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 [3]. 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 Protein kinases

Protein kinases act as molecular switches. They regulate a large variety of biological processes and represent amenable targets for drug discovery and development.

Cells receive signals from outside via receptors. When a ligand binds to its receptor, a signal is transmitted as part of the signalling cascade, which may alter the function of the cell. A central element of these signalling cascades are protein kinases. Protein kinases are regulatory enzymes that covalently transfer a γ-phosphate group from ATP onto a residue of a target protein. Biochemically, protein kinases catalyse the following reaction

$$\mathrm{MgATP^{1-}} + \text{protein–O:H} \longrightarrow \text{protein–O:}\mathrm{PO_3^{2-}} + \mathrm{MgADP} + \mathrm{H^+},$$

commonly known as *phosphorylation*. Protein phosphatases are enzymes that catalyze the opposite reaction to phosphorylation, called *dephosphorylation*. Phosphorylation is thus a reversible protein modification that is regulated by a balance of kinase and phosphatase activity. The addition of a phosphate group is a powerful mechanism that mediates several cellular pathways through functional modification of the target protein — in the form of activation, deactivation, cellular localization and association with other proteins — , especially signalling cascades.
So far, more than 500 protein kinases have been identified. They are all assigned into a complex system of groups, families and subfamilies, following a classification method that takes into account both structure and function. Protein kinases can thus be broadly classified according to the target protein's amino acid residue they phosphorylate, which is the free hydroxyl group of one among tyrosine (Tyr, Y), serine (Ser, S) and threonine (Thr, T). However, kinases that are able to phosphorylate a serine residue are also able to do the same for a threonine residue, thus are usually grouped together as serine/threonine kinases. Kinases that phosphorylate tyrosine form another large group identified as threonine kinases. Serine/threonine kinases and tyrosine kinases form the two largest groups, but other minor groups consisting of kinases acting on different amino acid residues exist (e.g. histidine kinases).
The deregulation of protein kinase function plays an important role in cancer as well as immunological, inflammatory, neurodegenerative, metabolic, cardiovascular and infectious diseases. For these reasons, protein kinases have become one of the top priority clinical targets and, as a consequence, the development of kinase inhibitors is invested on the most in the pharmaceutical industry[6]. Pharmacological strategies targeting protein kinases with small molecule drugs most frequently propose ATP-competitive inhibitors. Indeed, many protein kinases have been demonstrated to have a crucial role in carcinogenesis and metastases of various types of cancers, their activity often promoting cell proliferation, survival and migration. When constitutively overexpressed or active, protein kinases can be associated with oncogenesis. Several efforts have been directed towards the study of kinases as drug targets, with the intention of unravelling more personalised and less cytotoxic strategies for treating cancer. One of the most undertaken paths in this direction is the discovery of small molecule kinase inhibitors; however, despite being theoretically promising, they prove to be disappointing when it comes to clinical trials, revealing several unpredicted toxicities and side effects. In addition to that, the major obstacle in the way of a successful treatment of cancers with small molecule kinase inhibitors is acquired pharmacological resistance. Kinase inhibition appears to trigger a strong discerning pressure for cells to acquire resistance to chemotherapy through kinase mutations. The majority of the kinase resistance cases fall into the category of acquired resistance[7], which refers to the progression of a tumour that initially responds to treatment and subsequently becomes resistant to it, without variation in the administration of the inhibitor. Interestingly, a single point mutation in a kinase gatekeeper residue is sufficient to trigger drug resistance, because it prevents the drug inhibitor to form hydrogen bonds with deeper hydrophobic regions of the kinase binding pocket which would only be accessible

[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).
[6] Aurora kinase, [2]
[7] As opposed to *de novo* resistance

via "permission" of these residues. Overcoming gatekeeper-mutation-induced drug resistance requires meticulous structural fine-tuning of the drug candidate. Several follow-up strategies have been proposed to work around the problem, notably: i) designing inhibitors that can tolerate a number of different amino acid residues at the kinase gatekeeper position; ii) exploit alternative binding sites of the kinase for inhibitor binding; iii) focus on targeting other pathways that may be required for kinase transformation [4].

1.2.1 Aurora Kinases

This section contains a description of the chosen biological entity related to the previously described problem. The Aurora kinase protein will be presented in terms of its genetic, structural, functional and interaction characteristics.

The Aurora kinases are a subfamily of protein kinases the first specimen of which was discovered in 1995 by David Glover's laboratory during studies of mutant alleles associated with abnormal spindle pole formation in *Drosophila melanogaster* [7]. Since their discovery, they increasingly received attention due to their crucial role in cancer. Aurora kinases (AKs) are a family of serine/threonine kinases primarily active during mitosis and their dysregulation is responsible of several issues during the cell cycle [9].

Sequence The genes encoding the three human paralogue Aurora kinases map to regions of the DNA that are frequently affected by chromosomal abnormalities in different cancer types. A validation of this statement is provided by the recorded overexpression of each of the three human Aurora genes in tumour cell lines [5]. The chromosomal loci of the three Homo Sapiens genes are A:20q13, B:17p13 and C:19q13 [12], with genomic sequences available on UCSC Genome Browser at loci chr20:56369390-56392215, chr17:8204734-8210575, chr19:57231009-57235417, respectively, according to human genome assembly GRCh38/hg38 of 2013[8].

Structure Eukaryotic protein kinases are structurally related via sharing of a conserved catalytic core called the *kinase domain*, responsible for the catalytic activity of this group of proteins. The first crystal structure of a protein kinase, PKA, dates back to the 1990s [8]. It revealed the structural features of protein kinases as we know them today: a conserved core consisting of an N-lobe and a C-lobe region connected by a hinge and creating a cleft that gently envelopes an ATP substrate molecule. The N-lobe has a 5-stranded β-sheet (β1-β5) and at least one α-helix; the C-lobe is mostly α-helical but has a small yet important β-sheet (β6-β7) [2]. Generally speaking, all protein kinase structures have three distinguishing and conserved groups: residues that bind ATP, residues that bind the substrate and residues that carry out phosphate transfer. Figures **??**, **??**, **??** and **??** highlight on two different Aurora kinase crystal structures the conserved residues of any protein of the kinase family:

- The glycine (G)-rich loop is colored in purple; it extends over the top of ATP;
- The conserved lysine (K) residue is colored in green; it binds the α- and β-phosphates of ATP, keeping it in place. By mutating these amino acids, the kinase will not be able to bind ATP anymore and thus looses its activity;
- The conserved glutamate (E) is colored in blue; it forms a salt bridge with the conserved lysine and it is important for keeping the shape of the kinase stable;
- The DFG (aspartate, phenylalanine, glycine) motif is colored in orange; it binds metal ions that are needed to transfer the phosphate from ATP and is also the start of the large activation loop;
- The activation loop is colored in yellow; its amino acid sequence determine whether the kinase recognized tyrosines or serines and threonines;
- The HRD/YRD (histidine/tyrosine, arginine, aspartate) motif is colored in pink; this motif contains the catalytic aspartate residue that actually transfers the γ phosphate from ATP to the tyrosine or serine/threonine substrate.

Some studies [11] have claimed and demonstrated that the protein kinase conserved structure is organized in dynamic communities, identified as structurally contiguous regions of the protein that exhibit correlated motions. Each of these communities has a structurally conserved scaffold and large groups of residues that stick together in all four possible conformations of the protein kinase enzyme: closed (bound to ATP and 2 Mg^2+ ions), closed (bound to ATP and 1 Mg^2+ ion), open (bound to ATP and 1 Mg^2+ ion) and apo (not bound to ATP nor Mg^2+ ions). The found concept of communities helps to better understand the long-distance allosteric communication of residues on different sites of the kinase.
By viewing the Aurora kinase from a one more level higher perspective, several conserved structural blocks have been identified [5]: Aurora A has a silent carboxy-terminal destruction box (D-box), which is also present in Aurora B, but that is only functional in the presence of an amino-terminal A-box (also called the D-box-activating-domain (DAD)); the A-box/DAD is absent from Aurora B and C. Phosphorylation of the A-box seems to make Aurora A resistant to anaphase-promoting-complex/cyclosome (APC/C)-mediated degradation,

[8] https://genome-euro.ucsc.edu/index.html

suggesting possible novel mechanisms for targeting the protein with small molecule drugs.
When it comes to analyzing structural data, crystal structures for virtually any protein known to date can be found on the Protein Data Bank (PDB) website[9]. The PDB website is centrally maintained and curated by the Research Collaboratory for Structural Bioinformatics (RSCB) group. Every entry in PDB is designated with a four-character alphanumeric code called the PDB identifier or PDB ID. In this work we focus on the Aurora kinase A structure corresponding to PDB ID 4ZTQ[10] and on the Aurora kinase B structure corresponding to PDB ID 4AF3[11]. 4ZTQ represents the crystal structure of the catalytic domain of Aurora kinase A bound to inhibitor FK932 at 2.80Åresolution, while structure 4AF3 identifies a human Aurora B kinase in complex with INCENP and VX-680 at resolution 2.75Å. Another extremely useful resource is the KLIFS database[12].

Subcellular localization Surprisingly, despite the higher level of structural similarity of the three mammalian Aurora paralogues, they have very distinct subcellular localizations and functions [5].

Aurora A, *the polar Aurora*: it is normally associated with the centrosome from the time of centrosome duplication through the mitotic exit, and it is also associated with regions of microtubules that are proximal to centrosomes in mitosis. The defining characteristic of the Aurora A subfamily has been its association with centrosomes and regions of microtubules that are proximal to the centrosome itself, and it associates with the centrosomes that are separating during late S/early G2 phase. This localization is dynamic and the protein exchanges continuously with the cytoplasmic pool. The association with the centrosome is directed independently by both the amino-terminal region and the carboxy-terminal catalytic domain, but it does not require kinase activity. The protein TPX2 (targeting protein for XKLP2), which has been implicated in Aurora A activation, is required for the localization of the kinase to spindle microtubules, but not to spindle poles.

Aurora B, *the equatorial Aurora*: it forms a complex with two other proteins, called inner centromere protein (INCENP) and survivin, and behaves as a chromosomal passenger protein. Passenger proteins normally associate with centromeric heterochromatin early in mitosis, transfer to the central spindle in anaphase and are amongst the first proteins to localize at the cell cortex where the contractile ring subsequently forms. Chromosomal passenger proteins remain associated with the midbody during cytokinesis. The dramatic movements of passenger proteins during mitosis led to the proposal that they might have a role in the coordination of chromosomal and cytoskeletal events during the cell cycle. Human Aurora B was first identified in a polymerase chain reaction screen for kinases that were overexpressed in tumours. Aurora B expression and activity in proliferating tissues are cell-cycle regulated: expression peaks at the G2-M transition, and kinase activity is maximal during mitosis. More recent studies showed that the association of the kinase with centromeres during metaphase is dynamic: the protein exchanges continuously with the surrounding cytoplasmic pool just like Aurora A. Once the kinase associates with central spindle microtubules during anaphase (which requires kinase activity), its mobility is highly reduced. A subpopulation of Aurora B also seems to be transported by astral microtubules to the equatorial cell cortex.

Aurora C: much less is known about Aurora kinase C. They are specifically expressed at high levels in the testis and show centrosomal localization from anaphase to telophase. Challenge for the future: dissect out the features that define the substrate specificity, regulation and localization of each of the Aurora kinases. The expression levels of human auroras and some of their associated polypeptides are elevated in certain types of cancer, and overexpression of Aurora A can induce transformation.

Function Aurora A and B share significant sequence similarity, particularly within their kinase domains; however, each kinase exhibits unique precise temporal and spatial control by dynamic association with accessory proteins. Both Aurora A and B are mitotic regulators that are often found to be aberrantly expressed in tumour cells. Their activity is dependent on a number of cofactors; both demonstrate increase in kinase activity during mitosis as a result of association with accessory proteins that generate fully competent kinase complexes and by autophosphorylation of critical residues (T288 AurA and T232 AurB) in the kinase activation (T) loop [10].
The gene *AURKA* is ubiquitously expressed and regulates cell cycle events occurring from late S phase though the M-phase. Both activity and protein levels of Aurora A increase from late G2 phase through the M phase, with peak activity in prometaphase. The kinase activity of Aurora A is tightly regulated throughout the cell cycle. It is activated through the phosphorylation of T288 (human sequence) on its activation loop. It can be inactivated through dephosphorylation of T288 by protein phosphatase 1 (PP1). Beyond phosphorylation and dephosphorylation, its activity is also regulated by its expression and degradation.
Each Aurora kinase has specific functions during the cell cycle that are concordant with its subcellular localization [5].

Aurora A has a role in centrosome maturation and separation: the high frequency of monopolar mitotic figures in certain *Drosophila aurora* mutants indicate a potential role for the kinase in centrosome separation. In the absence of Aurora A, recruitment of several components of the pericentriolar matrix to the centrosome is deficient, and the microtubule mass of spindles is decreased by about 60%. Furthermore, the morphology of the astral microtubule array is also aberrant. This might partly reflect an impaired function of factors that

[9] https://www.rcsb.org/
[10] https://www.rcsb.org/structure/4ZTQ
[11] https://www.rcsb.org/structure/4AF3
[12] https://klifs.net/index.php

regulate microtubule dynamics, such as *Drosophila* transforming-acidic-coiled-coil protein (D-TACC), and/or Eg5, a kinesin-like protein that is involved in spindle assembly. Both of these proteins are substrates of Aurora A *in vitro*. TPX2 binds Aurora A at the centrosome and targets it to the microtubules proximal to the pole. TPX2 also regulates the kinase activity of Aurora A, both by counteracting the activity of protein phosphatase PP1 and stimulating Aurora A autophosphorylation at Thr295—the residue in the activation loop of Aurora A that is essential for kinase activity.

Aurora B has a role in chromosome biorientation: after nuclear envelope breakdown, prometaphase chromosomes rapidly establish attachments to a nearby spindle pole, destabilising syntelic attachments of sister chromatids. This might be especially important in *S. cerevisiae*, in which chromosomes are attached to nuclear microtubules for most of the cell cycle and replicated sister kinetochores enter mitosis attached to the same spindle-pole body. How the kinase recognizes syntelic attachments is not clear, but it has been proposed that tension between amphitelically oriented sister kinetochores stretches them apart enough to separate microtubule-binding sites from Aurora B that is sequestered in the inner centromere, which thereby limits the accessibility of the kinase to its substrate. Aurora B also seems to have an important role in regulating kinetochore-microtubule interactions in higher eukaryotes. Interference with its function by RNAi, microinjection of function-blocking antibodies or treatment with small-molecule inhibitors all cause defects in chromosome congression. The mammalian kinetochore-specific histone-H3 variant CENP-A is a substrate of Aurora B in mammalian cells. Phosphorylation of CENP-A by Aurora B peaks in prometaphase. Surprisingly, phosphorylation-site mutants show a delay in the late stages of cytokinesis. Why a kinetochore protein should show defects in completion of cytokinesis, a cytoskeleton/membrane event, is unclear.

Aurora C is still an obscure topic: very little is known about this Aurora kinase, except that it seems possible that it might act similarly to Aurora B.

Inhibitors An important step in this computational drug design pipeline is to try to describe the common structural features that underpin protein kinase regulatory mechanisms and how they relate to kinase inhibitor discovery [2].

In the hunt for a novel class of anticancer treatments, substantial research is being done on the inhibition of crucial regulatory mitotic kinases by using ATP-competitive small-molecules [10]. The inhibition of critical regulatory mitotic kinases using ATP-competitive small molecules is an active area of research in the quest for a new class of anticancer therapeutics. Numerous compounds targeting key cell cycle kinases including Cyclin-dependent kinases (Cdk), Aurora (Aur), Polo-like kinases (Plk) and the kinesin-5 molecular motor have been advanced into the clinical testing. The clinical rationale for targeting mitosis to treat cancer is provided by Taxol, a highly successful anticancer agent that arrests cell division by stabilizing microtubule polymers and thereby disrupting the cellular machinery required for mitotic spindle assembly. Unfortunately, to date most of the small molecules targeting cell cycle kinases have displayed limited clinical efficacy and have suffered from dose-limiting bone marrow toxicity.

Protein kinases are high-priority targets for drug discovery in oncology and other disease settings, and kinase inhibitors have transformed the outcomes of specific groups of patients. Most kinase inhibitors are ATP competitive, deriving potency by occupying the deep hydrophobic pocket at the heart of the kinase domain. Selectivity of inhibitors depends on exploiting differences between the amino acids that line the ATP site and exploring the surrounding pockets that are present in inactive states of the kinase. More recently, allosteric pockets outside the ATP site are being targeted to achieve high selectivity and to overcome resistance to current therapeutics. It is important to keep in mind what are the key regulatory features of the protein kinase family, describe the different types of kinase inhibitors, and highlight examples where the understanding of kinase regulatory mechanisms has gone hand in hand with the development of inhibitors.

The structural similarity of the protein kinase ATP binding site has allowed the identification of common hinge-binding scaffolds and enabled the prediction of binding modes and key interactions of novel ligands providing a rationale for inhibitor design. Approaches for developing ATP-competitive inhibitors are therefore relatively mature and can be used to rapidly generate libraries of compounds for inhibitor hit discovery.

The gatekeeper residue is a frequent site of point mutations in kinases that confer resistance, and therefore, it is a logical strategy to use this property to finetune kinase inhibitor selectivity. ATP-competitive kinase inhibitors are usually classified into three "types", all of which mimic ATP and make interactions with the hinge region within the kinase ATP site but vary in the conformational state of the kinase they interact with, and the additional interactions they make:

- Type I inhibitors are defined as small molecules that bind to the active conformation of the kinase (DFG-in and C-in);
- Type I12 inhibitors bind to the inactive DFG-in, C-out conformation of the kinase;
- Type II inhibitors bind to the inactive DFG-out conformation of the kinase (C-in or C-out).

While type I inhibitors are mostly confined to the ATP site, type I12 and type II inhibitors extend into the distinctive pockets that are opened up in the specific inactive conformations they bind. Inhibitors that do not compete with ATP provide the opportunity to gain greater selectivity over other kinases by binding to allosteric sites that are less conserved than the ATP-binding site. Allosteric inhibitors that bind adjacent to the ATP site

are classified as type III and those that bind to other more distant sites as type IV.
Another important class of kinase inhibitors are covalent kinase inhibitors or type VI inhibitors. These compounds exploit the presence of reactive amino acids in the ATP site, typically cysteine, although other residues such as lysine have also been successfully used. These cysteines are not required for catalysis, but they can be used to irreversibly block the ATP site through attachment of a small molecule [2].
The members of the Aurora kinase family play critical roles in the regulation of the cell cycle and mitotic spindle assembly and have been intensively investigated as potential targets for a new class of anticancer drugs. We describe a new highly potent and selective class of Aurora kinase inhibitors discovered using a phenotypic cellular screen. Optimized inhibitors display many of the hallmarks of Aurora inhibition including endoreduplication, polyploidy and loss of cell viability in cancer cells. Structure-activity relationships with respect to kinome-wide selectivity and guided by an Aurora B co-crystal structure resulted in the identification of key selectivity determinants and discovery of a subseries with selectivity toward Aurora A [10].

1.2.2 Aurora Kinase Pathway

This section introduces the resources that could be useful to investigate the biological pathways in which the Aurora kinase protein is involved.

The catalytic activity of protein kinases is regulated, and they can be thought of as molecular switches that are controlled through protein-protein interactions and post-translational modifications [2]. Aurora kinases are involved in cell cycle regulation and therefore are inserted in cell cycle pathways. The most useful resources for gaining insight, visualizing and extracting information about biological pathways are The Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Reactome pathway knowledgebase. Pathways in which either Aurora kinase A or Aurora kinase B are involved can be visualized at the following links: KEGG[13] and Reactome[14]. Analyzing and inspecting the biological pathways in which a biological entity is involved is of fundamental importance to understand the framework the protein acts in. The importance of such analysis is equivalent also for small-molecule drugs: indeed, several computational methods have been developed and deployed to analyze biological pathways and the possible interactions that drugs can have when inserted in a biological context relative to the protein they should act on.

[13] https://www.kegg.jp/kegg-bin/search_pathway_text?map=map&keyword=aurora+kinase&mode=1&viewImage=true
[14] https://reactome.org/PathwayBrowser/#/R-HSA-453276

1.3 Predicting ADME-Tox properties

1.3.1 Metabolism

1.3.2 Excretion

1.4 Methods

1.4.1 Machine Learning

1.4.2 Graph Neural Networks

1.4.3 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 [13]. 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.

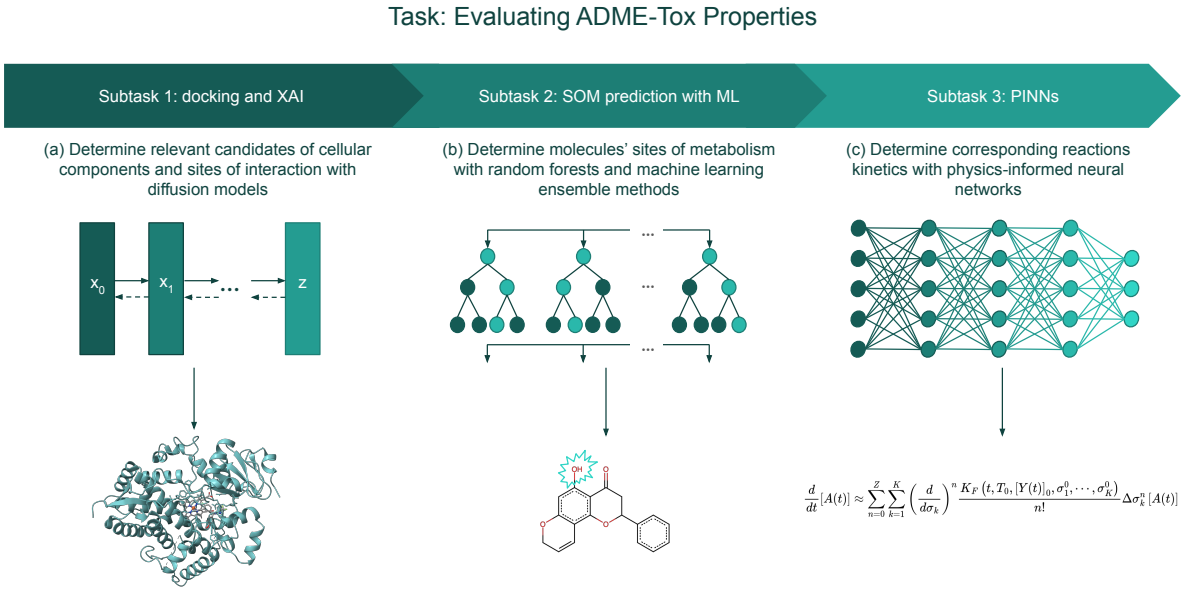

Figure 1: graphical representation of subtasks

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 [6], 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 [15]. We complete the subtask by verifying the biological plausibility of the complex using AutoDock Vina for classical molecular dynamics simulations [14] 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] Chris Arter et al. *Structural features of the protein kinase domain and targeted binding by small-molecule inhibitors*. Aug. 2022. DOI: 10.1016/j.jbc.2022.102247.

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

[4] Khushwant S. Bhullar et al. "Kinase-targeted cancer therapies: Progress, challenges and future directions". In: *Molecular Cancer* 17 (1 Feb. 2018). ISSN: 14764598. DOI: 10.1186/s12943-018-0804-2.

[5] Mar Carmena and William C. Earnshaw. "The cellular geography of Aurora kinases". In: *Nature Reviews Molecular Cell Biology* 4 (11 Nov. 2003), pp. 842–854. ISSN: 14710072. DOI: 10.1038/nrm1245.

[6] 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.

[7] David M Glover et al. *Mutations in aurora Prevent Centrosome Separation Leading to the Formation of Monopolar Spindles*. 1995, pp. 95–105.

[8] Daniel R Knighton et al. "Crystal Structure of the Catalytic Subunit of Cyclic Adenosine Monophosphate-Dependent Protein Kinase". In: *Science* 253 (July 1991), pp. 407–414. URL: www.sciencemag.org.

[9] Madhu Kollareddy et al. "Aurora kinase inhibitors: Progress towards the clinic". In: *Investigational New Drugs* 30 (6 Dec. 2012), pp. 2411–2432. ISSN: 01676997. DOI: 10.1007/s10637-012-9798-6.

[10] Nicholas Kwiatkowski et al. "Selective aurora kinase inhibitors identified using a taxol-induced checkpoint sensitivity screen". In: *ACS Chemical Biology* 7 (1 Jan. 2012), pp. 185–196. ISSN: 15548929. DOI: 10.1021/cb200305u.

[11] Christopher L. McClendon et al. "Dynamic architecture of a protein kinase". In: *Proceedings of the National Academy of Sciences of the United States of America* 111 (43 Oct. 2014), E4623–E4631. ISSN: 10916490. DOI: 10.1073/pnas.1418402111.

[12] Patrick Meraldi, Reiko Honda, and Erich A. Nigg. "Aurora kinases link chromosome segregation and cell division to cancer susceptibility". In: *Current Opinion in Genetics and Development* 14 (1 2004), pp. 29–36. ISSN: 0959437X. DOI: 10.1016/j.gde.2003.11.006.

[13] 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.

[14] 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.

[15] 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