Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Rewiring Drug Research and Development through Human Data Driven Discovery (HD³).

David B. Jackson [1], Rebecca Racz[2], Sarah Kim[3], Stephan Brock[1], Keith Burkhart[2].

[1] Molecular Health GmbH, Heidelberg, Germany; david.jackson@partner.molecularhealth.com
[2] Division of Applied Regulatory Science, Office of Clinical Pharmacology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA
[3] Center for Pharmacometrics and Systems Pharmacology, Department of Pharmaceutics, University of Florida, Orlando, FL, USA
* Correspondence: david.jackson@partner.molecularhealth.com; Tel.: +4917619181713

Abstract: In an era of unparalleled technical advancement, the pharmaceutical industry is struggling to transform data into increased research and development efficiency, and as a corollary, new drugs for patients. Here, we briefly review some of the commonly discussed issues around this counterintuitive innovation crisis. Looking at both industry and science related factors, we posit that traditional preclinical research is front-loading the development pipeline with data and drug candidates that are unlikely to succeed in patients. Applying a first principles analysis, we highlight the critical culprits and provide suggestions as to how these issues can be rectified through pursuit of a Human Data-driven Discovery (HD³) paradigm. Consistent with other examples of disruptive innovation, we propose that new levels of success are not dependent on new inventions, but rather the strategic integration of existing data and technology assets. In support of these suggestions, we highlight the power of HD³, through recently published proof-of-concept applications in the areas of drug safety analysis and prediction, drug repositioning, rational design of combination therapies and the global response to the Covid19 pandemic. We conclude that innovators must play a key role in expediting the path to a largely human focused, systems-based approach to drug discovery and research.

Keywords: Systems Pharmacology, Polypharmacology, Adverse Events, Drug Discovery, Functional genomics, Disease Modeling, Network analysis, Innovation

1. Introduction

The past three decades have provided some of the greatest technological advancements in human history. Never before have we had such a wealth of data and knowledge about living systems. The sequencing of the human genome, our ability to routinely probe the diverse molecular composition of living cells, the practical realization of Moore's Law and the rapid evolution of artificial intelligence, are but a few of the many advances that helped spawn a new era of hope for the betterment of human health. When viewed against the background of an ever-aging society and debilitating global pandemics of infectious agents and chronic disease, such hope is more than warranted. However, an unbiased appraisal of modern drug discovery suggests that commensurate impact on drug development success is still wanting [1]. Indeed, empirical evidence demonstrates that drug development success rates have decreased over time, while attrition rates, development times, and required investment have all increased considerably [2]. On the surface, it looks as though most companies are finding it difficult to transform these technological advancements into improvements in research and development (R&D) efficiency and as a corollary, new and effective medicines. As this trend continues, critical questions for the industry and patients remain, why is this the case and how might we rectify it?

© 2023 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

2. Eroom's law and the innovation crisis.

Several authors have sought to dissect the roots of this apparent "innovation crisis", quantifying both the costs of developing new drugs and the diverse factors impinging on success rates; two critical elements in the overall efficiency of the R&D process. While there is limited transparency regarding the cost of innovation in the pharmaceutical industry, a variety of studies have been reported. Using only publicly available data, Wouters *et al*., analyzed the costs associated with the drugs approved by the US Food and Drug Administration between 2009-2018 [3]. Out of 355 potential candidates, financial details were available for a total of 63 drugs (18%), coming from 47 mostly small to mid-sized companies. Median capitalized R&D investment was estimated at $985.3 million and a mean investment of $1335.9 million was estimated as required to bring a new drug to market. Median estimates also varied according to therapeutic area, ranging from $765.9 million for neurology drugs, to $2771.6 million for oncological and immunomodulating agents.

Complementary studies by DiMasi *et al*. used confidential data for 106 products developed by 10 large pharma companies to propose a base case mean cost estimate of around $2800 million [4]. These higher estimates potentially reflect a combination of higher clinical costs incurred by larger developers and lower estimates of trial success for each stage of development. Despite their differences, such studies emphasize the sheer scale of investment required to bring a new drug to market. In fact, no other industry invests so much in product development as the pharmaceutical industry. This is perhaps not surprising given the trajectory of market growth, with global prescription drug sales increasing from $649 to $768 billion between 2008 and 2016, with estimates prior to the COVID-19 pandemic predicting $1060 billion in sales by 2022 [5].

Given such enormous financial commitment, and opportunity, one might naturally expect parallel improvements in the overall rate of innovation and R&D efficiency – as measured by the number of newly approved drugs divided by the total overall costs of development. Unfortunately, however, empirical evidence again suggests that we have witnessed a declining rate of overall R&D efficiency over the past number of decades. Scannell *et al*. reported that the all-in cost of R&D on new drugs approved by the US FDA has risen exponentially for 60 years [6], leading to an associated exponential decrease in the overall rate of R&D efficiency. More specifically, the authors showed that the number of new FDA approved drugs per billion US dollars of R&D spend was halved approximately every 9 years between 1950-2010. They characterized this phenomenon as 'Eroom's Law' i.e. the inverse of Moore's Law. The potential causes for Eroom's Law have been widely discussed in the literature, but before we examine the details, it is instructive to first review the high-level details of the current drug R&D paradigm.

2.1. The Traditional Drug Discovery Paradigm

The drug discovery process typically begins with the search for target molecules that are either directly or indirectly associated with a disease of interest [7]. Once identified, this target must be validated by demonstrating a potential therapeutic effect through modulation [8]. Then, an assay is developed and screening is performed, in search of one or more lead compounds that can modulate its function in a therapeutically desirable manner, often with the additional help of structure-based drug design and virtual screening [9]. These lead compounds undergo preclinical evaluation in *in vitro* and/or *in vivo* model systems, to characterize potential pharmacological and toxicological behaviors. After further lead optimization through medicinal or computational chemistry [10], a single compound is usually selected, patent applications are filed and an Investigational New Drug Application (IND) is submitted to regulators. If successful, the sponsor can then proceed to the first of three phases of clinical testing in humans.

Development begins in phase Ia, where a single dose is administered to a group of healthy volunteers, followed by phase Ib, where increasing doses are used to assess safety, pharmacokinetics, and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

pharmacodynamics parameters. In an optimal case, additional studies are performed at this stage to decipher and/or confirm the precise mechanism of action. If no critical safety issues are identified the compound may move to phase II, where it is tested directly in several hundred patients in an attempt to glean further insights into its clinical efficacy, adverse event profile and optimal dosing regimen. What ultimately emerges is a clearer picture of the risk-benefit profile. In the final phase III study, the effectiveness of the compound is examined, normally in a randomized control trial (RCT) setting with thousands of patients, where its effectiveness is compared to either a placebo or a current standard of care. Once successfully completed, the sponsor submits a New Drug Application (NDA) to the regulatory authorities, and if successful the drug can be launched on the market. Thereafter, further Phase IV clinical trials can be used to demonstrate real-world safety and effectiveness of the drug or to compare it to other treatment modalities, in so-called pragmatic clinical trials.

The current drug R&D paradigm can therefore be viewed as a set of contiguous steps, where the financial outlays are compounded at each stage, leading many to adopt the "fail early and fail fast" approach. One can also view the process as a hypothesis filtration funnel, where new innovations have effectively increased the width of the funnel by, for example, increasing the efficiency of target identification e.g. through application of functional genomics methodologies, such as RNAi and the CRISPR-cas9 system [11]. Through widening of the funnel, the cumulative effect of all these advances should in theory lead to the negation of Eroom's Law, but compelling evidence of this is still wanting. The issues involved can be divided into two general categories of challenge a) Industry related challenges, pertaining to the nature and regulation of the pharmaceutical industry, and b) scientific challenges, pertaining to prevailing discovery modalities and accepted research norms. In the following section we review some of these issues, and then hone-in on what we believe to be the most fundamental, first-principle causes.

2.2 Industry level challenges

Industry-related challenges are often discussed, though in our opinion are probably not the primary drivers of the innovation crisis. In this section we introduce three of the most commonly quoted factors.

2.2.1. Regulatory oversight

Post-marketing safety issues surrounding drugs like Vioxx [12] and Avandia [13] have encouraged greater regulatory stringency in the pre- and post-approval assessment of drugs. In response to such problems, the FDA issued the Amendments Act which enabled the agency to stipulate risk evaluations prior to approval and additional clinical studies when post-marketing safety problems emerge [14]. While regulatory stringency is sometimes quoted as a challenge, it is clear that the remit of agencies is to ensure an appropriate risk:benefit ratio for each treatment and target population. As such, the process can be viewed as an essential element of the drug development process and is unlikely to play a significant role in perpetuating the innovation crisis. On the contrary, high safety standards can only be good for patients. There is also evidence that higher standards raise incentives to develop more innovative drugs and are generally associated with higher international market success [15].

2.2.2 The drug "innovation chasm"

Another issue that has certainly impacted the overall return on investment is the number of "me-too" and "follow-on" drugs. Developing a drug for an indication where an effective therapy already exists means that companies must demonstrate that the new drug has credible advantages over the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

approved standard. This has been referred to as the "Better than the Beatles" problem [6], but one can also think of it as an ever-expanding "innovation chasm", where with every newly approved drug in a specific indication, it becomes more and more difficult for newer drugs to cross the chasm to broader market acceptance. To do so often requires larger investment, for example, due to larger and/or longer clinical trials. The Tufts Center for the Study of Drug Development (CSDD), for example looked at 377 drugs and biologics approved by the FDA between 2008-2018 and found that while an average of 83.1 clinical trial months was required between 2008-2013, this number rose to 89.8 months in the following five years (2014-2018) [16]. With increasing time, comes greater financial outlay, thus exacerbating the R&D efficiency challenge.

2.2.3 Mergers and Acquisitions

High R&D costs, expiring patents and impending generic competition, have also forced larger companies to pursue a mergers and acquisitions strategy. Many empirical studies suggest that M&A's can have negative effects on the innovative freedoms of the companies involved [17]. Financial pressures can force closure of entire R&D projects, or even entire research sites, especially those of the acquired party. This may be particularly damaging if companies use M&A to escape innovation competition by buying others with related competing pipelines and terminating their development. Furthermore, important intangible resources may be irreparably damaged, through loss of 'knowledge capital' with the exit of key scientists. Thus, by restricting the R&D freedoms of smaller more innovative firms, there is evidence that M&A activity can compound the innovation crisis.

2.3 Science & technology challenges

Preclinical drug discovery is both dichotomous and reductionist in nature. The overall process was transformed in the 1990's, when the strategy of low-throughput *in vivo* physiological screening and medicinal chemistry optimization (i.e. phenotype-based drug discovery) [18], became largely supplanted by target-focused high-throughput screening (HTS) of large compound libraries (i.e. target-based drug discovery) [19]. Several radical innovations in molecular and computational technologies have recently transformed the scale and specificity of both approaches. Advances in the automation and parallelization of *in vitro* and *in vivo* models, improvements in imaging technologies and artificial intelligence (AI)-driven image analysis, have all contributed to the re-emergence of Phenotype-based screening as an effective discovery strategy. The target-based approach has also been revolutionized through advances in genomics, transcriptomics, proteomics, transgenic/humanized animal models and CRISPR/RNAi libraries [20]. Combinatorial chemistry, which led to a proliferation in the size and composition of chemical libraries for target-focused HTS, has also been complemented by the advent of structure-based drug design and related approaches such as fragment-based drug design [21]. Taken together, these advances should have revolutionized the return-of-investment in drug discovery research. However, while transforming it quantitatively speaking, important qualitative issues prevail, many of which are likely debilitating.

2.3.1 Target-based discovery

While the single target focus holds many practical efficiencies, it is clearly of greater relevance to the 'one gene-one disease' theory, including both inherited diseases (e.g. the delta508Phe deletion in CFTR associated with cystic fibrosis) [22] and certain cancer driver mutations (e.g. BCR-ABL in Chronic Myeloid Leukemia (CML)) [23]. Today, less than 700 human proteins have been established

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

as targets of approved drugs, most of which belong to either the kinase, GPCR or ion channel protein families [24]. Given that this represents less than 1% of the human proteome, the "low-hanging fruit" hypothesis has been proposed, suggesting that most of the obvious discoveries and easy to validate targets have already been found, leaving only more complex biomedical challenges for the industry to solve. This might certainly be true for complex neurological disorders like Alzheimers or metabolic syndrome. One of the most fundamental problems with the single target strategy is that we know that most diseases are multifactorial in nature and that single targets can play different functional roles across different cellular systems (see figure 1). Targets exist in highly structured and tightly integrated cellular information transduction systems that extend from nucleic acids to proteins to diverse biochemical molecules (e.g. lipids, carbohydrates, metabolites). Moreover, complex disease etiology likely arises because both protein function and expression are controlled by the quaternary structure and/or network context in which the protein exists.

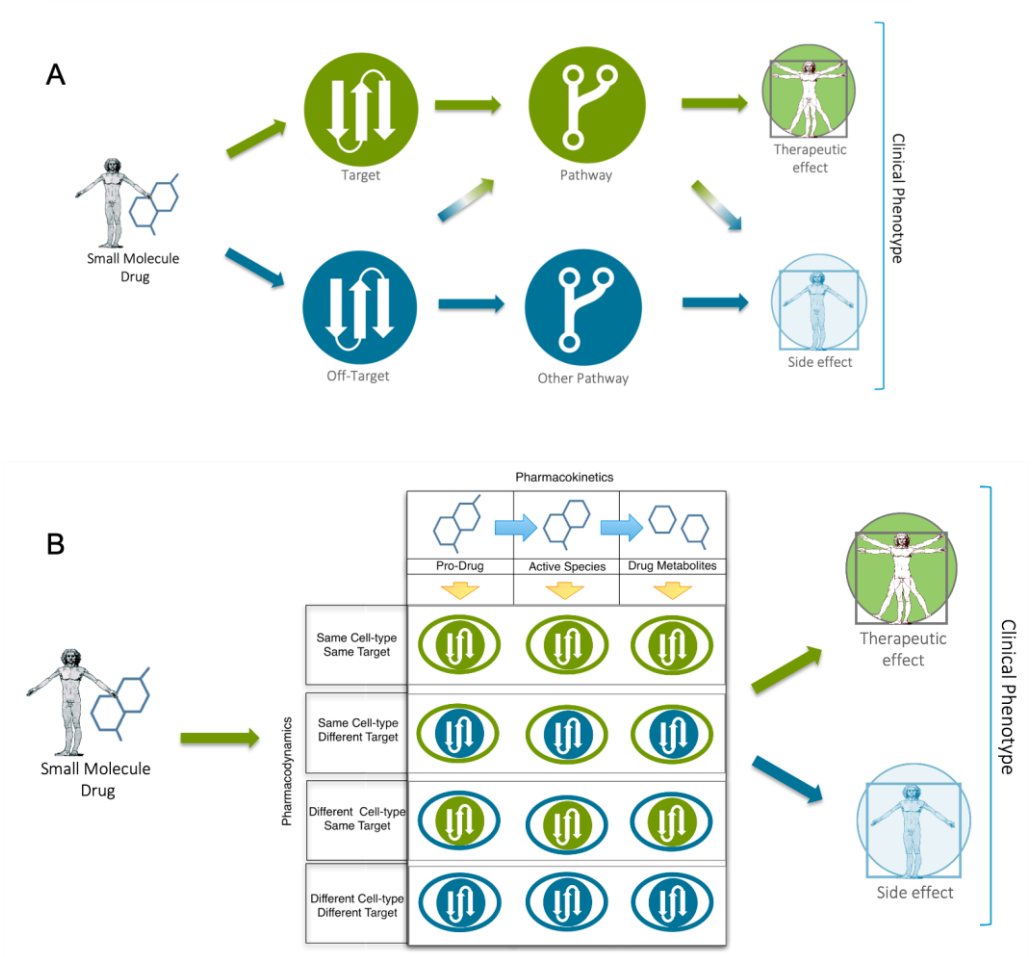

Figure 1. A. Schematic overview of drug MOA according to the Target Based Discovery paradigm. B. Schematic overview of drug MOA according to the HD³ paradigm. Here, single drugs and their metabolites can interact with targets proteins in the target cell/tissue, or in other tissues. In addition the drug can interact with different targets, in the target tissue and/or in other tissues. This complexity highlights the need for cell-type specific drug MOA models.

2.3.2 Drug Promiscuity

We also know that small molecule drugs, such as protein kinase inhibitors, tend to act at the level of several targets and that such molecular promiscuity (also termed polypharmacology) is critical to their clinical effects. To estimate the scale of this phenomenon, Mestres *et al*. analysed various drug:target databases to estimate the average number of targets per drug, identifying a total of 1.8 for the DrugBank database of curated therapeutic targets, and 2.7 for the WOMBAT database of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

drug:target affinities. When both validated targets and affinity-based targets were considered, the number increased to an average of 6.3 targets per drug. [25] Thus, in the absence of broader knowledge about the systemic interactions of targets and small molecules, it is unsurprising that many promising paths of therapeutic development, eventually fall victim to the poorly predictable intricacies of human molecular diversity. Moreover, such knowledge is key to understanding the entire clinical utility and risk profile of all small molecule drugs.

2.3.3 The reproducibility crisis

Another critical issue relates to the robustness and reproducibility of preclinical results. It is often assumed that the peer-review process ensures that only scientifically valid results get published and that they are described in enough detail to allow reproduction in other laboratories. Drug discovery is dependent on accurate and relevant results, as these form the basis of prudent decision-making along the discovery & development value-chain. However, just as our ability to generate massive tomes of new data has increased exponentially, so too have concerns about the reliability of most preclinical findings. Empirical studies have estimated that 75-90% of preclinical results from high-profile journals are not reproducible in independent experiments [26-28] These findings are not specific to *in vitro* studies. For example, Hackham *et al*. reported [29] found that only about a third of highly cited animal model research actually translated in randomized human trials. In some specific therapeutic areas, such as stroke, no evidence of translatability to patients has been found, despite massive investment [30].

2.3.4 The problem with model systems

Congruence between *in vitro* and animal models and the corresponding human disease is still accepted as a fundamental tenet of biomedical research. Yet this is an assumption that receives relatively little objective analysis at the research strategy level. Much pre-clinical pharmaceutical research is still today strongly dependent on target-focused *in vitro* assays and animal models of human disease. While this provides us with a wealth of new data and peer-reviewed insights, the question remains as to whether this new information is in any way representative of the human disease state? Patient cells do not normally grow on plastic, there is no circulation *in vitro*, nutrient/oxygen supplies are different and the cellular microenvironments are completely artificial (e.g. presence of antibiotics, different endocrine concentrations). In fact the list of differences between cell culture and *in vivo* systems is so extensive, that it warrants broader objective analysis and critique. Animal models too have clearly significant problems. For example, it is commonly observed that the disruption of a gene in one genetic background of mouse might cause severe phenotypes such as lethality, whilst in another strain be phenotypically innocuous. If such discrepancies can occur at the level of a single model organism, how can they reliably translate to the development of drugs for human disease?

To better understand this issue, it is instructive to consider the distinction between the "structural biochemistry" of the cell and its "functional biochemistry" [31]. Structurally speaking, we can reasonably assume that if a particular biochemical reaction is possible *in vitro*, then it is also likely to be possible *in vivo*. By examining such biochemical events *in vitro*, it is possible to construct a network of connected biochemical events *in vivo*. However, an *in vitro* system can tell us little about intricacies of the functional biochemistry within different cellular systems. Given that it is the functional biochemistry that is perturbed in most human diseases, all model systems hold critical weaknesses in the study of human disease. While such fundamental problems appear to be largely ignored due to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

underlying practicalities, they are likely having negative impacts on the overall return of investment. This begs the important question; are we relying on false principles, when we should instead be turning to the first principles of human biology?

3 The "First Principles" case for a Human Data Driven Discovery (HD^3) paradigm

The aforementioned issues, though incomplete, provide a good indication of the scale and diversity of the innovation challenge. However, in our opinion, it is the scientific issues that deserve immediate analysis and introspection. To do this, we can borrow from one of the numerous innovation frameworks that have successfully been applied in to solve complex social, engineering and technology challenges. These include design thinking, computational thinking, analytical thinking, lateral thinking and first principles thinking. Arguably, the most successful is First Principles Thinking (FPT), an approach first practiced by Aristotle who defined it as the search for "the fundamental basis from which a thing is known" [32]. In practice, innovators are encouraged to critically question every assumption of the challenge and break it down into basic components and to search for solutions from "first principles". From this, new learnings and opportunities are defined. This approach is distinct from reasoning by analogy (e.g. using competitive analysis), which builds knowledge and solves problems based on prior assumptions, beliefs and widely held 'best practices' approved by the majority of people.

Applying FPT to the drug discovery process, we have seen that much of preclinical research is built upon the inherent practicalities of model systems, as opposed to the functional realities of human biology. At the data level, such an approach has quantitative advantages, but not qualitative. Therefore, failure to expeditiously balance investment towards human-specific approaches, risks us applying new technologies to generate even more incongruent data. As a corollary, we can posit that current standards of preclinical research are frontloading the drug discovery process with data that lacks human specificity and technical reproducibility for routinely developing successful new drugs. This is ultimately what may be responsible for long development times and high attrition rates. Thus, pivoting discovery to a largely human-focused approach using human-specific data and *in silico* disease models (i.e. Human Data Driven Discovery: HD^3), has the potential to provide radical improvements in drug research efficiency.

FPT analysis also reminds us that while the target-based discovery paradigm provides a useful simplification, it is only rarely consistent with biological reality. When administered to a patient, drugs interact with a plethora of molecular entities across different cellular systems; a profile we refer to as the drugs "biotype". Understanding the relationship between a drugs biotype, it's chemical features (here termed "chemotype") and it's resulting clinical effects (i.e. phenotype), is essential to understanding the molecular basis of drug mode-of-action. FPT helps us to reduce the challenge to an even more simplistic fact. Fundamentally, it is the three-dimensional arrangement of atoms in a small molecule drug (i.e. the chemotype) that interact with complimentary binding pockets in targets throughout several cellular systems (i.e. the drug biotype) to induce network level changes that result in emergent clinical effects (i.e. the phenotype) (fig 4 and 5). Instead of thinking of the lock-and-key analogy for drug and target, the system-based perspective is more akin to that of a combination lock, or a "dial on a safe", where only a certain chemotypes possess 'the code' (i.e. biotype) to induce specific phenotypic effects. This simple perspective also emphasizes the key weakness of model systems. Structural and network-based differences in target paralogues within the model, mean they are unlikely to induce the same phenotypic response to a chemotype as observed in human – the model system biotype is simply too different. With this simple perspective in mind, we now examine some of the factors that can help us transition to a more "humanized" drug discovery paradigm, with the power and relevance of human data at the core of our focus.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

2.3 Human data as a driver for systems-based discovery

Today, the molecular characterization of human populations plays a critical role in target-driven discovery. In oncology, for example, the application of omics technologies has helped characterize the key molecular differences between treatment responders and non-responders, thereby enabling the development of efficacious therapies targeting specific driver mutations, such as imatinib [33] and crizotinib [34]. However, one of the primary advantages of model systems is that they also allow us to prospectively perturb the function of genes in the study of disease specific targets and to perform *in vivo* physiological screening with small molecule chemistries. While it is generally understood that such studies are unethical in patients, the reality is that doctors have for decades been performing physiological phenotypic studies on their patients with prescribed drugs. Once administered, the drug and/or its metabolites interact with targets in different cellular systems. The phenotypic readout is a consequence of the combined molecular interactions across the entire patient system (see fig 1B). From this perspective, we can view a side effect as drug-induced disease phenotype, whose molecular etiology tells us something not only about the drugs mode-of-action, but also about the targets/pathways associated with the observed phenotype. Thus, if we can accurately define a drugs interaction partners/biotype (targets, off-targets, metabolizing enzymes, transporters) across all cellular systems, we can learn more about the human-specific molecular networks involved in human phenotypes. The advantage of this approach, in comparison to model systems, is that the clinical observations and connected molecular knowledge are directly defined by the human condition.

An important facilitating factor in the pivot to a human-focused discovery paradigm is therefore the utility of treatment and outcome information from vast tomes of existing real-world data (RWD) and clinical trial results. RWD exists as a spectrum of different qualities, typically defined by both the context of assessment from which the data was derived, and the degree to which the data was generated to answer specific research questions (see figure 2). Data pertaining to treatments and clinical outcomes, whether positive (e.g. drug-induced disease remission) or negative (e.g. disease recurrence or adverse reactions) are of primary importance. Such data is widely available in research databases, large-scale clinical registries, EMR-linked sources, Administration/claims sources, facilitated networks and regulatory databases. Spontaneous Reporting System databases are particularly interesting and although redacted in terms of clinical narratives, they offer a highly valuable window into observed drug-induced adverse event (AE) phenotypes for millions of patients. Major sources include the FDA's Adverse Event Reporting System (FAERS) [35] and Sentinel Initiative [36, 37], together with the European Medicines Agency EudraVigilance system [38] and the global database of individual case safety reports (ICSRs) called VigiBase, maintained by the World Health Organization's (WHO) Uppsala Monitoring Center (UMC). [39] The data contained in these databases are analogous to chemical phenotype screening data from model systems, only this time specific to humans. By providing insights into the phenotypic effects of drug-induced perturbation on targets within the patient system, they allow us to capitalize on publicly available treatment and outcome data for tens of millions of patients.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

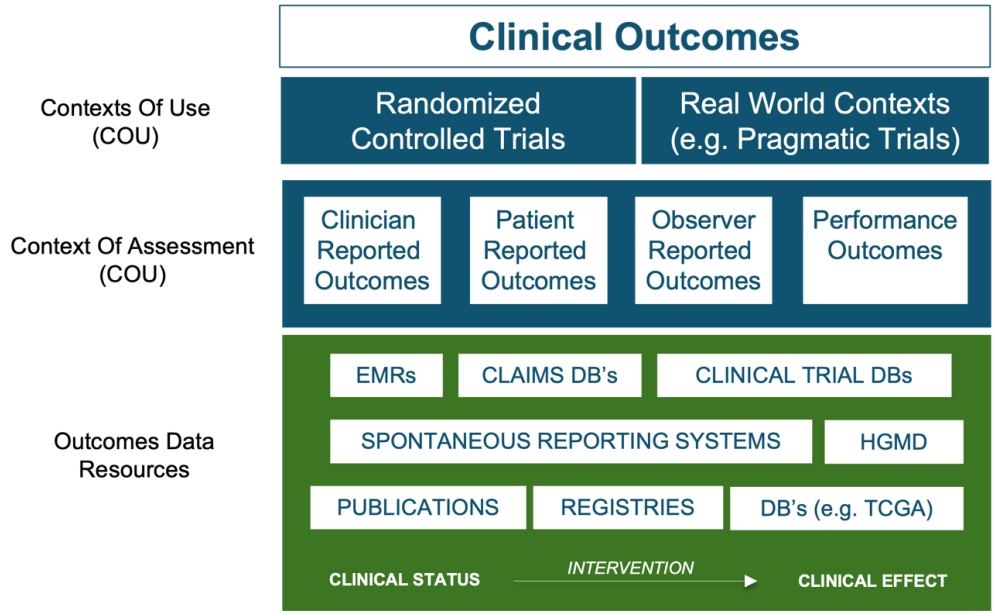

Figure 2. Overview of key sources of clinical outcomes data. Beyond Randomized Controlled Trial (RCT's) results, there are four types of outcome reports, otherwise known as "contexts of assessment" associated with RWD: 1) Clinician reported outcomes (e.g. from Electronic Medical Records (EMR's)), 2) Patient reported outcomes, 3) Observer reported outcomes and 4) Performance outcomes. While clinician reported outcomes are the most reliable source of outcomes data types, there is an ever-growing realization of the value of patient reported outcomes. Nevertheless, for the purposes of the HD[3] approach it is treatment outcomes data that provides the most valuable datapoints, with Spontaneous Reporting Systems such as FAERS and Vigibase, providing a highly accessible source of this information for tens of millions of patients. We can also extend the concept of "outcome" to include the phenotypic consequences of genetic aberrations, with such data being available in disease agnostic databases such as the Human Gene Mutation Database (HGMD) and OMIM, or disease specific databases, such as The Cancer Genome Atlas project (TCGA).

Treatment outcomes data can also be integrated with data emerging from the multi-omics characterization of human populations. A vast wealth of data resources and public genomics data initiatives have become available over the years, with general, organ-specific and disease-specific data now globally accessible (for a list of 86 different globally available resources, see supplementary Table 1). Here, disease-associated genetic data from resources such as Online Mendelian Inheritance in Man (OMIM) database [40] or the phenotype associated NHGRI-EBI Genome Wide Association Study (GWAS) catalog [41] are of particular interest. Updated daily, the OMIM database catalogues information for more than 15000 genes with a core emphasis on the molecular relationship between genetic and phenotypic diversity, especially around human disorders. The data also lends itself to organ specific analyses (and modelling), with Parsa *et al*. for example compiling a list of the 258 OMIM genes responsible for kidney related diseases, including renal hypoplasia, dysplasia or agenesis, end-stage renal disease and proteinuria [42]. By aligning this data with GWAS data from the CKDGen Consortium, they were further able to characterize the potential association of genetic polymorphisms and kidney function within the general population [43]. Although not reported by the authors, such data can also be further contextualized with kidney specific disease pathway information, such as the Kidney and Urinary Pathway Knowledge Base (KUPKB) (http://www.kupkb.org/) [44] or the Chronic Kidney Disease database (CKDdb) (http://www.padb.org/ckddb) [45], or with data from drugs used to treat kidney diseases and outcomes data related to kidney specific side-effects from other drug treatments.

The power of such strategies was also demonstrated by results from the TCGA project, where molecular data and phenotypic information have been analyzed to decipher novel targets and prognostic classifiers. Analysis of the TCGA endometrial carcinoma dataset, for example, has brought important

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

new insights into the molecular nature of this disease, including the discovery of a new classification system based on four prognostically significant subgroups [46]. Indeed, it can be argued that most recent clinical advancements emerge from analysis of patient-derived molecular data. However, vast tomes of clinical information available throughout the healthcare system remain underutilized for discovery purposes.

Two other types of data resource are fundamental to systems-based discovery, a) network models of disease systems and signaling pathways, defined initially at the level of proteins (nodes) and their interactions (edges) and b) accurate and comprehensive drug-to-target knowledge. The network view provides a basic proteo-anatomical structure of the system, upon which additional molecular data sources can be added. The scale-free and redundant characteristics of these networks often permit perturbation without a complete loss of function, implying that multiple perturbations, at nodes and/or edges, are typically involved in the emergence of disease phenotypes. Recent computational work by Zhong *et al*., for example, examined the system-level mutational features of heritable disease and found that they were more likely to be caused by mutations at edges, as opposed to nodes. [47]. Interestingly, edge-based perturbations, typically involving in-frame mutations of (near) full-length protein, were more commonly observed across multiple diseases. Such mutations tend to abrogate interaction with one or more neighboring nodes. Moreover, different disease phenotypes may be caused by different mutations in a single edge. Nodal mutations on the other hand typically involved truncated proteins and did not necessarily affect the interaction with other proteins nodes in a signaling network. Drug-to-Target knowledge is also a key prerequisite in systems-based discovery endeavors: In addition to the aforementioned DrugBank [48] and WOMBAT [49] databases, the Therapeutic Targets Database (TTD) [50], Pubchem [51] and ChEBI [52], all provide critical knowledge for systems-based analyses of drug targets.

This brings us to the next level of challenge. How do we optimally structure this drug and network knowledge to facilitate drug discovery? Ontologies will certainly play a critical role in making data not just machine readable, but also machine actionable. A key element here will be ontology interoperability and robust ontology applications to help make data Findable, Accessible, Interoperable and Reusable (i.e., FAIR compliant)[53]. Beyond the ontological challenges associated with knowledge modelling, the biological accuracy of systems-based models is critical, especially if we are aiming for whole patient models. Today, most pathway models represent an amalgamation of biochemical findings across a multitude of different cellular systems, under various physiological conditions. This leads to generic model representations that are likely inaccurate at the cell-type specific level. We must therefore aim for cell-type and tissue-specific representations of core biological mechanisms, which can then be mapped at the whole-patient level. Several existing resources can aid this endeavor, including important work by Jiang *et al*. who reported a quantitative proteome map of the human body, with expression data for 12,000 genes across 32 normal human tissues [54]. Other examples include The Human Protein Atlas (HPA), which presents the spatial distribution of proteins in 44 different human tissues and 20 cancer types [55]. Organ specific sources are also widely available, such as The Brain Protein Atlas [56] and the Human Kidney and Urine Proteome Project (HKUPP) (http://www.hkupp.org/) [57]. The resultant networks facilitate the mapping of drugs to phenotypes across all levels of the patient system and provide a powerful basis for hypothesis generation and AI-driven discovery. They also allow us to add additional data-types (e.g. genomic, transcriptomic, drug-binding constants and target activity data) that may later facilitate more functional analyses using, for example, simulation algorithms. Such an approach would largely meet the requirements for developing an effective *in silico* model system, since the human disease/systems and the *in silico* models should be substantially congruent with respect to structure and composition.

Finally, from a technical perspective, outcomes data such as spontaneous AE reports are typically stored in relational databases (RDBs), in multiple tables with information pertaining to the case demography, drugs (medications) given, reported AEs etc. Such data structure allows easy integration of new reports, distribution and sharing of data and a relatively straightforward retrieval of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

specific/individual information. However, the need to join a large number of tables to combine information for each outcome, AE, medication and co-medications, indications and demography, can result in rather complex queries and long computational times. Moreover, thorough understanding of the underlying data structure is required to avoid potential pitfalls, such as the multi-axiality of data, leading to erroneous results that are not always directly apparent. These characteristics can make advanced analyses of data in RDB structure more cumbersome and unpractical. One of practical solutions is to convert the RDB into a graph database for the purpose of data analysis. Graph databases (GDBs) store information in the form of nodes (entities) and their properties as edges (relationships) instead of tables with rows and columns [58]. As each node is directly connected to all the other relevant nodes, the queries to retrieve more 'distant' information are much simpler and faster, as there is no need to join multiple tables. More importantly, the GDB structure enables use of efficient algorithms from simple random walks to graph convolutional networks to facilitate discovery of hidden relationships between entities and preparation/transformation of data for machine learning/AI approaches, such as feature engineering. Several reports show that use of GDBs in the analysis of safety signals or predicting safety of drugs show that such approaches have the potential to outperform the current approaches [59-63]. Future development of knowledge graphs integrating full outcomes and spontaneous AE report databases, together with other information will not only improve the performance of drug toxicity predictions but also help to uncover hitherto unknown interactions between drugs and co-morbidities. From this perspective, the approach may prove quite effective in target deconvolution and drug (re-)positioning studies.

4 Current applications of the HD³ approach

4.1 Application to the analysis and prediction of Adverse Events

Some of the most successful proof-of-concept applications of HD³ systems-based discovery have come from the analysis and prediction of adverse event data. This is likely driven by the public availability of AE data for millions of patients, with Vigibase alone containing over 30 million case reports. The data itself is de-identified and typically comprised of basic patient information (e.g. gender, weight and age), the drugs they were prescribed, the adverse events that were observed, and the associated clinical outcomes. Traditional analysis of this data by pharmacovigilance (PV) teams, uses disproportionality statistics such as proportional reporting ratio's (PRR), to quantify the relative congruence of binary drug to AE relationships. Bayesian correlation models and a variety of machine Learning algorithms have also successfully been applied, but all such approaches lack molecular insight. From a more conceptual standpoint, we can view AE data as a form of human phenotypic screening data, where doctors administer drugs that impinge on the activity of several proteins within the patient system, eliciting AEs. Such AEs can be considered drug-induced diseases, whose molecular mechanisms tells us something about the underlying disease etiology. Given that such resources cover millions of patient lives, they are highly valuable resources for HD³ based approaches, such as those described next.

Work by Soldatos *et al*. demonstrated the feasibility of transforming 8.2 million FAERS case reports into as many patient-specific systems pharmacology models [64]. Given that drug names are reported without a controlled dictionary, the authors first implemented mapping of reported drugs to unique chemical ID's, resulting in the unequivocal integration of about 95% of all FAERS reports. Drugbank was then used to link these drugs to their targets, with further levels of molecular integration achieved by mapping all target proteins to their respective pathways within the Reactome [65]; and PID (NCI-Nature & BioCarta) [66] databases. The network was further extended through integration of the ATC ontology [67] and text-mined literature pertaining to all binary relationships captured within the network. Using this process, patient case reports detailing drug-induced phenotypes were linked to knowledge from nine entity types: 2600 drugs were linked to 1800 targets, 201 metabolizing enzymes, 103 transporter proteins, >1000 pathways, and 881 ATC drug classes, all of which were connected to clinical data pertaining to 15400 indications, 19300 adverse reactions and seven clinical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

outcomes. As a result, circa 8 million systems-based patient models were generated that could be queried at the level of any entity and/or compared either directly or through user-defined patient cohorts. For example, instead of just comparing the AE's of drug A versus drug B, it is possible to query AE's from the perspective of target A versus Target B, or Pathway A versus Pathway B. Beyond the analysis of potential off-target effects, the authors also demonstrated the utility of the platform (MH Effect) in predicting adverse events for developmental drugs, the rational design of combination therapies and target-based drug interactions. Similar platforms have since emerged, including AbbVie's Off-Target Safety Assessment (OTSA) technology [68] and Clarivates OFF-X system [69], with OFF-X and MH Effect being utilized by regulatory scientists.

Building on this general approach, Kim *et al.* recently reported two interesting proof-of-concept studies. Firstly, they explored the underlying molecular mechanisms causing trastuzumab-induced cardiotoxicity and how the rate of toxicity may change due to drug-drug interactions at the molecular pathway and target levels [70]. Trastuzumab is a targeted therapy drug targeting HER2 which is overexpressed in 25% of all breast cancer patients. HER2 blockade can cause cardiotoxic side effects because the HER2 receptors are present in not only breast cancer cells but also normal cardiomyocytes. Among ~750 molecular mechanisms found by mapping the FAERS data to chemical and biological databanks, the mechanisms related to apoptosis regulator proteins of BCL-2 members were found to have a statistically significant association with trastuzumab-induced cardiotoxicity, which align with other reports [71, 72]. They further explored how the concomitant use of other drugs affects the possible molecular mechanisms related to mitochondria dysfunction in cardiomyocytes. Doxorubicin, which is often given with trastuzumab, also has a high risk of cardiotoxicity. The PRRs between cardiotoxicity and each of trastuzumab and doxorubicin were higher than 40. The combination cohort of these two drugs indicated a ~2.5-fold increase in the PRR. Other potential different pathways related to mitochondrial biogenesis and function (i.e., peroxisome proliferator-activated receptor-γ coactivator 1-α (PGC-1α or PPARα) and PPARβ) were found, which might induce a synergistic effect leading to the increased risk of developing cardiotoxicity. In addition to the combination of trastuzumab and doxorubicin, drug-drug interactions that mitigate trastuzumab-induced cardiotoxicity were also explored for concomitant use of other drugs, i.e., tamoxifen, paroxetine, and lapatinib, with trastuzumab. For the mechanisms decreasing the toxicity, the highlighted molecular mechanisms were the anti-apoptotic effect through calcineurin-dependent pathways, the antioxidant activity of glutathione, and the adenosine monophosphate-activated protein kinase activation.

Secondly, the approach was also applied to untangle the complexity of the underlying molecular pathways and targets of adverse events of immune checkpoint inhibitors [73]. Their study focused on how mechanistic differences between cytotoxic T-lymphocyte antigen-4 (CTLA-4) and programmed death-1 (PD-1) affect colitis by mapping the FAERS data to the associated molecular pathway and target levels. The PRR indicating the statistical association between a drug and colitis was ~3 times higher for ipilimumab (anti-CTLA-4 drug), compared to the PRRs for nivolumab and pembrolizumab (anti-PD-1 drugs). They hypothesized that the severer toxicity of the anti-CTLA drug would be because of its early responses in the sequence of boosting T-cell activation. While there were limited data to drive statistically significant evidence with respect to the mechanistic causality, their work elucidated how to apply the reverse translational systems-based approach to predict the drug safety profile.

4.1.1 Examples from the FDA's Division of Applied Regulatory Science.

Clinical Pharmacologists at the FDA's Division of Applied Regulatory Science (Center for Drug Evaluation and Research (CDER)) are often tasked with examining potential molecular mechanisms to support or negate emergent safety signals from Pharmacovigilance (PV) endeavors. In contrast to the traditional PV, these so-called "Biological plausibility consults" utilize a palette of bioinformatics and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

chemoinformatics applications to examine both target- and systems-based mechanisms. Target-focused assessment, has been termed "Target Adverse Event" (TAE) analysis to emphasize that real-world safety data is being integrated and analyzed at the level of the target/system, as opposed to the traditional drug-based view. The combined output of PV and TAE analyses are then used to determine whether enough mechanistic evidence exists to support a label change for marketed drugs.

In one example, researchers at CDER analyzed the biological plausibility of Montelukast inducing neuropsychiatric events such as hallucinations, suicidal thoughts, depression, and sleep disturbances. Target analyses for neuropsychiatric events found drugs that bind serotonin, dopamine, and acetylcholine muscarinic receptors or act on protein targets that increase or decrease neurotransmission for these neurotransmitters to be highly associated with neuropsychiatric events. Montelukast binds to serotonin receptor 5-HT2B, dopamine receptor 3, muscarinic receptors 1 and 3, and dopamine and norepinephrine reuptake transporters. Structurally, montelukast has a quinoline moiety that is found in other drugs that cause neuropsychiatric events (see tafenoquine label [74]). This analysis provided biological plausibility for the FDA's decision to add neuropsychiatric events to the montelukast label.

TAE and systems-based analyses have provided signal strengthening to support safety label changes for other drugs. In evaluating a signal of seizures and gadolinium contrast agents, infusion reactions or anaphylactoid reactions were also noted. Histamine release enhances cholinergic neurotransmission which can include seizures. Further analyses noted that other drugs associated with infusion reactions also had reports for seizures in FAERS and the literature. Monoclonal antibodies that bind to proprotein convertase subtilisin kexin type 9 (PCSK9) were associated with flu-like illness. Analyses found disproportionality for flu-like symptoms in this class of drugs and other monoclonal antibodies that bind cytokines but not the statin drugs that are taken concomitantly. Finally, as pimavanserin is a highly specific 5-HT2a receptor antagonist, this target and comparator drugs supported the addition of 'falls' to the label.

Multiple case reports also suggested a relationship between second generation antipsychotics, such as risperidone and olanzapine, and serotonin syndrome. Therefore, FAERS was utilized in combination with molecular target information within the MH Effect system to develop a hypothesis of the mechanism for serotonin syndrome. Based on this combined data set, 5-HT2A antagonism and 5-HT1A agonism were identified as common mechanisms for second generation antipsychotic-associated serotonin syndrome. Additionally, FAERS and several case reports supported that interactions between second generation antipsychotics and other serotonergic agents may increase the risk for serotonin syndrome. These hypotheses were further supported by a literature search. Therefore, this study demonstrated that computational analyses of FAERS data using molecular knowledge, can easily generate hypotheses for adverse event mechanisms [75].

While improving the overall evidence-based used for drug re-labelling decisions, the above examples are still retrospective in nature. The FDA has, however, also expressed a strong interest in developing more proactive methods to predict AEs before they happen. Such a capability can assist both in post-market surveillance of AEs, and in the risk assessment during and after IND application. In recently reported proof-of-concept work, scientists used TAE profiles for six test drugs to aggregate data from FAERS and FDA drug labels based on drug targets. The goal of this study was to predict the evolution of the drug's label for four years post-approval. A genetic algorithm was utilized to set thresholds for features in these datasets to maximize performance. Utilizing on safety data available prior to the approval of each drug, the method correctly identified 78% of postmarket label changes and had a precision of 67%, recall of 81%, and specificity of 71% [76].

In follow-on work, the authors attempted to improve the performance of the previous adverse event prediction model utilizing target-adverse event profiles, additional drugs, learning features, and algorithms were evaluated [77]. In addition to FAERS data and drug labels, the new model

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

incorporated data from EMBASE, a biomedical literature database, as a feature. A larger set of 55 drugs were utilized in the study, and predictions were made for 36 groups of adverse events. Finally, an ensemble model containing Naïve Bayes, K Nearest Neighbor, Support Vector Machine, and C4.5 was applied to make predictions. Utilizing this updated model, overall precision improved to 72%, recall decreased to 70%, and specificity increased to 86%, although these metrics can be tuned to a user's specifications. Additionally, 69% of label changes were correctly identified. The results of these studies demonstrated promise for the target-adverse event profile approach of predicting adverse events.

Concurrent with development of the ensemble model, a Naïve Bayes approach was utilized to predict 135 individual adverse events for 54 drugs. Similar to the ensemble model study, adverse event information from FDA product labels and scientific literature was utilized, and additional features such as structural and target similarities and duration of post-market experience were incorporated. As the evaluation was performed using a probabilistic approach over 10,000 iterations, 53 of the 135 events demonstrated a high probability of having high positive predictive value, and many of these events had well-characterized target-event relationships. Additionally, 32% of the predicted drug label changes occurred. The ensemble model and Naïve Bayes model have been used concurrently to identify potential safety events for new drugs. [78]

4.2 Application of HD³ to drug repositioning and combinatorial therapy design.

Polypharmacology seeks to capitalize on systems-level knowledge about disease etiology to define either a single drug that binds two or more targets, or two (or more) drugs that bind to two (or more) targets. The mechanistic goal of these strategies is to recalibrate the perturbed network processes underlying the disease state. Polypharmacology is not only key to exploiting off-target effects in drug repurposing endeavours, but also to addressing more complex disease etiologies and adaptive resistance mechanisms. With molecular analysis of patient tumours revealing extraordinary levels of genetic complexity and clonal adaptability, oncology has provided an optimal testing bed for the development of these strategies,. To redress such challenge, we require systems level knowledge as to how multiple nodes in a signaling network combine to produce the pathophysiology of disease. Such models, while typically "proteo-centric", also include disease specific molecular perturbations, such as copy number variations, point mutations and differential expression. Combining empirical molecular knowledge through a clinically focused network perspective, allows us to identify combinatorial opportunities in a physiologically relevant manner. Beyond drug discovery, the strategy also holds particular promise for personalised oncology, with the mapping of molecular data from a patients tumor to system specific models, allowing the identification of personalized drug combinations.

The targeted design of combination therapies provides an interesting strategy for drug repositioning. This was elegantly demonstrated by recent discoveries around Beta Adrenergic Receptor (BAR) inhibition, via the beta-blocker class of drugs. First hints for a role of BAR in cancer came from studies indicating that psycho-social factors such as stress, depression, and lack of community might promote tumor growth and progression. In ovarian cancer in particular, such effects were found to directly enhance tumor pathogenesis by protecting tumor cells from anoikis, promoting tumor cell invasion and tumor-associated angiogenesis. The molecular mechanism was found to be mediated via tumor cell ADRB2 (Adrenergic Receptor), with phospho-proteomic analysis demonstrating that ADRB signaling leads to Src activation via a unique PKA-mediated mechanism [79]. This network was found to be critical to the regulation of phospho-proteomic signaling associated with ovarian cancer progression. Importantly, these observations were combined with real world data from FAERS, to examine the death rate of patients treated with ADRB2 inhibitors, compared to the death rate in a cohort without. Here, "death rate" was used as surrogate marker for treatment efficacy, with a lower death rate implying greater efficacy. Analysis revealed that mortality was reduced across major cancer types in patients where ADRB2 signaling was inhibited with beta-blockers, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

together with the lab findings identified BAR inhibition as a potential combinatorial route to anti-cancer treatment.

Several clinical studies have now provided evidence for repositioning beta-blockers, either alone or in combination with approved cancer therapies. For example, a phase-2 pilot-study (NCT01265576) examined ADRB2 inhibition in hepatocellular carcinoma (HCC) patients, where sorafenib was combined with the beta-blocker Propranolol and a COX2 inhibitor. The combination was found to increase therapy duration and overall survival, compared to Sorafenib alone [80]. Propranolol is a non-selective beta blocker that was first approved in 1967 for cardiovascular indications. Since then it has been repositioned in a wide range of indications, including essential tremor, prophylaxis of migraines, and as a first-line treatment for problematic infantile hemangiomas. [81] More recently, Fjaestad *et al*. reported that co-administration of propranolol significantly enhanced the efficacy of anti-CTLA4 therapy [82], whilst Amaya *et al*. found that beta-blockers increase progression free- and overall-survival in patients diagnosed with metastatic angiosarcoma [83].

Several other promising examples of positioning non-oncology drugs towards the treatment of cancer have been reported, including itraconazole, statins [84], metformin, aspirin, digoxin, and pantoprazole. Itraconazole is a potent anti-fungal drug that inhibits lanosterol-14α-demethylase in the cell membrane of the fungus [85]. Interestingly, it also shows anticancer activity across a number of solid and haematological cancer types. While the mechanism remains unclear, evidence points to the modulation of several components of signaling pathways involved in apoptosis and cell cycle arrest, including AMPK, mTOR, Wnt/β-catenin and Hedgehog [86]. Interestingly, Cheng *et al*. used complex network theory, to develop three supervised inference methods for drug repositioning, namely target-based similarity inference (TBSI), drug-based similarity inference (DBSI) and network-based inference (NBI) [87]. With its superior performance, NBI predicted potent polypharmacological effects of itraconazole, ketoconazole montelukast, diclofenac, simvastatin on estrogen receptors and dipeptidyl peptidase-IV enzyme, that were later validated in *in vitro* binding assays. Related studies showed that itraconazole synergizes with paclitaxel in the inhibition of endothelial cells, raising hopes for its use in combination therapy for endothelial cancer [88].

In other work, Huang *et al*. used multiomic patient datasets and the Broad Institute's Connectivity Map (CMap) database to identify members of the cardiac glycoside family, including digoxin, as potential treatments for medulloblastomas [89] CMap (available as CLUE for commercial users) is a cloud-based data analysis platform for perturbational datasets generated using gene expression (L1000) and proteomic (P100 and GCP) assays. Several different forms of *in silico* analyses have also pointed to the potential clinical utility of metformin in cancer. Clinical meta-analyses and *in vitro* studies have together pointed to metformin as a potential treatment option in several solid tumor types, including breast, pancreas and lung. Sun *et al*. [90], developed a metformin specific signaling pathway network (SPNetwork) using their Drug specific Signaling Pathway Network resource (DSPathNet). Using gene enrichment analysis from type 2 diabetes and cancer GWAS studies, network analyses and literature mining, the authors identified seven enriched genes (PPARGC1A, CDKN1A, MYC, ESR1, MAX, STK11, and SP1) impinging on a novel Myc-associated pathway that they proposed to play a key role in metformin's anti-cancer MOA.

Beyond oncology, HD^3 also played a critical role in the global response to the recent COVID-19 pandemic. As highlighted by Sheridan *et al*., numerous large-scale data and AI initiatives were launched to provide a testbed for pandemic forecasting and response [91]. At the time, knowledge connecting intricate clinical manifestations of COVID-19 with molecular underpinnings was both fragmented and new findings were being reported daily, thus exacerbating the challenge to stay abreast of emergent knowledge. In response, several groups developed COVID-19 knowledge models. For example. Domingo-Fernández *et al*. used knowledge graphs (KGs) to develop an integrated COVID-19 model. KG's have several advantages for systems-modelling, as they provide a means to capture, represent and formalize structured information. KG's are also complemented by a broad range of algorithms

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

that partially automate the process of knowledge discovery. Their open-source model could be used as a framework for target identification and drug repositioning efforts [92].

In more recent work, researchers from the ETH in Zurich reported the generation of a whole patient knowledge model of COVID-19 symptomatology. The authors used the commercially developed Dataome technology to extract and combine emergent clinical and molecular data around COVID-19 [93]. 332 high confidence virus-host interactors from SARS-CoV-1 were used as a seed for knowledge expansion. Next, this "seed model" was further expanded to include protein interactors, associated pathways and regulatory information. The resultant "base model" centered on a converging molecular phenotype that included the host proteins responsible for virus entry, TMPRSS2 and ACE2, together with significantly differentially down-regulated components of the interferon stimulated genes (ISG) induced by the virus infection (ACE2 and SERPING1). This base model was then further expanded through integration of genes associated with COVID-19 pathophysiology and symptomatology including a) common disease symptoms (e.g. fatigue), b) severe manifestations (e.g. acute respiratory distress syndrome (ARDS)), c) outcome- and severity-associated risk-factors (e.g. age) and d) diseases with COVID-19 like symptoms. Symptomatology associated 'pre-models' were then structured by domain experts and then linked through related molecular protagonists (see Figure 3).

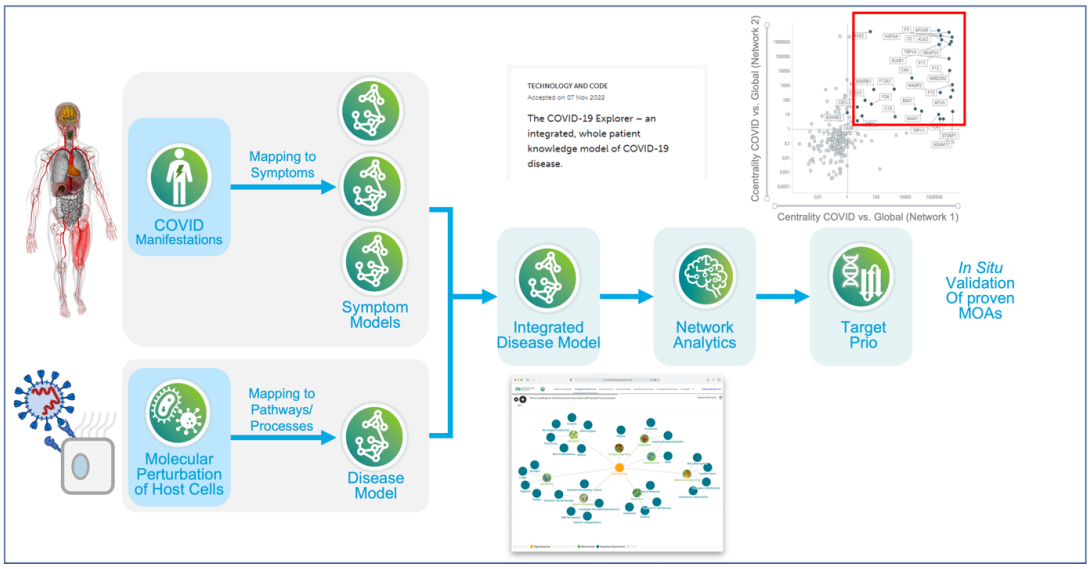

Figure 3. Schematic of the knowledge modeling process used by Brock *et al*. to produce a whole patient knowledge model of COVID19 symptomatology. The process used a previously reported SARS-CoV protein interaction map1 of 332 high confidence interactors as a seed for knowledge expansion. This was achieved through inclusion of factors from the host cell, host-specific response (e.g. innate immune response) and associated phenotypes. Next, this "seed model" was further expanded to include protein interactors, associated pathways and regulatory information. This "base model" was further expanded through integration of key molecular protagonists associated with COVID-19 pathophysiology and symptomatology including a) common disease symptoms (e.g. anosmia), b) severe manifestations (e.g. acute respiratory distress syndrome (ARDS)), c) outcome- and severity-associated risk-factors (e.g. age) and d) diseases with COVID-19 like symptoms. Symptomatology associated 'pre-models' were then manually curated. Linkage of these pre-models through related molecular protagonists resulted in a comprehensive and fully interactive whole patient COVID-19 disease model representation, providing a link between key molecular disease mechanisms and eight core pathogenic processes: inflammatory signaling, coagulation, barrier permeability, senescence, autoimmunity, fibrogenic signaling, nociception and exocytosis. Finally, these mechanisms were linked with respective symptoms and associated pathogenic pathways and affected organ-systems. Results were then made available via the open source COVID19 explorer (https://covid19.molecularhealth.com) and used as a basis for drug repurposing endeavors.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

The resultant whole patient model (available at http://covid19.molecularhealth.com) revealed that SARS-CoV-2 perturbs eight core pathogenic processes: inflammatory signaling, coagulation, barrier permeability, senescence, autoimmunity, fibrogenic signaling, nociception and exocytosis [94]. Together, these mechanisms were proposed to be responsible for unleashing a pathogenesis spectrum, ranging from 'a perfect storm' triggered by acute hyper-inflammation to chronic fatigue in protracted COVID-19 syndromes. Importantly, when viewed at the whole patient level, it appears that the ever-growing list and complexity of seemingly unrelated clinical symptoms, can be related to a limited set of molecular mechanisms. Although the model was initially based on data-mining hypotheses, many of the predicted clinical phenotypes (e. g. Kawasaki-like syndromes) and molecular mechanisms were subsequently reported in the peer-reviewed literature, thus providing both a real-time and real-world validation of their knowledge model. The work also presented several drug repositioning opportunities, including spironolactone, atorvastatin, losartan, many of which have since entered clinical trial.

5. Discussion

The persistence of Eroom's Law and associated innovation crisis runs counter to levels of technical invention that have been achieved in recent decades. Reviewing the evidence, we posit that currently accepted norms in drug discovery are causing us to frontload the drug development pipeline with preclinical data that bears little congruence to the human *in vivo* condition. Here, we emphasize the need to "Humanize" the entire drug discovery process and propose Human Data Driven Discovery (HD^3) as a focal point for these endeavors. HD^3 clearly provides a rational framework for structuring clinical and molecular evidence around drug polypharmacology and mode-of-action (MOA). By providing more physiologically representative *in silico* models of current knowledge, it provides a way to improve the accuracy of both expert-driven hypothesis generation and data-/AI-driven discovery. Today, this *in silico* model-based framework is already augmenting a variety of existing research avenues, including the rational design of combination therapies, drug (re-) positioning, and the analysis and prediction of adverse events. Indeed, with the Covid19 pandemic providing excellent validation of the power of human focused systems-based discovery, compelling evidence that HD^3 approaches might improve the quality of candidates entering the development pipeline is starting to emerge.

In an update to their initial description of Eroom's Law, Ringel *et al*. recently revisited the issue to examine the trajectory of R&D efficiency over the last decade [95]. Their results provide at least initial validation that HD^3-based approaches may now be starting to positively impact drug R&D efficiency. By 2018, the downward trendline in R&D efficiency, as measured by the number of New Molecular Entities (NME's) approved per $Billion spent, appeared to have broken out to the upside, with an additional 0.7 NME's approved between 2010-2018. Two major factors seem to be driving this change; a) the availability of better information for decision-making and b) the improved utility of that information. Importantly, they hypothesize that better information appears to derive from increasing focus on targets that have been validated using human derived data (e.g. from GWAS studies). This position is supported by other recent studies suggesting the positive impact of human data on drug R&D efficiency [96]. A common theme appears to be that data quality (i.e. human-derived data) is a superior success factor to data quantity (e.g. high-throughput data from model systems). Interestingly, Astra Zeneca also recently reported the strategic 5R framework (Right Target, Right Tissue, Right Safety, Right Patient, Right Commercial), that helped transform their R&D productivity and supported a return to business growth. [97]. The central focus of the 5R initiative was a move to increased utility of "humanized models", such as patient-derived xenograft models, organs-on-a-chip technology, humanized miniature organs, and 3D bioprinting. Interestingly, the strategy also appears to have had important intangible effects on personal biases, with researchers tending to adopt a more "truth-seeking" behavior in contrast to over-optimism or fear of failure, that so often

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

accompanies research endeavors. Their transformative success is further evidence that the innovation crisis is best tackled by efforts to "humanize" our current drug discovery paradigm; a premise also supported by the many proof-of-concept examples presented in this review.

We have provided a very simple first principles perspective on why we must expeditiously pivot to a primarily HD^3 drug research paradigm. At the most fundamental level, it is the three-dimensional arrangement of atoms in a small molecule drug (i.e. the chemotype) that interact with typically numerous proteins in biological systems (i.e. the biotype) to induce network-based changes that ultimately result in clinical effects (i.e. phenotype). Defining the relationship between these three core components not only provides a focus for re-wiring the utility of current data and technology assets, but also a stark reminder of the shortcomings of non-human model systems. Differences in structural complementarity between a human drug and paralogous ligand binding domains within model systems, likely become amplified at the whole system level, which can seriously obfuscate the relevance and utility of any phenotypic observation or result. Even small structural changes in the complementarity between drug and paralogous target can have profound functional consequences. Thus, while model systems are extremely helpful in deciphering the structural biochemistry of the cell, from a functional perspective, they are simply too far removed from the *in vivo* realities of human disease. Notwithstanding, before we can consider an *in silico* whole patient systems medicine approach, we must first focus on getting the simplest elements of the modeling process right, particularly comprehensive drug:target knowledge and system specific disease models.

We described several recent examples of the HD^3 approach in action, both at the level of drug discovery and applied regulatory science, that emphasize the power of deciphering the relationship between a drugs chemotype, biotype and resultant clinical phenotypes. We believe the tripartite nature of the challenge (Chemotype -> Biotype -> Phenotype), provides a useful framework for technical focus. For example, with respect to chemotypes, we know that available chemistry space is astronomical in dimension, with estimates ranging from 10^{26} and 10^{62} [98] chemotypes that would comply with the Lipinski guidelines for oral drugs. Borrowing from James Floods famous quote that "the best way to discover a new drug is to study an old one", a rational starting point for rewiring the drug discovery process is to focus modelling efforts on all chemistries ever tested in humans. While we could not find any peer-reviewed estimate of precise numbers, they are likely to be in the tens of thousands. Often associated with phenotype data (e.g. side effect information), these chemistries represent the most feature-/knowledge-rich part of chemistry space and thus lend themselves to the power of AI-based discovery methodologies. Later, other important components, such as the human metabolome and bioactive chemistries from natural food sources will also be important in extending the power of the approach. For all components of this "Human Chemome", broader efforts are required to capture the full extent of their biotypes i.e. their polypharmacology and molecular paths across different cellular systems within the patient. This requires focus on experimental and curation efforts towards defining accurate and comprehensive target information for each drug. Publicly available drug-to-target knowledge resources still have significant weaknesses, including, a) limited target coverage, b) false positives (e.g. genes regulated by drug), c) metabolizing enzymes and transporters classified as targets, and d) limited information on the effect of mutations on target binding affinities. It also requires a clear focus on the generation of cell-type specific knowledge models that accurately capture the structural biochemistry of cellular systems across all organs. As emphasized by the COVID-19 knowledge modelling work, individual cell and tissue models can later be combined at the level of whole patient knowledge models – enabling us to optimally link drug treatments with observed clinical symptomatology and outcomes. A summary of the associated data priorities for these endeavors are provided in Figure 4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

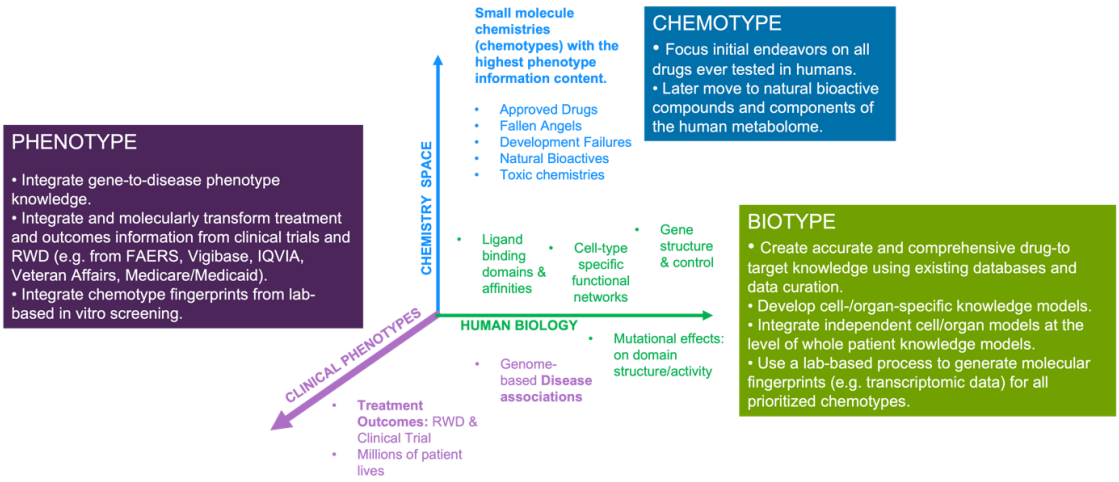

Figure 4. Schematic overview of the components and priorities associated with drug chemotypes, biotypes and associated phenotypes. We believe that this simple categorization can facilitate the structuring of efforts around the HD³ approach. While key datapoints are highlighted in the graph, data priorities are provided in the boxes.

A key shortcoming of traditional approaches to disease-modelling is the assumption that cellular organisation occurs primarily at the proteomic level, with proteins acting often sequentially to provide a structural framework for cellular response to diverse environmental cues. While this convenient simplification allows us to build useful models of protein pathways, it obviates a critical question. Where does the specificity of pathway/protein complex membership come from? This cannot alone be explained by complementarity at the level of protein:protein interaction domains. With many proteins possessing the same interaction domain, how is it that the right protein partner is found within the vast expanse of the cellular milieu? The answer no doubt lies within poorly appreciated genomic signatures that regulate the co-transcription and co-translation of otherwise distant genes coding for protein members of the same pathway/quaternary structure. Here, features such as gene promoters and enhancers will likely play a role, but the fact remains that such "signatures" have yet to be properly characterized or integrated within cellular models. From an *in silico* modelling perspective, this argues for the need to move beyond "proteo-centric" models of human disease, to include signatures at the level of diverse biological features, from genomic to proteomic to cellular to human phenotype levels.

Outcomes data from clinical trial and real-world contexts is of prime importance to the HD³ paradigm. Beyond the vast availability of disease and organ specific multi-omics data, we have also highlighted how treatment outcomes data can provide important insights into a drugs molecular mode of action. By matching a patients prescribed drugs to their associated targets, we can transform raw clinical data into *in silico* patient models [63]. The targets thus provide the biological integration point for the systems level analyses of observed clinical outcomes, further emphasizing the importance of comprehensive drug to target knowledge. Beyond the plethora of RWD sources mentioned, outcomes data is also available from commercial entities such as IQVIA, and other governmental agencies such as the Center for Medicare/Medicade Services (CMS), which alone hold data for hundreds of millions of treated patients, covering thousands of drugs and diseases. Importantly, as patient empowerment continues to grow, so too does the availability of patient reported outcomes, with social media platforms, digital therapeutics, and online forums such as Patients-like-me providing a wealth of potential insight. Physiological parameters are also being captured through wearable mhealth solutions and bringing us closer to an era where every patient will be defined and analyzed in the context of their own big data – our so-called digital twin. In this context, we can envisage the advent of whole patient disease models as not only a framework for AI-driven drug discovery, but also their utility in precision medicine where systems-based analysis of patient data can be used to prescribe or contraindicate treatments.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

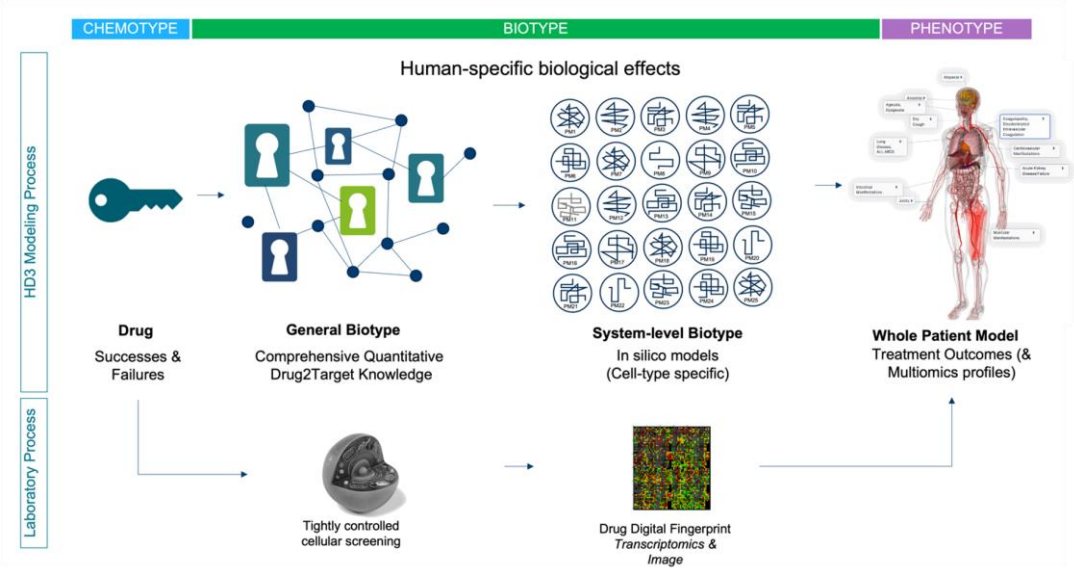

Figure 5. High-level schematic showing how HD3 systems-based discovery can be integrated with tightly controlled drug fingerprinting assays, to evolve whole patient molecular models for hundreds of millions of individual patients. Key to this approach is a) prioritized drug space, b) complete drug-to-target knowledge, c) network models (i.e. cell-type or tissue/organ specific), integrated onto whole patient knowledge frameworks. Using this approach, new uses for old drugs may be found, the area of failed drug space may again be explored and novel chemistries can be analyzed. .

It is important to emphasize that it is not our intention to negate the critical role that *in vitro/in vivo* model system research has played in drug discovery, but rather to reiterate the important existing and too often ignored weaknesses. The human body is a universe of molecular complexity and although HD³ data can provide only a snapshot, the information content is still directly congruent with the goal of conquering human disease. Notwithstanding, model systems will likely continue to play an important role in drug discovery, for example in the study of infectious disease and toxicology, but their general relevance will likely become increasingly dependent on the strategic alignment with the principles of the HD³ approach. For example, focusing on the aforementioned "Human Chemome", we could use cell-based screening to evolve drug-specific molecular fingerprints of drug action; a "molecular phenotype" of sorts (see fig.5). The goal of these fingerprints is not to help us define the drugs MOA, but rather the pattern of molecular effects within a standardized biological system. In an integrated strategy, well-controlled and replicated transcriptomics studies could be used to produce such fingerprints, which would then be mapped to the associated whole patient knowledge models for each drug. The same fingerprinting process can be applied to new molecular entities, which can then be analyzed and interpreted against the combined HD³ knowledge framework. Similar strategies are being applied by innovative new AI-driven drug discovery companies that have evolved over the past five years (for a list of 230 AI-driven drug discovery start-ups, see [99]). Using their current pipelines as a surrogate for R&D success, these innovators provide hope that "humanizing" the drug research and discovery paradigm can indeed expedite the path to important new therapies for patients. We can further hope that their success might expedite the transformation of the entire biopharmaceutical industry to a primarily human focused discovery paradigm.

6. Conclusions

Borrowing from learnings across other industries, we now know that most disruptive innovations do not necessarily derive from new inventions, but rather from the strategic integration of existing data and technology assets. Embracing this reality, we believe that all the right elements are now in place to expeditiously pivot our drug discovery endeavors towards a primarily Human Data Driven Discovery (HD³) paradigm. Advances in Omics technologies, computing, and AI analytics can now

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

interface with enormous tomes of human multi-omics datasets and under-utilized sources of treatment outcomes data for hundreds of millions of patients, to make this a viable proposition today. Combined with more comprehensive definitions of drug polypharmacology, and the structuring of data within cell/tissue/organ-specific *in silico* models, the HD³ strategy will allow us to evolve whole patient models of disease symptomatology and drug MOA, initially for hundreds of millions of patients. These can serve as the perfect knowledge framework for technologies such as GraphDB and the many exciting machine learning methodologies that continue to evolve apace. This can also eventually provide the basis for more accurate dynamic simulations of drug activity and disease mechanisms. Emergent evidence from early adopters validates the effectiveness of the HD³ approach, particularly in the areas of drug safety analysis/prediction and drug (re-)positioning. In contexts where it has been applied, intangible benefits have also been observed, such as encouraging "truth seeking" behavior amongst scientists, as opposed to over-optimism and fear of failure. Still, much remains to be done to close the gaps on even the simplest of challenges including the generation of comprehensive drug:target knowledge, starting with all approved drugs and the generation of accurate systems-specific disease models. Finally, it may ultimately serve the ethically important goal of decreasing our reliance on animal model research. From these perspectives, the HD³ paradigm stands to not only "Humanize" drug discovery, but also to make it more humane. This is the kind of disruptive innovation that is required to tackle and reverse Eroom's law, for once and for all; but most importantly, for patients.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1 Table S1: 86 Human Genomics Data Initiatives from around the world

Funding: This review received no external funding

Acknowledgments: The authors would like to thank Dr. Martin Kos for their helpful comments around GraphDB technologies in the analysis of outcomes data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hutchinson L, Kirk R. High drug attrition rates--where are we going wrong? Nat Rev Clin Oncol. 2011 Mar 30;8(4):189-90. doi: 10.1038/nrclinonc.2011.34.
2. Barton P, Riley RJ. A new paradigm for navigating compound property related drug attrition. Drug Discov Today. 2016 Jan;21(1):72-81. doi:10.1016/j.drudis.2015.09.010.
3. Evaluate Vantage 2020 Preview. Evaluate Ltd. https://www.evaluate.com/thought-leadership/vantage/evaluate-vantage-2020-preview#download. (Accessed Jan 15th, 2023)
4. Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. JAMA. 2020 Mar 3;323(9):844-853. doi: 10.1001/jama.2020.1166.
5. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. 2016 May;47:20-33. doi: 10.1016/j.jhealeco.2016.01.012.
6. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. 2012 Mar 1;11(3):191-200. doi: 10.1038/nrd3681.
7. Davis RL. Mechanism of Action and Target Identification: A Matter of Timing in Drug Discovery. iScience. 2020 Aug 21;23(9):101487. doi:10.1016/j.isci.2020.101487.
8. Wyatt PG, Gilbert IH, Read KD, Fairlamb AH. Target validation: linking target and chemical properties to desired product profile. Curr Top Med Chem. 2011;11(10):1275-83. doi: 10.2174/156802611795429185.

9. Morra G, Genoni A, Neves MA, Merz KM Jr, Colombo G. Molecular recognition and drug-lead identification: what can molecular simulations tell us? Curr Med Chem. 2010;17(1):25-41. doi: 10.2174/092986710789957797.
10. Wang S, Dong G, Sheng C. Structural simplification: an efficient strategy in lead optimization. Acta Pharm Sin B. 2019 Sep;9(5):880-901.doi:10.1016/j.apsb.2019.05.004
11. Haley B, Roudnicky F. Functional Genomics for Cancer Drug Target Discovery. Cancer Cell. 2020 Jul 13;38(1):31-43. doi: 10.1016/j.ccell.2020.04.006. Epub 2020 May 21.
12. Williams M. Editorial overview: from Vioxx to Luckenbach: drug discovery at a crossroads. Curr Opin Investig Drugs. 2005 Jan;6(1):17-20.
13. Nissen S. Rosiglitazone: a disappointing DREAM. Future Cardiol. 2007 Sep;3(5):491-2. doi: 10.2217/14796678.3.5.491.
14. Kaitin K, DiMasi J (2011) Pharmaceutical innovation in the 21st century: new drug approvals in the first decade, 2000–2009. Clin Pharmacol Ther 2011 Feb;89(2):183-8. doi: 10.1038/clpt.2010.286.
15. Thomas L (1996) Industrial Policy and International Competitiveness in the Pharmaceutical Industry. In: Helms R (ed) Competitive Strategies in the Pharmaceutical Industry. The American Enterprise Institute, Washington, D.C., pp 107–129.
16. Tufts Center Report on Trial timelines: https://www.centerwatch.com/articles/25033-trend-of-longer-trial-timelines-is-likely-to-continue Accessed November 19th 2022.
17. LaMattina JL. The impact of mergers on pharmaceutical R&D. Nat Rev Drug Discov. 2011 Aug 1;10(8):559-60. doi: 10.1038/nrd3514.
18. Szabo M, Svensson Akusjärvi S, Saxena A, Liu J, Chandrasekar G, Kitambi SS. Cell and small animal models for phenotypic drug discovery. Drug Des Devel Ther. 2017 Jun 28;11:1957-1967. doi: 10.2147/DDDT.S129447.
19. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: applications to targets and beyond. Br J Pharmacol. 2007 Sep;152(1):21-37. doi: 10.1038/sj.bjp.0707306.
20. Luo J. CRISPR/Cas9: From Genome Engineering to Cancer Drug Discovery. Trends Cancer. 2016 Jun;2(6):313-324. doi: 10.1016/j.trecan.2016.05.001.
21. Bon M, Bilsland A, Bower J, McAulay K. Fragment-based drug discovery-the importance of high-quality molecule libraries. Mol Oncol. 2022 Nov;16(21):3761-3777. doi: 10.1002/1878-0261.13277.
22. Tewkesbury DH, Robey RC, Barry PJ. Progress in precision medicine in cystic fibrosis: a focus on CFTR modulator therapy. Breathe (Sheff). 2021 Dec;17(4):210112. doi: 10.1183/20734735.0112-2021.
23. Carofiglio F, Lopalco A, Lopedota A, Cutrignelli A, Nicolotti O, Denora N, Stefanachi A, Leonetti F. Bcr-Abl Tyrosine Kinase Inhibitors in the Treatment of Pediatric CML. Int J Mol Sci. 2020 Jun 23;21(12):4469. doi: 10.3390/ijms21124469.
24. Santos, R et al. A comprehensive map of molecular drug targets. Nat. Rev. Drug Discov. 16, 19-34 (2016)
25. Mestres J, Gregori-Puigjane E, Valverde S, Sole RV. Data completeness – The Achilles heel of drug-target networks. Nat Biotechnol. 2008; 26(9):983–984. doi: 10.1038/nbt0908-983.
26. Begley CG, Ellis L. Drug development: raise standards for preclinical research.Nature. 2012; 483:531–533. doi: 10.1038/483531a.
27. Peers IS, Ceuppens PR, Harbron C. In search of preclinical robustness.Nat Rev Drug Discov. 2012; 11:733–734. doi: 10.1038/nrd3849.
28. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets?Nat Rev Drug Discov. 2011; 10:712. doi:10.1038/nrd3439-c1.
29. Hackam DG, Redelmeier DA. Translation of research evidence from animals to humans. JAMA. 2006; 296:1731–1732 doi: 10.1001/jama.296.14.1731.
30. Kaste M. Use of animal models has not contributed to development of acute stroke therapies: pro. Stroke. 2005;36:2323-2324. doi: 10.1161/01.STR.0000179037.82647.48.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

31. Horrobin DF. Modern biomedical research: an internally self-consistent universe with little contact with medical reality? Nat Rev Drug Discov. 2003 Feb;2(2):151-4. doi: 10.1038/nrd1012.
32. First Principles Thinking - https://www.csc.edu/media/website/content-assets/documents/pdf/tlpec/First-Principles-Thinking.pdf
33. Workman P, Al-Lazikani B, Clarke PA. Genome-based cancer therapeutics: targets, kinase drug resistance and future strategies for precision oncology. 2013 Aug;13(4):486-96. doi: 10.1016/j.coph.2013.06.004.
34. Cui JJ, Tran-Dubé M, Shen H, et al. Structure based drug design of crizotinib (PF-02341066), a potent and selective dual inhibitor of mesenchymal-epithelial transition factor (c-MET) kinase and anaplastic lymphoma kinase (ALK). J Med Chem 2011; 54(18): 6342-63. doi: 10.1021/jm2007613.
35. FDA Adverse Events Reporting System (FAERS) Public Dashboard. https://fis.fda.gov/sense/app/d10be6bb-494e-4cd2-82e4-0135608ddc13/sheet/7a47a261-d58b-4203-a8aa-6d3021737452/state/analysis.
36. Sentinel Initiative. https://www.sentinelinitiative.org/. Accessed January 12th , 2023.
37. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative-A comprehensive approach to medical product surveillance. 2016 Mar;99(3):265-8. doi: 10.1002/cpt.320.
38. European database of suspected adverse drug reaction reports. https://www.adrreports.eu/en/index.html. (accessed December 20th 2022)
39. UMC | VigiBase. https://www.who-umc.org/vigibase/vigibase/. Accessed December 28th, 2022
40. Amberger J.S., Bocchini C.A., Schiettecatte F., Scott A.F., Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. 2015 Jan;43(Database issue):D789-98. doi: 10.1093/nar/gku1205.
41. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, Ibrahim A, Ji Y, John S, Lewis E, MacArthur JAL, McMahon A, Osumi-Sutherland D, Panoutsopoulou K, Pendlington Z, Ramachandran S, Stefancsik R, Stewart J, Whetzel P, Wilson R, Hindorff L, Cunningham F, Lambert SA, Inouye M, Parkinson H, Harris LW. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 2023 Jan 6;51(D1):D977-D985. doi: 10.1093/nar/gkac1010.
42. Parsa A, Fuchsberger C, Kottgen A et al. . Common variants in Mendelian kidney disease genes and their association with renal function. J Am Soc Nephrol 2013; 24: 2105–2117 doi: 10.1681/ASN.2012100983.
43. Pattaro C, Kottgen A, Teumer A et al. . Genome-wide association and functional follow-up reveals new loci for kidney function. PLoS Genet 2012;8(3):e1002584. doi: 10.1371/journal.pgen.1002584.
44. Jupp S, Klein J, Schanstra J, Stevens R. Developing a kidney and urinary pathway knowledge base. J Biomed Semantics. 2011 May 17;2 Suppl 2(Suppl 2):S7. doi: 10.1186/2041-1480-2-S2-S7.
45. Fernandes M, Husi H. Establishment of a integrative multi-omics expression database CKDdb in the context of chronic kidney disease (CKD). Sci Rep. 2017 Jan 12;7:40367. doi: 10.1038/srep40367.
46. Cancer Genome Atlas Research Network; Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, Robertson AG, Pashtan I, Shen R, Benz CC, Yau C, Laird PW, Ding L, Zhang W, Mills GB, Kucherlapati R, Mardis ER, Levine DA. Integrated genomic characterization of endometrial carcinoma. Nature. 2013 May 2;497(7447):67-73. doi: 10.1038/nature12113.
47. Zhong Q, Simonis N, Li QR, Charloteaux B, Heuze F, Klitgord N, Tam S, Yu H, Venkatesan K, Mou D, Swearingen V, et al. Edgetic perturbation models of human inherited disorders. Mol Syst Biol. 2009; 5:321. doi: 10.1038/msb.2009.80.
48. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

DrugBank database for 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D1074-D1082. doi: 10.1093/nar/gkx1037.

49. Southan C, Várkonyi P, Muresan S. Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics. Curr Top Med Chem. 2007;7(15):1502-8. doi: 10.2174/156802607782194761.
50. Zhou Y, Zhang Y, Lian X, Li F, Wang C, Zhu F, Qiu Y, Chen Y. Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. Nucleic Acids Res. 2022 Jan 7;50(D1):D1398-D1407. doi: 10.1093/nar/gkab953.
51. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021 Jan 8;49(D1):D1388-D1395. doi: 10.1093/nar/gkaa971
52. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. 2016 Jan 4;44(D1):D1214-9. doi: 10.1093/nar/gkv1031.
53. He Y, Xiang Z, Zheng J, Lin Y, Overton JA, Ong E. The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. J Biomed Semantics. 2018 Jan 12;9(1):3. doi:10.1186/s13326-017-0169-2.
54. Jiang L, Wang M, Lin S, Jian R, Li X, Chan J, Dong G, Fang H, Robinson AE; GTEx Consortium; Snyder MP. A Quantitative Proteome Map of the Human Body. Cell. 2020 Oct 1;183(1):269-283.e19. doi: 10.1016/j.cell.2020.08.036.
55. Digre A, Lindskog C. The Human Protein Atlas - integrated omics for single cell mapping of the human proteome. Protein Sci. 2023 Jan 5:e4562. doi: 10.1002/pro.4562.
56. Lam KHB, Faust K, Yin R, Fiala C, Diamandis P. The Brain Protein Atlas: A conglomerate of proteomics datasets of human neural tissue. Proteomics. 2022 Dec;22(23-24):e2200127. doi: 10.1002/pmic.202200127.
57. Yamamoto T. The 4th Human Kidney and Urine Proteome Project (HKUPP) workshop. 26 September 2009, Toronto, Canada. Proteomics. 2010 Jun;10(11):2069-70. doi: 10.1002/pmic.201090041.
58. Lysenko A, Roznovăţ IA, Saqi M, Mazein A, Rawlings CJ, Auffray C. Representing and querying disease networks using graph databases. BioData Min. 2016 Jul 25;9:23. doi: 10.1186/s13040-016-0102-8.
59. Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34, i457–i466. 10.1093/bioinformatics/bty294.
60. Muñoz, E., Nováček, V., and Vandenbussche, P.-Y. (2019). Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models. Briefings in Bioinformatics 20, 190–202. 10.1093/bib/bbx099.
61. Bean, D.M., Wu, H., Iqbal, E., Dzahini, O., Ibrahim, Z.M., Broadbent, M., Stewart, R., and Dobson, R.J.B. (2017). Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. Sci Rep 7, 16416. 10.1038/s41598-017-16674-x.
62. Joshi, P., V, M., and Mukherjee, A. (2022). A knowledge graph embedding based approach to predict the adverse drug reactions using a deep neural network. Journal of Biomedical Informatics 132, 104122. 10.1016/j.jbi.2022.104122.
63. Bobed, C., Douze, L., Ferré, S., and Marcilly, R. (2018). PEGASE: A Knowledge Graph for Search and Exploration in Pharmacovigilance Data. EKAW Posters and Demonstrations. https://hal.inria.fr/hal-01976818
64. Soldatos TG, Taglang G, Jackson DB. In silico Profiling of Clinical Phenotypes for Human Targets Using Adverse Event Data. High Throughput. 2018 Nov 23;7(4):37. doi: 10.3390/ht7040037.
65. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. Nucleic Acids Res. 2020 Jan 8;48(D1):D498-D503. doi: 10.1093/nar/gkz1031.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

66. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009 Jan;37(Database issue):D674-9. doi: 10.1093/nar/gkn653. Epub 2008 Oct 2..
67. The ATC ontology [WHOCC—Structure and Principles. Available online: https://www.whocc.no/atc/structure_and_ principles/). Accessed December 19th, 2022.
68. Rao MS, Gupta R, Liguori MJ, Hu M, Huang X, Mantena SR, Mittelstadt SW, Blomme EAG, Van Vleet TR. Novel Computational Approach to Predict Off-Target Interactions for Small Molecules. Front Big Data. 2019 Jul 17;2:25. doi: 10.3389/fdata.2019.00025.
69. The OFF-X Platform - https://clarivate.com/products/biopharma/off-x Accessed January 8th 2023.
70. Kim S, Lahu G, Vakilynejad M, Soldatos TG, Jackson DB, Lesko LJ, Trame MN. A case study of a patient-centered reverse translational systems-based approach to understand adverse event profiles in drug development. Clin Transl Sci. 2022 Apr;15(4):1003-1013. doi: 10.1111/cts.13219.
71. Force T, Krause DS, Van Etten RA. Molecular mechanisms of cardiotoxicity of tyrosine kinase inhibition. Nat Rev Cancer. 2007 May;7(5):332-44. doi: 10.1038/nrc2106.
72. Grazette LP, Boecker W, Matsui T, Semigran M, Force TL, Hajjar RJ, Rosenzweig A. Inhibition of ErbB2 causes mitochondrial dysfunction in cardiomyocytes: implications for herceptin-induced cardiomyopathy. J Am Coll Cardiol. 2004 Dec 7;44(11):2231-8. doi: 10.1016/j.jacc.2004.08.066
73. Kim S, Lahu G, Vakilynejad M, Soldatos TG, Jackson DB, Lesko LJ, Trame MN. Application of a patient-centered reverse translational systems-based approach to understand mechanisms of an adverse drug reaction of immune checkpoint inhibitors. Clin Transl Sci. 2022 Jun;15(6):1430-1438. doi: 10.1111/cts.13254.
74. Tafenoquine label https://www.accessdata.fda.gov/drugsatfda_docs/label/2020/210795s001lbl.pdf
75. Association Between Serotonin Syndrome and Second-Generation Antipsychotics via Pharmacological Target-Adverse Event Analysis. Racz R, Jackson DB, Soldatos, Burkhart K. TG Clin Transl Sci. 2018 May;11(3):322-329 doi: 10.1111/cts.12543.
76. Target-Adverse Event Profiles to Augment Pharmacovigilance: A Pilot Study With Six New Molecular Entities. Schotland P, Racz R, Jackson DB, Strauss DG, Burkhart K. CPT Pharmacometrics Syst Pharmacol. 2018 Dec;7(12):809-817 doi: 10.1002/psp4.12356.
77. Schotland P, Racz R, Jackson DB, Soldatos TG, Levin R, Strauss DG, Burkhart K. Target Adverse Event Profiles for Predictive Safety in the Postmarket Setting. Clin Pharmacol Ther. 2021 May;109(5):1232-1243. doi: 10.1002/cpt.2074.
78. Daluwatte C, Schotland P, Strauss DG, Burkhart KK, Racz R. Predicting potential adverse events using safety data from marketed drugs. BMC Bioinformatics. 2020 Apr 29;21(1):163. doi: 10.1186/s12859-020-3509-7.
79. Armaiz-Pena GN, Allen JK, Cruz A, Stone RL, Nick AM, Lin YG, Han LY, Mangala LS, Villares GJ, Vivas-Mejia P, Rodriguez-Aguayo C, Nagaraja AS, Gharpure KM, Wu Z, English RD, Soman KV, Shahzad MM, Zigler M, Deavers MT, Zien A, Soldatos TG, Jackson DB, Wiktorowicz JE, Torres-Lugo M, Young T, De Geest K, Gallick GE, Bar- Eli M, Lopez-Berestein G, Cole SW, Lopez GE, Lutgendorf SK, Sood AK. Src activation by β-adrenoreceptors is a key switch for tumour metastasis. Nat Commun. 2013;4:1403. doi: 10.1038/ncomms2413.
80. De la Torre, A.N.; Castaneda, I.; Hezel, A.F.; Bascomb, N.F.; Bhattacharyya, G.S.; Abou-Alfa, G.K. Effect of coadministration of propranolol and etodolac (VT-122) plus sorafenib for patients with advanced hepatocellular carcinoma (HCC). J. Clin. Oncol. 2015, 33. DOI: 10.1200/jco.2015.33.3_suppl.390
81. Srinivasan AV. Propranolol: A 50-Year Historical Perspective. Ann Indian Acad Neurol. 2019 Jan-Mar;22(1):21-26. doi: 10.4103/aian.AIAN_201_18.
82. Fjæstad KY, Rømer AMA, Goitea V, Johansen AZ, Thorseth ML, Carretta M, Engelholm LH, Grøntved L, Junker N, Madsen DH. Blockade of beta-adrenergic receptors reduces

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 May 2023 doi:10.20944/preprints202305.0945.v1

cancer growth and enhances the response to anti-CTLA4 therapy by modulating the tumor microenvironment. Oncogene. 2022 Feb;41(9):1364-1375. doi: 10.1038/s41388-021-02170-0

83. Amaya CN, Perkins M, Belmont A, Herrera C, Nasrazadani A, Vargas A, Khayou T, Montoya A, Ballou Y, Galvan D, Rivas A, Rains S, Patel L, Ortega V, Lopez C, Chow W, Dickerson EB, Bryan BA. Non-selective beta blockers inhibit angiosarcoma cell viability and increase progression free- and overall-survival in patients diagnosed with metastatic angiosarcoma. Oncoscience. 2018 Apr 29;5(3-4):109-119. doi: 10.18632/oncoscience.413
84. Shaghaghi Z, Alvandi M, Farzipour S, Dehbanpour MR, Nosrati S. A review of effects of atorvastatin in cancer therapy. Med Oncol. 2022 Dec 2;40(1):27. doi: 10.1007/s12032-022-01892-9.
85. Weng N, Zhang Z, Tan Y, Zhang X, Wei X, Zhu Q. Repurposing antifungal drugs for cancer therapy. J Adv Res. 2022 Sep 5:S2090-1232(22)00199-0. doi: 10.1016/j.jare.2022.08.018.
86. Kim J, Tang JY, Gong R, Kim J, Lee JJ, Clemons KV, Chong CR, Chang KS, Fereshteh M, Gardner D, Reya T, Liu JO, Epstein EH, Stevens DA, Beachy PA. Itraconazole, a commonly used antifungal that inhibits Hedgehog pathway activity and cancer growth. Cancer Cell. 2010 Apr 13;17(4):388-99. doi: 10.1016/j.ccr.2010.02.027.
87. F. Cheng, C. Liu, J. Jiang, et al., Prediction of drug-target interactions and drug repositioning via network-based inference, PLoS Comput. Biol. 8 (5) (2012) e1002503, , https://doi.org/10.1371/journal.pcbi.1002503
88. C.H. Choi, J.Y. Ryu, Y.J. Cho, et al., The anti-cancer effects of itraconazole in epithelial ovarian cancer, Sci. Rep. 7 (1) (2017) 6552, https://doi.org/10.1038/s41598-017-06510-7.
89. L. Huang, S. Garrett Injac, K. Cui, et al., Systems biology-based drug repositioning identifies digoxin as a potential therapy for groups 3 and 4 medulloblastoma, Sci TranslMed. 10 (464) (2018), https://doi.org/10.1126/scitranslmed.aat0150 pii: eaat0150.
90. J. Sun, M. Zhao, P. Jia, et al., Deciphering signaling pathway networks to understand the molecular mechanisms of metformin action, PLoS Comput. Biol. 11 (6) (2015) e1004202, , https://doi.org/10.1371/journal.pcbi.1004202 eCollection 2015 Jun
91. Sheridan C. Massive data initiatives and AI provide testbed for pandemic forecasting. Nat Biotechnol. 2020 Sep;38(9):1010-1013. doi:10.1038/s41587-020-0671-4.
92. Domingo-Fernández D, Baksi S, Schultz B, Gadiya Y, Karki R, Raschka T, Ebeling C, Hofmann-Apitius M, Kodamullil AT. COVID-19 Knowledge Graph: a computable, multimodal, cause-and-effect knowledge model of COVID-19 pathophysiology. Bioinformatics. 2021 Jun 9;37(9):1332-1334. doi:10.1093/bioinformatics/btaa834
93. Brock, S., Soldatos, T. G., Jackson, D. B., Diella, F., Hornischer, K., Schäefer, A., et al. (2022). The COVID-19 Explorer - an integrated, whole patient knowledge model of COVID-19 disease. Front. Mol. Med. 2, 1035215. doi:10.3389/fmmed.2022.1035215.
94. Brock, S., Jackson, D. B., Soldatos, T. G., Hornischer, K., Schäfer, A., Diella, F., et al. (2022). Whole patient knowledge modeling of COVID-19 symptomatology reveals common molecular mechanisms. Front. Mol. Med. 2:1035290. doi:10.3389/fmmed.2022.1035290
95. Ringel MS, Scannell JW, Baedeker M, Schulze U. Breaking Eroom's Law. Nat Rev Drug Discov. 2020 Dec;19(12):833-834. doi: 10.1038/d41573-020-00059-3.
96. Scannell JW, Bosley J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. PLoS One. 2016 Feb 10;11(2):e0147215. doi: 10.1371/journal.pone.0147215.
97. Morgan P, Brown DG, Lennard S, Anderton MJ, Barrett JC, Eriksson U, Fidock M, Hamrén B, Johnson A, March RE, Matcham J, Mettetal J, Nicholls DJ, Platz S, Rees S, Snowden MA, Pangalos MN. Impact of a five-dimensional framework on R&D productivity at AstraZeneca. Nat Rev Drug Discov. 2018 Mar;17(3):167-181. doi: 10.1038/nrd.2017.244.
98. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996 Jan;16(1):3-50. doi: 10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6.
99. 230 AI-driven drug discovery start-ups https://blog.benchsci.com/startups-using-artificial-intelligence-in-drug-discovery