic uptake of NMs, the accuracy is 94%, as shown in Fig. 4C.

The composition of the protein corona is influenced by various factors, including the physicochemical properties of NPs, external environment, and protein information. However, due to limited data, most existing studies rely solely on hundreds to thousands of data points to train models, making it difficult to encompass all potential factors. Establishing a large, high-quality dataset can significantly enhance the predictive and analytical capability of ML in protein corona analysis.

2.2.1.2. Nanomedicine−cell interaction. Cell internalization is a dynamic process of recognition, adhesion and uptake of NMs by cells, and is the main interaction between NMs and cells. Influenced by physicochemical properties and cellular characteristics, most NMs have different internalization mechanisms,and the eventual internalization efficiency varies widely. Many papers have discussed the effect of NPs's size, shape, and chemical properties on cell internalization[14,81,82], but they still remain at a qualitative analysis of individual properties, lacking quantitative measurement, and furthermore, unable to conduct multi-factorial combinatory analysis. ML is a powerful tool suitable for multi-factor quantitative analysis. Boehnke et al.[83] developed a HTSc method to systematically evaluate the internalization behavior of 35 types of NMs in 448 cancer cell lines, using the RF model as a multivariable analysis tool to identify biomarkers related to cellular internalization capacity from biological features such as mRNA expression, mutation status, and gene copy number. They found that NPs' cellular internalization is widely affected by protein networks distributed across the plasma membrane, extracellular region, and extracellular matrix, further revealing the inverse regulatory role of SLC46A3 in NPs internalization. Compared with traditional laboratory experiments, ML method can effectively process large quantities of high-complexity data, and freely introduce genotomic and transcriptomic information, which is a powerful tool to analyze and understand the pharmacokinetic behavior of NMs.Yan et al.[86] integrated virtual modeling with ML and further developed various models, including KNN[84], RF[85], CNN[86,87], for predicting the internalization behavior and cytotoxicity characteristics of inorganic NMs. They focused on the bioactivity endpoints of various NMs based on cellular uptake, carried out four studies, and built an online prediction platform to achieve the prediction of cellular activity of NMs.

Single-particle tracking (SPT) is a powerful technique for revealing molecular and particle dynamics at the nanoscale. ML

can further enhance the potential of SPT image data, providing deeper insights into the cellular uptake mechanisms of NMs. Zhong et al.[88] first used NN to identify and distinguish 3D SPT images, and found that DL methods can effectively identify image patterns of fluorescent NPs, even for images with high signal-to-noise ratios. Subsequently, CV algorithms have been employed to analyze the positions and spatial orientations of NMs in 3D SPT images[89,90]. The research of Song et al.[91] provides a new way to analyze the spatio-temporal distribution of NPs. They first built a CNN model based on reinforcement learning (RL) to automatically the 3D posture of NPs and successfully investigated the dynamic process of motor proteins transporting cargo along the microtubule skeleton in living cells. The technique cleverly combines automated experimental strategies with DL to automatically track the dynamic changes of NMs in living cells, providing a new way to understand the relevant biophysical mechanisms.

2.2.1.3. Nanomedicine and hemodynamics. ML model enables the analysis of how blood circulation influences NMs distribution. In conditions such as tumors[92], infections[93], and cardiovascular diseases[94], vascular system alienation and remodeling often occur, resulting in abnormal blood perfusion, increased circulation resistance, and poor delivery outcomes for NMs. Computer simulation methods have become a powerful tool for analyzing abnormal blood processes by creating mathematical models to reproduce complex blood vessels and perfusions[95]. EVONANO platform combined with cell length scales and organization, for example, dynamic simulation and ML, optimization design and the effect of prediction of NMs[96].

In the field, however, strategies based on wet experiments are still dominant, For example, researchers have used transgenic zebrafish embryos to analyze the effects of hemodynamic factors on liposome accumulation[97], and investigated how the intratumoral coagulation environment regulates blood flow velocity, thereby influencing NMs accumulation[98]. However, ML models still face challenges in directly analyzing the impact of hemodynamics on NMs distribution. The main obstacles include: (1) Data scarcity. Hemodynamic characteristics in healthy animal models and humans are generally similar, while data from pathological conditions are more valuable for research but difficult to obtain. Patient-derived data are expensive, and data from animal models may suffer from poor reproducibility. (2) Modeling challenges. Hemodynamics are influenced by multiple factors, and small variations can lead to significant differences in outcomes. As a result, relying solely on conventional parameters is insufficient to fully capture the complexity of the physiological environment. And we think microfluidic chips[99] and organoids[100] offer effective tools for *in vitro* hemodynamic analysis. Additionally, CV algorithms and multidimensional feature fusion techniques provide potential solutions to overcome modeling challenges.

2.2.2. Distribution outcome of nanomedicine

NMs are considered to protect drugs from degradation, improve biodistribution, enhance tissue penetration, prolong half-life, and reduce off-target effects[101]. However, studies show that only 0.7% of NMs reach the tumor microenvironment after systemic administration[102]. NPs must overcome physiological barriers such as the blood circulation barrier, organ physical barriers, cellular and intracellular barriers to fulfill their mission[103,104]. The *in vivo* fate of NPs is complex and unpredictable. Analyzing their pharmacokinetics and summarizing structure−activity relationships are essential for accelerating their development and translation. ML can predict interactions between NPs, blood vessels, organs, and tumors, helping optimize their pharmacokinetics.

ML can be used to predict the accumulation of nanomedicines. For example, May et al.[105] utilized GBDT to analyze nanomedicine accumulation data in mouse tumor models and identified vascular density and tumor-associated macrophage density as key predictive features. Based on these two features, they developed a biomarker score that effectively predicts the accumulation of liposomal doxorubicin in tumors, providing potential biomarkers for patient stratification in cancer nanomedicine therapy.

In addition, DL has been used to visualize and quantitatively evaluate the spatial distribution characteristics of NMs in tissues. Nano-ISML[106] is a single vessel quantitative analysis method based on U-net, which can automatically identify vessels and NMs regions in confocal images of tumor frozen sections, so as to quantitatively characterize tumor heterogeneity related to NMs vascular permeability, which can be used for adjuvant pharmacokinetic analysis. By analyzing 67,530 blood vessels in 32 tumors, they found that different blood vessels within the same tumor have highly heterogeneous vascular permeability, and the distribution of dominant blood vessel groups varies greatly between different tumor types, emphasizing that vascular characteristics are an important determinant of NM endothelial penetration. Similarly, Dhaliwal et al.[107] developed NanoMASK, a 3D U-net-based tool for automated segmentation of 3D contours of the heart, lungs, liver, spleen, kidneys, and tumors from PET/CT multimodal images, it can be used to assist in analyzing the distribution of NPs *in vivo*. ML can also predict the distribution of NPs with vascular permeability information[108]. LungVis[109] is also a typical representative in this respect. By combining CV and active learning (AL), the algorithm was able to automatically, accurately and quickly segment the pulmonary bronchial tree from the alveoli, and found significant differences in NP distribution between different pulmonary drug delivery strategies. Based on the generative model, the distribution image of quantum dots (QD) in the tumor can even be directly obtained[110], and the image synthesized by AI can be used for downstream analysis such as QD density analysis, extravasation distance, and subregion distribution as shown in Fig. 5A.

ML can also be used to directly predict the pharmacokinetic parameters of NPs. Lin et al.[111] has conducted a number of studies on the prediction of tumor delivery with ML-based NPs. They use meta analysis to build the NPs pharmacokinetic data set,combines the AI and PBPK model realized the DL model was established for NPs organ distribution of the high precision of prediction[24,32,112] (Fig. 5B), and developed a platform website (https://nanoqsar.phhp.ufl.edu), which marking a key step toward practical AI-nano applications. Meanwhile, as shown in Fig. 5C, based on the above data set, Ma et al.[113] integrated tumor genome profile information into this data set and proposed an interpretable XGBoost-SHAP model, finding that most genes affecting NPs delivery encode transporters, scaffold/adaptor proteins, transmembrane signaling receptors and other proteins related related to cellular uptake, angiogenesis, inflammation, and EGF receptor and integrin signaling, which indicates that tumor heterogeneity has a strong influence on NPs delivery, emphasizing the need to develop precision nanomedicine.

Although a wide range of studies have been conducted to predict the *in vivo* distribution of NMs, the current results are difficult to directly predict the whole process of absorption of

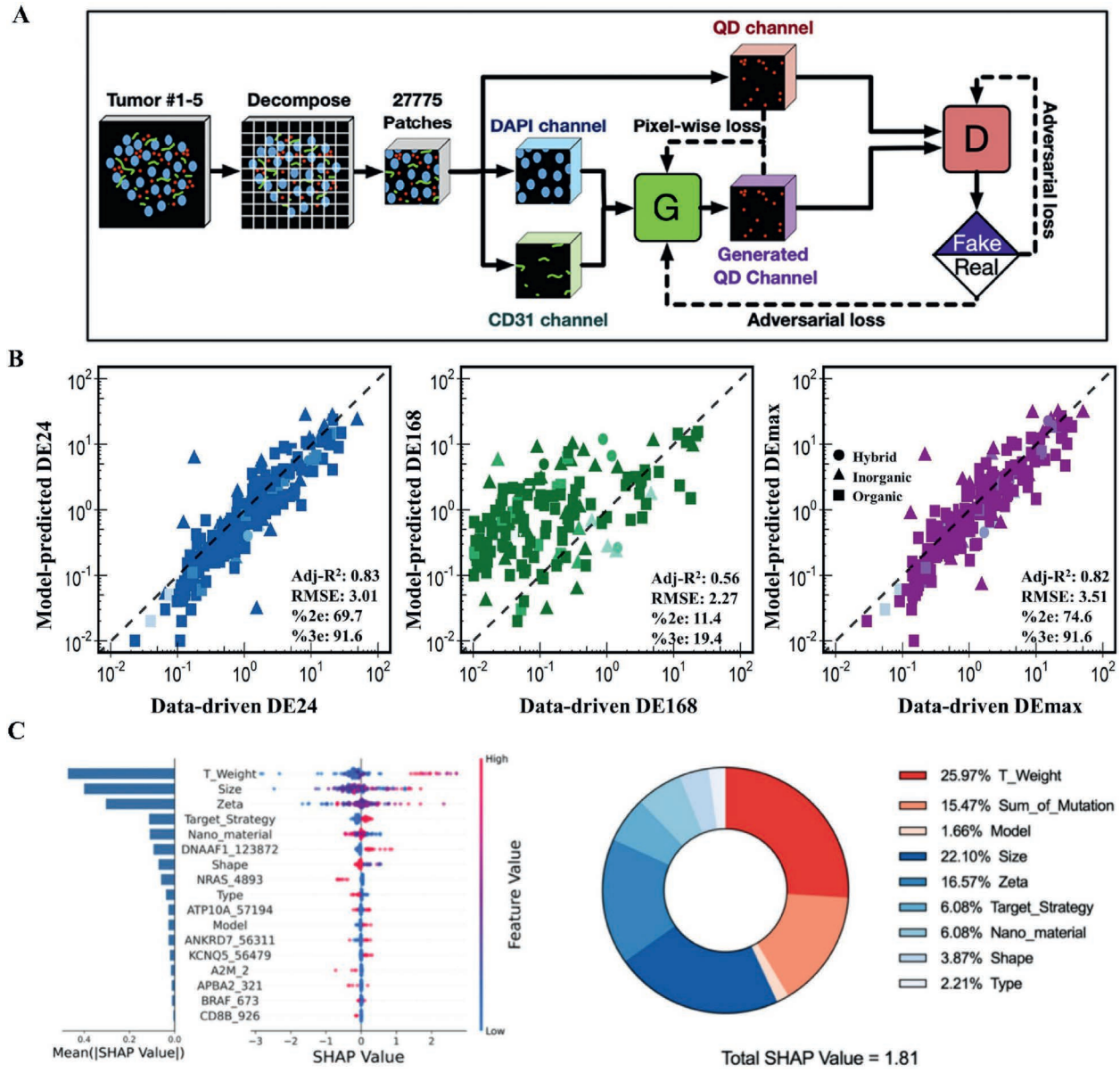

Figure 5 ML in distribution outcome of nanomedicine. (A) The framework of GANDA. In the training phase, the whole-slide 4T1 tumor section images are decomposed into patches. The DAPI and/or CD31 channels of tumor No.1 to 5 are learned by the generator network to generate the corresponding QD channel. A discriminator network is trained to distinguish between the generated and real QD channels. The adversary and pixel-wise losses propagate backward to upgrade the generator until the discriminator could not distinguish the generated and real QD channels patches anymore. Reprinted with the permission from Ref. 110. Copyright © 2021 Elsevier. (B) A global evaluation of goodness of fit by AI-PBPK model for delivery efficiency (DE) at 24, 168 h and the maximum DE. Reprinted with the permission from Ref. 112. Copyright © 2019 American Chemical Society. (C) Summary plots, beeswarm plots and dount plots of the feature importance analysis by SHAP for predicting 24-h tumor delivery efficiency. Reprinted with the permission from Ref. 113. Copyright © 2023 American Chemical Society.

complex NMs, and the corresponding models for predicting the metabolism and excretion of NMs are more scarce.

2.3. *Nanotoxicology*

After entering the biological system, due to their unique properties compared to conventional materials, nanomedicines trigger various biochemical activities, leading to complex and diverse migration and transformation pathways, potentially resulting in continuous accumulation in closed biological systems, thereby causing unpredictable toxicity, all of which severely hinder the investment and clinical translation of NMs[114]. ML is inherently suited for handling high-dimensional and complex data, making it valuable for exploring the intrinsic relationships between the physicochemical properties of NMs and their toxicity. Nanotoxicology is also one of the most extensively researched fields in ML-drived nanomedicine. Various ML models, such as DT[115], RF[116], and association rule mining[117], have been applied to predict the cytotoxicity of NMs, and related studies have been reviewed in several papers[12,19,20,118–120]. However, most of them focus on a single material[121] or a single cell[116,122], which restricts their generalization capabilities. Based on the above limitations,

Vinogradov et al.[123] developed a relatively comprehensive nanotoxicity prediction model with good generalization performance. By summarizing all available data on the cytotoxicity of NMs, they generated a dataset covering 77 cell lines and 21 inorganic NMs, with a total of 3087 data pieces. LightGBM regression model demonstrated high accuracy in predicting the cytotoxicity of inorganic NMs, with $Q^2 = 0.86$ and RMSE = 12.2%. Compared to conventional binary classifiers, this model is capable of predicting concentration-dependent cytotoxicity of NMs. Moreover, due to the incorporation of nanodescriptors based on quantitative atomic features, it offers greater flexibility, allowing for the inference of cytotoxicity for unknown samples.

Compared with other downstream tasks in the field of nanomedicine, nanotoxicology data is relatively rich thanks to the establishment of public nanotoxicity databases and the development of HTSc techniques, which provide conditions for the use of complex models for toxicity prediction. Traditional ML methods rely on manual feature selection and are susceptible to human bias and the properties of the NMs themselves. To overcome this limitation, researchers have for the first time attempted to use DL to predict the cytotoxicity of NMs[124], with the construction of the DeepScreen model based on CNN which allows automatic extraction of features from single-cell images derived from flow cytometry to evaluate the influence of NMs on cell viability。This model is capable of capturing changes in plasma membrane structure from single-cell images, reflecting possible cell damage or apoptosis induced by NPs. Compared to conventional methods of cytotoxicity detection like MTT or lactate dehydrogenase (LDH) release assays, it can avoid operator-induced errors or inaccuracies inherent in experimental assays, offering greater accuracy, convenience, and efficiency. However, this model relies on single cell images from experimental sources, which can only be used as an auxiliary tool for cytotoxicity assessment, and cannot completely replace experimental means.

Although NMs show great potential in immunotherapy, they may also cause immunotoxicity[30], including immune and inflammatory responses. For example, some NMs can nonspecifically activate the immune system, leading to excessive immune reactions, and their prolonged presence in the body may trigger chronic immunotoxicity, such as persistent low-grade inflammation, organ burden, and autoimmune responses. Due to the complexity of the immune system and NPs' structures, traditional QSARs or molecular dynamics simulations (MDS) cannot predict immune responses or organ burden accurately. Yu et al.[125] proposed a tree-based RF feature importance and feature network interaction analysis framework (TBRFA). This approach uses multi-path importance analysis to reduce bias caused by imbalanced small datasets. By constructing feature interaction networks, the model is given interpretability, providing guidance for the design and application of ideally safe NMs, and discovering feature interaction networks that are helpful for complex systems with small-scale data in various fields.

Nanotoxicology has accumulated a large amount of data; however, establishing appropriate standards to integrate this data into high-quality databases is essential. Only with such databases can high-performance ML models be developed to uncover universal patterns hidden within complex datasets. Currently, ML is not only used to predict the effects of NMs on cell viability but also applied to various aspects, including biomolecule interactions, cellular responses, mammalian toxicity, and ecotoxicity. These applications have been summarized in relevant reviews[118]. Nevertheless, data diversity and quality remain key factors limiting model performance. Increasing the standardization of data and enhancing model interpretability are urgent challenges that must be addressed in ML-driven nanotoxicology research.

2.4. *Clinical application prediction*

In the previous sections, we focused on the laboratory stage of nanomedicine development, aiming to optimize nanomedicine characterization, *in vivo* behavior, and low toxicity using ML. These studies primarily address the fundamental optimization of nanomedicines, including their physicochemical properties, pharmacokinetics, and toxicological evaluation, with the goal of ensuring good performance and minimal side effects under laboratory conditions. However, these laboratory-stage studies do not directly consider the actual clinical value of nanomedicines or their performance in practical application scenario.

In this section, we shift our focus to the clinical application of nanomedicines, which differs from the laboratory stage. Once nanomedicines enter the clinical stage, they encounter a more complex biological environment and stricter safety requirements[126]. Their development must take into account patient-specific conditions, disease characteristics, and the feasibility of implementation during treatment. The application scenario of nanomedicines play a role in imaging examination, tumor staging diagnosis, immune agent delivery, microbial classification and identification relying on their special size, crystal structure and surface chemical properties. Based on the combination of structural information and *in vivo* activity data of the NMs, ML can be used to predict or optimize clinical therapeutic indicators of the NMs.

However, ML in these areas is still in its infancy, which we suspect may be due to the fact that each study was conducted on a practical application problem, the purpose of which was to solve a practical problem and achieve application landing, rather than expecting high performance using ML; It is also shows that although ML is in its infancy in this field, it has penetrated into all aspects of the application of NMs, and has achieved fruitful results, not only limited to the preparation and evaluation of the characterization and pharmacodynamic properties of NMs. With advancements in technology and the accumulation of data, ML is expected to facilitate the translation of nanomedicines from the laboratory to clinical applications, enabling more precise therapeutic outcomes in the future.

Additionally, the research progress of ML in the diagnosis and treatment of diseases using nanomedicines, as discussed in this section, is summarized in Table 1.

2.4.1. *Disease diagnosis*

As the functionality of NMs increases, their application in diagnosis is not limited to being an alternative to traditional probes, but also realizing the accurate capture and amplification of molecular signals. ML has accelerated the development of nanomedicines in disease diagnostics by extracting potential information from vast datasets, improving diagnostic sensitivity and accuracy[138]. The following will use imaging agent design and tumor diagnosis as examples to demonstrate the groundbreaking achievements of nanomedicines and ML in disease diagnosis.

2.4.1.1. Design imaging agent. Many nanocarriers, such as graphene, carbon nanotubes, and metal nanoclusters, have been successfully used in advanced medical technologies like

Table 1 The research progress of ML in clinical application using nanomedicines.

Application field		Sources	Description	Type of NMs	Type of ML model
Disease diagnosis	Design imaging agent	Copp et al.[127]	Using integrated SVM to predict the spectral characteristics of Ag_N-DNAs based on the DNA structure, which enables the design of Ag_N-DNAs with 800 nm near-infrared emission.	Silver clusters	Integrated SVM
		Kelich et al.[128]	Based on CNN processing of DNA sequences, the optical response values of NPs formed by DNA and carbon nanotubes are predicted from the DNA sequence.	Carbon nanotubes	CNN/SVM
	Tumor diagnostics	Alafeef et al.[129]	Analyzing the cellular internalization of CNPs using ResNet to predict cancer cell types and assist in the grading diagnosis of triple-negative breast cancer.	CNPs	ResNet
		Shin et al.[130]	Using AuNPs to acquire SERS of exosomes, and employing ANN model for lung cancer diagnosis.	AuNPs	ANN
Disease treatment	Immunotherapy	Wang et al.[131]	Screening ionizable lipids based on ML to optimize the delivery efficiency of LNPs.	LNP	LightGBM
		Li et al.[35]	Based on high-throughput experiments and ML, three ionizable lipids with excellent mRNA delivery efficiency were identified.	LNP	XGBoost
		Hunter et al.[132]	Exploring the nonlinear relationship between LNP and mRNA delivery efficiency based on ML to reveal the delivery mechanism of mRNA.	LNP	CNN/RF
		Ma et al.[113]	Through computer-aided molecular design and ML, new AuNPs adjuvants were developed.	AuNPs	RF
		Chandler et al.[133]	Predicting the immune response level of NANPs using a transformer-based DL model.	NANP	Transformer
		Yamankurt et al.[134]	Using an ML model to predict the ability of SNA structural composition to activate TLR9 and trigger an immune response.	SNA	XGBoost
	Antibiotic therapy	Na et al.[135]	Establishing a nano-QSAR model to predict the bacterial toxicity of MONPs.	MONPs	RF/SVM/MLR/BRNN
		Mirzaei et al.[136]	Using ML tools to predict the antibacterial effects of NMs.	Various NMs	RF/SVM/LASSO/RR/ENR
		Yu et al.[137]	Predicting the microbial type based on the characterization results of biosynthesized AuNPs.	AuNPs	PCA/LDA/RF

photodynamic therapy (PDT) and photothermal therapy (PTT) due to their unique optical properties[139−144]. On this basis, unique DNA sequences can "program" nanocarriers, giving them specific optical responses and enabling nanoclusters to function in biosensing and bioimaging[145].

However, the vast sequence space of DNA makes it challenging to screen DNA sequences with specific imaging functions. ML and bioinformatics tools are able to revealing the relationship between DNA sequences and optical activity rapidly. For example, Copp et al.[127] developed a data-driven approach to design DNA templates that can stabilize silver nanoclusters. They synthesized and characterized 1432 DNA oligonucleotides consisting of 10 bases and used this data to train a support vector machine (SVM) classifier to predict DNA template sequences for specific fluorescence emission bands. Previous work was limited to AgN-DNAs with fluorescence emission in the 450−800 nm range, restricting their application in the near-infrared region within the tissue transparency window. By combining high-throughput experiments and ML, the team identified significant DNA sequence features that determine the color of AgN-DNAs, leading to the design of AgN-DNAs with 800 nm near-infrared emission[146].

Similarly, inspired by the nucleic acid aptamer screening technology SELEX (Systematic Evolution of Ligands by Exponential Enrichment), Jeong et al.[147] developed SELEC, a system

evolution ligand and exponential enrichment technique targeting single-walled carbon nanotubes (SWNTs) and 5-hydroxytryptamine (5-HT). This technique generates a large dataset containing rich DNA sequence information, which is closely related to both the selectivity of the analyte and the binding affinity of SWNTs. Based on the above achievement, Kelich et al.[128] established a ML model using a CNN classifier to train classify DNA ligands as high or low responders to 5-HT based on nucleotide sequence features. Subsequently, they trained an SVM regression model to predict the relative optical response values of DNA sequences. Their research demonstrates that an integrated prediction system combining a neural network classifier and an SVM regression model can effectively identify high-response and low-response DNA sequences. ML methods have accelerated the identification of efficient DNA-SWNT conjugates, significantly advancing the development of the technology in biosensors, bioelectronics, and SWNT chiral separation.

2.4.1.2. Tumor diagnostics. Nanomedicines play an irreplaceable role in tumor diagnosis. Current NMs-based diagnostic approaches primarily include using nanocarriers to deliver optical or magnetic probes, directly overcomes the limitation of the short blood elimination half-life of traditional molecular probes[148,149]. Additionally, due to high surface area, tunable particle size, and excellent biocompatibility which could confer NMs unique physicochemical properties and interactions with biomolecules *in vivo*, tumors can also be indirectly diagnosed and classified by exploring the intrinsic relationships between nanomedicine structures and various *in vivo* endpoints.

However, the heterogeneity and complexity of the tumor microenvironment pose challenges to achieving safe, sensitive, and non-invasive personalized clinical detection for each patient. Therefore, based on NMs surface properties and the biological characteristics of cells or tumors, Alafeef et al.[129] employed ML to predict the internalization pathways of carbon nanoparticles (CNPs) and further used these internalization pathways to classify cancer cell types and grade triple-negative breast cancer. This unique approach provides a new perspective for cancer grading and diagnosis. As shown in Fig. 6A. Shin et al.[130] utilized GNP to enhance the surface-enhanced Raman spectroscopy (SERS) of exosomes, which are nanoscale extracellular vesicles found in blood and proposed as promising biomarkers for liquid biopsy, and further explored the similarity and intrinsic relationship between cellular exosomes and human plasma exosomes by analyzing SERS data with DL methods. This process allows early lung cancer diagnosis without re-acquiring plasma exosome from patients. These AI tools, combined with large-scale multi-source data, are capable of comprehensively and multi-dimensionally interpreting the structure−property relationship of NMs and have the potential to revolutionize current medical practices by providing efficient and accurate tumor cell identification and early diagnosis methods.

2.4.2. Disease therapy

The primary goal of using nanomedicines for drug delivery is to enhance therapeutic efficacy by optimizing their intrinsic properties and biological processes, with disease therapy as the ultimate objective. ML provides innovative approaches to uncovering the complex relationships between nanomedicine structures and therapeutic outcomes, significantly improving research efficiency and advancing their applications in cancer therapy[62], immunotherapy, and antimicrobial treatment.

2.4.2.1. Immunotherapy. NMs interact with the immune system through various mechanisms, triggering complex immune responses. These responses can enhance immunotherapy efficacy or cause immunotoxicity. In therapeutic applications, NMs are often used as drug carriers and immune enhancers to achieve precise regulation of the immune system and strengthen immune responses. ML can utilize features like nanomaterial formulations and administration methods to predict their effects in immunotherapy, enabling more efficient design and optimization. mRNA vaccines are considered an ideal option for tackling emerging infectious diseases, influenza, and other viral illnesses. BNT162b2, the mRNA-based COVID-19 vaccine developed by Pfizer and BioNTech in 2020, was among the first COVID-19 vaccines approved worldwide. A successful mRNA vaccine relies on an effective delivery system.

Currently, ML focuses on material design to enhance the efficiency of LNPs in mRNA delivery, improve the efficiency of LNP mRNA delivery, develop efficient mRNA vaccine, and overcome the problem of low mRNA delivery efficiency[33,35,150].For example, as shown in Fig. 6B, the study by Wang et al.[131] is a representative case of integrating ML with pharmaceutical experiments to design LNPs with high delivery efficiency. They trained a simple LightGBM model using data from publicly available patents and literature to predict the pK_a of ionizable lipids and the mRNA delivery efficiency of LNPs. Through this approach, they successfully identified six ionizable lipids, among which LQ089 exhibited mRNA delivery efficiency comparable to commercially available reagents, with acceptable safety and stability. Their work demonstrated the potential of ML-driven methods in NMs design, providing a reference for traditional pharmaceutical scientists. In addition to using publicly available data, high-throughput experimental methods can also be employed for data collection. Another example is Li et al.[35] developed a new HTSc platform based on a four-component reaction (4CR), enabling the large-scale synthesis of 584 novel ionizable lipids. They then trained an XGBoost model with this dataset and successfully identified three ionizable lipids with excellent mRNA transfection efficiency from a lipid library containing 40,000 candidates.

However, Hunter et al.[132] analyzed the basis and principles of LNP design from a biological perspective. By studying the nonlinear relationship between LNP and mRNA delivery efficiency, we can understand the delivery mechanism of mRNA, aiming to improve mRNA transport efficiency and therapeutic effect. Besides LNPs, nanocarriers such as nucleic acid nanoparticles (NANPs) and spherical nucleic acids (SNAs) can also overcome the limitations of traditional nucleic acid therapies. Chandler et al.[133] developed a Transformer-based computational model, "AI-cell", using a NANPs database to predict immune response levels of NANPs. Similarly, Yamankurt et al.[134] built a QSAR model to predict the ability of SNAs to activate immune responses *via* TLR9 based on their structural composition. These studies constructed high-quality nucleic acid nanomaterial databases and demonstrated the potential of ML in predicting immune efficacy and guiding nucleic acid delivery system design. This approach addresses key public health challenges related to nucleic acid therapy safety standards and accelerates the development of novel biomedical tools.

2.4.2.2. Antibiotic therapy. As a data-driven algorithm, ML is easily influenced and limited by data size, so its research and application in nanomedicine are primarily focused on

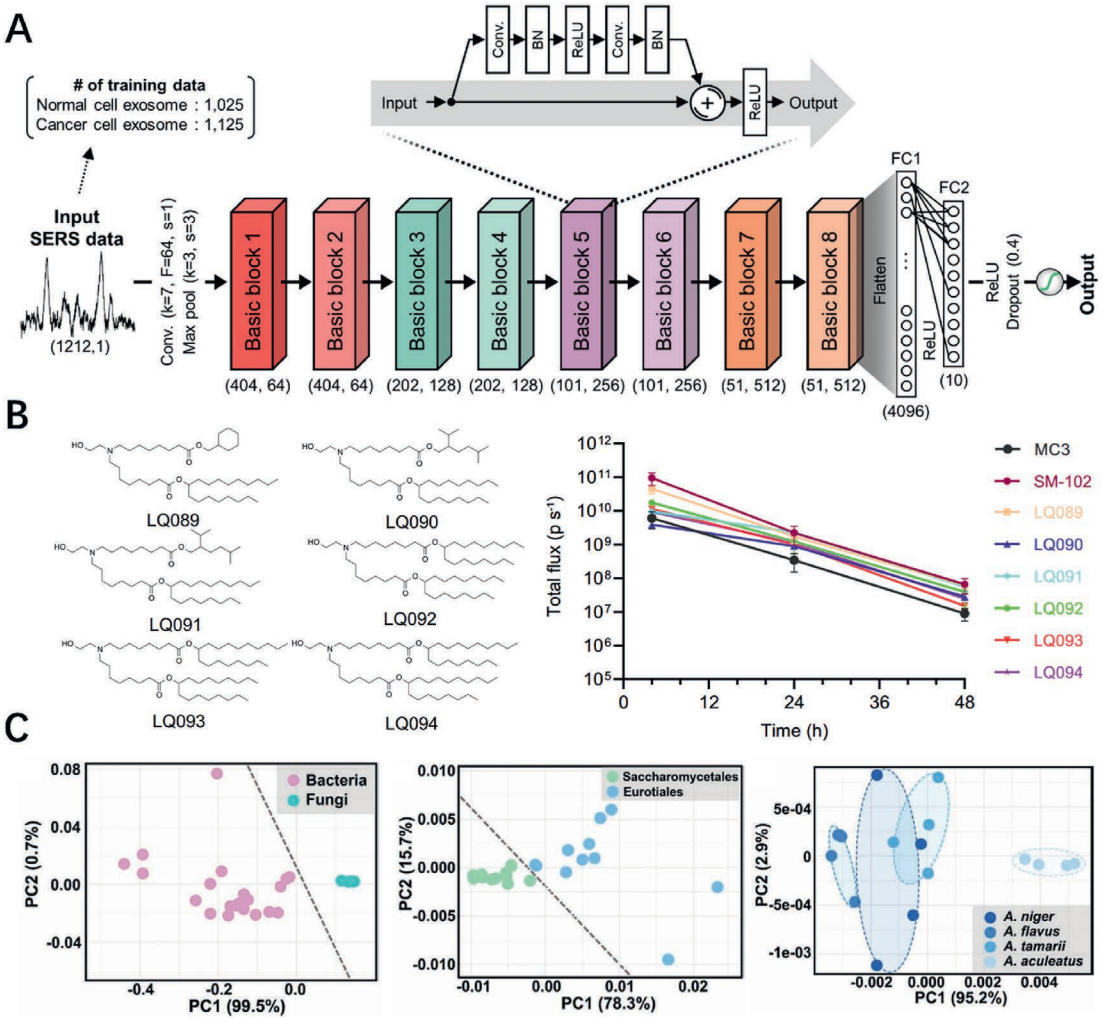

Figure 6 The application of ML in other clinical scenarios of nanomedicine. (A) Architecture of the Resnet-based deep learning model for cell exosome classification. One-directional input data extended their channel by initial convolutional layer, and reduced data length by the pooling layer. After basic blocks, two FC layers with a ReLU activation layer and a 40% dropout layer were connected. Reprinted with the permission from Ref. 130. Copyright © 2020 American Chemical Society. (B) Six ionizable lipids identified through machine learning. Female BALB/c mice (6−8 weeks old) were intravenously injected with LNPs loaded with luciferase mRNA at a dose of 5 μg per mouse, and at certain time points, total luminescence was detected after injection of D-luciferin. Reprinted under the terms of the CC-BY ND 4.0 Creative Commons Attribution license from Ref. 131. Copyright © 2024 The Authors. Published by Springer Nature. (C) Identification results of bacteria and fungi at the kingdom level, *Bacillus* and *Enterobacteriaceae* at the order level, and four species of *Staphylococcus* at the genus level using microbial classification models. Reprinted with the permission from Ref. 137 Copyright © 2022 Wiley.

oncology[151]. While some studies have applied ML to microbiology-oriented nanomedicine using high-quality self-constructed datasets, such as assisting in the detection of antimicrobial activity of metal/metal oxide NMs[135,136]. Another notable study predicted the type of microorganism based on the characterization of GNPs synthesized by microorganisms[137] as shown in Fig. 6C. Microorganisms can serve as biological factories for synthesizing inorganic NMs. The GNPs synthesized by different microorganisms exhibit variations in particle size, surface plasmon resonance (SPR) spectra, and surface potential. They established a ML classification model that utilizes the relationship between the characteristics of microbial-synthesized GNPs and microbial species at multiple levels, providing a new strategy for microbial taxonomy and expanding the application of inorganic materials.

3. Nanoinformatics in drug delivery: available tools and strategy

Nanoinformatics is an emerging discipline that integrates computer science, information technology, nanotechnology and medicine[152,153], which mainly focuses on the collection and processing strategies of nano-data. In recent years, the accumulation of extensive experimental and molecular simulation data has driven rapid advancements in nanoinformatics. Data from PubMed indicates a significant surge in publications featuring keywords such as “nanoformulation” and “nanomedicine” over the past decade. However, in the early stages of research, such datasets were often not standardized and unsuitable for immediate use in ML modeling. Specific challenges included the lack of quality control in nanomedicines, the absence of standardized

protocols for characterization, missing data features, non-uniform annotation formats for nanostructures, and the difficulty of quantitatively or qualitatively defining endpoints.

To address these issues, many recent studies have focused on systematically integrating existing datasets and developing rational characterization strategies. The database of NMs based on the idea of nanoinformatics can be used as a "reference book" for scientists to consult and retrieve, and the data can be transformed into information for computer learning through reasonable characterization strategies.

Here, we integrate existing nano-related databases or datasets, and collect existing characterization strategies for NMs for use in ML modeling. The database/dataset is equivalent to the experimental raw material, and the data characterization strategy is equivalent to the experimental method, which must be simultaneously possessed and complementary to each other, so as to build a high-performance ML prediction platform, and transform the field of nanomedicine from an experience-driven research paradigm to a data-driven research paradigm.

3.1. Existing data

Prompted by the improvement of drug databases and protein databases, nanoinformatics researchers have recognized the necessity of establishing centralized nanomedicine databases to enhance the storage and sharing of analytical data. At present, ML has been covered in various NMs, and the representative NMs types used in ML research are shown in Fig. 7.

Through text mining, high-throughput synthesis(HTSy), and HTSc, numerous databases and datasets have been developed across diverse fields, including the QnSAR dataset, the quantitative nano-structure property relationships (QnSPR) dataset, nanotoxicology database, and nanomaterial characterization database, etc. Most of these databases and datasets are open source and can be downloaded directly for subsequent research.

Databases tend to be larger and more comprehensive than datasets. Many databases provide user-friendly, open-access web platforms that allow users to query data, upload data, and realize data sharing. Establishing a comprehensive database for nanodrug delivery is the ultimate goal of nanoinformatics. Such a platform would enable researchers to integrate and utilize data more effectively, driving innovation and progress in nanodrug delivery systems (NDDS). Herein, we summarize the databases and datasets related to NMs science, nanomedicine, and nanodrug delivery established over the past 20 years in Table 2.

caNanoLab[154]: in 2004, the National Cancer Institute's Office of Cancer Nanotechnology Research (OCNR) launched the Alliance for Nanotechnology in Cancer (ANC) program, marking a milestone in the field of nanotechnology. As a core product of the ANC, caNanoLab was established as a data-sharing and mining platform aimed at providing characterization and biological activity data for NMs in the biomedical nanotechnology domain. The database primarily focuses on collecting biological experimental data, including cytotoxicity, genotoxicity, oxidative stress, immunotoxicity, and pharmacokinetics. As of 2024, caNanoLab has curated nearly 2000 samples. As one of the longest-standing NMs databases, it continues to expand its collection.

NBIK[155]: Oregon State University has developed the Nanomaterial−Biological Interactions Knowledgebase (NBIK), an integrated repository that combines data on NMs characterization, synthesis methods, and annotations of nanomaterial−biological interactions. The primary objective of NBIK is to understand the exposure risk of NMs by exploring the relationship between their physicochemical properties and the biological interactions they cause. The database, comprising over 100 samples, focuses primarily on biological interaction data. By integrating QnSAR, NBIK provides regulatory agencies with impartial scientific insights necessary to support regulations aimed at protecting human health and the environment.

ISA−TAB−Nano[156]: to promote the standardization of NMs descriptions and the sharing of characterization data among individual researchers and platforms like caNanoLab and the NBI Knowledgebase, the National Cancer Institute's (NCI) Nanotechnology Working Group (Nano WG) developed ISA−TAB−Nano in 2013. This framework, based on the newly proposed "ISA−TAB" format, provides a universal structure for representing and integrating diverse types of data related to the description of NMs using spreadsheets. By enabling researchers to adopt standardized approaches for data representation in nanotechnology publications, ISA−TAB−Nano aims to facilitate the derivation of QnSAR using harmonized datasets.

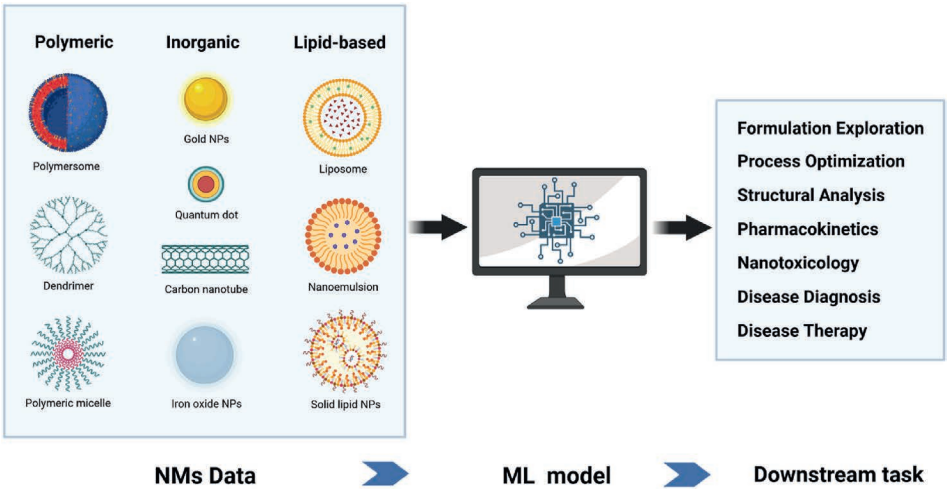

Figure 7 Representative types of nanomedicines (polymeric, inorganic, and lipid-based) used in ML research and their applications in formulation design, process optimization, and therapeutic development (Created with bioRendor.com).

Tabel 2 Summary of open-source datasets published in the literature.

Datasets	Data points	Source	Remark	Material type	URL
caNanoLab[154]	1779	Provided by laboratories	Comprehensive database of cancer nanomedicine	Various NMs	https://cananolab.cancer.gov/
NBIK[155]	147	—	A repository of data on synthesis, characterization, and biological interactions of NMs	Various NMs	http://nbi.oregonstate.edu/
ISA-TAB-Nano[156]	—	Provided by laboratories	A spreadsheet-based format for the exchange of a database of NMs description and characterization	Various NMs	https://wiki.nci.nih.gov/display/ICR/ISA-TAB-Nano
PubVINAS[157]	705	Internal data & literature	An online modeling tool for nanostructure analysis and visualization	Inorganic & organic NMs	http://www.pubvinas.com
ViNAS-Pro[158]	>750	Internal data & literature	A visual modeling tool that integrates entity prediction and virtual prediction	Inorganic & organic NMs	https://vinas-toolbox.com/
eNanoMapper[159]	6701	Provided by laboratories & literature	Toxicology database of engineered NMs	Inorganic NMs	https://www.enanomapper.net/
S^2Nano[160]	33393	Provided by laboratories & literature	Nano-security database	Oxide NMs	http://portal.s2nano.org/
Reker et al.[51] study	1440	High-throughput experiment	Small-molecule self-assembled NPs data based on high-throughput experiments	Self-assembled NPs	https://github.com/DanReker/CoAggregators/blob/master/data/screening_data.tsv
Azagury et al.[50] study	>150	High-throughput experiment & literature	Dataset of cancer self-assembled NPs from meta-synergistic drug pairs	Self-assembled NPs	Available upon request
Kimmig et al.[60] study	3753	High-throughput experiment	Prediction of polymer particle size utilizing compound characterization data	Polymethacrylate polymer NPs	https://github.com/JulianKimmig/nanoparticle_size_prediction/tree/master/data
Wang et al.[162] study	161	Literature	Metal−organic frame drug loading dataset	MOFs	https://ars.els-cdn.com/content/image/1-s2.0-S0378517324003624-mmc1.pdf
Labouta et al.[115] study	2896	Literature	Meta-analysis of NPs toxicity *via* data-mining the literature	Inorganic & organic NPs	https://pubs.acs.org/doi/suppl/10.1021/acsnano.8b07562/suppl_file/nn8b07562_si_001.xlsx
Martin et al.[121] study	5029	Internal data & literature	Cytotoxicity dataset of amorphous SiO_2-NPs	SiO_2-NPs	https://pubs.acs.org/doi/suppl/10.1021/acsnano.2c11968/suppl_file/nn2c11968_si_003.xlsx
Yu et al.[125] study	1620/301	Literature	Datasets on pulmonary immune responses and lung burden of NPs	Inorganic & organic NMs	https://doi.org/10.5281/zenodo.4661099
Shirokii et al.[123] study	3087	Literature	Cytotoxicity dataset of inorganic NMs	Inorganic NMs	https://github.com/acid-design-lab/DiTox_v1/blob/main/final_model/dataset_3090.csv
Chen & mi et al.[111,24] study	534	Literature	Nano-tumor database describing the distribution of NPs	Inorganic & organic & Hybrid NMs	https://pubs.acs.org/doi/suppl/10.1021/acsnano.3c04037/suppl_file/nn3c04037_si_002.xlsx
Kingston et al.[163] study	1301	High-throughput experiment	Image-based prediction of intratumor NPs distribution	GNPs	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6660785/bin/pnas.1907646116.sd01.xlsx
Ban et al.[78] study	652	Literature	Protein corona composition dataset of NPs	Inorganic & organic & Hybrid NMs	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7229677/bin/pnas.1919755117.sd01.xlsx
Mendes et al.[34] study	745	Literature	Data of preclinical study of cancer NMs	Inorganic NMs	https://github.com/RekerLab/NanoAnalysis/tree/main/data
Yamankurt et al.[134] study	960	High-throughput experiment	Dataset on how SNA structure activates immune response	SNAs	Available upon request
Boehnke et al.[83] study	17080	High-throughput experiment		LIPO/PLGA/PS	https://zenodo.org/records/6642633
Mirzaei et al.[136] study	436	Literature	Dataset on antimicrobial activity of MNPs and MO-NPs	MNPs & MO-NPs	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8308172/bin/nanomaterials-11-01774-s001.zip

PubVINAS[157] and ViNAS-Pro[158]: many existing databases are not suitable for directly computational modeling, making them challenging to use in ML research. Zhu et al. from the Center for Computational and Integrative Biology at Rutgers University, in collaboration with Guangzhou University, developed an online modeling tool for nanostructure analysis and visualization: PubVINAS. PubVINAS encompasses 11 material types with over 10 computational endpoints, storing each NMs in PDB structure file format. Additionally, its nanostructure annotation program generates 2142 nanodescriptors for all 705 nanostructures in the database, enabling researchers to directly use these comprehensive descriptors for computational modeling. However, PubVINAS only stores existing NMs data and cannot provide new nanostructures to guide NMs design. Based on this dilemma, the team released the Virtual Nanostructure Simulation Professional (ViNAS-Pro) platform in 2024. Compared to PubVINAS, ViNAS-Pro expands its coverage to include 14 types of NMs, providing detailed structural information and bioactivity data. Furthermore, ViNAS-Pro integrates a virtual database for predicting the properties and bioactivity of virtual NMs, which greatly realizes the integration of data processing, visualization, model construction, and virtual prediction of NMs.

eNanoMapper[159]: in response to the urgent need for managing multi-mechanistic data on engineered nanomaterials (ENMs) in European nanotechnology research, eNanoMapper was developed. This initiative leverages existing standardization tools, such as ISA−TAB−Nano, to systematically acquire experimental metadata and protocols. By establishing a modular and scalable infrastructure, eNanoMapper supports collaborative safety assessments of ENMs. Its primary goals are to enable transparent data sharing, data analysis, and the creation of toxicological ML models for ENMs.

S^2NANO[160,161]: S^2NANO is a large-scale nano-safety database constructed from experimental and literature data related to NMs. By employing text mining, it has amassed 33,393 raw samples, containing detailed physicochemical properties and cytotoxicity information of NMs. Furthermore, S^2NANO has developed 14 nano-safety prediction models to uncover the connection between the physicochemical attributes and cytotoxicity of NMs by employing ML and DL algorithms. This integration of data and predictive modeling facilitates a deeper understanding of NMs safety and supports the rational design of safer nanostructures.

In addition to databases, many studies using ML for task-specific analysis often build task-specific datasets based on literature mining or HTSy[151], some of these datasets are open-source and available for download. However, due to the limited scope of these datasets, they are typically designed for specific research tasks and have not evolved into user-oriented open-source databases, making their application scenarios more limited. Nonetheless, certain datasets have high follow-up value, offering insights and foundational data for other tasks.

For instance, in the field of NMs preparation, Reker et al.[51] combined MD simulations and ML based on a HTSy platform. Based on 16 small molecule drugs with self-aggregation ability and 90 excipients in the FDA inactive ingredient list, they tested the ability of 1440 formulations to form self-assembled NPs through the high-throughput experimental platform. This is the largest dataset of drug self-assembly reported so far. The dataset reports the particle size and PDI of NPs after self-assembly, and gives labels for ML training, which can be directly used in ML. Similarly, Azagury et al.[50] established an open-source dataset for small molecule drug self-assembly, aiming to develop biologically synergistic self-assembled nanomedicines based on chemical synergy. These studies have advanced the application of ML in the drug self-assembly field.

A highly representative work in pharmacology is carried by Mendes et al.[34], who created a large inorganic NMs dataset by combining text mining and data organization. This dataset compiled experimental data from 745 preclinical cancer nanomedicine studies conducted between 2006 and 2021. It deeply explored the physicochemical properties of NMs (such as morphology and surface functionalization), animal model setups (such as cancer types and administration routes), and measurement endpoints (such as tumor size reduction and tumor/organ biodistribution). Based on this dataset, the mechanisms and optimization strategies of inorganic NMs can be deeply explored, providing a design paradigm for inorganic NMs in cancer treatment research.

3.2. *Nanodescriptor*

When using ML for NMs screening, design and performance prediction, systematic characterization data is the core basis for model building and algorithm optimization. After collecting samples, in order for these samples to be interpreted by the computational system, each sample is given a specific description, called a "representation" or "feature vector", which is able to match the corresponding predicted endpoints. Varieties of characterization methods can provide key information on the morphology, size, crystal structure, surface chemistry and optical properties of NMs. However, unlike small-molecule drugs, obtaining meaningful descriptors for NMs is highly challenging due to their varying sizes, shapes, and surface properties. The structure and physicochemical properties of NMs are also easily influenced by biological media, such as protein adsorption, aggregation, or degradation. Their structural characteristics often change after entering the body. For certain tasks, considering only the intrinsic features of nanomedicines while ignoring their interactions with biomolecules makes it difficult to analyze complex structure−activity relationships. Moreover, not all features are useful for model predictions. Therefore, appropriate feature engineering is needed to extract features and identify descriptors that best represent the structure and properties of nanomedicines.

3.2.1. *Experiment descriptor*

In the field of ML-drived nanomedicines development, the most widely used nanodescriptors are experimental descriptors. These low-dimensional feature vector can summarize multiple aspects of NMs preparation and activity evaluation. which is suitable for the training of simple ML models. Additionally, experimental descriptors support the subsequent Explainability analysis of ML models, providing insights into how specific features of NMs contribute to their performance and activity, such as Tree-based feature importance analysis[122], permutation importance, SHapley additive exPlanations (SHAP)[164−166], which contributed to identify key features.

Nanomedicine can be characterized by its composition, structure, and physicochemical properties, the composition of the NMs is determined by the preparation method. For instance, the concentration of cholesterol in liposomes[66], mRNA sequence[33], type of metal ions in MOFs[166], and the degree of polymerization of polymers[60]. Nanostructures can be represented by size, shape, composition, and surface area, and these can be measured using techniques such as transmission electron microscopy (TEM), scanning electron microscopy (SEM), X-ray diffraction (XRD),

dynamic light scattering (DLS), etc. which are determined by the preparation method. For instance, the concentration of cholesterol in liposomes[66], mRNA sequence[33], type of metal ions in MOFs[166], and the degree of polymerization of polymers[60]. The structures can be represented by size, shape, composition, and surface area, and these can be measured by transmission electron microscopy (TEM), scanning electron microscopy (SEM), X-ray diffraction (XRD), dynamic light scattering (DLS) and other techniques. Since the structural characterization is influenced by preparation conditions, preparation parameters such as excipient concentration, preparation temperature, and machine rotation speed can also serve as descriptors. Furthermore, the physicochemical properties of NMs, such as Log*P*, surface charge, zeta potential, isoelectric point, and solubility, are important experimental descriptors for ML modeling. *In vivo* process or therapeutic effect prediction, relevant cell line, animal, tumor and other information is also widely used to elucidate the interaction between nanostructures and biomolecules or organisms, such as cell line types[123], cancer types[32,112], tumor size, and tumor type[34,113] are also commonly used.

Experimental descriptors often involve both continuous variables and discrete variables. Therefore, it is common to use one-hot encoding[78] to convert discrete variables into binary vectors, which can be recognized by ML algorithms. If the feature distribution range of the original data is large, feature scaling is needed, such as normalization[167], regularization[135], or *z*-score[107] methods, which can effectively eliminate the weight differences between features.

However, obtaining experimental descriptors for nanomedicines is challenging, it requires complex experimental procedures and characterization processes. Additionally, the academic community has not yet established standardized protocols for NMs production and preparation. This lack of standardization leads to significant differences in characterization results between laboratories and introduces batch effects. Moreover, the physicochemical properties of nanomedicines are influenced by factors such as encapsulated APIs and preparation conditions, but not all studies account for these factors, resulting in low data reliability. Therefore, researchers aim to use computational methods such as image analysis, molecular modeling, and MD to extract critical information from nano-data, facilitating the development of universal nanodescriptors.

3.2.2. *Computational descriptors*

Compared with experiment-based nanodescriptors, computational descriptors are significantly more efficient and flexible. Computational descriptors extract features directly from data through algorithms, enabling large-scale extraction and avoiding experimental errors. In addition, computational descriptors can integrate multi-dimensional information of NMs, such as chemical structure, functional modification, dynamic behavior, etc., to provide comprehensive support for QnSAR modeling.

At present, computational descriptors are divided into 1D, 2D, 3D and 4D according to the input data dimension. From simple symbolic sequences (*e.g.* SMILES, InChI) to complex descriptors based on images and MD, these methods are able to capture the static structure and dynamic behavior of NMs, especially when combined with DL algorithms. Compared with experimental descriptors, computational descriptors can achieve diversified characterization, rapid modeling and interdisciplinary applications, and will provide more effective tool support for the design and optimization of NMs. Fig. 8 details the computational nanodescriptors in four dimensions.

3.2.2.1. *Descriptors based on 1D test.*

Currently, 1D text representation has been widely used and formed a complete system in the field of characterization of small molecule compounds. With the continuous development of computational analysis tools in the field of materials science modeling, similar methods have also emerged for polymers, mixtures, and reactions, etc. SMILES (simplified molecular input line entry system) can express the chemical structure of compounds clearly. In order to integrate more structural information of NMs into SMILES, Toropov et al.[168] developed a new symbolic sequence called quasi-SMILES, which integrates molecular structure, experimental conditions, and physicochemical properties into traditional SMILES by combining string characters and numbers.

In addition to SMILES, InChl[169], as an open source structure-based chemical symbol that can encode chemical characteristics with a hierarchical, layered structure, has been included in PubChem, ChEMBL and other commonly used large compound databases. To accommodate more complex molecular systems, InChI has been extended for use in polymers[170] (PInChI), mixtures[171] (MInChI), and reactions[172] (RInChI). In February 2020, the Transatlantic Workshop in Iceland proposed the development of NInChI[173], a format focused on NMs. On the basis of traditional InChI, hierarchical information including core chemical composition, size and morphology modification, functionalization information, etc. On the basis of traditional InChI, NInChI uses a modular hierarchical structure to add information such as core chemical composition, size, morphology modification, and functionalization details This ensures that NMs data of varying complexity can be captured and directly used for ML modeling. Future research will aim to integrate NInChI into the InChI standard, establishing it a universally used NMs identifier for nanosafety, material production and material modeling.

3.2.2.2. *Descriptors based on 2D image.*

Image-based DL techniques, such as CNN and Fully Convolutional Network (FCN), have introduced new models for NMs characterization[174]. For example, Yan et al.[86] used LeNet, a CNN based on CNN, to directly learn nanostructure features from multi-angle images of NMs obtained in VMD. This approach avoids complex descriptor calculations. DeepScreen[124] uses CNN trained on single-cell images from flow cytometry. It processes millions of images, overcoming the limitations of traditional ML in data volume and manual feature extraction. This approach enables fast and accurate evaluation of nanomedicine efficacy and cellular states.

In 2015, the end-to-end DL network U-net[175,176] demonstrated exceptional performance in medical image classification. Addressing issues like simple semantics and limited data in biomedical images, U-net introduced a novel network architecture and training method. It relys on data enhancement to make effective labeling data more effective. According to the achievements of U-net in the field of medical image segmentation, many researches have applied U-net to the field of nanomedicine, which provides a new paradigm for the characterization of nanomedicine. Zhu et al.[106] constructed Nano-ISML, a single-vessel quantitative model based on CNN U-net. Using U-net as the basic framework, two independent segmentation models were developed for the two channels of tumor images (vascular and FTn penetration) respectively. Based on two U-net segmentation models, 4 basic image features and 9 blood vessel features were extracted, which realized the leap from qualitative to high-throughput quantitative analysis of blood vessel permeability.

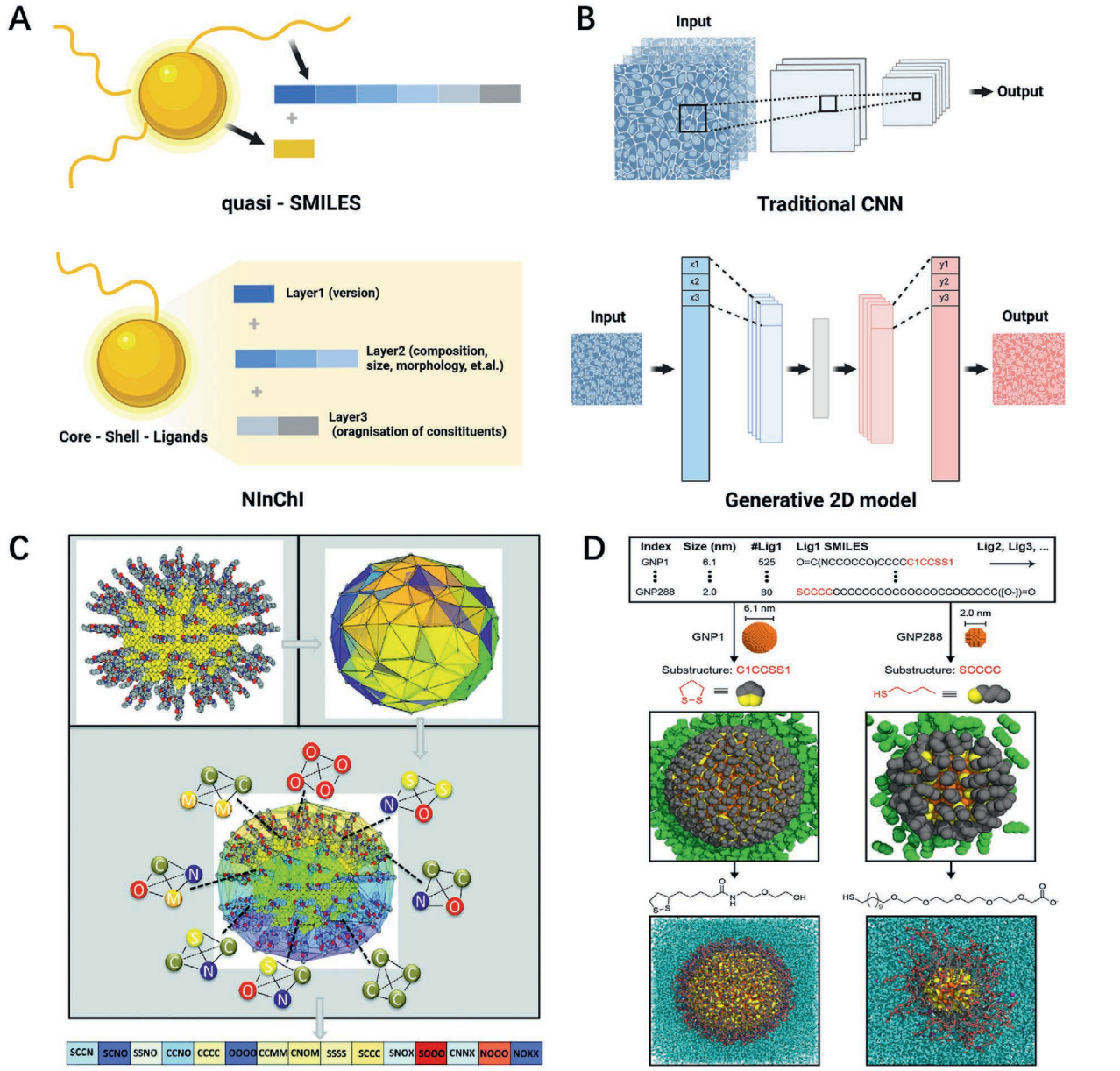

Figure 8 Four classes of computational descriptors divided by dimension. (A) 1D nanodescriptor. Using text information to extract NMs representation characterization by quasi-SMILES or NInChI. (Created with bioRendor.com). (B) 2D nanodescriptor. Using CNN or Generative 2D model to construct 2D nanodescriptor. (Created with bioRendor.com). (C) 3D nanodescriptor. Novel tetrahedron nanodescriptors calculated from atomic electronegativity and the Delaunay tessellation of virtual nanostructures. Reprinted with the permission from Ref. 85. Copyright © 2019 Royal Society of Chemistry. (D) 4D nanodescriptor. Nanodescriptors generated from MD. Reprinted with the permission from Ref. 165. Copyright © 2022 American Chemical Society.

In addition to directly using CNN to extract image features, some researches rely on advanced image analysis software, such as Columbus[132], to pre-transform images into computer-recognizable feature vectors, such as embedding, which represents the information of cell number, cell morphology, and fluorescence intensity, and then input it into the ML model for analysis. Generative models also provide a promising image-based characterization of NMs. Using an FCN-based generative adversarial network (GAN), researchers generated intratumoral distribution images of quantum dots (QDs) based on CD31 and DAPI staining of tumors[110]. This generative approach realized virtual quantitative analysis of tumor−nanomaterial interactions, even without real QDs distribution data.

3.2.2.3. Descriptors based on 3D atomic position. The nanodescriptors mentioned above can partially characterize the structure and function of nanomedicines. However, they simplify the structure, overlooking details such as the atomic spatial positions and ligand structures of NMs. Additionally, batch effects inevitably impact the overall characterization[84]. Therefore, developing nanodescriptors based on 3D atomic positions and properties can compute the complete nanostructure, including the nano-core and surface ligands.

This approach fully represents the structural details of nanomedicines and facilitates the development of QnSAR models.

In response to these ideas, Yan et al.[86] reported a series of work on virtual full nanostructure descriptors. They created an in-house GNPrep program, which inputs three basic structural parameters: particle size, surface ligand structure, and ligand density. Using MOE software, they optimized the 3D structure and virtual GNPs (vGNPs) in batches. They generated 34 GNP descriptors based on the simulated surface information of vGNPs, successfully predicting biological activity indicators such as cellular uptake[86]. They also developed a refined surface chemistry simulation method to evaluate the hydrophobicity of GNPs based on this vGNP library[177]. To address issues related to computational resources and commercial software, the team used Pauling electronegativity and Delaunay tessellation methods to develop a novel tetrahedral nanostructure descriptor for computational modeling based on 17 properties such as atomic electronegativity and atomic spatial location. This tetrahedral descriptor allowed faster and more comprehensive nnotation of nanostructures[85], and this advanced descriptor can be used for immune adjuvant screening[150] and prediction of graphene toxicity focus[178]; The research results were collected in PubVINAS[157] and VINAS-Pro[158].

3.2.2.4. Descriptors based on 4D molecular dynamics simulation. Compared with static structural characterization, four-dimensional MD descriptors can simulate the behavior of NMs in complex environments and capture information such as molecular conformation change, dispersion and solubility of NMs under aqueous solution, membrane environment or stress conditions, which helps to understand the dynamic structure of NMs and identify the surrounding supramolecular environment[135,179−181].

For example, Reker et al.[51] generated 19 MD descriptors from short MD simulations to characterize drug−excipient interaction potentials, and proved the powerful ability of MD-derived descriptors in self-assembly prediction compared with other types of molecular descriptors through RF-based feature importance analysis. In another study, Chew et al.[165] performed MD simulations on 154 GNPs and calculated a small descriptor library. These descriptors captured structural and chemical properties of NMs in aqueous solutions and were parameterized into QNAR models using interpretable regression algorithms to predict various biological and chemical endpoints of NPs.

Additionally, SOAP descriptors[182] and radial distribution function (RDF) [183] descriptors derived from MD trajectories can be used to simulate the structure and local supramolecular environment of NMs through unsupervised computational analysis. However, MD descriptors often require significant computational resources and are limited by the size of NMs. As a result, their application scenarios are restricted, making them challenging to generalize to other materials and prediction tasks.

In summary, both experimental and computational descriptors have achieved significant progress. However, different methods may be required for different data forms, target task types and feature engineering requirements. For example, if the study pays more attention to the influence of experimental conditions on the pharmacodynamic endpoint of NMs, experiment descriptor may be more suitable for modeling. However, if the image data of NMs is provided in the study, it may be necessary to use 2D image-based neural network for feature extraction. We cannot judge which descriptor is better, but it is important to choose the appropriate descriptor acquisition method according to the target data and target task.

4. Challenge and overlook

The development and implementation of ML have begun to optimize and transform the R&D processes of nanomedicine, demonstrating substantial advantages. From delivery system design to *in vivo* process evaluation, ML methods have made remarkable progress in accelerating material screening and optimizing drug performance. However, ML may also have negative implications for the development of nanomedicine:

(1) Generation of potentially erroneous predictions: the accuracy of ML in nanomedicine remains insufficient, possibly due to the poor reproducibility of pharmaceutical experiments and inherent subjectivity in experimental operations. These factors compromise data reliability, leading to reduced predictive accuracy and poor generalization. Consequently, ML models may produce biased or incorrect predictions, making it difficult to provide stable and accurate guidance for experimental research.

(2) Diminished creativity of biologists[184]: ML in nanomedicine primarily relies on existing data for optimization and prediction, excelling at pattern recognition within known knowledge but lacking creativity and imagination, which makes it difficult to break traditional design frameworks and explore novel formulations. In contrast, human researchers possess interdisciplinary thinking and are adept at breaking away from traditional research paradigms. Over-reliance on ML for formulation design may not only confine research to incremental extensions of existing data but also diminish potentially groundbreaking innovations, hindering the development of truly disruptive nanomedicines.

(3) Limited interpretability of biological mechanisms[122,132,185]: AI and ML models in nanomedicine suffer from limited interpretability and struggle to elucidate reaction mechanisms and structure−activity relationships, leading to predictions that often lack clear scientific rationale. Therefore, formulation scientists must still rely on extensive biochemical experiments to validate predictions, rather than directly gaining mechanistic insights from the models. This not only increases the experimental costs and resource consumption but may also slow the development of novel nanomedicines, limiting the practical utility of ML in the field.

(4) Mismatch between model complexity and data scale may undermine practical utility: the application of ML in nanomedicine is constrained by the balance between model complexity and data scale. Complex models typically require large datasets, but the limited data in nanomedicine increases the risk of overfitting, reducing predictive performance in real-world applications and potentially leading to misleading results that affect experimental decisions. Conversely, overly simplistic models may fail to capture the intricate structure−activity relationships of nanomedicines, also limiting their predictive power and providing insufficient guidance for formulation optimization. Striking an optimal balance between model complexity and data scale remains challenging, often requiring significant time investment from researchers. This difficulty may also undermine trust in ML, hindering its long-term adoption in formulation development and limiting its intended supportive role.

Moreover, the application of ML in nanomedicine still faces significant challenges that require further optimization. Currently, most ML-assisted nanomedicine development studies are limited by sparse datasets, simplistic characterization methods, and basic algorithms. In contrast, other biomedical fields have been transformed by large-scale datasets, biologically meaningful representations, and advanced DL and generative AI models. Large-scale pre-trained models, such as AlphaFold[186], ChatGPT[187], Sora[188], and Pangu-Weather[189], have revolutionized research from molecular structure prediction to disease diagnosis and treatment optimization. Therefore, fully realizing the potential of ML in the development of nanomedicine will entail overcoming significant challenges in the future.

4.1. Data

Firstly, the effectiveness of ML is highly dependent on the quality of data. Compared to small molecule databases such as PubChem and ZINC, or the Protein Data Bank (PDB), the field of nanodrug delivery lacks a comprehensive and well-established database, which significantly restricts the development of ML applications in nanomedicine. While existing datasets from more mature fields such as engineered NMs and nanotoxicology can provide some insights, they are not specifically designed for NDDS. Differences in data types, features, and applications make these datasets insufficient to fully address the needs of this particular field. These difficulties not only limit the performance of ML models but also hinder the application of complex algorithms. Developing a larger, more comprehensive, and richer nanomedicine database, which will require not only the collection of vast amounts of data but also the formulation of detailed data standardization protocols, is a primary challenge for future development.

Data mining and data creation: since the first introduction of liposomes in the 1960s, the achievements in nanomedicines have been steadily increasing year by year. In this context, there has been a vast accumulation of nanomedicine formulation structure and activity data. However, much of this data is buried in scientific literature, books, and other public publications, making it difficult to fully mine manually. Text mining tools based on fine-tuned large language models, which demonstrate the ability to effectively extract information from complex and diverse chemical texts[190], hold promise for the design of tailored tools that can automatically gather nanomaterial information from both text and images, thereby enriching existing nanomedicine databases. Additionally, high-throughput experiments, robotic automated synthesis, and microfluidic reaction systems have undergone rapid development in recent years,the integration of which with NMs will inevitably lead to the continuous creation of high-quality data. Finally, if necessary, reliable computational methods such as MDS can also be used to supplement nanomaterial data.

Data standardization:the lack of standardized nanomaterial data storage protocols has made it difficult to integrate existing nanomaterial databases, and their long-term absence will indeed further hinder the establishment of large-scale nanomaterial database platforms in the future. caNanoLab, NBIK, eNanoMapper and other databases have realized data sharing to a certain extent, but it is challenging to integrate their data into a large-scale data sharing platform due to the different data formats, standards, and storage methods; Regarding the aforementioned issues, on one hand, it is necessary to establish standardized experimental methods and evaluation criteria for the preparation and characterization of nanodrug formulations during the wet-lab data generation phase. On the other hand, it is necessary to standardize nanomaterial databases following the FAIR principles[191] to regulate the data format of nanodrug formulations within these databases and, as much as possible, convert the data into information that can be recognized by ML models, which would allow researchers to directly use the data for ML training without the need for complex data transformation, ultimately creating a Findable, Accessible, Interoperable, and Reusable (FAIR) nanomaterial database.

Feature engineering: in data characterization, current strategies primarily rely on experimental descriptors, which use quantitative or qualitative experimental data, resulting in only a rough representation of the shape, structure, and surface properties of nanodrug formulations, and failing to accurately characterize and annotate their 3D structures. Existing strategies for full nanostructure characterization are infrequent and typically limited to modeling simple metal or metal oxide NMs, while organic NMs, such as liposomes and LNPs, are challenging to model due to their variable structures, greater flexibility, instability, and larger size. Establishing in-depth and universal characterization strategies for nanodrug formulations is the second major challenge for future development.

4.2. Representation

To this end, through fundamental research and computational modeling, efforts should be made to jointly develop nanomaterial structural analysis tools, obtain precise structural information, and thereby drive the establishment of universal and efficient nanodrug descriptors, which should be open-access, computationally feasible, and interpretable, capable of effectively capturing the structural and physicochemical properties of nanodrug formulations. However, developing new nanodescriptors is extremely challenging. Converting the 3D/4D structural information of NMs (such as microscopy images, spectral data, MD data, etc.) into graph or image data, and employing DL strategies to automatically extract representations, is a viable subordinate alternative strategy.

Appropriately integrating advanced computational strategies is another major challenge. Prevailing research primarily focuses on using ML as a data analysis tool, aiming to explore the structure−activity relationship of nanomedicine and identify the key features that influence their performance, rather than focusing on building high-performance ML prediction platforms. To fully unleash the potential of ML, future improvements and optimizations could focus on the following areas.

4.3. Algorithm

A key challenge in developing ML models for nanomedicines is the rational integration of advanced computational strategies. Current studies primarily focus on utilizing ML as a data analysis tool to interpret the structure−activity relationships of nanomedicines and identify key features influencing their performance, rather than on constructing high-performance DL-based prediction platforms. To fully unlock the potential of ML, future improvements and optimizations can be approached from the following aspects:

Advanced deep learning predictive algorithm: most existing studies primarily use simple ML models, while some overlooked algorithms may be more suitable for addressing nanomedicine issues. For example, Few-shot learning is another DL strategy that

enables modeling with a limited number of samples and has been employed to address data scarcity in fields like CV and natural language processing. Nanomedicine modeling similarly faces challenges of limited data, where few-shot learning strategies can maximize the potential of available data and offer the possibility of constructing more efficient and robust prediction platforms for nanomedicine.

Generative model. Currently, ML and DL in nanomedicine focus on predictive models rather than innovative generative models. Researchers mainly use existing ML models to predict the structure or properties of nanomedicines. However, these models rely on pattern recognition within available data and hard to break existing design paradigms. In contrast, generative models have the potential to discover nanomedicines with novel structures, compositions, and structure−activity relationships. By integrating self-supervised learning (SSL), physics-informed DL and RL, specialized generative models for nanomedicine can be developed. This approach may enable a change from data-driven prediction to intelligent design, offering new solutions for personalized nanomedicines and precision medicine.

Interpretable algorithm: interpretability refers to assessing the extent to which the internal mechanisms of AI models in data analysis and decision-making can be understood by humans, significantly influencing researchers' trust and acceptance of AI models. Conducting interpretability analyses not only unveils the decision-making processes of models but also aids in optimizing them to construct higher-performing AI systems. General-purpose interpretability tools, such as SHAP analysis for interpreting predictions from ML models and techniques like attention mechanisms and Integrated Gradients for DL interpretability, are commonly employed for such analyses. On the other hand, using inherently interpretable models, such as tree-based models like DT or RF, can also enhance model interpretability. It is worth noting, however, that there is often a trade-off between accuracy and interpretability in ML. For example, linear regression models are easier to interpret but lack fitting capacity, whereas DL models like NNs achieve higher prediction accuracy but suffer from poor interpretability. Therefore, model construction often requires balancing these two aspects.

Applicability domain analysis: the applicability domain (AD) defines the potential material space that a model can reliably predict, reflecting its generalization capability and providing critical guidance for its real-world deployment. However, most existing models are built on datasets with limited types and structures of NMs, leading to low generalization ability. These models often have lower prediction accuracy for data types not seen in the training set, limiting their application scope. Moreover, universal methods based on distance, geometry, and probability density struggle to distinguish the AD of complex ML models, further exacerbating researchers' neglect of applicability domain analysis. Aiming at the above problems, it is essential to develop large, diverse databases of NMs with enriched structural and compositional variety for training high-generalization ML models. Additionally, AD analysis methods tailored to the characteristics of specific datasets and models need to be developed for domain-specific applications[192].

Multi-property predictive algorithms. Current studies often collect data and build ML models for a single task, making it difficult to achieve unified predictions across multiple downstream tasks. However, many of these tasks share common features and interconnections. For example, predicting formulation behavior in blood, pharmacokinetic parameters, and nanomedicine−protein interactions all focus on *in vivo* processes and are influenced by similar factors. Moreover, ML algorithms in formulation research are primarily used for early-stage optimization, helping researchers narrow experimental conditions rather than fully guiding nanomedicine development. As a result, the demand for extremely high model accuracy is relatively low, it might not be necessary to develop an algorithm for a specific task. By integrating the universal nanomedicine characterization strategies discussed in section 4.2., multi-attribute prediction algorithms can be further developed. This approach not only reduces computational costs but also facilitates the creation of ML platforms with broad applicability, strong generalization, and high practical value.

5. Conclusions

With the progressive accumulation of data and the rapid advancements in algorithms, AI and ML have demonstrated significant potential in advancing nanomedicine research, encompassing the entire nanomedicine development process. However, they remain in their infancy, and significant efforts are still required at the computational level, including in data, characterization, and algorithms, as well as at the nanomedicine level, encompassing formulation design, *in vivo* processes, and downstream applications. The integration of advanced computational strategies and experimental techniques will significantly accelerate ML-driven nanomedicine research, resulting in the safe, effective, stable, and quality-controlled nanotechnologies.

Acknowledgments

This work was supported by the Research Foundation of the Education Bureau of Hunan Province (No. 23A0002, China) and National Key R&D Program of China (No. 2023YFC2604400 to Song He).

Author contributions

Ziye Wei, Shijie Zhuo and Yixin Zhang designed and wrote the paper and prepared Figures. Lianlian Wu participated part of the edit. Xiang Gao, Song He, Xiaochen Bo and Wenhu Zhou edited and revised the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Yao S, Ren P, Song R, Liu Y, Huang Q, Dong J, et al. Nanomaterial-enabled flexible and stretchable sensing systems: processing, integration, and applications. *Adv Mater* 2020;**32**:e1902343.
2. Zhang A, Lieber CM. Nano-bioelectronics. *Chem Rev* 2016;**116**: 215−57.
3. Bilal M, Iqbal HMN, Adil SF, Shaik MR, Abdelgawad A, Hatshan MR, et al. Surface-coated magnetic nanostructured materials for robust bio-catalysis and biomedical applications—a review. *J Adv Res* 2022;**38**:157−77.
4. Wang C, O'Hagan MP, Li Z, Zhang J, Ma X, Tian H, et al. Photoresponsive DNA materials and their applications. *Chem Soc Rev* 2022;**51**:720−60.

5. Peer D, Karp JM, Hong S, Farokhzad OC, Margalit R, Langer R. Nanocarriers as an emerging platform for cancer therapy. *Nat Nanotechnol* 2007;**2**:751−60.
6. Tang L, Li J, Pan T, Yin Y, Mei Y, Xiao Q, et al. Versatile carbon nanoplatforms for cancer treatment and diagnosis: strategies, applications and future perspectives. *Theranostics* 2022;**12**:2290−321.
7. Quader S, Kataoka K. Nanomaterial-enabled cancer therapy. *Mol Ther* 2017;**25**:1501−13.
8. Vail DM, Amantea MA, Colbern GT, Martin FJ, Hilger RA, Working PK. Pegylated liposomal doxorubicin: proof of principle using preclinical animal models and pharmacokinetic studies. *Semin Oncol* 2004;**31**:16−35.
9. Kaplon H, Crescioli S, Chenoweth A, Visweswaraiah J, Reichert JM. Antibodies to watch in 2023. *mAbs* 2023;**15**:2153410.
10. Keam SJ. Ozoralizumab: first approval. *Drugs (Basel)* 2023;**83**: 87−92.
11. Sun Q, Zhou Z, Qiu N, Shen Y. Rational design of cancer nanomedicine: nanoproperty integration and synchronization. *Adv Mater* 2017;**29**.
12. Figueiró Longo JP, Narcizo de Souza PE, Morais PC. Artificial intelligence and machine learning in nanomedicine. What do we expect for 2030?. *Nanomedicine* 2023;**18**:1041−3.
13. Cheng Z, Li M, Dey R, Chen Y. Nanomaterials for cancer therapy: current progress and perspectives. *J Hematol Oncol* 2021;**14**:85.
14. Blanco E, Shen H, Ferrari M. Principles of nanoparticle design for overcoming biological barriers to drug delivery. *Nat Biotechnol* 2015;**33**:941−51.
15. Tahmasbi Rad A, Chen CW, Aresh W, Xia Y, Lai PS, Nieh MP. Combinational effects of active targeting, shape, and enhanced permeability and retention for cancer theranostic nanocarriers. *ACS Appl Mater Interfaces* 2019;**11**:10505−19.
16. Wu L, Wen Y, Leng D, Zhang Q, Dai C, Wang Z, et al. Machine learning methods, databases and tools for drug combination prediction. *Briefings Bioinf* 2021;**23**:bbab355.
17. Shao K, Zhang Y, Wen Y, Zhang Z, He S, Bo X. DTI-HETA: prediction of drug−target interactions based on GCN and GAT on heterogeneous graph. *Briefings Bioinf* 2022;**23**:bbac109.
18. Sarkar C, Das B, Rawat VS, Wahlang JB, Nongpiur A, Tiewsoh I, et al. Artificial intelligence and machine learning technology driven modern drug discovery and development. *Int J Mol Sci* 2023;**24**: 2026.
19. Singh AV, Ansari MHD, Rosenkranz D, Maharjan RS, Kriegel FL, Gandhi K, et al. Artificial intelligence and machine learning in computational nanotoxicology: unlocking and empowering nanomedicine. *Adv Healthcare Mater* 2020;**9**:1901862.
20. Tan P, Chen XT, Zhang H, Wei Q, Luo K. Artificial intelligence aids in development of nanomedicines for cancer management. *Semin Cancer Biol* 2023;**89**:61−75.
21. Lv H, Chen X. Intelligent control of nanoparticle synthesis through machine learning. *Nanoscale* 2022;**14**:6688−708.
22. Bonaccorso A, Ortis A, Musumeci T, Carbone C, Hussain M, Di Salvatore V, et al. Nose-to-brain drug delivery and physico-chemical properties of nanosystems: analysis and correlation studies of data from scientific literature. *Int J Nanomed* 2024;**19**:5619−36.
23. Arshad I, Kanwal A, Zafar I, Unar A, Mouada H, Razia IT, et al. Multifunctional role of nanoparticles for the diagnosis and therapeutics of cardiovascular diseases. *Environ Res* 2024;**242**:117795.
24. Mi K, Chou WC, Chen Q, Yuan L, Kamineni VN, Kuchimanchi Y, et al. Predicting tissue distribution and tumor delivery of nanoparticles in mice using machine learning models. *J Control Release* 2024;**374**:219−29.
25. Zhang J, Mucs D, Norinder U, Svensson F. LightGBM: an effective and scalable algorithm for prediction of chemical toxicity-application to the tox21 and mutagenicity data sets. *J Chem Inf Model* 2019;**59**: 4150−8.
26. Su R, Wu H, Liu X, Wei L. Predicting drug-induced hepatotoxicity based on biological feature maps and diverse classification strategies. *Briefings Bioinf* 2021;**22**:428−37.
27. Xu Y, Yao H, Lin K. An overview of neural networks for drug discovery and the inputs used. *Expet Opin Drug Discov* 2018;**13**: 1091−102.
28. Chen C, Yaari Z, Apfelbaum E, Grodzinski P, Shamay Y, Heller DA. Merging data curation and machine learning to improve nanomedicines. *Adv Drug Deliv Rev* 2022;**183**:114172.
29. Wang H. Prediction of protein−ligand binding affinity *via* deep learning models. *Briefings Bioinf* 2024;**25**:bbae310.
30. Furxhi I, Murphy F, Mullins M, Arvanitis A, Poland CA. Practices and trends of machine learning application in nanotoxicology. *Nanomaterials* 2020;**10**:116.
31. Singh AV, Varma M, Laux P, Choudhary S, Datusalia AK, Gupta N, et al. Artificial intelligence and machine learning disciplines with the potential to improve the nanotoxicology and nanomedicine fields: a comprehensive review. *Arch Toxicol* 2023;**97**:963−79.
32. Lin Z, Chou WC, Cheng YH, He C, Monteiro Riviere NA, Riviere JE. Predicting nanoparticle delivery to tumors using machine learning and artificial intelligence approaches. *Int J Nanomed* 2022; **17**:1365−79.
33. Wang W, Feng S, Ye Z, Gao H, Lin J, Ouyang D. Prediction of lipid nanoparticles for mRNA vaccines by the machine learning algorithm. *Acta Pharm Sin B* 2022;**12**:2950−62.
34. Mendes BB, Zhang Z, Conniot J, Sousa DP, Ravasco JMJM, Onweller LA, et al. A large-scale machine learning analysis of inorganic nanoparticles in preclinical cancer research. *Nat Nanotechnol* 2024;**19**:867−78.
35. Li B, Raji IO, Gordon AGR, Sun L, Raimondo TM, Oladimeji FA, et al. Accelerating ionizable lipid discovery for mRNA delivery using machine learning and combinatorial chemistry. *Nat Mater* 2024:1−7.
36. Torng W, Altman RB. Graph convolutional neural networks for predicting drug-target interactions. *J Chem Inf Model* 2019;**59**:4131−49.
37. Zhao T, Hu Y, Valsdottir LR, Zang T, Peng J. Identifying drug-target interactions based on graph convolutional network and deep neural network. *Briefings Bioinf* 2021;**22**:2141−50.
38. Zhang Y, Hu Y, Han N, Yang A, Liu X, Cai H. A survey of drug-target interaction and affinity prediction methods *via* graph neural networks. *Comput Biol Med* 2023;**163**:107136.
39. Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, Omori S, et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature* 2024;**626**:177−85.
40. Choudhary K, DeCost B, Chen C, Jain A, Tavazza F, Cohn R, et al. Recent advances and applications of deep learning methods in materials science. *npj Comput Mater* 2022;**8**:1−26.
41. Tripathy A, Patne AY, Mohapatra S, Mohapatra SS. Convergence of nanotechnology and machine learning: the state of the art, challenges, and perspectives. *Int J Mol Sci* 2024;**25**:12368.
42. Mei H, Cai S, Huang D, Gao H, Cao J, He B. Carrier-free nanodrugs with efficient drug delivery and release for cancer therapy: from intrinsic physicochemical properties to external modification. *Bioact Mater* 2021;**8**:220−40.
43. Wei Y, Wu J, Wu Y, Liu H, Meng F, Liu Q, et al. Prediction and design of nanozymes using explainable machine learning. *Adv Mater* 2022;**34**:e2201736.
44. Zhou K, Tian B, Lu J, Dong B, Xu H. Machine learning-guided synthesis of nanomaterials for breast cancer therapy. *Sci Rep* 2024; **14**:25795.
45. Yang MY, Zhao RR, Fang YF, Jiang JL, Yuan XT, Shao JW. Carrier-free nanodrug: a novel strategy of cancer diagnosis and synergistic therapy. *Int J Pharm* 2019;**570**:118663.
46. Zhang X, Li N, Zhang S, Sun B, Chen Q, He Z, et al. Emerging carrier-free nanosystems based on molecular self-assembly of pure drugs for cancer therapy. *Med Res Rev* 2020;**40**:1754−75.
47. Zhong YT, Cen Y, Xu L, Li SY, Cheng H. Recent progress in carrier-free nanomedicine for tumor phototherapy. *Adv Healthcare Mater* 2023;**12**:2202307.
48. Fu S, Li G, Zang W, Zhou X, Shi K, Zhai Y. Pure drug nanoassemblies: a facile carrier-free nanoplatform for efficient cancer therapy. *Acta Pharm Sin B* 2022;**12**:92−106.

49. Shamay Y, Shah J, Işık M, Mizrachi A, Leibold J, Tschaharganeh DF, et al. Quantitative self-assembly prediction yields targeted nanomedicines. *Nat Mater* 2018;**17**:361−8.
50. Azagury DM, Gluck BF, Harris Y, Avrutin Y, Niezni D, Sason H, et al. Prediction of cancer nanomedicines self-assembled from meta-synergistic drug pairs. *J Control Release* 2023;**360**:418−32.
51. Reker D, Rybakova Y, Kirtane AR, Cao R, Yang JW, Navamajiti N, et al. Computationally guided high-throughput design of self-assembling drug nanoparticles. *Nat Nanotechnol* 2021;**16**:725−33.
52. Albanese A, Tang PS, Chan WC. The effect of nanoparticle size, shape, and surface chemistry on biological systems. *Annu Rev Biomed Eng* 2012;**14**:1−16.
53. Cabral H, Matsumoto Y, Mizuno K, Chen Q, Murakami M, Kimura M, et al. Accumulation of sub-100 nm polymeric micelles in poorly permeable tumours depends on size. *Nat Nanotechnol* 2011;**6**:815−23.
54. Chithrani BD, Ghazani AA, Chan WC. Determining the size and shape dependence of gold nanoparticle uptake into mammalian cells. *Nano Lett* 2006;**6**:662−8.
55. Yue ZG, Wei W, Lv PP, Yue H, Wang LY, Su ZG, et al. Surface charge affects cellular uptake and intracellular trafficking of chitosan-based nanoparticles. *Biomacromolecules* 2011;**12**:2440−6.
56. Hoshyar N, Gray S, Han H, Bao G. The effect of nanoparticle size on *in vivo* pharmacokinetics and cellular interaction. *Nanomedicine* 2016;**11**:673−92.
57. Fahmy OM, Eissa RA, Mohamed HH, Eissa NG, Elsabahy M. Machine learning algorithms for prediction of entrapment efficiency in nanomaterials. *Methods* 2023;**218**:133−40.
58. He Y, Ye Z, Liu X, Wei Z, Qiu F, Li HF, et al. Can machine learning predict drug nanocrystals?. *J Control Release* 2020;**322**:274−85.
59. Bayat F, Dadashzadeh S, Aboofazeli R, Torshabi M, Baghi AH, Tamiji Z, et al. Oral delivery of posaconazole-loaded phospholipid-based nanoformulation: preparation and optimization using design of experiments, machine learning, and TOPSIS. *Int J Pharm* 2024;**653**:123879.
60. Kimmig J, Schuett T, Vollrath A, Zechel S, Schubert US. Prediction of nanoparticle sizes for arbitrary methacrylates using artificial neuronal networks. *Adv Sci (Weinh)* 2021;**8**:2102429.
61. Stiepel RT, Pena ES, Ehrenzeller SA, Gallovic MD, Lifshits LM, Genito CJ, et al. A predictive mechanistic model of drug release from surface eroding polymeric nanoparticles. *J Control Release* 2022;**351**:883−95.
62. Xia J, Gan Z, Zhang J, Dong M, Liu S, Cui B, et al. Geometric-aware deep learning enables discovery of bifunctional ligand-based liposomes for tumor targeting therapy. *Nano Today* 2025;**61**:102668.
63. Jo J, Kwak B, Choi H-S, Yoon S. The message passing neural networks for chemical property prediction on SMILES. *Methods* 2020;**179**:65−72.
64. Sato K, Sasaki N, Svahn HA, Sato K. Microfluidics for nano-pathophysiology. *Adv Drug Deliv Rev* 2014;**74**:115−21.
65. Gimondi S, Ferreira H, Reis RL, Neves NM. Microfluidic devices: a tool for nanoparticle synthesis and performance evaluation. *ACS Nano* 2023;**17**:14205−28.
66. Rebollo R, Oyoun F, Corvis Y, El-Hammadi MM, Saubamea B, Andrieux K, et al. Microfluidic manufacturing of liposomes: development and optimization by design of experiment and machine learning. *ACS Appl Mater Interfaces* 2022;**14**:39736−45.
67. Di Francesco V, Boso DP, Moore TL, Schrefler BA, Decuzzi P. Machine learning instructed microfluidic synthesis of curcumin-loaded liposomes. *Biomed Microdevices* 2023;**25**:29.
68. Xie M, Shimogawa R, Liu Y, Zhang L, Foucher AC, Routh PK, et al. Biomimetic control over bimetallic nanoparticle structure and activity *via* peptide capping ligand sequence. *ACS Nano* 2024;**18**:3286−94.
69. Pan K, Bricker WP, Ratanalert S, Bathe M. Structure and conformational dynamics of scaffolded DNA origami nanoparticles. *Nucleic Acids Res* 2017;**45**:6284−98.
70. Zhu G, Mei L, Vishwasrao HD, Jacobson O, Wang Z, Liu Y, et al. Intertwining DNA−RNA nanocapsules loaded with tumor neoantigens as synergistic nanovaccines for cancer immunotherapy. *Nat Commun* 2017;**8**:1482.
71. Wang S, Liu T, Jiang D. Locating hydrides in ligand-protected copper nanoclusters by deep learning. *ACS Appl Mater Interfaces* 2021;**13**:53468−74.
72. Dijkstra M, Luijten E. From predictive modelling to machine learning and reverse engineering of colloidal self-assembly. *Nat Mater* 2021;**20**:762−73.
73. Pink DL, Loruthai O, Ziolek RM, Terry AE, Barlow DJ, Lawrence MJ, et al. Interplay of lipid and surfactant: impact on nanoparticle structure. *J Colloid Interface Sci* 2021;**597**:278−88.
74. Hajipour MJ, Safavi-Sohi R, Sharifi S, Mahmoud N, Ashkarran AA, Voke E, et al. An overview of nanoparticle protein corona literature. *Small* 2023;**19**:2301838.
75. Saldinger JC, Raymond M, Elvati P, Violi A. Domain-agnostic predictions of nanoscale interactions in proteins and nanoparticles. *Nat Comput Sci* 2023;**3**:393−402.
76. Pihlajamäki A, Matus MF, Malola S, Häkkinen H. GraphBNC: machine learning-aided prediction of interactions between metal nanoclusters and blood proteins. *Adv Mater* 2024;**36**:2407046.
77. Findlay MR, Freitas DN, Mobed-Miremadi M, Wheeler KE. Machine learning provides predictive analysis into silver nanoparticle protein corona formation from physicochemical properties. *Environ Sci Nano* 2018;**5**:64−71.
78. Ban Z, Yuan P, Yu F, Peng T, Zhou Q, Hu X. Machine learning predicts the functional composition of the protein corona and the cellular recognition of nanoparticles. *Proc Natl Acad Sci U S A* 2020;**117**:10492−9.
79. Sengottiyan S, Mikolajczyk A, Jagiełło K, Swirog M, Puzyn T. Core, coating, or corona? The importance of considering protein coronas in nano-QSPR modeling of zeta potential. *ACS Nano* 2023;**17**:1989−97.
80. Lazarovits J, Sindhwani S, Tavares AJ, Zhang Y, Song F, Audet J, et al. Supervised learning and mass spectrometry predicts the *in vivo* fate of nanomaterials. *ACS Nano* 2019;**13**:8023−34.
81. Behzadi S, Serpooshan V, Tao W, Hamaly MA, Alkawareek MY, Dreaden EC, et al. Cellular uptake of nanoparticles: journey inside the cell. *Chem Soc Rev* 2017;**46**:4218−44.
82. Almeida MS de, Susnik E, Drasler B, Taladriz-Blanco P, Petri Fink A, Rothen-Rutishauser B. Understanding nanoparticle endocytosis to improve targeting strategies in nanomedicine. *Chem Soc Rev* 2021;**50**:5397−434.
83. Boehnke N, Straehla JP, Safford HC, Kocak M, Rees MG, Ronan M, et al. Massively parallel pooled screening reveals genomic determinants of nanoparticle delivery. *Science* 2022;**377**:eabm5551.
84. Wang W, Sedykh A, Sun H, Zhao L, Russo DP, Zhou H, et al. Predicting nano−bio interactions by integrating nanoparticle libraries and quantitative nanostructure activity relationship modeling. *ACS Nano* 2017;**11**:12641−9.
85. Yan X, Sedykh A, Wang W, Zhao X, Yan B, Zhu H. *In silico* profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale* 2019;**11**:8352−62.
86. Yan X, Zhang J, Russo DP, Zhu H, Yan B. Prediction of nano−bio interactions through convolutional neural network analysis of nanostructure images. *ACS Sustainable Chem Eng* 2020;**8**:19096−104.
87. Russo DP, Yan X, Shende S, Huang H, Yan B, Zhu H. Virtual molecular projections and convolutional neural networks for the end-to-end modeling of nanoparticle activities and properties. *Anal Chem* 2020;**92**:13971−9.
88. Zhong Y, Li C, Zhou H, Wang G. Developing noise-resistant three-dimensional single particle tracking using deep neural networks. *Anal Chem* 2018;**90**:10748−57.
89. Zhang P, Liu S, Chaurasia A, Ma D, Mlodzianoski MJ, Culurciello E, et al. Analyzing complex single-molecule emission patterns with deep learning. *Nat Methods* 2018;**15**:913−6.
90. Hu J, Liu T, Choo P, Wang S, Reese T, Sample AD, et al. Single-nanoparticle orientation sensing by deep learning. *ACS Cent Sci* 2020;**6**:2339−46.

91. Song D, Zhang X, Li B, Sun Y, Mei H, Cheng X, et al. Deep learning-assisted automated multidimensional single particle tracking in living cells. *Nano Lett* 2024;**24**:3082−8.
92. Liu J, Zhang J, Gao Y, Jiang Y, Guan Z, Xie Y, et al. Barrier permeation and improved nanomedicine delivery in tumor microenvironments. *Cancer Lett* 2023;**562**:216166.
93. Colling ME, Tourdot BE, Kanthi Y. Inflammation, infection and venous thromboembolism. *Circ Res* 2021;**128**:2017−36.
94. Zhuang J, Zhang X, Liu Q, Zhu M, Huang X. Targeted delivery of nanomedicines for promoting vascular regeneration in ischemic diseases. *Theranostics* 2022;**12**:6223−41.
95. Vuong TNAM, Bartolf Kopp M, Andelovic K, Jungst T, Farbehi N, Wise SG, et al. Integrating computational and biological hemodynamic approaches to improve modeling of atherosclerotic arteries. *Adv Sci* 2024;**11**:2307627.
96. Stillman NR, Balaz I, Tsompanas MA, Kovacevic M, Azimi S, Lafond S, et al. Evolutionary computational platform for the automatic discovery of nanocarriers for cancer treatment. *npj Comput Mater* 2021;**7**:1−12.
97. Gomez-Garcia MJ, Abdelkarim M, Cramb DT, Childs SJ, Rinker KD, Labouta HI. Blood vessel wall shear stress determines regions of liposome accumulation in angiogenic vasculature. *Drug Deliv and Transl Res* 2024;**14**:3608−20.
98. Wang S, Zhang J, Liu H, Guo X, Lu Y, Guan G, et al. The effect of blood velocity in solid tumor on intratumorally accumulation and penetration of nanocarriers and drugs. *Nano Today* 2023;**50**:101870.
99. Chen YY, Syed AM, MacMillan P, Rocheleau JV, Chan WCW. Flow rate affects nanoparticle uptake into endothelial cells. *Adv Mater* 2020;**32**:1906274.
100. Lim J, Ching H, Yoon JK, Jeon NL, Kim Y. Microvascularized tumor organoids-on-chips: advancing preclinical drug screening with pathophysiological relevance. *Nano Convergence* 2021;**8**:12.
101. Li X, Peng X, Zoulikha M, Boafo GF, Magar KT, Ju Y, et al. Multifunctional nanoparticle-mediated combining therapy for human diseases. *Signal Transduct Targeted Ther* 2024;**9**:1−33.
102. Wilhelm S, Tavares AJ, Dai Q, Ohta S, Audet J, Dvorak HF, et al. Analysis of nanoparticle delivery to tumours. *Nat Rev Mater* 2016;**1**:1−12.
103. Sindhwani S, Syed AM, Ngai J, Kingston BR, Maiorino L, Rothschild J, et al. The entry of nanoparticles into solid tumours. *Nat Mater* 2020;**19**:566−75.
104. Mitchell MJ, Billingsley MM, Haley RM, Wechsler ME, Peppas NA, Langer R. Engineering precision nanoparticles for drug delivery. *Nat Rev Drug Discov* 2021;**20**:101−24.
105. May JN, Moss JI, Mueller F, Golombek SK, Biancacci I, Rizzo L, et al. Histopathological biomarkers for predicting the tumour accumulation of nanomedicines. *Nat Biomed Eng* 2024;**8**:1366−78.
106. Zhu M, Zhuang J, Li Z, Liu Q, Zhao R, Gao Z, et al. Machine-learning-assisted single-vessel analysis of nanoparticle permeability in tumour vasculatures. *Nat Nanotechnol* 2023;**18**:657−66.
107. Dhaliwal A, Ma J, Zheng M, Lyu Q, Rajora MA, Ma S, et al. Deep learning for automatic organ and tumor segmentation in nanomedicine pharmacokinetics. *Theranostics* 2024;**14**:973−87.
108. MacMillan P, Syed AM, Kingston BR, Ngai J, Sindhwani S, Lin ZP, et al. Toward predicting nanoparticle distribution in heterogeneous tumor tissues. *Nano Lett* 2023;**23**:7197−205.
109. Yang L, Liu Q, Kumar P, Sengupta A, Farnoud A, Shen R, et al. LungVis 1.0: an automatic AI-powered 3D imaging ecosystem unveils spatial profiling of nanoparticle delivery and acinar migration of lung macrophages. *Nat Commun* 2024;**15**:10138.
110. Tang Y, Zhang J, He D, Miao WF, Liu W, Li Y, et al. GANDA: a deep generative adversarial network conditionally generates intratumoral nanoparticles distribution pixels-to-pixels. *J Control Release* 2021;**336**:336−43.
111. Chen Q, Yuan L, Chou WC, Cheng YH, He C, Monteiro Riviere NA, et al. Meta-analysis of nanoparticle distribution in tumors and major organs in tumor-bearing mice. *ACS Nano* 2023;**17**:19810−31.
112. Chou WC, Chen Q, Yuan L, Cheng YH, He C, Monteiro Riviere NA, et al. An artificial intelligence-assisted physiologically-based pharmacokinetic model to predict nanoparticle delivery to tumors in mice. *J Control Release* 2023;**361**:53−63.
113. Ma X, Tang Y, Wang C, Li Y, Zhang J, Luo Y, et al. Interpretable xgboost-shap model predicts nanoparticles delivery efficiency based on tumor genomic mutations and nanoparticle properties. *ACS Appl Bio Mater* 2023;**6**:4326−35.
114. Winkler DA. Role of artificial intelligence and machine learning in nanosafety. *Small* 2020;**16**:2001883.
115. Labouta HI, Asgarian N, Rinker K, Cramb DT. Meta-analysis of nanoparticle cytotoxicity *via* data-mining the literature. *ACS Nano* 2019;**13**:1583−94.
116. Huang Y, Li X, Cao J, Wei X, Li Y, Wang Z, et al. Use of dissociation degree in lysosomes to predict metal oxide nanoparticle toxicity in immune cells: machine learning boosts nano-safety assessment. *Environ Int* 2022;**164**:107258.
117. Gul G, Yildirim R, Ileri Ercan N. Cytotoxicity analysis of nanoparticles by association rule mining. *Environ Sci Nano* 2021;**8**:937−49.
118. Yan X, Yue T, Winkler DA, Yin Y, Zhu H, Jiang G, et al. Converting nanotoxicity data to information using artificial intelligence and simulation. *Chem Rev* 2023;**123**:8575−637.
119. Li J, Wang C, Yue L, Chen F, Cao X, Wang Z. Nano-QSAR modeling for predicting the cytotoxicity of metallic and metal oxide nanoparticles: a review. *Ecotoxicol Environ Saf* 2022;**243**:113955.
120. Ji Z, Guo W, Wood EL, Liu J, Sakkiah S, Xu X, et al. Machine learning models for predicting cytotoxicity of nanomaterials. *Chem Res Toxicol* 2022;**35**:125−39.
121. Martin null, Watanabe R, Hashimoto K, Higashisaka K, Haga Y, Tsutsumi Y, et al. Evidence-based prediction of cellular toxicity for amorphous silica nanoparticles. *ACS Nano* 2023;**17**:9987−99.
122. Meneses J, González-Durruthy M, Fernandez-de-Gortari E, Toropova AP, Toropov AA, Alfaro-Moreno E. A nano-QSTR model to predict nano-cytotoxicity: an approach using human lung cells data. *Part Fibre Toxicol* 2023;**20**:21.
123. Shirokii N, Din Y, Petrov I, Seregin Y, Sirotenko S, Razlivina J, et al. Quantitative prediction of inorganic nanomaterial cellular toxicity *via* machine learning. *Small* 2023;**19**:e2207106.
124. Zhu Y, Huang R, Zhu R, Xu W, Zhu R, Cheng L. DeepScreen: an accurate, rapid, and anti-interference screening approach for nanoformulated medication by deep learning. *Adv Sci* 2018;**5**:1800909.
125. Yu F, Wei C, Deng P, Peng T, Hu X. Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles. *Sci Adv* 2021;**7**:eabf4130.
126. Guo P, Huang J, Moses MA. Cancer nanomedicines in an evolving oncology landscape. *Trends Pharmacol Sci* 2020;**41**:730−42.
127. Copp SM, Gorovits A, Swasey SM, Gudibandi S, Bogdanov P, Gwinn EG. Fluorescence color by data-driven design of genomic silver clusters. *ACS Nano* 2018;**12**:8240−7.
128. Kelich P, Jeong S, Navarro N, Adams J, Sun X, Zhao H, et al. Discovery of DNA−carbon nanotube sensors for serotonin with machine learning and near-infrared fluorescence spectroscopy. *ACS Nano* 2022;**16**:736−45.
129. Alafeef M, Srivastava I, Pan D. Machine learning for precision breast cancer diagnosis and prediction of the nanoparticle cellular internalization. *ACS Sens* 2020;**5**:1689−98.
130. Shin H, Oh S, Hong S, Kang M, Kang D, Ji Y-G, et al. Early-stage lung cancer diagnosis by deep learning-based spectroscopic analysis of circulating exosomes. *ACS Nano* 2020;**14**:5435−44.
131. Wang W, Chen K, Jiang T, Wu Y, Wu Z, Ying H, et al. Artificial intelligence-driven rational design of ionizable lipids for mRNA delivery. *Nat Commun* 2024;**15**:10804.
132. Hunter MR, Cui L, Porebski BT, Pereira S, Sonzini S, Odunze U, et al. Understanding intracellular biology to improve mRNA delivery by lipid nanoparticles. *Small Methods* 2023;**7**:e2201695.

133. Chandler M, Jain S, Halman J, Hong E, Dobrovolskaia MA, Zakharov AV, et al. Artificial immune cell, AI-cell, a new tool to predict interferon production by peripheral blood monocytes in response to nucleic acid nanoparticles. *Small* 2022;**18**:2204941.
134. Yamankurt G, Berns EJ, Xue A, Lee A, Bagheri N, Mrksich M, et al. Exploration of the nanomedicine-design space with high-throughput screening and machine learning. *Nat Biomed Eng* 2019;**3**:318−27.
135. Na M, Nam SH, Moon K, Kim J. Development of a nano-QSAR model for predicting the toxicity of nano-metal oxide mixtures to *Aliivibrio fischeri*. *Environ Sci Nano* 2023;**10**:325−37.
136. Mirzaei M, Furxhi I, Murphy F, Mullins M. A machine learning tool to predict the antibacterial capacity of nanoparticles. *Nanomaterials* 2021;**11**:1774.
137. Yu T, Su S, Hu J, Zhang J, Xianyu Y. A new strategy for microbial taxonomic identification through micro-biosynthetic gold nanoparticles and machine learning. *Adv Mater* 2022;**34**:e2109365.
138. Kenry. Machine-learning-guided quantitative delineation of cell morphological features and responses to nanomaterials. *Nanoscale* 2024;**16**:19656−68.
139. Krasley AT, Li E, Galeana JM, Bulumulla C, Beyene AG, Demirer GS. Carbon nanomaterial fluorescent probes and their biological applications. *Chem Rev* 2024;**124**:3085−185.
140. He X, Htoon H, Doorn SK, Pernice WHP, Pyatkov F, Krupke R, et al. Carbon nanotubes as emerging quantum-light sources. *Nat Mater* 2018;**17**:843.
141. Sistemich L, Galonska P, Stegemann J, Ackermann J, Kruss S. Near-infrared fluorescence lifetime imaging of biomolecules with carbon nanotubes. *Angew Chem Int Ed Engl* 2023;**62**:e202300682.
142. Danai L, Rolband LA, Perdomo VA, Skelly E, Kim T, Afonin KA. Optical, structural and antibacterial properties of silver nanoparticles and DNA-templated silver nanoclusters. *Nanomedicine* 2023;**18**: 769−82.
143. Chinen AB, Guan CM, Ferrer JR, Barnaby SN, Merkel TJ, Mirkin CA. Nanoparticle probes for the detection of cancer biomarkers, cells, and tissues by fluorescence. *Chem Rev* 2015;**115**: 10530−74.
144. Sobhanan J, Anas A, Biju V. Nanomaterials for fluorescence and multimodal bioimaging. *Chem Rec* 2023;**23**:e202200253.
145. Beyene AG, Alizadehmojarad AA, Dorlhiac G, Goh N, Streets AM, Král P, et al. Ultralarge modulation of fluorescence by neuromodulators in carbon nanotubes functionalized with self-assembled oligonucleotide rings. *Nano Lett* 2018;**18**:6995−7003.
146. Mastracco P, Gonzàlez-Rosell A, Evans J, Bogdanov P, Copp SM. Chemistry-informed machine learning enables discovery of DNA-stabilized silver nanoclusters with near-infrared fluorescence. *ACS Nano* 2022;**16**:16322−31.
147. Jeong S, Yang D, Beyene AG, Del Bonis-O'Donnell JT, Gest AMM, Navarro N, et al. High-throughput evolution of near-infrared serotonin nanosensors. *Sci Adv* 2019;**5**:eaay3771.
148. Yu M, Zheng J. Clearance pathways and tumor targeting of imaging nanoparticles. *ACS Nano* 2015;**9**:6655.
149. Pallares RM, Mottaghy FM, Schulz V, Kiessling F, Lammers T. Nanoparticle diagnostics and theranostics in the clinic. *J Nucl Med* 2022;**63**:1802.
150. Ma J, Wang S, Zhao C, Yan X, Ren Q, Dong Z, et al. Computer-aided discovery of potent broad-spectrum vaccine adjuvants. *Angew Chem Int Ed Engl* 2023;**62**:e202301059.
151. Vora LK, Gholap AD, Jetha K, Thakur RRS, Solanki HK, Chavda VP. Artificial intelligence in pharmaceutical technology and drug delivery design. *Pharmaceutics* 2023;**15**:1916.
152. Panneerselvam S, Choi S. Nanoinformatics: emerging databases and available tools. *Int J Mol Sci* 2014;**15**:7158−82.
153. Sadan T, Cormode DP, Popovtzer R. Nanoinformatics revolutionizes personalized cancer therapy. *Trends Cancer* 2018;**4**:397−9.
154. Ke W, Crist RM, Clogston JD, Stern ST, Dobrovolskaia MA, Grodzinski P, et al. Trends and patterns in cancer nanotechnology research: a survey of NCI's caNanoLab and nanotechnology characterization laboratory. *Adv Drug Deliv Rev* 2022;**191**:114591.
155. Ji Z, Guo W, Sakkiah S, Liu J, Patterson TA, Hong H. Nanomaterial databases: data sources for promoting design and risk assessment of nanomaterials. *Nanomaterials* 2021;**11**:1599.
156. Thomas DG, Gaheen S, Harper SL, Fritts M, Klaessig F, Hahn Dantona E, et al. ISA−TAB−Nano: a specification for sharing nanomaterial research data in spreadsheet-based format. *BMC Biotechnol* 2013;**13**:2.
157. Yan X, Sedykh A, Wang W, Yan B, Zhu H. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat Commun* 2020;**11**:2519.
158. Wang T, Russo DP, Demokritou P, Jia X, Huang H, Yang X, et al. An online nanoinformatics platform empowering computational modeling of nanomaterials by nanostructure annotations and machine learning toolkits. *Nano Lett* 2024;**24**:10228−36.
159. Jeliazkova N, Chomenidis C, Doganis P, Fadeel B, Grafström R, Hardy B, et al. The eNanoMapper database for nanomaterial safety information. *Beilstein J Nanotechnol* 2015;**6**:1609−34.
160. Trinh TX, Ha MK, Choi JS, Byun HG, Yoon TH. Curation of datasets, assessment of their quality and completeness, and nanoSAR classification model development for metallic nanoparticles. *Environ Sci Nano* 2018;**5**:1902−10.
161. Baker NA, Klemm JD, Harper SL, Gaheen S, Heiskanen M, Rocca-Serra P, et al. Standardizing data. *Nat Nanotechnol* 2013;**8**:73−4.
162. Wang Y, He L, Wang M, Yuan J, Wu S, Li X, et al. The drug loading capacity prediction and cytotoxicity analysis of metal-organic frameworks using stacking algorithms of machine learning. *Int J Pharm* 2024;**656**:124128.
163. Kingston BR, Syed AM, Ngai J, Sindhwani S, Chan WCW. Assessing micrometastases as a target for nanoparticles using 3D microscopy and machine learning. *Proc Natl Acad Sci U S A* 2019;**116**:14937−46.
164. Hassanzadeh P, Atyabi F, Dinarvand R. The significance of artificial intelligence in drug delivery system design. *Adv Drug Deliv Rev* 2019;**151**−**152**:169−90.
165. Chew AK, Pedersen JA, Van Lehn RC. Predicting the physicochemical properties and biological activities of monolayer-protected gold nanoparticles using simulation-derived descriptors. *ACS Nano* 2022;**16**:6282−92.
166. Liu X, Wang Y, Yuan J, Li X, Wu S, Bao Y, et al. Prediction of the ibuprofen loading capacity of MOFs by machine learning. *Bioengineering-Basel* 2022;**9**:517.
167. Choi JS, Trinh TX, Yoon TH, Kim J, Byun HG. Quasi-QSAR for predicting the cell viability of human lung and skin cells exposed to different metal oxide nanomaterials. *Chemosphere* 2019;**217**:243−9.
168. Toropov AA, Toropova AP. Quasi-SMILES as a basis for the development of models for the toxicity of ZnO nanoparticles. *Sci Total Environ* 2021;**772**:145532.
169. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC international chemical identifier. *J Cheminf* 2015;**7**:23.
170. Lin T-S, Coley CW, Mochigase H, Beech HK, Wang W, Wang Z, et al. BigSMILES: a structurally-based line notation for describing macromolecules. *ACS Cent Sci* 2019;**5**:1523.
171. CurlySMILES AD. A chemical language to customize and annotate encodings of molecular and nanodevice structures. *J Cheminf* 2011;**3**:1.
172. Grethe G, Blanke G, Kraut H, Goodman JM. International chemical identifier for reactions (RInChI). *J Cheminf* 2018;**10**:22.
173. Lynch I, Afantitis A, Exner T, Himly M, Lobaskin V, Doganis P, et al. Can an InChI for nano address the need for a simplified representation of complex nanomaterials across experimental and nanoinformatics studies?. *Nanomaterials* 2020;**10**:2493.
174. Hassanpour S, Tomita N, DeLise T, Crosier B, Marsch LA. Identifying substance use risk based on deep neural networks and Instagram social media data. *Neuropsychopharmacology* 2019;**44**:487−94.
175. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *Med Image Comput Comput Assist Interv − MICCAI* 2015:234−41. Cham; 2015.
176. Mill L, Wolff D, Gerrits N, Philipp P, Kling L, Vollnhals F, et al. Synthetic image rendering solves annotation problem in deep learning nanoparticle segmentation. *Small Methods* 2021;**5**:2100223.

177. Wang W, Yan X, Zhao L, Russo DP, Wang S, Liu Y, et al. Universal nanohydrophobicity predictions using virtual nanoparticle library. *J Cheminf* 2019;**11**:6.
178. Wang T, Russo DP, Bitounis D, Demokritou P, Jia X, Huang H, et al. Integrating structure annotation and machine learning approaches to develop graphene toxicity models. *Carbon N Y* 2023;**204**:484−94.
179. Riniker S. Molecular dynamics fingerprints (MDFP): machine learning from MD data to predict free-energy differences. *J Chem Inf Model* 2017;**57**:726−41.
180. Fourches D, Ash J. 4D- quantitative structure−activity relationship modeling: making a comeback. *Expet Opin Drug Discov* 2019;**14**: 1227−35.
181. Gu S, Shen C, Yu J, Zhao H, Liu H, Liu L, et al. Can molecular dynamics simulations improve predictions of protein-ligand binding affinity with machine learning?. *Briefings Bioinf* 2023;**24**:bbad008.
182. Gabellini C, Şologan M, Pellizzoni E, Marson D, Daka M, Franchi P, et al. Spotting local environments in self-assembled monolayer-protected gold nanoparticles. *ACS Nano* 2022;**16**:20902.
183. Telari E, Tinti A, Settem M, Maragliano L, Ferrando R, Giacomello A. Charting nanocluster structures *via* convolutional neural networks. *ACS Nano* 2023;**17**:21287−96.
184. Tao H, Wu T, Aldeghi M, Wu TC, Aspuru Guzik A, Kumacheva E. Nanoparticle synthesis assisted by machine learning. *Nat Rev Mater* 2021;**6**:701−16.
185. Gong D, Ben Akiva E, Singh A, Yamagata H, Est Witte S, Shade JK, et al. Machine learning guided structure function predictions enable *in silico* nanoparticle screening for polymeric gene delivery. *Acta Biomater* 2022;**154**:349−58.
186. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**:493−500.
187. Thorp HH. ChatGPT is fun, but not an author. *Science* 2023;**379**:313.
188. O'Callaghan J. How OpenAI's text-to-video tool Sora could change science − and society. *Nature* 2024;**627**:475−6.
189. Bi K, Xie L, Zhang H, Chen X, Gu X, Tian Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* 2023;**619**:533−8.
190. Zhang W, Wang Q, Kong X, Xiong J, Ni S, Cao D, et al. Fine-tuning large language models for chemical text mining. *Chem Sci* 2024;**15**: 10600−11.
191. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**: 160018.
192. Liu W, Wang Z, Chen J, Tang W, Wang H. Machine learning model for screening thyroid stimulating hormone receptor agonists based on updated datasets and improved applicability domain metrics. *Chem Res Toxicol* 2023;**36**:947−58.