Pipeline report
PK-sim ADME-Tox

Luna Pianesi

February 2025

Contents

Abstract

The countless applications of Artificial Intelligence to the biological domain are still underexplored. The hybrid field known as Computational Biology offers the possibility of applying a large variety of methods and perspectives from the Artificial Intelligence world to classically challenging biological problems. This project fits in the Drug Design field and aims at investigating the possible existence of novel drugs for certain families of proteins. Computational methods have great potential to develop new techniques and speed up the process of inventing new small-molecule drugs. Drug design already heavily relies on automation, but it could benefit even more when coupled to artificial intelligence methods. This project aims to investigate the possibility of completely automating a new drug design pipeline infused with artificial intelligence and biological domain expertise: it will address topics such as the generation process of new ligand small-molecule drugs, the prediction of activity of such molecules, the analysis of their docking, the prediction of their molecular properties, the explainability of their interaction with the receptors, and ultimately the analysis of the pathways and the prediction of the polypharmacy of such drugs. Artificial intelligence, deep learning, in particular, can bring great benefits to drug discovery and development thanks to its different approach to the matter. The opportunity of speeding up the process of drug design, and consequently that of development and approval of new drugs for diseases that classically plague our society is appealing and can be made real with collaborations between domain experts in computer science and biology. e propose an initial approach towards this direction that exploits Machine Learning, Graph Neural Networks, Diffusion Models, and Explainable Artificial Intelligence for producing novel small molecule inhibitors.

Keywords— drug discovery, deep learning, ADMET properties, cancer

1 Introduction

This project sits at the intersection between two research fields: Artificial Intelligence and Biology. The use of Artificial Intelligence (AI) in our society has now become pervasive and is bound to keep growing during the next few years. AI tools represent an incredibly valuable partner for humans in terms of decision-making tasks: whether it is to validate a medical diagnosis or to predict market trend under certain conditions, the ability of learning from data, hierarchically representing it through features and crafting decisions upon this built mechanism, is proving to be an increasingly essential need. On the other hand, biology has existed for much longer than artificial intelligence. The biological field is fraught with challenging problems still unsolved after decades of efforts: the functioning of the brain, therapies for cancer, or the role that genomic heterochromatin has in cells, among many others. Despite the amount of effort put into researching these topics and experimenting solutions for them, in most of the cases nobody was even remotely close to fully comprehending them. But as biology progressed towards the handling of large quantities of data—as the composition of the human genome can have—it became even more obvious that manually dealing with such quantities was hardly possible. One step at a time, computational tools entered the biological research, firstly as mere helper machines but lately almost as colleagues on par. Computational biology is the discipline that incorporates the challenges and domain knowledge of the biological field and the tools and strategies of the computer science field. It is a spectrum of topics and methods that is able to cover the intersection between those two apparently distant and unrelated domains. In recent years, computational biology achieved great breakthrough discoveries, some of which were from the medical field and provided new perspectives for patients affected by diseases that were thought to be untreatable. Achievements such as computationally predicting the three-dimensional structure of proteins, performing computationally-aided gene editing and designing and developing new drugs for treating diseases are results that represent only the beginning of what will certainly be one of the most fruitful fields hybridizations of the century. Top priority challenges are awaiting to be addressed; global burdens like climate change, novel pandemics and wealth redistribution have now the possibility of being tackled by approaches that benefit from the collection of efforts of many different domain experts, providing never-before-seen collaborative frameworks. This work is an attempt towards a contribution to computational biology. Computational biology is one of the fields that in the near future will benefit most from the introduction of artificial intelligence and deep learning in place of classical methods. Here, we try to demonstrate that a drug design pipeline can be fully automated by using a specific type of deep learning architecture, called the graph neural network. The chosen task is to find a novel inhibitor for the Aurora kinase protein, a protein that when dysregulated, has a major impact on the onset and development of several types of cancer. The problem of finding small-molecule inhibitors for the Aurora kinase goes back about thirty years, with both successes and failures. Just like any other biological problem, there is an extremely large number of variables to be taken into account; in this work, we try to pay attention to some of these variables by combining the domain knowledge and AI expertise at our disposal. We seek to demonstrate that a drug discovery pipeline for designing a novel small-molecule inhibitor for the Aurora kinase protein family can be completely automated via machine learning methods, particularly relying on the graph neural network architecture. A standard drug discovery pipeline could last decades from conceptualization to novel drug approval, but through deep learning, some steps of this long and excruciating road could benefit from a great speed-up, but not only: the ability of deep learning of learning from data can be exploited to find previously unseen perspectives about the same problem. Novel molecule generation and drug repositioning are instances of these perspectives.

1.1 Drug discovery

Drug discovery is a long, costly and multi-stage research process that is carried out in two major phases: preclinical and clinical studies. Each of these two phases is regulated by tight scientific and law standards, the latter ones being notoriously enforced in the European Union and the United States of America by respectively the European Medicines Agency (EMA)[1] and the Federal Drug Administration (FDA)[2]. The scientific method followed in the discovery process is finely paced and only in exceptional cases it gets disrupted: during the SARS-CoV-2 pandemic outburst in 2020, all pharmaceutical efforts were directed towards the development of a new vaccine, implying non-compliance with default process checkpoints [2]. Under normal conditions, the discovery process might take decades—as well as billions of dollars of investments—to be completed, since it must be complemented with development and subsequent approval by EMA or FDA of the new candidate drug. But the road to approval is already troubled per se: preclinical studies are preceded by preliminary steps that include selection and validation of a target, identification and optimization of a lead, and proposal of a new candidate drug, which is only then tested for safety and efficacy in preclinical development stages. A target is any system that can potentially be modulated by a molecule to produce a beneficial effect for a specific disease; generally, it is a protein, but it could also be a nucleic acid, a carbohydrate, a lipid, and so on[3]. Identifying the "perfect" target is a long research process that requires domain knowledge, druggable potential and market interest. When a target gets identified, screening or design studies are enacted in order to find a hit, a small molecule or biologics capable of acting as drug on that target and displaying promising biocompatibility characteristics (medium-to-high specificity, zero-to-low toxicity). A lead is a prototype compound with desired and improved biological/pharmacological activity and selectivity with respect to the hit, activity in in vitro assays and a favourable predicted ADME (adsorption, distribution, metabolism, excretion) profile5. The lead is then optimized to obtain a candidate drug, having actual desired activity, safety and large-scale developability. Only at this point preclinical studies can begin: the candidate drug is synthesized in a laboratory; it is tested for efficacy and safety in test tubes and animals; assessment of formulation, stability, scale-up synthesis and chronic safety in animals is carried out; eventually, the pharmaceutical company files an Investigational New Drug (IND) application with the drug administration organ of the reference country and the proposed novel drug passes on to clinical studies. Clinical studies are further divided in phases: phase I studies are used to determine the toleration of the drug in healthy humans; phase II studies assess the efficacy of the drug in non-healthy humans (patients); phase III studies consist in large clinical trials in many patients; if the drug successfully overcomes all phases, the pharmaceutical company files for a request of drug application that is then potentially approved by the drug administration organ of the country and lately approved for marketing and commercialization. An alternative to this strategy is represented by drug repurposing: it is an approach for investigating new applications for licensed or experimental drugs that go beyond their original medical purpose6. In any case, the process through which new drugs are discovered is excruciating, disseminated with failure cases and spans decades before a new molecule can be considered safe and effective for the treatment of a disease. There is an urge to speed up this lengthy process somehow; computational tools already provide an appealing alternative to human-effort-intensive tasks and Artificial Intelligence (AI) in particular can be a valuable partner in the discovery operation. The biological domain knowledge can be exploited to identify a druggable target, while the AI expertise can rescue the process to perform high-throughput screening of databases or de novo design of potential drugs. This work is an attempt at investigating the benefits that these two components can bring when used together. We defined and tested all that we could automate of a drug design pipeline from the hybrid perspective of AI and biology, starting from a chosen target, the Aurora kinase protein. Proteins have a high druggability: they are biological targets that can be modulated by a drug altering their function with a beneficial effect for the patient. The concept of druggability is most often restricted to protein-small molecule interactions, but it can easily be extended to include biological medical products such as therapeutic monoclonal antibodies. Post-translational modifications (PTMs) of proteins are often involved in disease development. They are numerous, diverse and occur at specific protein sites. PTMs are able to change the charge, conformation or size of the protein molecule, therefore they are a fundamental factor determining the stability of the protein, its biochemical activity, its localization and signalling. Technologies of differential expression proteomics are able to detect protein changes associated with a disease or drug treatment in a specific cell type or tissue; the description of such techniques is beyond the scope of this research, but their use is complementary to in silico methods for assessing validation of proposed novel targets. Following this section there will be a detailed dissertation about the importance of the Aurora kinase protein and its involvement in cancer, as well as an essay from the methodological perspective of the computational techniques used. What is fundamental to introduce before the treatise of the matter is that, when practically implementing

[1] https://www.ema.europa.eu/en

[2] https://www.fda.gov/

[3] Dr. Stephen Carney, Managing Editor, Drug Discovery Today

a computational drug design pipeline, there are three possible scenarios: the ground truth 3D coordinates of both the receptor's and ligand's binding atoms might be low-resolution, might be completely unknown or might not be of interest7. The first scenario is the most frequent: the availability of 3-dimensional structural data suffers from scarcity and sometimes low quality[4]. The second scenario refers to the novel drug discovery process, in which we aim at generating or repurposing molecules that hopefully treat some disease; it is clear that in this case we completely lack the knowledge of the atoms' locations involved in binding. The third scenario refers to the case in which it is worth exploring the "binding possibilities" of a ligand-receptor complex that can already be known; this investigation is generally directed towards the exploration of allosteric binding sites rather than orthosteric binding locations[5].

1.1.1 Biological target

1.2 Predicting ADME-Tox properties

1.2.1 Metabolism

1.2.2 Excretion

1.3 Methods

1.3.1 Machine Learning

1.3.2 Graph Neural Networks

1.3.3 Diffusion models

2 Task: Evaluating ADME-Tox properties of compounds

Computationally predicting all ADME-Tox properties of compounds is becoming an increasingly more concrete reality. A large number of methods with vast and comprehensive capabilities are now available as open-source tools and can easily be accessed by everyone; most of these tools make use of Artificial Intelligence concepts, specifically building on top of Deep Learning and/or Machine Learning methods [4]. In particular, predicting the "M" ("Metabolism") and "E" ("Excretion") in ADME-Tox, can be broken down in several steps, each of them easily approachable in a fully computational way, without the need of human or animal experimentation.

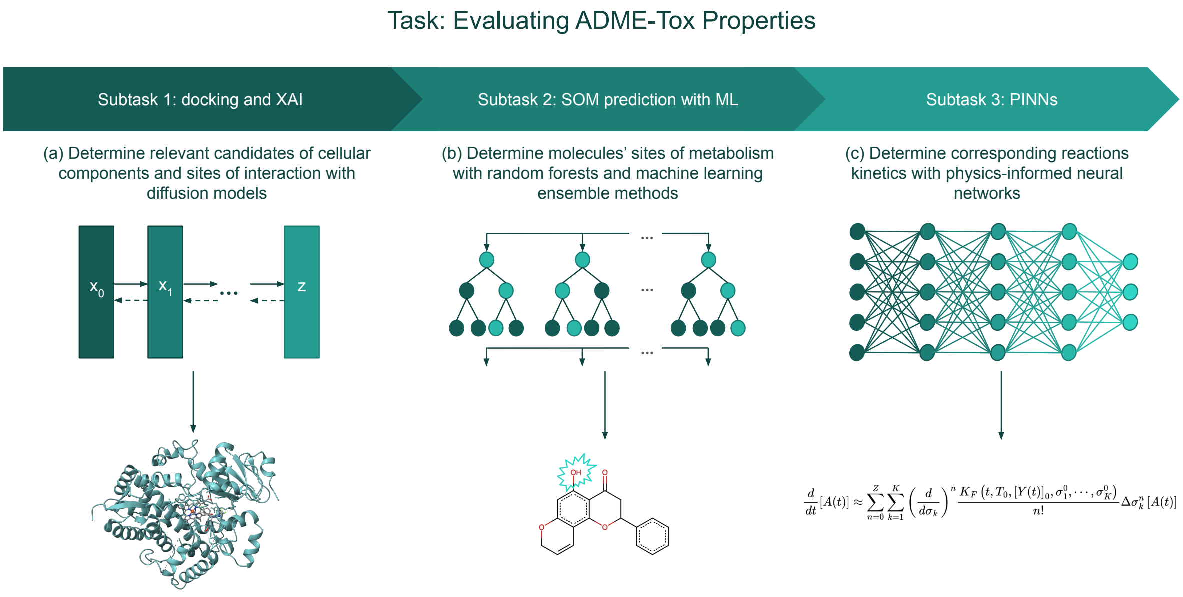
Task: Evaluating ADME-Tox Properties
Subtask 1: docking and XAI
Subtask 2: SOM prediction with ML
Subtask 3: PINNs
(a) Determine relevant candidates of cellular components and sites of interaction with diffusion models
(b) Determine molecules' sites of metabolism with random forests and machine learning ensemble methods
(c) Determine corresponding reactions kinetics with physics-informed neural networks
x_0
x_1
...
z
...
OH
O
O
O
$\frac{d}{dt}[A(t)] \approx \sum_{n=0}^{Z}\sum_{k=1}^{K}\left(\frac{d}{d\sigma_k}\right)^n \frac{K_F\left(t, T_0, [Y(t)]_0, \sigma_1^0, \cdots, \sigma_K^0\right)}{n!}\Delta\sigma_k^n\,[A(t)]$

Figure 1: graphical representation of subtasks

[4] Only around 19 thousand experimental complexes are publicly available in the PDBBind database.

[5] An orthosteric ligand binds its target receptor at the active site, competing with the natural substrate or ligand of that receptor; an allosteric ligand instead binds elsewhere on the protein surface with respect to the active site, and allosterically changes the conformation of the protein binding site10(where allosteric regulation refers to the regulation enacted by a ligand on a receptor to which it bound on a site different than the active site).

3 Subtask 1: Determining relevant candidates of cellular components and sites of interaction

After having obtained a set of promising molecules from the previous steps,we proceed by checking which are the cellular components (metabolizing enzymes, transporters and protein binding partners) that are participating in the drug interaction. We focus on considering cellular components that the literature has already documented as relevant for the metabolic process (the Cytochrome P450 (CYP) system and the UDP-glucuronosyltransferases (UGTs) system), and we use a structure-based virtual screening approach: protein-ligand docking. Docking methods based on equivariant Graph Neural Networks have been surpassed by diffusion models, which are the new state-of-the-art. We thus build on in-house preliminary work and use DiffDock [3], and Boltz-1, an open-source alternative to AlphaFold3 [1], for predicting the joint-structure of the interaction between the drug of interest and the relative biologic [6]. We complete the subtask by verifying the biological plausibility of the complex using AutoDock Vina for classical molecular dynamics simulations [5] and borrowing from concepts of Explainable Artificial Intelligence (XAI). Figure 1a reports a standard visualization of diffusion models and an example of the subtask's relative output (PDB ID: 4NYA, crystal structure of CYP3A4 in complex with an inhibitor).

4 Subtask 2: Determining molecules' sites of metabolism

Sites of metabolism (SOMs) are the atomic hotspots of interest for the lead optimization process in drug design. Predicting SOMs is of high importance, because it is the first step towards the prediction of potential drug-derived metabolites. The determination of SOMs and metabolite structures will be computationally carried out by assuming the mediation of CYP enzymes; indeed, the processes mediated by this group account for approximately 75% of the overall drug metabolism in humans (Tran et al., 2023). We thus proceed to use CypReact, an open-source online tool for the reaction prediction of ligands with nine critical CYP isozymes, based on machine learning methods (random forest, support vector machines, logistic regression, and decision trees) (Tian et al., 2018). Figure 1b reports a visualization of the random forest machine learning method along with an example of the subtask's relative output (cytochrome P450 1A2's (CYP1A2) inhibitor 5H78PF).

5 Subtask 3: Determining corresponding reaction kinetics

PBPK (physiologically based pharmacokinetic modelling) is a mathematical modelling technique for predicting ADME properties of synthetic or natural molecules in humans and other species. Frequently, PBPK modelling is based on QSAR (quantitative structure-activity relationship) models, whose parameters can in turn be estimated by artificial neural networks. Therefore, computationally determining reaction kinetics for compound-protein pairs amounts to estimating parameters of a differential equation that describes the system's kinetics. As a final subtask, after determination of promising hits and relative interaction biologics, we use physics-informed neural networks (PINNs) as a data-driven approach for studying complex kinetic problems. Building on Adebar et al., 2024, we adapt a modified form of PINNs for the study and prediction of concentration profiles under the influence of reaction parameter variations in our context. Figure 1c reports a visualization of a standard deep neural network together with an example of the subtask's relative output (change of reaction species concentration as approximated by a truncated Taylor's series expansion).

References

[1] Josh Abramson et al. “Accurate structure prediction of biomolecular interactions with AlphaFold 3”. In: *Nature* 630.8016 (2024), pp. 493–500. DOI: 10.1038/s41586-024-07487-w.

[2] Philip Ball. “What the lightning-fast quest for Covid vaccines means for other diseases”. In: *Nature* 589 (Jan. 2021), pp. 16–18.

[3] Gabriele Corso et al. *DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking*. 2023. arXiv: 2210.01776 [q-bio.BM]. URL: https://arxiv.org/abs/2210.01776.

[4] Thi Tuyet Van Tran, Hilal Tayara, and Kil To Chong. “Artificial Intelligence in Drug Metabolism and Excretion Prediction: Recent Advances, Challenges, and Future Perspectives”. In: *Pharmaceutics* 15.4 (2023). ISSN: 1999-4923. DOI: 10.3390/pharmaceutics15041260. URL: https://www.mdpi.com/1999-4923/15/4/1260.

[5] Oleg Trott and Arthur J. Olson. “AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading”. In: *Journal of Computational Chemistry* 31.2 (2010), pp. 455–461. DOI: https://doi.org/10.1002/jcc.21334. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21334. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21334.

[6] Jeremy Wohlwend et al. “Boltz-1 Democratizing Biomolecular Interaction Modeling”. In: *bioRxiv* (2024). DOI: 10.1101/2024.11.19.624167. eprint: https://www.biorxiv.org/content/early/2024/11/20/2024.11.19.624167.full.pdf. URL: https://www.biorxiv.org/content/early/2024/11/20/2024.11.19.624167.

A Deep learning concepts

A.1 Graph generative models