End-to-End AI-based Drug Formulation Optimization

AI 2025
SUMMER SCHOOL
ai.uni-jena.de

Saba Seifi

Friedrich Schiller University Jena

INTRODUCTION

AI technology and machine learning present a transformative opportunity in the drug discovery, formulation, and testing of pharmaceutical dosage forms. By utilizing AI algorithms that analyze extensive biological data, including genomics and proteomics, researchers can identify disease associated targets and predict their interactions with potential drug candidates. This enables a more efficient and targeted approach to drug discovery, thereby increasing the likelihood of successful drug.

The adoption of new technologies is helpful to address global healthcare challenges and medical emergencies, such as the recent COVID pandemic with minimal face-to-face contact and reducing the need for extensive and costly animal testing.

METHODOLOGY

1. DATA COLLECTION:

1.1. ENCODING

Drug molecule structures can be represented as sequences using various molecular descriptors. Simplified Molecular Input Line Entry System (SMILES) notation is the most widely used descriptor in AI-driven drug discovery.

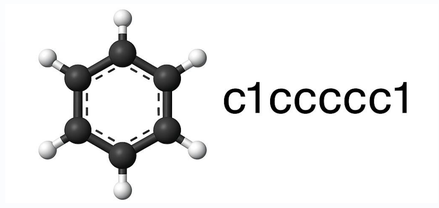


Figure 1- Benzene (C_6H_6) SMILES descriptor (generated by ChatGPT)

1.2. DECODING

The SMILES is passed through a cheminformatics toolkit (RDKit, Open Babel), which interprets the atom symbols, bond types, branches, and ring closures.

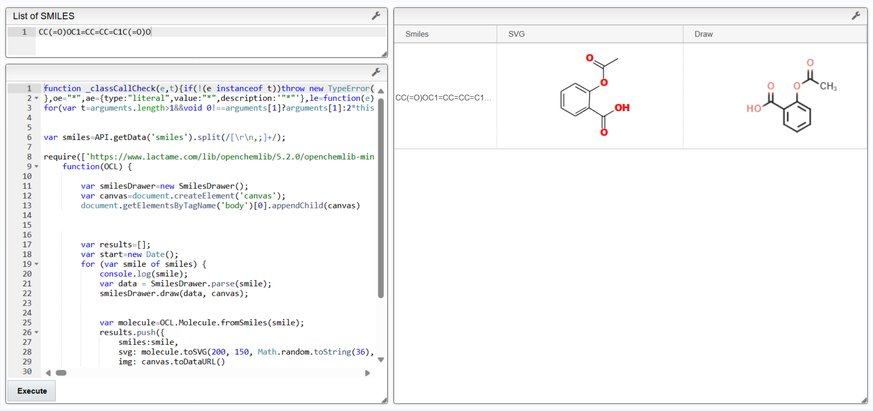


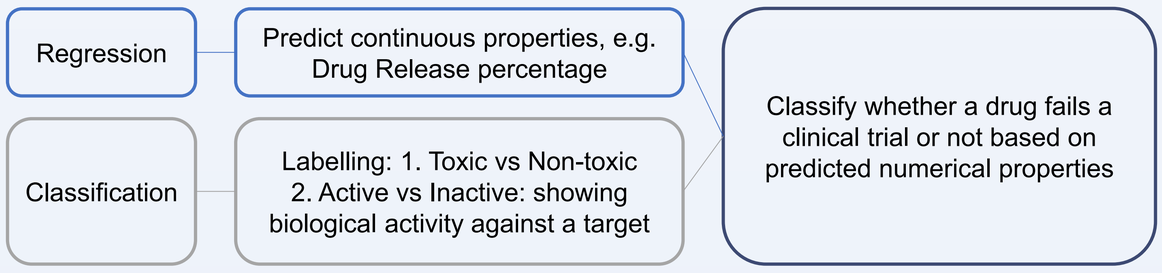
Figure 2- Aspirin decoded by Open Bebel Cheminformatics [6]

Alternatively, Natural Language Processing (NLP) techniques can treat SMILES strings as sentences and extract meaningful patterns through tokenization.

2. MODEL TRAINING:

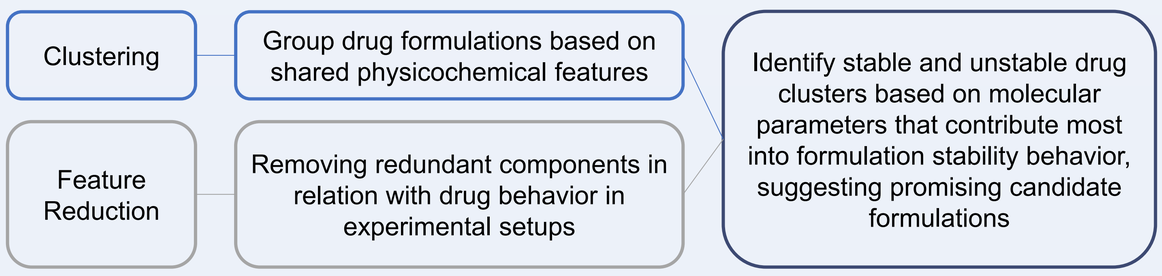
2.1. SUPERVISED LEARNING

It involves training models on labelled datasets—where each input (e.g., molecular descriptor, excipient ratio, pH level) is associated with a known output (e.g., solubility, release rate, toxicity level). These models learn to find meaningful linear (through Support Vector Machine (SVM) regressor and Decision Trees) or complex nonlinear (through Neural Networks (NN) and Random Forest (RF)) relationships between formulation components and therapeutic outcomes.



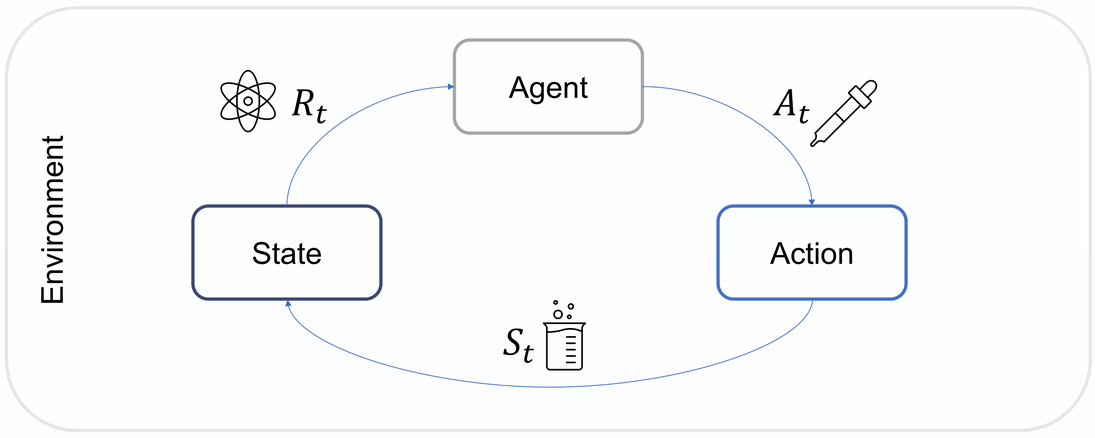
2.2. UNSUPERVISED LEARNING

Unsupervised learning is used when the data lacks predefined labels. Instead of predicting a specific outcome, these algorithms explore hidden structures or patterns within complex datasets. In drug formulation, unsupervised learning helps identify clusters of similar compounds and reduce dimensionality for drug interactions. Unlike supervised learning, its outputs are most actionable with interpretation.



2.3. (DEEP) REINFORCEMENT LEARNING

It is used for sequential decision-making tasks such as adaptive formulation design. An agent learns optimal actions (e.g., ingredient ratios) through trial and error, guided by reward signals (e.g., solubility improvement). By generating molecular structures step by step in a simulated chemical space, Drug-Target Binding Affinity (DTBA) can be optimized through interactions with biological targets.



2.4. Generative AI (GenAI)

Generative AI (GenAI) refers to AI systems that can create novel outputs – ranging from text and images to molecular designs – by learning patterns from large datasets.

3. NEW DRUG EVALUATION

Evaluation of newly formulated drugs involves a combination of physicochemical, microbiological, and stability assessments to ensure quality, efficacy, and safety. Some commonly analyzed parameters include:

- PH of the Solution
- Thermodynamics Stability
- Particle Size Distribution
- Antioxidant/Antimicrobial Preservative Content

Additional tests and acceptance criteria generally should be applied, depending on the type of drug, e.g. hard capsules, soft capsules, oral liquids.

CONCLUSION & FUTURE PROSPECTS

AI is revolutionizing drug formulation by enabling data-driven experimentation. In healthcare sector, leading biotech companies such as Sanofi, J&J, and Eli Lily are currently embedding AI into their R&D workflows. As an example, Lipid Nanoparticles (LNPs) underpinning mRNA COVID-19 vaccines (e.g. Pfizer-BioNTech) are outcomes of optimized AI-informed formulation strategies. In summary, AI accelerates drug development and enables personalization of pharmaceutical products—marking a new era of intelligent, efficient, and patient-centric formulation.

Figure 3- Getty Images for Pfizer/BioNTech

SCAN ME

REFERENCES

1. Vora, L. K., Gholap, A. D., Jetha, K., Thakur, R. R. S., Solanki, H. K., & Chavda, V. P. (2023). Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design. *Pharmaceutics*, *15*(7), 1916.
2. Hoseini, B., Jaafari, M. R., Golabpour, A., Rahmatinejad, Z., Karimi, M., & Eslami, S. (2024). Machine Learning-Driven Advancements in Liposomal Formulations for Targeted Drug Delivery: A Narrative Literature Review. *Current drug delivery*, 10.2174/0115672018302321240620072039. Advance online publication.
3. Ros, H., Abdalla, Y., Cook, M. T., & Shorthouse, D. (2025). Efficient discovery of new medicine formulations using a semi-self-driven robotic formulator. Digital Discovery.
4. Sharma, R., Saghapour, E., & Chen, J. Y. (2024). An NLP-based technique to extract meaningful features from drug SMILES. *Iscience*, *27*(3).
5. CPMP/ICH/367/96 (www.ema.europa.eu)
6. https://www.cheminfo.org/Chemistry/Cheminformatics/SMILES_to_svg/index.html
7. https://intuitionlabs.ai/articles/generative-ai-mrna-vaccine-covid19-case-study

ACKNOWLEDGEMENT & CONTACT

I sincerely express my gratitude to my mentor and all the organizers of AI Summer School 2025 for the opportunity to explore new AI-driven research streams. For future collaborations, reach me at: saba.seifi@uni-jena.de