

Artificial Intelligence and Design of Experiments in the Development of Pharmaceutical Formulations: A Taxonomy of Approaches*

Michał HOFFMAN and Marcin MRUGALSKI¹

University of Zielona Góra, Zielona Góra, Poland

Correspondence should be addressed to: Michał HOFFMAN, 107711@stud.uz.zgora.pl

*Presented at the 47th IBIMA International Conference, 29-30 June 2026, Madrid, Spain

Abstract

Traditional Design of Experiments (DoE) methods in pharmaceutical formulation design, which rely mainly on static linear regression, have limitations when modelling multidimensional and non-linear biological systems. The growing complexity of modern overcoming these limitations using AI methods was highlighted. Furthermore, AI methods used in the design and delivery of pharmaceutical formulations methods necessitates the search for highly flexible alternatives for multi-criteria optimisation of substances. Despite the advanced capabilities of Artificial Intelligence (AI), its widespread adoption in research and development stages is still heavily inhibited by concerns regarding a lack of transparency in the pharmaceutical formulations design process and shortages in the availability of representative clinical data. Furthermore, a very clear gap can be observed in the current literature, manifested by the lack of a coherent, universal conceptual framework linking complex algorithms with conventional DoE schemes. To address this issue, the article introduces a new taxonomy categorising AI methodology in the field of DoE into five key groups: Predictive Machine Learning (PML), Deep Learning (DL), Adaptive DoE (ADoE), Global Optimisation (GO), and Generative and Support AI (G&S AI). Analysis of the applications presented clearly demonstrates that the fusion of adaptive DoE with DL or PML enables the precise development of pharmaceutical formulations.

Keywords: Artificial intelligence, designs of experiments, pharmaceutical formulations design, predictive machine learning, deep learning, adaptive DoE, global optimisation, and generative and support AI.

Introduction

AI is relatively often used in planning traditional processes or products, but sporadically in pharmaceutical production. The limitations on the use of AI methods stem mainly from the pharmaceutical industry's considerable scepticism and conservatism regarding the generation of pharmaceutical formulations using techniques that lack determinism, predictability and transparency. During the development of pharmaceutical formulations, classical DoE methods based on linear regression are often used. However, these methods are characterized by certain limitations, namely they can generate biased results when certain assumptions of the linear model are not met: linearity of the relationship between variables, randomness of the sample, independence and lack of collinearity of variables, as well as the assumption of constant variance of errors/noises and normality of the distribution of errors/noises. In practice, such limitations very often lead to the ineffectiveness of classical DoE methods based on linear regression. To solve this problem, one can try to develop DoE methods using methods belonging to the Bounded-Error Approach (BEA) class (Mrugalski, 2014). In these methods, it is necessary to assume only the knowledge of the limit values of errors/noises, but it is not necessary to know their distribution. Nevertheless, the area of application of the BEA method is also limited to linear models. To solve DoE problem for non-linear

¹ This research was funded in whole by the National Science Centre, Poland, grant number 2025/57/B/NZ7/04293

systems, it is possible to use methods belonging to the AI, among which we can mention Artificial Neural Networks (ANNs), Neuro-Fuzzy Networks (NF) and Takagi-Sugeno approach (T-S) (Mrugalski, 2014; Mrugalski, 2023). These methods are gaining popularity since their use is possible in the case of modelling nonlinear dependencies even in the absence of knowledge of the physical laws describing the functioning of the modelled system. In the case of AI-based methods, a model is created, e.g. in the form of an artificial neural network describing only the behaviour of the modelled system. To create the model, learning methods and sets of training data are used. These methods have proven to be effective in many areas of science and practical applications, such as control, technical diagnostics or manufacturing industry (Mrugalski, 2014; Mrugalski, 2023). However, it should be emphasized that there are very few research works devoted to the application of AI techniques in planning experiments in the field of pharmaceutical formulations development (Ferreira, 2025; Kleandrova et al., 2024; Lim et al., 2018). This is mainly due to the fact that the effectiveness of this type of method requires the availability of very large and representative data sets, which in practice are often unavailable in medicine and pharmacy, and the opinion that the results of this type of application can be difficult to analyse. Nevertheless, combining AI methods with the DoE seems to be a very promising research area due to the potential cognitive and financial benefits. The use of AI techniques in DoE can lead to obtaining optimal results for new classes of nonlinear problems for which traditional methods based on linear regression do not allow finding an optimal solution. The aim of this article is to provide an overview of AI methods which, in combination with DoE methods, may have potential applications in the development of new pharmaceutical formulations.

AI methods in pharmaceutical formulations design

The application of AI methods in pharmacy allows to move away from traditional analytical chemistry methodologies which rely on a laborious, empirical search for molecules in favour of automated, mathematical, multi-criteria experiments aimed at identifying new pharmaceutical formulations. These technologies centre on two main pillars of modern pharmacy: the discovery of active pharmaceutical formulations ingredients and the engineering of vectors for their transport within the human body. Among the key areas and AI methods enabling pharmaceutical formulations design, the approaches described below can be highlighted:

- Virtual screening and molecular docking performed using Global Optimisation (GO) (Rudrapal and Chetia, 2020). Traditional computer-aided docking methods often got stuck in local minima, leading to false assessments of pharmaceutical formulation-receptor bond stability. The use of GO algorithms, such as Particle Swarm Optimisation (PSO) and Evolutionary Algorithms (EA) are used in fragment-based pharmaceutical formulation design by efficiently searching through a large number of energy states to find the absolute, stable structural conformation within the target protein.
- Generative pharmaceutical formulations design based on the Deep Learning (DL) (Bian and Xie, 2021). Traditional searches through databases of millions of available molecules have confined scientists solely to compounds already known to chemistry. DL models break down this barrier by learning the rules of chemical topology from vast datasets. The system can generate, from scratch, an innovative structure with specified solubility and activity within a fully synthesizable chemical space, ensuring that the generated compound is technically feasible to produce in the laboratory.
- Protein engineering and rational peptide design using Predictive Machine Learning (PML) (Hunklinger and Ferruz, 2025). Large-molecule biological pharmaceutical formulations are frequently used in modern advanced oncology and immunology. The design of highly specific proteins is now achieved using predictive machine learning models. Large language model architectures adapted to read amino acid sequences can, without knowledge of the physical three-dimensional structure, predict with the highest accuracy how a planned sequence will affect the desired stability in the bloodstream, immunogenicity and the avoidance of side effects.
- Multidimensional modelling of toxicology and ADMET profiles using the Perturbation Theory Machine Learning (PTML) method (Kleandrova et al., 2024). Before any in vivo animal experiments are conducted, AI provides advanced safety filters. PTML-class systems and Quantitative Structure-Activity Relationship (QSAR) algorithms decode hidden properties based on simple topological parameters of the lead compound. Such algorithms enable critical conclusions to be drawn regarding which specific structures or molecular micro-fragments are responsible for the toxic profile allowing for their rapid correction or elimination from the experiment.
- Active learning in lead structure optimisation (Reker and Schneider, 2015). The method assumes that promising compounds from the early stages of pharmaceutical formulation discovery are refined using Bayesian Optimisation (BO) methods within the framework of algorithms based on adaptive DoE. This method overcomes the limitations of traditional approaches, which involve testing numerous variants using a trial-and-error approach. In the proposed method, the agent network evolves only that specific compound which, according to the generated error curve, will mathematically yield the most useful

knowledge and minimise pharmaceutical formulation design time saving clinical trial resources.

Another area of research into AI methods is their application in the development of new pharmaceutical formulation delivery methods to the body. In these methods, the pharmaceutical formulation substance must overcome physiological barriers to reach the affected cell safely. The use of nanotechnology supported by AI methods has led to the creation of new pharmaceutical formulation delivery systems. Designing smart nanocarriers that respond to stimuli using DL models (Gao et al., 2024) is one of such methods. Conventional carriers are often characterised by so-called burst release, which is toxic to healthy organs. The application of AI methods, including advanced DL, enables the engineering of nanocarriers that respond to very specific local environmental stimuli. An example is the computer simulation of polymeric nanoparticles, which, as a result of these calculations, are designed to undergo natural degradation and release the pharmaceutical formulation only upon detection of an acidified microenvironment typical of tumorous lesions. Predicting the physicochemical parameters of delivery systems using PML (Abdel-Rehim et al., 2025) is another example of the use of AI in pharmaceutical formulation delivery design. In this method the transition from creating random suspensions on a laboratory bench to rational design is achieved using predictive tools. These algorithms assimilate materials science, pharmacokinetic and clinical data to predict the safety profile, the optimal size of the liposomal nanoparticle, the circulation rate and its physical pharmaceutical formulation loading efficiency. Next method is called maximising the efficiency and quality of large-scale formulation using the Adaptive Design of Experiment method (ADE) (Otto, 2024). Complex pharmaceutical formulation delivery systems, such as modified carriers for gene therapy or engineered cells in Chimeric Antigen Receptor T-cell (CAR-T) therapies, require flawless optimisation of cultivation processes. To ensure an ideal distribution ratio or maximise fraction activity, machine learning incorporates adaptive design of experiment. Neural networks generate models which, during iterative changes to the culture medium, identify the best combinations of variables, minimising the wastage of biological raw material.

AI methods in DoE

Modern approaches to DoE increasingly utilise AI methods, which enable both the modelling of complex relationships and the effective exploration of the design space. These methods have been divided into five main groups: Predictive Machine Learning (PML), Deep Learning (DL), Adaptive DoE (ADoE), Global Optimisation (GO), and Generative and Support AI (G&S AI) (c.f. Fig. 1). Each group plays a specific role in experiments (c.f. Table 1). Understanding their properties makes it possible to select the most suitable one for the pharmaceutical formulations design process.

Table 1: Detailed characteristics of AI methods in DoE

Method / Algorithm	Group	Specific Application and Significance	Reference
Artificial Neural Networks (ANNs)	PML	An advanced surrogate model that maps complex, nonlinear relationships between input variables and critical product quality characteristics	(Joshi and Sheth, 2025; Shen et al., 2026)
Support Vector Machines (SVM)	PML	Classification of formulation stability; determining optimal boundaries between stable and unstable systems.	(Mehrpour et al., 2022; Ferreira and Carneiro, 2025; Joshi and Sheth, 2025; Shen et al., 2026)
Random Forest (RF)	PML	Variable significance analysis – determines which factors have the greatest impact on process performance.	(Mehrpour et al., 2022; Ferreira and Carneiro, 2025; Joshi and Sheth, 2025; Shen et al., 2026)
XGBoost	PML	A surrogate model for mapping complex experimental spaces. It narrows the scope of the study to the most influential variables.	(Shen et al., 2026)
Multi-Layer Perceptron (MLP)	DL	Used to model the correlation between multiple process variables and critical product quality characteristics	(Mehrpour et al., 2022; Ferreira and Carneiro, 2025; Shen et al., 2026)

Graph Neural Networks (GNN)	DL	A direct representation of molecular structures as graphs, where atoms serve as nodes and chemical bonds as edges.	(Ferreira and Carneiro, 2025; Joshi and Sheth, 2025)
Recurrent Neural Nets (RNN)	DL	Modeling sequential and time-series data, which is used in drug stability studies.	(Ferreira and Carneiro, 2025)
Transformers (e.g., Mol-BERT)	DL	Smart screening filters that identify the most promising molecules prior to the laboratory formulation optimization stage.	(Ferreira and Carneiro, 2025; Wang et al., 2025)
Bayesian Optimization (BO)	ADoE	Intelligent planning of experimental sequences to minimize the number of laboratory trials.	(Joshi and Sheth, 2025; Shen et al., 2026)
Gaussian Processes (GP)	ADoE	Estimating expected value and prediction uncertainty.	(Shen et al., 2026)
Thompson Sampling (TS)	ADoE	A method for selecting successive formulation variants or parameters based on sampling from the model's posterior distribution.	(Kandasamy et al., 2018)
BALD / BADGE / Coreset	ADoE	Strategies for selecting the most informative experimental points.	(Mehrpour et al., 2022)
Ant Colony Optimization (ACO)	GO	Solving chemical synthesis route optimization problems based on ant colony behavior.	(Li et al., 2023)
Simulated Annealing (SA)	GO	Searching for the global optimum in multidimensional surrogate models.	(Ghannadi et al., 2023; Kuo et al., 2022)
Generative Adversarial Nets (GAN)	G&S AI	Creation of new, synthetic chemical structures and images for training other models.	(Joshi and Sheth, 2025)
Variational Autoencoders (VAE)	G&S AI	Mapping molecular space onto vectors; generating molecules with desired properties.	(Lim et al., 2018)
Reinforcement Learning (RL)	G&S AI	Supporting the generation of de novo structures and optimizing multi-step technological processes.	(Mehrpour et al., 2022; Joshi and Sheth, 2025; Shen et al., 2026)
Fuzzy Logic (FL)	G&S AI	A method for transforming linguistic concepts into mathematical values using membership functions.	(Shen et al., 2026)

The PML comprises a set of artificial intelligence techniques that predict unknown future events based on the analysis of historical data. It is based on the creation of approximation models, the purpose of which is to replace computationally expensive and time-consuming simulations with a simplified mathematical approximation. They enable the creation of robust models with a limited number of trials. The DL is an extension of classical machine learning methods. It uses multi-layer neural networks (Mrugalski, 2014) to analyze complex, unstructured data. The models automatically detect patterns and features, enabling tasks such as image recognition and speech processing. The algorithms learn from training data, automatically discovering relationships without the need to manually define features. In DoE, it enables the modeling of complex dynamic processes, the identification of hidden patterns, the processing of molecular graph topologies, chemical sequences, and high-resolution microscopic images. It is useful for processing large datasets, such as large chemical libraries, or for monitoring complex production processes in real time. The ADoE involves sequentially selecting the most informative subsequent experimental points based on information from previous experiments, thereby maximizing the knowledge gained for the model. The key idea is to maximize the utility of subsequent experiments by considering model uncertainty and potential gain of information. This approach significantly reduces the number of experiments, which allows saving on costly components or accelerating the development of new products. The GO methods include heuristic and metaheuristic algorithms, which are used to search for the global optimum in complex, multimodal solution spaces. They can avoid local optima and have the advantage of being able to search for large decision spaces without requiring knowledge of the objective function's gradient. These methods allow us to both optimize the DoE and find optimal parameter values for the systems under study. The G&S AI methods

enable models to be fine-tuned and allow machines to operate in a more intelligent and human-understandable manner. Generative methods make it possible to create entirely new products that are not present in the training database. They also allow for the creation of synthetic data, which enables models to be fine-tuned when there is a shortage of real-world data. Support methods make work easier for both machines and humans; for example, they allow data to be fed into the model in a form that is more human-readable, rather than relying solely on numbers, or they help humans understand why the algorithm made a particular decision.

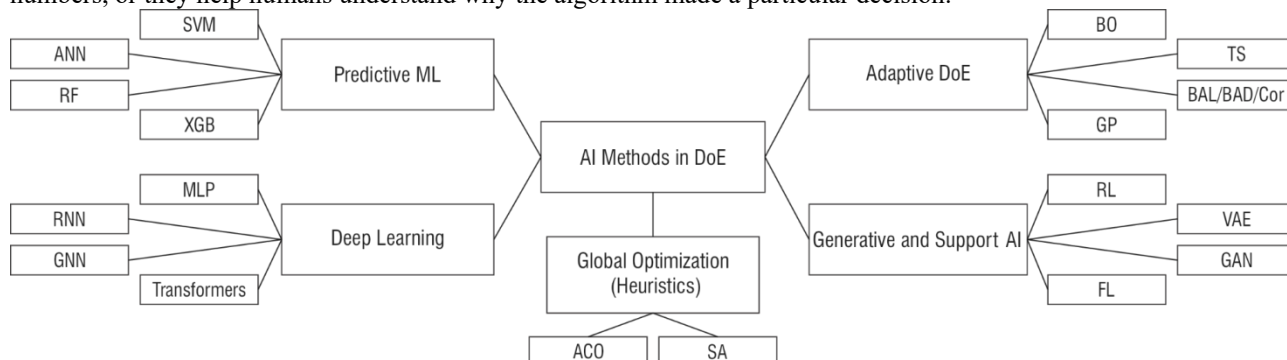


Fig 1. Taxonomy of AI Methods in DoE

Conclusions

This article provides an overview of the AI methods used in DoE. It also outlines the areas of application for AI in pharmaceutical formulation design and delivery to the human body. The combination of AI methods and DoE in pharmaceutical formulation design appears attractive because of the ability to perform simultaneous multi-criteria optimisation during their design. Goal-oriented optimisation AI-based tools can quickly balance a molecule's physicochemical parameters alongside its toxicological profile and synthesizability. Moreover, application of the AI and DoE in pharmaceutical formulation design allows for prediction of hidden non-linear interactions. Classical statistical systems fail when confronted with the multidimensionality of the immune system's biological networks. Neural networks can predict with a high degree of probability that a given protein-based pharmaceutical formulation will be rejected even before it is translated into the genetic code of the protein being tested. Thus, the application of AI and DoE in pharmaceutical formulations design represents an interesting direction for future research.

References

- Abdel-Rehim, A., Orhobor, O., Griffiths, G., Soldatova, L. and King, R.D. (2025), 'Establishing predictive machine learning models for drug responses in patient derived cell culture,' *NPJ precision oncology*, 9 (1), 180.
- Bian, Y. and Xie, X.Q. (2021), 'Generative chemistry: drug discovery with deep learning generative models,' *Journal of Molecular Modeling*, 27 (3), 71.
- Ferreira, F.J. and Carneiro, A.S. (2025), 'Ai-driven drug discovery: a comprehensive review,' *ACS omega*, 10 (23), 23,889–23,903.
- Gao, X.J., Ciura, K., Ma, Y., Mikołajczyk, A., Jagiello, K., Wan, Y., Gao, Y., Zheng, J., Zhong, S., Puzyn, T., et al. (2024), 'Toward the integration of machine learning and molecular modeling for designing drug delivery nanocarriers,' *Advanced Materials*, 36 (45), 2407,793.
- Ghannadi, P., Kourehli, S.S. and Mirjalili, S. (2023), 'A review of the application of the simulated annealing algorithm in structural health monitoring (1995-2021),' *Fracture and Structural Integrity*, 17 (64), 51–76.
- Hunklinger, A. and Ferruz, N. (2025), 'Toward the explainability of protein language models for sequence design,' *arXiv e-prints*, arXiv–2506.
- Mehrjou, A., Soleymani, A., Jesson, A., Notin, P., Gal, Y., Bauer, S., & Schwab, P. (n.d.). 'ICML2022 Workshop on Adaptive Experimental Design and Active Learning in the Real World GeneDisco: A Benchmark for Experimental Design in Drug Discovery.' (2022), Tech. rep.
- Joshi, S. and Sheth, S. (2025), 'Artificial intelligence (ai) in pharmaceutical formulation and dosage calculations,' *Pharmaceutics*, 17, 1440.
- Kandasamy, K., Krishnamurthy, A., Schneider, J. and Póczos, B. (2018), 'Parallelised bayesian optimisation via thompson sampling,' *International conference on artificial intelligence and statistics*, PMLR, 133–142.

- Kleandrova, V.V., Cordeiro, M.N.D. and Speck-Planche, A. (2024), 'Perturbation theory machine learning model for phenotypic early antineoplastic drug discovery: design of virtual anti-lung-cancer agents,' *Applied Sciences*, 14 (20), 9344.
- Kuo, C.L., Kuruoglu, E.E. and Chan, W.K.V. (2022), 'Neural network structure optimization by simulated annealing,' *Entropy*, 24 (3), 348.
- Li, G., Liu, C., Wu, L. and Xiao, W. (2023), 'A mixing algorithm of aco and abc for solving path planning of mobile robot,' *Applied Soft Computing*, 148, 110,868.
- Lim, J., Ryu, S., Kim, J.W. and Kim, W.Y. (2018), 'Molecular generative model based on conditional variational autoencoder for de novo molecular design,' *Journal of cheminformatics*, 10 (1), 31.
- Mrugalski, M. (2014), *Advanced neural network-based computational schemes for robust fault diagnosis*, vol. 510. Springer.
- Mrugalski, M. (2023), 'Identification and health-aware economic control of production systems: A fuzzy logic max plus algebraic approach,' *Eng. Appli. of Artificial Intelligence*, 120, 105,802.
- Otto, D.P. (2024), 'Experimental designs in pharmaceutical formulation development,' *Principles of Research Methodology and Ethics in Pharmaceutical Sciences*, CRC Press, 137–197.
- Reker, D. and Schneider, G. (2015), 'Active-learning strategies in computer-assisted drug discovery,' *Drug discovery today*, 20 (4), 458–465.
- Rudrapal, M. and Chetia, D. (2020), 'Virtual screening, molecular docking and qsar studies in drug discovery and development programme,' *Journal of drug delivery and therapeutics*, 10 (4), 225–233.
- Shen, C., Zhang, M., Lu, M., Chang, E., Gao, Z., Ban, W., Liu, Q., Zuo, Z. and Jiang, C. (2026), 'Machine learning empowered formulation design, optimization and characterization of nanoparticulate drug delivery systems: Current applications, challenges, and future perspectives,' *Acta Pharmaceutica Sinica B*, 16, 665–685.
- Wang, S., Allauzen, A., Nghe, P. and Opuu, V. (2025), 'A guide for active learning in synergistic drug discovery,' *Scientific Reports*, 15.