Taylor & Francis
Taylor & Francis Group

Expert Opinion on Drug Discovery

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/iedc20

Industrializing AI-powered drug discovery: lessons learned from the *Patrimony* computing platform

Mickaël Guedj, Jack Swindle, Antoine Hamon, Sandra Hubert, Emiko Desvaux, Jessica Laplume, Laura Xuereb, Céline Lefebvre, Yannick Haudry, Christine Gabarroca, Audrey Aussy, Laurence Laigle, Isabelle Dupin-Roger & Philippe Moingeon

To cite this article: Mickaël Guedj, Jack Swindle, Antoine Hamon, Sandra Hubert, Emiko Desvaux, Jessica Laplume, Laura Xuereb, Céline Lefebvre, Yannick Haudry, Christine Gabarroca, Audrey Aussy, Laurence Laigle, Isabelle Dupin-Roger & Philippe Moingeon (2022) Industrializing AI-powered drug discovery: lessons learned from the *Patrimony* computing platform, Expert Opinion on Drug Discovery, 17:8, 815-824, DOI: 10.1080/17460441.2022.2095368

To link to this article: https://doi.org/10.1080/17460441.2022.2095368

EXPERT OPINION ON DRUG DISCOVERY
2022, VOL. 17, NO. 8, 815–824
https://doi.org/10.1080/17460441.2022.2095368

Taylor & Francis
Taylor & Francis Group

TECHNOLOGY EVALUATION

Industrializing AI-powered drug discovery: lessons learned from the *Patrimony* computing platform

Mickaël Guedj [a], Jack Swindle[b], Antoine Hamon[b], Sandra Hubert[a], Emiko Desvaux[a], Jessica Laplume[a], Laura Xuereb[a], Céline Lefebvre[a], Yannick Haudry[a], Christine Gabarroca[a], Audrey Aussy[a], Laurence Laigle[a], Isabelle Dupin-Roger[a] and Philippe Moingeon [a]

[a]Servier, Research & Development, Suresnes, France; [b]Lincoln, Research & Development, Boulogne-Billancourt, France

ABSTRACT

Introduction: As a mid-size international pharmaceutical company, we initiated 4 years ago the launch of a dedicated high-throughput computing platform supporting drug discovery. The platform named '*Patrimony*' was built up on the initial predicate to capitalize on our proprietary data while leveraging public data sources in order to foster a Computational Precision Medicine approach with the power of artificial intelligence.

Areas covered: Specifically, *Patrimony* is designed to identify novel therapeutic target candidates. With several successful use cases in immuno-inflammatory diseases, and current ongoing extension to applications to oncology and neurology, we document how this industrial computational platform has had a transformational impact on our R&D, making it more competitive, as well time and cost effective through a model-based educated selection of therapeutic targets and drug candidates.

Expert opinion: We report our achievements, but also our challenges in implementing data access and governance processes, building up hardware and user interfaces, and acculturing scientists to use predictive models to inform decisions.

ARTICLE HISTORY
Received 19 April 2022
Accepted 24 June 2022

KEYWORDS
Drug discovery; target identification; data integration; artificial intelligence; multi-omics; computing platform; Computational Precision Medicine

1. Introduction

In the last decades, the pharmaceutical industry has faced a continuous decrease in productivity. R&D efficiency, measured by the number of new drugs brought to patients per dollar spent, has halved approximately every 10 years since 1950. This trend is often referred to as the Eroom's Law, i.e. a reverse of the well-known Moore's Law reflecting the exponential growth observed over time for numbers of transistors on a microchip [1]. As of today, it takes on average 12 years and 2.6 billion U.S. dollars to bring a new drug to patients, with a probability of success of around 5–10%[2]. Root causes encompass an increase in regulatory requirements, the lack of sufficient validation in the selection of both therapeutic targets and drug candidates prior to setting up costly and time-consuming clinical trials, and a need to optimize organizational processes to better integrate scientific knowledge within R&D [3–5]. Arguably, the most actionable lever for the pharmaceutical industry in order to increase the probability of success during drug development is to strengthen the rationale behind decision-making, most particularly as it relates to the choice of the therapeutic target and the selection of the drug candidate.

To this aim, recent breakthroughs in both biomedical and computational sciences create new opportunities to inform drug development through computer-based approaches [6–13]. Rapid advances in omics, imaging, and electronic capture technologies make it now possible to characterize individuals at both molecular, cellular, and clinical levels in a cost and time effective way [14]. Those advances occur in parallel with the exponential accumulation of information and knowledge accessible through hundreds of structured biomedical databases, such as those managed by the *European Bioinformatics Institute* (EBI) or the US *National Center for Biotechnology Information* (NCBI). An effective use of these massive amounts of data is facilitated by new computational approaches including artificial intelligence (AI) and machine-learning (ML), thus creating an unprecedented opportunity to better inform decision-making and decrease both costs and attrition rates at all stages of drug discovery and development [5]. This ongoing revolution toward AI-powered drug discovery already translates into concrete successes, with machine-designed anticancer molecules reported to reach Phase 1 in less than 2 years in contrast to the 5–7 years commonly needed for the discovery phase [15].

An emerging strategy for the pharmaceutical industry to capitalize on those new approaches is to build-an internal computing platform to support the identification of disease targets as well as the repurposing of existing drugs in a systematic and efficient way. To this aim, a technical challenge is to integrate the ever-increasing number of data sources compiling massive and multidimensional information coming from genetics, multi-omics, molecular interactions,

CONTACT Mickaël Guedj mickael.guedj@gmail.com Servier, Research & Development, Suresnes Cedex, France

Supplemental data for this article can be accessed online at https://doi.org/10.1080/17460441.2022.2095368

Article highlights

- *Patrimony* is a computing platform implemented by the pharmaceutical company Servier to capture the value of massive biomedical data (proprietary and public) with machine-learning techniques to support drug discovery.
- The platform is based on a knowledge-graph connecting biomolecular, pharmacological, and clinical domains, which can be mined to generate new therapeutic hypotheses.
- Hypotheses are assessed and prioritized by aggregating information around five summary criteria (Biological Relevance, Causality, Tractability, Safety, and Innovativeness).
- The platform was implemented following three iterations (proof-of-concept, structuration, and industrialization), while relying upon an Agile operating model and FAIR guiding principles for data management.
- Reported applications encompass immuno-inflammatory diseases and COVID-19; from these successes, the use of the platform is being extended to oncology and neurology.
- *Patrimony* fosters an open innovation mindset, requiring both agility and transversality across a broad range of existing and emerging expertises associated with AI.

preclinical experiments, clinical as well as real-life evidence data. A dedicated and specifically designed computational framework is necessary to capture the value of all these information in order to guide decision-making and increase the probability of success during drug development.

In light of the rapidly evolving environment driven by AI and digital technologies, a decision was made in early 2018 by the pharmaceutical group Servier to implement within the R&D new data processes and computational methods through a dedicated high-throughput computing platform. Given that a primary objective of this initiative was to valorize existing data, knowledge, and assets produced by the company during previous or ongoing R&D projects, the platform was named '*Patrimony*.' The latter was built up with the intent to capitalize on both proprietary and public data to drive innovation. After 4 years of implementation, the *Patrimony* platform has transformed very significantly Servier's approach to drug discovery and development. Herein, we share the main lessons learned from this initiative during its conception and implementation, as well as the challenges that were faced to make it impactful on our drug development processes.

2. Overview of the market

Several public or public–private initiatives have emerged with this orientation such as *Open Targets* [16,17]. In parallel, the perspective of substantial cost savings offered to pharmaceutical industries by substituting computerized modeling systems to traditional wet-bench biology, prompted the emergence of numerous start-ups aiming to reinvent drug discovery (**Supp. Table 1**). In this context, in parallel to initiating this project, we performed a benchmark analysis of external solutions. We concluded that given its foreseen strategic relevance for the company, there was a strong added value to implement the project internally for gaining reactivity, flexibility, as well as for a better integration of internal and external data sources.

3. How the *Patrimony* computing platform works

3.1. General framework

A computing platform comprises a set of hardware infrastructures, software, and user interface components that allow storing data and running algorithms in order to achieve a set of well-defined tasks within a specific field of application. The methodology we applied to build up such a computing platform for drug discovery is summarized in Figure 1, combining a range of computational approaches in workflows to integrate, analyze, and interpret data. A first step was to identify all relevant existing biomedical databases and knowledge sources, both structured and unstructured, public or internal. A second step was to curate and integrate data sources into a knowledge graph (or network) as defined below. Lastly, algorithms were developed to mine this graph for a specific application (related to a given disease or a therapeutic area) in order to generate and prioritize hypotheses regarding new therapeutic targets or opportunities for drug repurposing. Analyzing such a volume and diversity of data turned out to be challenging, thus requiring an adapted computational framework to ensure both a good integration, appropriate use, and traceability. This methodology was concretized into a high-throughput scalable process encompassing all steps from data acquisition, hypothesis generation, prioritization of outcomes, and experimental validation.

We separated data sources based upon whether they were either in-house or public, as well as application-agnostic versus application-related. From a core set of in-house/public sources shared for all applications, we subsequently added a data package specific to each single application of interest, making the *Patrimony* framework both robust and highly flexible. In-house sources broadly used to support Servier's R&D projects comprised data related to both therapeutic targets being investigated, potentially relevant proprietary drug(s), the phase within research or development as well as the therapeutic area. Core public sources used such as *DrugBank* or *UniProt* listed in Supp. Table 2 encompass the existing public knowledge on multiple domains of interest (e.g. biomolecular, pharmacological, clinical) [16–52]. The application-related data package further assembled focused on the disease or set of diseases of interest, which for Servier relates to immuno-inflammatory, oncological, and neurological disorders. The level of implication of various genes and proteins in pathophysiological processes was obtained from multi-omics patient profiling data, by comparing cases and controls in different relevant conditions; either from aggregated statistics or derived from sample-level molecular data (subsequently turned internally into aggregated statistics) retrieved from both public repositories such as *GEO* or *UK Biobank* listed Supp. Table 2, partnerships such as public–private IMI projects as well as proprietary experimental data [53–57].

3.2. Building up a proprietary and adaptable knowledge graph

Knowledge graphs are network-like digital structures representing knowledge as a set of concepts and their relationships. As such, they facilitate the interface between humans and machines to analyze their content and support complex

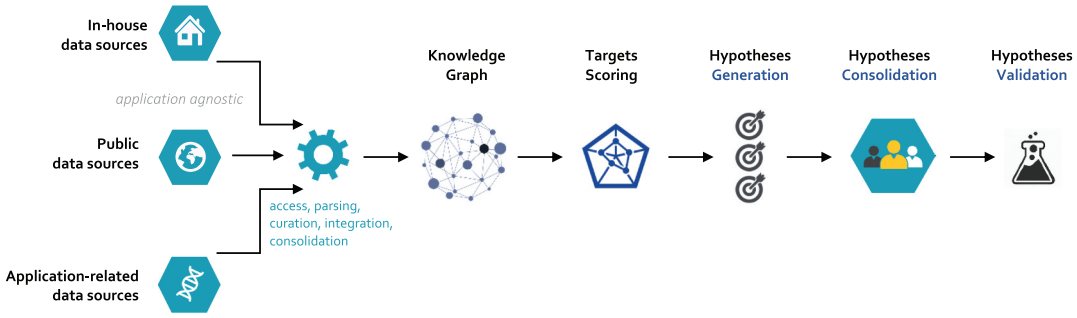

Figure 1. Patrimony general framework. Data sources are curated and integrated into a knowledge graph. For each application starting from one disease or a set of diseases, the knowledge graph is mined in order to evaluate the putative targets, generate hypotheses, and assist the consolidation.

decision-making. They are well adapted to model the interconnectivity of biomedical systems. Numerous detailed and formal introductions to their application in Medicine have been published elsewhere [58–60].

From the data sources described above, we built a knowledge graph connecting an unbiased mapping of all known molecular interactions (i.e. the Human Interactome) to pharmacological and clinical domains (Figure 2a). The resulting graph was made of overall 50k+ nodes and 200k+ relationships, combined with all information we could gather or generate as attributes on each node or interaction, and further enriched with semantics or ontologies to help navigating through the concepts (e.g. *GO, ChEBI*, EFO, *MedDRA*) [29–31,36,48]. For each application, specific related data sources were used to complement the core knowledge graph with a set of additional nodes, interactions and attributes. Aggregate statistics resulting from multi-omics analyses (e.g. fold-changes, *p*-values, etc.) were mapped to the knowledge graph as attributes of gene or protein nodes. Interactions inferred from patient-level data such as gene–gene co-expression values were used to weight or complete the known interactions between them.

At the heart of *Patrimony* is the relationship between the biomolecular, pharmacological, and clinical spaces in order to identify the most relevant therapeutic targets with regard to the measured pathophysiological manifestations of a disease or a set of diseases (Figure 2a). The resulting knowledge graph is unique to the owner and proprietary in the sense that it integrates public data with internal inputs. Consequently, it represents a very strategic entry point to capture the value of the scientific knowledge accumulated over time by the company.

In order to mine this knowledge graph and extract the most relevant information, some specific methodological expertise has been developed (Figure 2b). As main examples, the identification of nodes exhibiting more frequent interactions with other nodes (i.e. hubs), or nodes having frequent interactions with each other (i.e. clusters) were deemed of

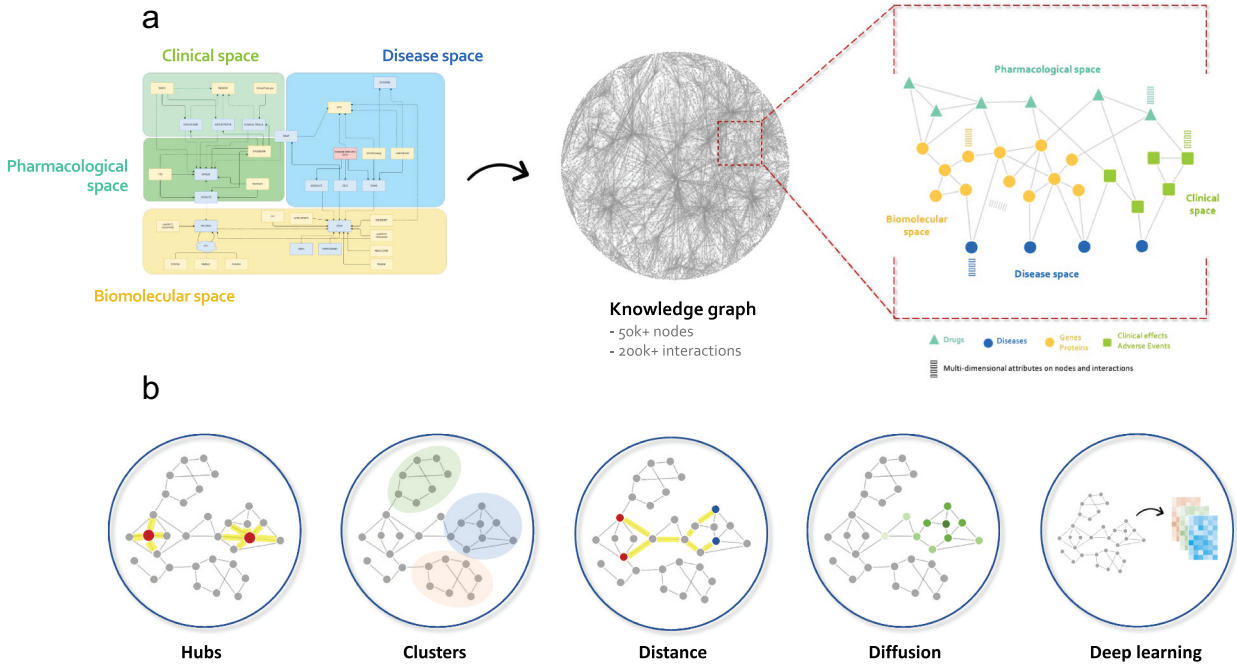

Figure 2. Patrimony knowledge graph. (a) Data sources are integrated through a relational diagram to form the Patrimony knowledge graph covering and interconnecting biomolecular, pharmacological, disease, and clinical domains. (b) Different metrics and approaches are used to mine the knowledge graph.

particular interest as they tend to correspond to molecules predicted to play a key role in biological or pathological processes. The concept of distance within networks was also found to be critical, albeit not easy to quantify. Whereas the distance between two nodes can appear trivial to determine, assessing the distance between two sets of nodes raises many questions depending upon the use case. We generally use a topological distance corresponding to the shortest path length in the graph between the nodes of interest. As the significance of the distance depends on the density of the graph, we generate a distribution from bootstrapping similar nodes defined by same degree in the graph; from this given distribution, we derive a strandardized *z*-score and a corresponding *p*-value. Diffusion/propagation algorithms were selected as particularly useful because – as their name infer – they diffuse or propagate the information along the links of the graph. It assesses the impact of a perturbation starting from a given node on the network by use of radom-walk probabilities. As such, they allow to capture the information on both the nodes and their neighborhood [61]. Furthermore, the graph structure turned out to be very adapted to facilitate the direct application of graph-based deep learning approaches such as Graph Convolutional Networks (GCNs) that aggregate features of the different types of nodes and their relationships in order to predict new associations between drugs, targets, and diseases [60,62]. The general principle is to learn how to represent the graph and map its nodes into a compact embedding space. As a result, it embeds diseases associated with similar genes or drugs whose target proteins have similar local neighborhoods close together in the embedding space. Altogether, one advantage of the knowledge graph supporting the *Patrimony* computing platform turned out to allow initiating investigations from any entry point among genes, diseases, or drugs depending on the data sources already integrated within the platform. Potential queries related for instance to the most relevant targets for a given disease, or to potential indications in which a drug could be repurposed. As such, the *Patrimony* computing platform has now been well established within Servier as a versatile tool to create and maintain a competitive advantage when developing drugs against diseases of interest.

3.3. Target hypothesis assessment and prioritization

A central application of *Patrimony* is to identify novel therapeutic target candidates from the modeling of a disease of interest. To capitalize on the high quantity of information contained in its knowledge graph, some metrics have been established in order to rationalize the assessment and prioritization of actionable therapeutic targets. Inspired by principles originally found in *Open Targets*, we eventually selected strategic summary criteria along five distinct dimensions (Figure 3a).

The first and most important criterion relates to *Biological Relevance*. For a given use, it summarizes all the activities contributing to the understanding of the pathophysiology of a disease from multi-omics data. A score is generated to quantify the level of cumulated evidence predicting a gene or a protein to be a highly relevant target because of biomolecular associations or dysregulations, by itself and/or when considering its neighborhood within the graph. Genes with a high *Biological Relevance* are referred to as disease-related genes, in that they are likely to contribute to the pathophysiology as a cause or a consequence. Also, they tend to cluster and form identifiable disease modules within the knowledge

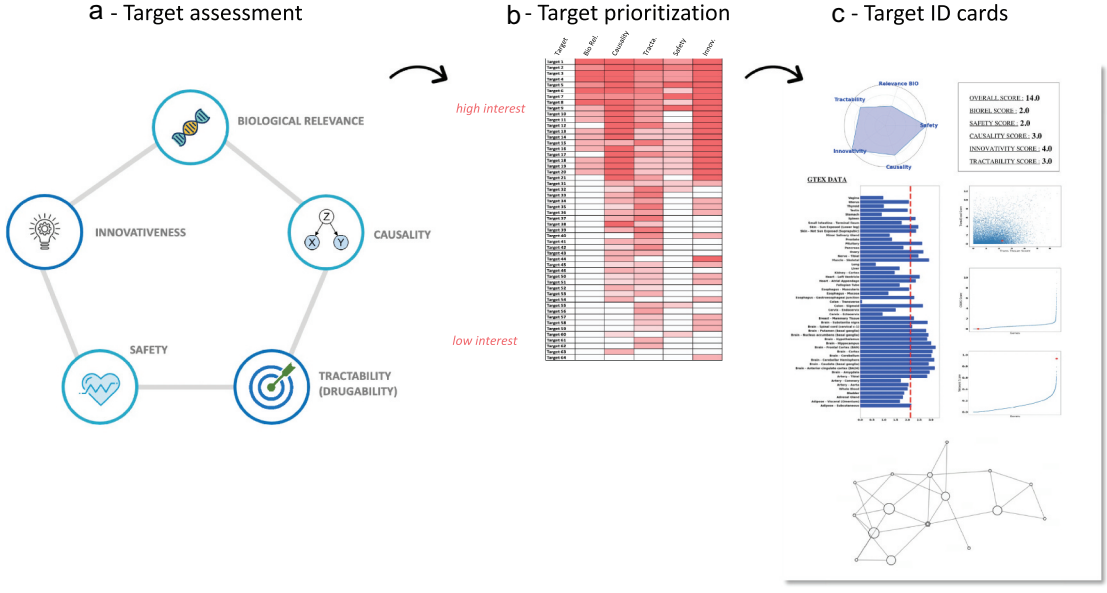

Figure 3. Target assessment and prioritization. (a) Targets are assessed based on strategic summary criteria along five distinct dimensions. (b)The five criteria are individually quantified and a global scoring is then computed in order to prioritize the top targets for which the overall rationale is the highest. (c) Individual target ID cards are generated to represent the global assessment in a visual and easily interpretable format to end users.

graph. Computing techniques described previously including AI/ML support here the prediction of new disease–gene associations as well as disease modules. A second criterion, related to biological relevance but nevertheless assessed independently, is *Causality*, which we consider as a highly critical dimension. The latter is being assessed to discriminate between causes and consequences in the observed pathophysiology. As such and depending on the information available, it can be derived from (i) the force of genetic associations around the target (taking into account that genetic signals are rather causal compared to transcriptomic profiles which often result from downstream regulations), (ii) the expression of the target in cells or tissues relevant to the disease of interest (documented for instance from *GTeX*), and (iii) AI/ML predictions of what is likely to be a real target based on graph features and known approved targets as training. On a concrete basis, *Causality* may for instance refer to master regulators or driver mutations [32,33]. A third criterion relates to *Tractability* (also referred to as *Druggability*), which represents the likelihood to modulate the function of a target with either small synthetic or biological drugs. To date, we have been using the measure proposed by *Open Targets* based on both precedence of the target in clinical trials, discovery experiments and computational predictions. In the future, this assessment will be supplemented with our own measures of Tractability to include additional modalities of interest, e.g. RNA-based antisense oligonucleotides. The fourth criterion is linked to potential *Safety* implications when interfering with a given target. It is assessed by the number of safety events associated with drugs known to bind the target reported in databases such as *SIDER* [44]. In subsequent developments, we will consider not only the number but also the severity of adverse events. The last criterion considers *Innovativeness* in relationship with the application on a disease of interest, documented from either clinical trials, as well as patent or literature-mining by using *Natural Language Processing* (NLP) [63]. Given the flexibility of the *Patrimony* computational framework, any other relevant summary criterion could be easily included in the future. For instance, assessing the *Feasibility* of drug development could be added to support target ranking, as suggested by others [64].

The above mentioned five criteria are individually quantified and subsequently computed in a global scoring in order to prioritize the top targets for which the overall rationale is the highest (Figure 3b). For any given target, individual 'target ID cards' are generated to represent the overall assessment in a visual and easily interpretable manner to end users (Figure 3c).

3.4. From targets to drugs

Following the prioritization of top hypotheses on candidate targets with the *Patrimony* scoring system, the rationale is subsequently consolidated through a deep-dive investigation by biologists and pharmacologists. During this consolidation phase, both an extensive literature review and in house translational data analyses are performed to confirm that identified target candidates are involved in specific disease pathways and are druggable with a specific compound modality.

Researchers can then validate target hypotheses, through an experimental confirmation that disease activity is impacted following perturbation of the target of interest with a drug or a tool compound. Conducting wet-lab gene inhibition (e.g. via CRISPR-Cas9 deletion or RNA silencing) or preclinical experiments by using cellular assays or animal models are commonly implemented to corroborate the hypothesis that drugs interacting with the target exhibit the anticipated pharmacological activity. Once a therapeutic target has been selected, multiple processes streamlined by the pharmaceutical industry can be used to identify small molecules or biologicals interacting with it. For instance, High-Throughput Screening (HTS) can be implemented to test the company's proprietary compound library in various molecular or cell-based assay systems in order to identify drug candidates [65]. As of today, another strategy relies upon dedicated computational methods to select in silico drugs predicted to engage the target of interest [15].

One specific use of the *Patrimony* platform consists in identifying existing drugs modulating the target, e.g. within the R&D pipeline of the company or among marketed compounds. Referred to as drug repurposing, drug rescuing or indication extension, this approach has generated a growing interest during the last decade, with evidence that mining available biomedical data with proper algorithms can generate fast and valid innovative hypotheses [41,66–68]. The knowledge graph used in *Patrimony* has thus also been applied to such drug repurposing approaches, by using both distances between drug targets and disease-related proteins [69], connectivity maps [23,24], and deep-learning methodology to identify in silico new targets for known drugs [70].

3.5. Implementation

We defined three iterations to implement the *Patrimony* computing platform: a *proof-of-concept*, a *structuration,* and an *industrialization* step. The *proof-of-concept* was pilot initiative performed within 3 months aiming for quick-wins. Based on a minimal set of datasets and algorithms, the aim was to position a given set of targets of interest into one disease (namely Sjögren syndrome). Then, the *structuration* step aimed to list, retrieve, and implement all the necessary datasets and algorithms along with conceiving a first dedicated and adapted computational infrastructure. This step was performed around a well-defined application focused on immuno-inflammatory diseases. Finally, the *industrialization* step aimed to transfer the existing *Patrimony* platform into a more scalable architecture in order to pave the ground for a subsequent application to all therapeutic areas of interest for the company. Within each iteration, we adopted an Agile operating model for software development with alternated sequences of brainstorming, implementation (sometimes in precisely defined sprints), generation of results, consolidation, and feedback [71]. Technical choices for IT infrastructure, software and algorithms have proven to be challenging to ensure a robust and flexible solution in a fast-evolving field. Codes were developed in *Python* and *R*. For the two first iterations, we put in place a *Microsoft Azure* cloud-based sandbox with *MongoDB* for database management. We subsequently moved

to *Google Cloud Platform (GCP)* and *BigQuery* for scalability. We also used a mixture of *Neo4j, Cytoscape, Python graph-tool,* and *R igraph* to support graph storage, mining, and visualization [72]. We followed FAIR guiding principles to enable findability, accessibility, interoperability, and reusability of the data when building up our new data governance. Integrating large-scale and multidimensional data generated from multiple technologies with proper quality attributes in terms of consistency and reliability remained a significant difficulty throughout data life-cycle management. Assessing the right to use data from public sources also turned out to be complex. Whereas some data sources apply a clear 'no restriction' policy for any use of the data, e.g. under a *Creative Common* license for sharing, others make a distinction between the type of requesting institution (profit versus nonprofit) or intended use (research versus commercial purposes). Of note, it has been decided not to include any patient-level data in order to overcome the regulatory complexity linked to their use. Also, the possibility to include data from partnerships such as IMI projects needs to be evaluated case by case.

4. Applications

The two first iterations (*proof-of-concept* and *structuration*) of the *Patrimony* initiative were built while focusing on immuno-inflammatory diseases as selected indications. They were specifically designed to evaluate the capacity of the *Patrimony* platform to support two direct applications, i.e. identifying therapeutic targets and generating hypotheses for drug repurposing.

As a first run, we mapped multi-omics profiling data from the PRECISESADS cohort of patients with various autoimmune diseases into the knowledge graph to support drug development against primary Sjögren's syndrome [73]. A particular challenge in this indication was to rationally design immunotherapeutic approaches acting at a systemic level and/or target organs (i.e. salivary and lachrymal glands). As concrete outputs, *Patrimony* helped to identify and prioritize several innovative therapeutic targets, supported by a robust and multidimensional set of evidence. Most of the targets identified as of interest in primary Sjogren Syndrome were confirmed to be valid as well in several autoimmune diseases sharing common pathophysiological mechanisms, e.g. Systemic Lupus Erythematosus. Furthermore, *Patrimony* was very useful not only to identify new therapeutic targets but also to validate other ones for which our company had already initiated the development of drug candidates. Specifically, two monoclonal antibodies at an early clinical development stage for autoimmune diseases, including anti-type 1 interferon and anti-IL7R antibodies (ClinicalTrial.gov NCT04605978) were confirmed as valid therapeutic options in Sjogren Syndrome and Systemic Lupus Erythematosus. Lastly, the availability of disease models providing emphasis on specific therapeutic targets as being relevant in various autoimmune diseases has been a powerful tool to support Servier's assessment of external licensing opportunities for drug candidates.

As an effort to contribute to the global fight against the COVID-19 pandemics, a second application of the *Patrimony* platform aimed at identifying existing drugs that could be repurposed to treat those patients infected by the SARS-CoV -2 virus who develop severe forms of the disease requiring hospitalization. Specifically, we modeled the severe lung inflammation associated with the life-threatening acute respiratory distress syndrome, which affects up to 75% of COVID-19 patients transferred to intensive care units [74]. From data available in the scientific literature documenting differences at a molecular level in both immune responses and tissue inflammation between patients with either mild or very severe forms of the disease, an interactome of proteins predicted to contribute to lung inflammation in severe COVID-19 was produced. The latter was used to confirm the interest of several drugs already used in this indication such as dexamethasone, anti-IL6R antibodies or JAK2 inhibitors. It further identified additional drugs, either available in other indications or in development, as being relevant for repurposing in severe COVID-19, such as inhibitors of alarmins and their receptors [75].

Based on these promising pilot applications to support our drug development in immuno-inflammation, we have now initiated an industrialization phase to extend its application to oncology and neurology. Whereas the methodology and processes behind were broadly applicable in those additional indications, new challenges emerged reflecting the specificities in terms of categories of data predominantly used in distinct therapeutic areas. For instance, a complexity in neurology is to get access to data related to biological processes occurring in the brain beyond solely postmortem samples. This hurdle induces as a consequence an emphasis in this field in genetics and animal data when compared with other therapeutic areas. In contrast, disease modeling in oncology can benefit from an abundance of molecular data from a great variety of sources, encompassing both constitutional and tumor-specific molecular alterations. In the latter field, documenting gene essentiality through CRISPR-Cas9 deletion experiments, e.g. as the ones centralized in the *Depmap* project, are critical to prioritize among a wealth of hypotheses [54].

Various additional applications could be considered in the future, as an extension to further support the identification and optimization of drug candidates, beyond the repurposing of existing drugs. The current models produced by *Patrimony* leading to therapeutic target identification could be nicely extended by machine-learning approaches to perform multi-task parallel prediction of drug candidate characteristics. The latter include training artificial neural networks to select suitable therapeutic modalities for engaging a given target, predicting both binding characteristics as well as pharmacological and ADMET properties of virtual compounds, and even creating new molecules by using generative adversarial networks [6]. Other applications requiring specific computational methods include generating hypotheses on potential combination therapies to address complex diseases [76]. Whereas the interest of the latter approach is well identified in oncology, we have as well recently documented how it could be

implemented in autoimmune diseases, and arguably in many other therapeutic areas [77].

5. Challenges

Implementing ex nihilo in a pharmaceutical environment at an industrial scale, an initiative such as *Patrimony* has been disruptive. A major challenge was that it required a high level of transversality between multidisciplinary teams assembling numerous expertises encompassing computational hardware, cloud computing, network computing, machine-learning and AI, statistics, bioinformatics, large-scale biology, pharmacology, clinical knowledge, but also legal skills to assess data accessibility and use. Aligning those very diverse human expertises necessitated continuous training and internal communication to foster acculturation to computational modeling as it applies to drug discovery and development. Another important challenge was the consolidation step needed to validate the scientific rationale of predicted targets. Algorithms can generate many hypotheses in a short time frame. The extensive literature search by human experts remained time consuming in order to corroborate or refute the hypotheses generated. In this exercise, we found it critical for each given application to a disease of interest to evaluate the outputs with known targets and disease-modifying drugs. To that end, clinically approved drugs and their respective targets with established proof of efficacy have been assessed on the platform and their relevance verified. Altogether, during the consolidation step, we consistently observed that the relevance of the hypotheses generated depends on the quality, completeness and continuous updates of data sources. The interpretation of model outputs by human experts was also difficult, owing to the inherent complexity of the analysis of biomedical data, but also in light of the numerous existing gaps in the thesaurus of human knowledge. The current Human Interactome is estimated to cover only circa 25% of all possible molecular interactions [78]. As a consequence, many genes related to a given disease may appear to be dispersed in the interactomes making the identification of coherent disease modules challenging. In our experience, we found that protein expression, documented for example by proteomics or flow cytometry, was more translatable to disease pathological mechanisms than gene expression solely assessed by transcriptomics.

6. Conclusions

The knowledge-graph represented within *Patrimony* is a sophisticated and evolving way to represent a disease within a computer system. Disease modeling has recently emerged as a powerful mean to educate the design and development of drug candidates, in the context of Precision Medicine approaches aiming to offer therapies better tailored to the patients' specificities. Such strategies are thus being established to capitalize on a comprehensive knowledge of the disease and of patient population heterogeneity to select therapeutic targets and drug candidates predicted to be most suitable in this indication. This trend that we termed Computational Precision Medicine has become highly strategic for pharmaceutical industries to differentiate from the competition [6]. We thus emphasize the tremendous value of developing a proprietary computational platform to create a major competitive advantage for drug developers in their disease areas of interest.

Within Servier, we designed the *Patrimony* platform by combining innovative concepts, methodologies and supportive infrastructure to grow significantly our ability to integrate large-scale biomedical data. As of today, *Patrimony* allows to generate multidimensional models relating to diseases, therapeutic targets, and to some level drug candidates. Based on our positive experience in exploring pilot applications in immuno-inflammatory diseases, the use of this computational platform is now being deployed throughout all therapeutic areas of interest for the company.

7. Expert opinion

As AI has the potential to transform drug discovery and several AI-driven molecules have progressed into clinical trials (most often with accelerated timelines and reduced costs), its impact and remaining challenges need to be globally addressed [79,80]. As explainabilty is often invoked, we recommend paying attention to generating results that are both robust (e.g. to small changes in parameters and data), interpretable (e.g. avoiding applying transformation to variables that would disconnect them to their biological or clinical meaning), and reproducible (e.g. by applying good coding and data practices). Another important hurdle to implement disease modeling in support of drug discovery is the difficulty to distinguish causal from incidental genes or proteins in the pathophysiology. In future analyses, computational inferences of causality based for example on Bayesian networks represent an interesting option to shed light on disease-related master regulators or driver mutations [81,82]. The converging advances in high-throughput technologies such as single-cell RNA sequencing or deep immunophenotyping along with protein structure prediction with the AlphaFold algorithms will contribute to expand continuously our knowledge space [83,84]. Integrating this flow of new data implies to regularly update the knowledge graph, thus raising concerns in terms of trustworthiness of data sources and quality control. The analysis of disease-related processes requires complementing topology with dynamic properties at large scale, which remains a major challenge [85,86]. Also, as computational models are accumulating, there is an increasing need for rapid validation and consolidation with data generated on purpose or well-established external reference datasets such as those proposed in the frame of data challenges (e.g. *Kaggle*, Dream, or *GeneDisco*). Eventually, computational innovative developments should also encompass other therapeutic modalities beyond small molecules, such as biologics, antisense oligonucleotides, protein degradation targeting and nanoparticles as well as potential combination therapies.

Importantly, the *Patrimony* platform has become a highly transforming asset within our company's R&D. It was seminal to initiate a transition from classical drug development relying upon life sciences, chemical expertise, and biotechnologies to an educated computer-based approach powered by AI. As

such, it still challenges traditional organizational structures for drug discovery and imposes a close cooperation between multidisciplinary teams. It progressively fosters an open innovation mindset within the company, requiring both agility and transversality across a broad range of existing and emerging expertises. In light of this very significant cultural change, acculturation to help acceptance by human experts of computational approaches to drug R&D is presently one of the most critical issues to solve in order to capture their full benefit.

Acknowledgments

The authors would like to thank all members of Servier R&D who supported the initiation and implementation of *Patrimony*. In particular M Coste, S Ollivier, V Robert, C Hébert, A Saugeot, A Bril, R Jeggo, F Schmidlin, and C Bertrand. The authors also thank B Theofilopoulou and T Bolba for legal assistance.

Funding

This work was funded by Servier.

Declaration of Interest

All authors are employees at Servier, an international pharmaceutical company governed by a non-profit foundation. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Author contributions

Conceptualization and supervision: M Guedj, S Hubert, J Laplume, I Dupin-Roger, and P Moingeon

Implementation: M Guedj, J Swindle, A Hamon, J Laplume, L Xuereb, C Lefebvre, Y Haudry, and C Gabarroca

Data access, curation, and integration: J Swindle, A Hamon, and J Laplume

Design of applications: S Hubert, E Desvaux, A Aussy, L Laigle, I Dupin-Roger, and P Moingeon

Writing of the manuscript: M Guedj and P Moingeon

Review and editing: all the authors

URLs

Public sources

AlphaFold https://alphafold.ebi.ac.uk
BindingDB https://www.bindingdb.org
Blueprint https://www.blueprint-epigenome.eu
ChEBI https://www.ebi.ac.uk/chebi
ClinicalTrials.gov https://clinicaltrials.gov
CMAP https://clue.io/cmap
Depmap https://depmap.org
DisGenet https://www.disgenet.org
DrugBank https://www.drugbank.ca
EBI https://www.ebi.ac.uk
EFO http://www.ebi.ac.uk/efo
Ensembl https://www.ensembl.org
GEO https://www.ncbi.nlm.nih.gov/geo
GO http://geneontology.org
GTeX https://gtexportal.org
GWAS Catalog https://www.ebi.ac.uk/gwas
IMI https://www.imi.europa.eu
MedDRA https://www.meddra.org
MSigDB https://www.gsea-msigdb.org
NCBI https://www.ncbi.nlm.nih.gov
NCBI Genes (ex Entrez) https://www.ncbi.nlm.nih.gov/gene
OmicSoft DiseaseLand https://digitalinsights.qiagen.com
Open Targets https://www.opentargets.org
PharmGKB https://www.pharmgkb.org
Reactome https://reactome.org
SIDER http://sideeffects.embl.de
STRING https://string-db.org
TTD http://db.idrblab.net/ttd
UK Biobank https://www.ukbiobank.ac.uk
UniProt https://www.uniprot.org

Implementation

BigQuery https://cloud.google.com/bigquery
Creative Common https://creativecommons.org
Cytoscape https://cytoscape.org
FAIR https://www.go-fair.org/fair-principles
GCP https://cloud.google.com
Graph-tool https://graph-tool.skewed.de
Igraph https://igraph.org
Microsoft Azure https://azure.microsoft.com
MongoDB https://www.mongodb.com
Neo4j https://neo4j.com
Python https://www.python.org
R https://cran.r-project.org

Data challenges

Dream http://dreamchallenges.org
GeneDisco https://www.gsk.ai/genedisco-challenge
Kaggle https://www.kaggle.com

ORCID

Mickaël Guedj http://orcid.org/0000-0002-9742-957X
Philippe Moingeon http://orcid.org/0000-0002-2380-9983

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Ringel MS, Scannell JW, Baedeker M, et al. Breaking Eroom's law. Nat Rev Drug Discov. 2020;19(12):833–834.
2. Waring MJ, Arrowsmith J, Leach AR, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat Rev Drug Discov. 2015;14(7):475–486.
3. Paul SM, Mytelka DS, Dunwiddie CT, et al., How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov. 2010;9(3):203–214.

•• **A key paper describing the challenges of pharma R&D.**

4. Morgan P, Brown DG, Lennard S, et al. Impact of a five-dimensional framework on R&D productivity at AstraZeneca. Nat Rev Drug Discov. 2018;17(3):167–181.
5. Pammolli F, Righetto L, Abrignani S, et al. The endless frontier? The recent increase of R&D productivity in pharmaceuticals. J Transl Med. 2020;18(1):162.
6. Moingeon P, Kuenemann M, Guedj M. Artificial intelligence-enhanced drug design and development: toward a computational precision medicine. Drug Discov Today. 2022;27(1):215–222.

•• **An introduction to the concept Computational Precision Medicine.**

7. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. Nat Med. 2022;28(1):31–38.

•• **A key introduction to AI in Medicine.**

8. Chen H, Engkvist O, Wang Y, et al. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23(6):1241–1250.
9. Ekins S, Puhl AC, Zorn KM, et al. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater. 2019;18(5):435–441.
10. Greener JG, Kandathil SM, Moffat L, et al. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23(1):40–55.

•• **A didactic introduction to machine learning applied to biomedical data.**

11. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
12. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–477.
13. Apweiler R, Beissbarth T, Berthold MR, et al., Whither systems medicine? Exp Mol Med. 2018;50(3):e453–e453.

• **An introduction to the concept of Systems Medicine.**

14. Yu MK, Ma J, Fisher J, et al. Visible machine learning for biomedicine. Cell. 2018;173(7):1562–1565.
15. Savage N. Tapping into the drug discovery potential of AI. Biopharm Deal. 2021:d43747-021-00045–7
16. Koscielny G, An P, Carvalho-Silva D, et al. Open targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 2017;45(D1):D985–D994.
17. Carvalho-Silva D, Pierleoni A, Pignatelli M, et al. Open targets platform: new developments and updates two years on. Nucleic Acids Res. 2019;47(D1):D1056–D1065.
18. Wishart DS, Knox C, Guo AC, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008;36(suppl_1):D901–D906.
19. Apweiler R. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2004;32(90001):115D–119 .
20. The UniProt Consortium, Bateman A, Martin M-J, Orchard S, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480–D489.
21. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–D1082.
22. Liu T, Lin Y, Wen X, et al. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 2007;35(Database):D198–D201.
23. Lamb J. The connectivity map: a new tool for biomedical research. Nat Rev Cancer. 2007;7(1):54–60.
24. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171(6):1437–1452.e17.
25. Pinero J, Queralt-Rosinach N, Bravo A, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database. 2015;2015:bav028–bav028.
26. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2019;gkz1021. 10.1093/nar/gkz1021
27. Hubbard T. The Ensembl genome database project. Nucleic Acids Res. 2002;30(1):38–41.
28. Howe KL, Achuthan P, Allen J, et al. Ensembl 2021. Nucleic Acids Res. 2021;49(D1):D884–D891.
29. Malone J, Holloway E, Adamusiak T, et al. Modeling sample variables with an experimental factor ontology. Bioinformatics. 2010;26(8):1112–1118.
30. Ontology Consortium G, Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32(90001):258D–261 .
31. The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47(D1):D330–D338.
32. Lonsdale J, Thomas J, Salvatore M, et al. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45(6):580–585.
33. The GTEx Consortium, Ardlie KG, Deluca DS, Segrè AV, et al. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648–660.
34. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42(D1):D1001–D1006.
35. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896–D901.
36. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Saf. 1999;20(2):109–117.
37. Liberzon A, Subramanian A, Pinchback R, et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–1740.
38. Maglott D. Entrez gene: gene-centered information at NCBI. Nucleic Acids Res. 2004;33:D54–D58.
39. Maglott D, Ostell J, Pruitt KD, et al. Entrez gene: gene-centered information at NCBI. Nucleic Acids Res. 2011;39(Database):D52–D57.
40. Hewett M. PharmGKB: the pharmacogenetics knowledge base. Nucleic Acids Res. 2002;30(1):163–165.
41. Cheng F, Desai RJ, Handy DE, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun. 2018;9(1):2691.
42. Joshi-Tope G. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2004;33:D428–D432.
43. Fabregat A, Jupe S, Matthews L, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–D655.
44. Kuhn M, Letunic I, Jensen LJ, et al. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016;44(D1):D1075–D1079.
45. Snel B. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. 2000;28(18):3442–3444.
46. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49(D1):D605–D612.
47. Chen X. TTD: therapeutic target database. Nucleic Acids Res. 2002;30(1):412–415.
48. Degtyarenko K, de Matos P, Ennis M, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2007;36(Database):D344–D350.
49. Ross JS, Mulvey GK, Hines EM, et al. Trial publication after registration in clinicalTrials.gov: a cross-sectional analysis. PLoS Med. 2009;6(9):e1000144.
50. Zarin DA, Tse T, Williams RJ, et al. Trial reporting in ClinicalTrials.gov — the final rule. N Engl J Med. 2016;375(20):1998–2004.
51. Wang Y, Zhang S, Li F, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Res. 2019;gkz981. 10.1093/nar/gkz981.
52. Menche J, Sharma A, Kitsak M, et al. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347(6224):1257601.
53. Adams D, Altucci L, Antonarakis SE, et al. BLUEPRINT to decode the epigenetic signature written in blood. Nat Biotechnol. 2012;30(3):224–226.
54. Tsherniak A, Vazquez F, Montgomery PG, et al. Defining a cancer dependency map. Cell. 2017;170(3):564–576.e16.
55. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLOS Med. 2015;12(3):e1001779.
56. Bycroft C, Freeman C, Petkova D, et al. The UK biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–209.
57. Edgar R. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–210.

58. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68.
•• **A key introduction to networks in Medicine.**
59. Lee LY-H, Loscalzo J. Network medicine in pathobiology. Am J Pathol. 2019;189(7):1311–1326.
•• **A key introduction to networks in Medicine.**
60. Gaudelet T, Day B, Jamasb AR, et al. Utilizing graph machine learning within drug discovery and development. Brief Bioinform. 2021;22(6):bbab159.
•• **A comprehensive review of the application of graph machine-learning in drug discovery and development.**
61. Cowen L, Ideker T, Raphael BJ, et al. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet. 2017;18(9):551–562.
62. Sun M, Zhao S, Gilvary C, et al. Graph convolutional networks for computational drug development and discovery. Brief Bioinform. 2021;21(3):919–935.
63. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet. 2006;7(2):119–129.
64. Vergetis V, Liaropoulos G, and Georganaki M, et al. A machine learning approach for assessing drug development risk. bioRxiv. 2020.
65. Hughes J, Rees S, Kalindjian S, et al. Principles of early drug discovery: principles of early drug discovery. BrJ Pharmacol. 2011;162(6):1239–1249.
66. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004;3(8):673–683.
• **A key introduction to the concept of drug repurposing.**
67. Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015;47(8):856–860.
68. Pushpakom S, Iorio F, Eyers PA, et al., Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019;18(1):41–58.
• **A review of systematic drug repurposing approaches.**
69. Fiscon G, Conte F, Farina L, et al. SAveRUNNER: a network-based algorithm for drug repurposing and its application to COVID-19. PLOS Comput Biol. 2021;17(2):e1008686.
70. Zeng X, Zhu S, Liu X, et al. deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics. 2019;35(24):5191–5198.
71. Steinberg D, Horwitz G, Zohar D. Building a business model in digital medicine. Nat Biotechnol. 2015;33(9):910–920.
72. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–2504.
73. Soret P, Le Dantec C, Desvaux E, et al. A new molecular classification to drive precision treatment strategies in primary Sjögren's syndrome. Nat Commun. 2021;12(1):3523.
74. Tzotzos SJ, Fischer B, Fischer H, et al. Incidence of ARDS and outcomes in hospitalized patients with COVID-19: a global literature survey. Crit Care. 2020;24(1):516.
75. Desvaux E, Hamon A, and Hubert S, et al. Network-based repurposing identifies anti-alarmins as drug candidates to control severe lung inflammation in COVID-19. PLOS ONE. 2021;16(7):e0254374.
• **Application of the *Patrimony* platform to COVID-19.**
76. Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. Nat Commun. 2019;10(1):1197.
77. Desvaux E, Aussy A, and Hubert S, et al. Model-based computational precision medicine to develop combination therapies for autoimmune diseases. Expert Rev Clin Immunol. 2021;18(1):47–56.
78. Rolland T, Taşan M, Charloteaux B, et al. A proteome-scale map of the human interactome network. Cell. 2014;159(5):1212–1226.
79. Jayatunga MKP, Xie W, Ruder L, et al. AI in small-molecule drug discovery: a coming wave? Nat Rev Drug Discov. 2022;21(3):175–176.
80. Aittokallio T. What are the current challenges for machine learning in drug discovery and repurposing? Expert Opin Drug Discov. 2022;17(5):423–425.
•• **Introducing some current challenges for AI/MAL in drug discovery.**
81. Needham CJ, Bradford JR, Bulpitt AJ, et al. Inference in Bayesian networks. Nat Biotechnol. 2006;24(1):51–53.
82. Luo Y, Peng J, and Ma J. When causal inference meets deep learning. Nat Mach Intell. 2020;2(8):426–427.
83. Heath JR, Ribas A, Mischel PS. Single-cell analysis tools for drug discovery and development. Nat Rev Drug Discov. 2016;15(3):204–216.
84. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–589.
85. Liang Y, Kelemen A. Dynamic modeling and network approaches for omics time course data: overview of computational approaches and applications. Brief Bioinform. 2018;19(5):1051–1068.
86. Selevsek N, Caiment F, Nudischer R, et al. Network integration and modelling of dynamic drug responses at multi-omics levels. Commun Biol. 2020;3(1):573.