Published in final edited form as:

Expert Rev Precis Med Drug Dev. 2016 ; 1(3): 245–253. doi:10.1080/23808993.2016.1174062.

Use of big data in drug development for precision medicine

Rosa S. Kim^{1,*}, Nicolas Goossens^{1,2,*}, and Yujin Hoshida¹

¹Division of Liver Diseases, Department of Medicine, Liver Cancer Program, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, USA ²Division of Gastroenterology and Hepatology, Geneva University Hospital, Geneva, Switzerland

Summary

Drug development has been a costly and lengthy process with an extremely low success rate and lack of consideration of individual diversity in drug response and toxicity. Over the past decade, an alternative “big data” approach has been expanding at an unprecedented pace based on the development of electronic databases of chemical substances, disease gene/protein targets, functional readouts, and clinical information covering inter-individual genetic variations and toxicities. This paradigm shift has enabled systematic, high-throughput, and accelerated identification of novel drugs or repurposed indications of existing drugs for pathogenic molecular aberrations specifically present in each individual patient. The exploding interest from the information technology and direct-to-consumer genetic testing industries has been further facilitating the use of big data to achieve personalized Precision Medicine. Here we overview currently available resources and discuss future prospects.

Keywords

Big data; drug development; precision medicine; high-throughput screen; in silico drug discovery

1. Introduction

Drug development has been a protracted and costly multi-step process. Traditionally, candidate drug targets are defined and assessed based on a reductionist approach or a close-world assumption (CWA), relying on a highly simplified view of biology and restricted to manipulating one molecule or molecular pathway. Due to our limited understanding of biology at the systems level, each step in drug discovery and development has been suffering from a high level of uncertainty, resulting in an extremely low success rate. It takes billions of investment dollars and an average of about 9–12 years to bring a new drug to the market [1, 2]. Despite this, failures throughout the drug development pipeline make improving drug

Correspondence: Yujin Hoshida, MD, PhD, Division of Liver Diseases, Department of Medicine, Liver Cancer Program, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, Mail: 1470 Madison Ave, New York, NY 10029, Box 1123, USA, Telephone: 1-212-824-8862, Fax: 1-646-537-9576, yujin.hoshida@mssm.edu.

*Contributed equally to the work.

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

research and development (R&D) productivity an overarching priority. In fact, according to one study, the clinical approval success rate of drugs originating from pharmaceutical companies in the United States was 16% [3, 4].

Current challenges in drug development include (i) a limited capability to comprehensively characterize and/or monitor biological contexts of interest, (ii) a lack of adequate experimental models to test candidate drugs/perturbations and their clinical relevance, (iii) a range of accompanying multiple off-target effects of the drugs that are generally unknown (can be referred to as “polypharmacology”), and (iv) the genetic and environmental diversity across individuals in drug metabolizing enzymes, on/off-target molecules, and other drug response-related factors. The increasing rate of attrition at the late stages of drug development despite mounting R&D spending underscores the need for innovative and network-based approaches to drug discovery.

The recent emergence of “big data” in biology has fundamentally revolutionized how we study molecular biology and drug development. We are now at the tipping point to resolve some of the aforementioned drug discovery obstacles and facilitate the delivery of new drugs into clinical practice as evidenced by the initiation of the Precision Medicine Initiative [5]. In this review, we overview recent attempts of utilizing big data specifically in the field of drug development, and discuss exciting future prospects.

2. Big data

The word “big data”, generally refers to an approach where comprehensive data are first collected in an unbiased manner, without a priori hypotheses, and subsequently analyzed by data mining algorithms to formulate novel ideas [6]. In molecular biology, this approach has been led by the analysis of nucleic acid and protein sequences accumulated in public databases as well as DNA microarray-based genome-wide transcriptome and DNA structural variation data [7]. With unprecedented levels of development of genomic assay technologies and the massive growth of global information storage and sharing technologies, the big data approach has been further expanded and integrated with other data types such as epigenomic features, bio-ontologies, chemical structures, protein structures, electronic medical records, clinical trial registries, and clinical toxicity databases, at human population level. In parallel, data analysis algorithms and infrastructure tailored to analyze this type of data have been actively developed [8–10].

A major challenge in the analysis of big data has been to adapt statistical techniques to the added scale and complexity of datasets. For example, correction of p-values for multiple hypothesis testing (MHT) is a key issue in order to control false discovery, while minimizing false negatives [11]. In addition, high-dimensional data often requires, dimensionality reduction using data projection techniques such as extraction of principal components, non-negative matrix factorization, and molecular pathway-level abstraction of the data [12, 13]. A variety of unsupervised and supervised machine learning algorithms has been adapted to the analysis of biomedical big data to identify unknown subpopulations of diseases, elucidate novel disease targets, and predict clinical outcome based on genomic and clinical data [14].

3. Big data analysis for drug development

3.1 Therapeutic target discovery

Our rapidly expanding capability to comprehensively characterize biomolecules at the whole genome level has led to a paradigm shift in the field of therapeutic target discovery. Over the past decade, the analysis of genomic DNA structural alterations within or near protein-coding regions as well as transcriptome profiles have led to the development of novel methodologies to enable big data-based, unbiased target exploration. Now we can also measure genome-wide DNA methylation, histone protein modifications, splicing variants, transcription factor binding sites, protein abundance and modifications such as phosphorylation, non-coding RNA expression, and structural alterations in non-coding regions of genomic DNA. These technological developments have resulted in ever-growing big data repositories for biomedical research and drug development that allow hypothesis-free, unbiased target discovery (Table 1).

Multiple cancer genome sequencing projects have yielded a catalog of somatic DNA structural alterations such as single nucleotide variations, small insertions and deletions, copy number alterations, and genomic translocations for each cancer type as potential cancer drivers and new therapeutic targets [15, 16] (Table 1). Detailed catalogs of transcriptional regulation machinery such as genome-wide transcription factor binding sites for major transcription factors as well as epigenomic marks such as histone protein acetylation have been also developed to help identify biologically functional drug targets [17–21]. Transcriptome databases have been extensively utilized. Together with the rapid expansion of highly selective chemical libraries, these databases have substantially shortened the potential time period required from target discovery to clinical deployment (traditionally a few decades).

An example of such big data-driven translational success story pertains to the treatment of non-small cell lung cancer (NSCLC) patients. Approximately 5% of NSCLC cases were identified to harbor somatic rearrangements in the anaplastic lymphoma kinase (*ALK*) gene [22], and a selective *ALK* inhibitor, crizotinib, showed improved progression-free survival in the *ALK* gene rearrangement-positive NSCLC patients [23, 24]. Based on these results, crizotinib as well as companion biomarker assays detecting the gene fusion were FDA-approved and incorporated into clinical practice guideline, replacing the traditional combination chemotherapy only a few years after the initial discovery of the *ALK* rearrangements [25]. Similarly, pharmacologically targetable epidermal growth factor receptor (*EGFR*) gene mutations have been successfully utilized in the clinical deployment of selective *EGFR* inhibitors [26].

Despite these promising achievements, the relatively low prevalence of such somatic gene mutations in cancer often obscures the therapeutic benefit observed when all, unselected, patients enrolled in clinical trial are analyzed. Also, prediction of functional consequences of such genetic alterations remains challenging as recently highlighted in a follow-up case study of a patient with *ALK*-rearranged lung cancer, who developed a crizotinib resistance-conferring mutation, requiring the use of a third generation *ALK* inhibitor, but subsequently resensitizing to crizotinib when a further mutation developed negating the initial resistance

[27]. This example suggests that the therapeutic consequence of molecular targeted drug treatment depends on each specific mutation even if the same target gene is mutated. In other words, the true therapeutic effect of a molecular-targeted drug on a specific molecular aberration in the target will never be known unless it is tested in patients characterized for the molecular target. However, it is practically infeasible to test the astronomical number of potential combinations of a drug-target mutation in the traditional format of a clinical trial, testing each single therapy at a time. Several attempts have been proposed and/or tested to overcome the challenges with new clinical trial designs. For example, integrating information regarding predictive biomarker-based enrichment of potential drug responders to clinical trials is expected to improve statistical power, while reducing sample size [28]. This approach has been further expanded to a variety of specific scenarios such as two-stage patient enrichment to cope with biomarker misclassification [29] and enrichment for time to event endpoint [30]. Nevertheless, running a clinical trial for each one of drug-target pairs is unrealistic given the vast number of such pairs. In addition, evaluation of multiple drugs in parallel in a trial has been practically infeasible due to many reasons, including conflict in intellectual property. To address this long-standing challenge, the umbrella trial design, first screens multiple biomarkers (most of them are expected to be less prevalent markers) at a central platform/infrastructure, then distributes each patient to a matched sub-study, assessing an agent expected to target the positive biomarker, focusing on a single tumor type or histology. The basket trial design expands the strategy to multiple tumor types sharing the same molecular aberrations to increase the chance of finding therapies for rare genetic variants/diseases [31]. The National Cancer Institute-Molecular Analysis for Therapy Choice (NCI-MATCH) is a basket trial, in which up to 3,000 cancer patients will be screened for “actionable” DNA mutations and assigned to either of 20 treatments [32]. Exploration of “exceptional responders” followed by mechanistic interpretation is another alternative approach [33]. Co-clinical trial, simultaneously assessing therapeutic effect in human and mechanisms in animal model, was also proposed [34]. Yet another approach to account for intersubject variability in therapeutic response is the n-of-1 trial, where the study focuses on response of an intervention in one subject [35, 36]. Detailed data is collected for one participant during an intervention with the possibility of interrogating mechanisms of extreme drug response, testing another intervention after a suitable “wash-out” period, etc. This approach allows for example to explore molecular effects of a new compound at early stages of drug development, but also to draw inferences about population-wide effectiveness when combining multiple n-of-1 trials. In addition, this also allows taking into account individual factors linked to toxicity and may help to reduce attrition rate due to drug safety problems in subsequent traditional larger clinical trials. These new strategies have been rapidly evolving in the field of cancer therapeutics, which could be applicable to similarly heterogeneous polygenic diseases.

Further, the genome sequencing initiatives have now shifted the focus to population-based studies that provide deeper insight into rare pathogenic variants [37] (Table 1). The rapid dissemination of sequencing as part of routine clinical practice is further expanding the resource of population-level genomic variation data [38]. Lastly, there is a progressive transition to the use of electronic medical record (EMR) for unbiased mining of clinical phenotypic information, which has the potential to identify novel disease subtypes and

potential therapeutic targets [39], although the establishment of a unified and interchangeable infrastructure is still challenging [40].

One of the key challenges facing researchers will be to properly integrate the multiple layers of data into a manageable and systematic unit for better informing drug development and patient care. Integration of different “omics” data (for example, genomic, epigenomic, transcriptomic and proteomic) with clinical phenotype information recorded in EMR is particularly important to delineate clinically-relevant pathogenic molecular alterations as drug targets. For instance, topology-based patient-patient networks based on integrative clinical and genomic data from 11,000 individuals could identify 3 novel subgroups of type 2 diabetes [39]. Properly leveraged and harnessed, the integration of dynamic genomic and phenomic information, medical data and scans, as well as socioeconomic and environmental variables will revolutionize clinical medicine in the near future [41].

3.2 Drug discovery

Big data resources related to chemical substances have been utilized for computational virtual screening of compounds. Chemical structure-based exploration of small molecule compounds is a traditional way to identify refined compounds with similar biological activity. This approach is based on an assumption that compounds with similar structure share similar biological properties, although it does not hold in many cases (Table 2). Nevertheless, structure family-based prioritization and/or pooling of compounds to be screened is a commonly used strategy to reduce complexity and cost of drug screening. When structure of target protein is available together with biochemical properties of each amino acid residue, informatic estimation of physical interaction with small molecules has been widely performed as another popular strategy of drug discovery. Previously determined protein structures from individual studies and large-scale projects/initiatives such as the Protein 3000 Project [42], have been assembled and made available in public databases (Table 2).

Similarly, inhibition of protein-protein interaction can also be computationally explored [43]. Although X-ray crystallography was historically applicable to a limited classes of proteins, recent technical developments have been expanding our capability to crystalize membrane and soluble proteins and complexes proteins as well as small molecules [44]. Computational protein structure prediction from amino acid sequence has been an outstanding problem in biomedical research that has been tackled by long-standing community efforts such as The Critical Assessment of protein Structure Prediction (CASP) (www.predictioncenter.org). Imaging reconstruction-based methods such as cryo-electron microscopy (cryo-EM) are potential alternative approaches addressing the problem from a completely different angle [45]. Large-scale machine learning predictions of on/off-target effects based on structure of receptor ligand (instead of target receptor protein itself) was proposed to capture ligand-based similarities of otherwise unrelated proteins [46, 47].

Transcriptome is increasingly recognized as a common space, where the functional consequence of various types of molecular dysregulations and aberrations as well as drug treatment can be directly compared and linked based on the similarity of transcriptional changes measured as gene signatures [48]. This approach, systematically merging disease-

associated gene signature(s) and drug signatures, has enabled high-throughput search of compounds potentially reversing the disease gene signature and therefore treating the disease, as an alternative to the traditional target-based drug discovery [49]. The simple computational unbiased screening of gene signature similarity often identifies unexpected compounds outside known mechanisms of drug action and biology [50], and has been successfully utilized for a variety of diseases such as cancer and inflammatory diseases [51–53]. Application of this method for clinically approved compounds or compounds with halted development for certain indications at early stages could identify other indications of drugs already used in clinical practice. This “drug repurposing” approach has increasingly gained popularity [43, 54] and is getting more attention from pharmaceutical industry and regulatory authorities because it could substantially reduce the cost and time for drug development [55]. A database of transcriptome signatures of >1,000 Food and Drug Administration (FDA)-approved compounds (connectivity map) has enabled convenient compound search similar to the internet search engines [56]. The database is being expanded to bioactives and tool compounds beyond the FDA compounds as well as genetic manipulation (open reading frame [ORF] expression and shRNA-based knockdown) in multiple cellular contexts under the National Institute of Health (NIH) Library of Integrated Network-based Cellular Signatures (LINCS) program to further utilize the approach to probe the highly inter-connected wiring of cellular signaling. Addition of transcriptome profiles of other class of compounds such as diversity-oriented synthesis (DOS) library and collections of natural products could further increase the chance of novel drug discovery and facilitate drug development [57]. Another unique approach is to use similarity in side-effect profiles [58]. The assembly of large compound libraries has enabled high-throughput compound library screening, and yielded big data of phenotypic readouts of pharmacological interventions in addition to transcriptome such as cell survival, abundance of target molecules, enzymatic activity, and morphological changes, many of which are deposited in the public domain [59–61] (Table 2). Similar resources are expected to evolve into more integrated resources for systematic therapeutic discovery [62].

3.3 Drug toxicity

From a pharmaco-therapeutics perspective, one of the major goals of precision medicine is to improve the therapeutic index of drug products. However, a recent report has underlined that only approximately one in ten drugs entering phase I clinical trials will eventually be approved by the FDA, an approval rate that was even lower for new molecular entities reaching only 7.5% rate of FDA approval [4]. Even when examining the filings for new drug applications and biologic license application, a relatively advanced stage in drug development, safety concerns were the root cause for drug development suspension in 31% of programs, the second cause behind a lack of efficacy [4]. Therefore, there is an increasing interest to develop novel early detection systems for detecting drug toxicity and *in silico* approaches to model toxicity provide a high-throughput and low-cost alternative to currently accepted drug development processes. These methods aim to complement *in vitro* and *in vivo* toxicity tests to reduce the cost and time of toxicity testing, minimize the requirement to animal testing and improve overall safety assessment.

An ever-expanding array of methodologies is used for *in silico* toxicity modeling. Although it is beyond the scope of this text to give a comprehensive overview of the field, we underline major themes and methodologies currently being used in this dynamic field. Quantitative structure-activity relationship (QSAR) denotes a family of models seeking a relationship between molecular characteristics of a given molecule and its toxicity, under the assumption that chemicals that fit the same QSAR model may work through the same mechanism [63, 64]. Advantages of QSAR include ease of interpretation and modeling of categorical and continuous toxicity endpoints but it requires large datasets and cannot be extrapolated between species. QSAR models are implemented in a number of tools including OECD QSAR toolbox and TopKat (Table 3). Alternatively, “expert systems” are software programs codifying a set of rules such as Oncologic Cancer Expert System, Toxtree, and DEREK Nexus [65–67]. An example of expert systems includes structural alerts, a list of chemical structures that indicate or are associated to toxicity. Other models include read-across prediction, using compounds with known toxicity and similar chemical structures, to predict toxicity of investigational compounds.

Adsorption, distribution, metabolism and excretion (ADME) properties are key components to the drug for understanding therapeutic effects, but also for better understanding toxicity related, for example, to drug accumulation and metabolism. Among the components of ADME, metabolism has been among the most studied, due to its crucial role in drug toxicity [68]. Computational approaches to assess drug metabolism is a complex process, but can be generally classified as specific or comprehensive tools [69]. Specific approaches focus on specific molecules of pathways, while comprehensive tools take a more general approach and apply to a wider variety of biological systems. The methodologies used in modelling metabolic systems include expert systems, QSAR, data mining and others, but it has become clear that the key is the integration of a number of different methods and resources. Resources useful for predicting toxicological effects of metabolites include DEREK Nexus, ToxTree and HazardExpert among others [66, 67, 70]. Increasing efforts attempt to integrate ADME predictions into pharmacokinetic and pharmacodynamics models. These models attempt to calculate concentrations at a given time (pharmacokinetics) or the effect of a drug at a given concentration (pharmacodynamics) and therefore quantify ADME processes. Toxicokinetic studies attempt to correlate drug concentrations in various tissue compartments to the timing of toxic responses [71]. Although we do not directly consider these here, it is notable that modelling of off-target side-effects has also been attempted [72]. One of the recurring problems in *in silico* modelling of toxicity is the lack of data available to enter in these data-hungry models, in particular in the case of heterogeneous compound modelling. Further addition of experimental data and additional integration of these different approaches will ensure the growth of the field of drug toxicity modelling [73].

An additional toxicological concern that has raised attention in the past decades is the presence of pharmaceuticals in the environment [74]. For example, environmental toxicity of ethinylestradiol has been linked to feminization of male fish in effluent-dominated rivers [75] and altered behavior and feeding rate of the wild fish *Perca fluviatilis* was shown to be induced by oxazepam at environmentally relevant concentrations [76]. QSAR-based modelling and quantitative structure-toxicity relationships (QSTR) has been used to model ecotoxicity of organic chemicals and fungicides [77, 78]. In addition, the US Environmental

Protection Agency (EPA) has developed a high throughput screening program, ToxCast, to predict potential for toxicity and environmental hazard of various chemicals [79]. This approach has been used to develop a computational network model that integrates in vitro screening assays measuring estrogen receptor activation and applied to a library of 1,812 commercial and environmental chemicals identified 52 environmental chemicals predicted to be strongly activate or inhibit the estrogen receptor and therefore prioritize environmental chemicals for additional in vivo endocrine testing [80].

3.4 Combination therapy

Combination drug therapy was developed in the mid-20th century, in the context of cancer chemotherapy, to increase effectiveness and mitigate side effects of anti-cancer therapy [81] [82]. However, in recent decades, with an increased focus to treating disease of multifactorial etiology, such as diabetes, obesity, and cardiovascular disease it is becoming more apparent that the ‘one drug-one target’ approach may be over simplistic, and that combination drug therapy may be reexamined, in a broader perspective than only cancer therapy [83]. In addition, linked to the emergence of modern genomics and a systems-medicine approach, the classic “mechanism of action” has transitioned to a broader “signature”-based prediction achieving powerful insight to examine drug mechanism [84].

Potential mechanisms of action of combination drug therapy include complementary actions where 2 or more drugs target multiple targets along the same protein or pathway, anticounteractive actions where a drug targets the biologic response to a first drug and facilitating actions where the second drug promotes the activity of the first drug. Recent successful examples of combinatorial drug design include targeting multiple immune checkpoints in melanoma immunotherapy [85], the combination of two human epidermal growth factor receptor 2 targeting drugs, pertuzumab and trastuzumab with docetaxel in breast cancer [86] and the combination of a BRAF inhibitor (dabrafenib) and a MAPK/ERK kinase (MEK) inhibitor (trametinib) for melanoma [87].

Strategies for computational screening for combinatorial drug treatment are still being developed but include methods based on empirical models, for example network-based methods centered around the concept of gene neighborhood or methodologies based on physiological models in cases where the pathophysiology of a disease process is well understood [83]. Nevertheless, despite a better understanding of these approaches, multiple challenges remain. In particular our limited biological understanding of the complex biological systems involved and the increased potential of toxicity by using combinatorial drug therapy has in part limited this approach.

4. Expert commentary and five-year view

Big data-driven drug development has been gaining popularity and rapidly evolving over the past decades. The field started from proof-of-concept studies, followed by the development of new methodologies spanning multiple disciplines, accumulated experimental data for public use and testing, and is now further growing to cover more specific biological contexts and clinical scenarios for better applicability in real-world drug development challenges. Big data-driven drug development will facilitate identification of therapeutic strategies for rare

subtypes of common diseases, rare orphan diseases, and disease conditions specific to genetically minor populations. The direct-to-consumer genetic testing industry has generated multiple large-scale population-level genetic and clinical databases outside traditional medical institutions and clinical practice [88]. The ever-growing interest and involvement of the information technology (IT) industry is transforming the field by providing infrastructure for data storage, sharing, and analysis, even without the involvement of traditional medical institutions, via wearable devices directly connected to cloud-based commercial storage/analysis centers [89, 90]. For instance, as part of the Precision Medicine Initiative, the NIH is seeking to create a longitudinal cohort of 1 million or more Americans, with extensive characterization of biologic specimens and sophisticated exposure assessment using mobile devices or wearable sensors [91]. The integration of this data with the multitude of other data points generated for each individual may allow to achieve the goal of precision medicine, “to work together toward development of individualized care [92].”

Acknowledgments

Declaration of interest

This work was supported by the FLAGS foundation, the Nuovo-Soldati Cancer Research Foundation and an advanced training grant from Geneva University Hospital to N Goossens and NIH/NIDDK R01 DK099558 and the Irma T. Hirsch Trust to Y Hoshida.

References

1. Dickson M, Gagnon JP. Key factors in the rising cost of new drug discovery and development. *Nature reviews Drug discovery*. 2004; 3:417–429. [PubMed: 15136789]
2. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. *Nat Rev Drug Discov*. 2011; 10:428–438. [PubMed: 21629293]
3. DiMasi JA, Feldman L, Seckler A, Wilson A. Trends in risks associated with new drug development: success rates for investigational drugs. *Clinical pharmacology and therapeutics*. 2010; 87:272–277. [PubMed: 20130567]
4. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nature biotechnology*. 2014; 32:40–51.
5. Collins FS, Varmus H. A New Initiative on Precision Medicine. *N Engl J Med*. 2015
6. Golub T. Counterpoint: Data first. *Nature*. 2010; 464:679. [PubMed: 20360719]
7. Bender E. Big data in biomedicine. *Nature*. 2015; 527:S1. [PubMed: 26536215]
8. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. *Science*. 2011; 334:1518–1524. [PubMed: 22174245]
9. Good BM, Su AI. Crowdsourcing for bioinformatics. *Bioinformatics*. 2013; 29:1925–1933. [PubMed: 23782614]
10. Omberg L, Ellrott K, Yuan Y, Kandoth C, Wong C, Kellen MR, et al. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat Genet*. 2013; 45:1121–1126. [PubMed: 24071850]
11. Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical methods in medical research*. 2008; 17:347–388. [PubMed: 17698936]
12. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999; 401:788–791. [PubMed: 10548103]
13. Tamayo P, Scanfeld D, Ebert BL, Gillette MA, Roberts CW, Mesirov JP. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proc Natl Acad Sci U S A*. 2007; 104:5959–5964. [PubMed: 17389406]

14. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet.* 2015; 16:321–332. [PubMed: 25948244]
15. Cancer Genome Atlas Research N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine.* 2013; 368:2059–2074. [PubMed: 23634996]
16. Cancer Genome Atlas Research N, Brat DJ, Verhaak RG, Aldape KD, Yung WK, Salama SR, et al. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *The New England journal of medicine.* 2015; 372:2481–2498. [PubMed: 26061751]
17. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489:57–74. [PubMed: 22955616]
18. de Souza N. The ENCODE project. *Nature methods.* 2012; 9:1046. [PubMed: 23281567]
19. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. *Nature.* 2012; 489:75–82. [PubMed: 22955617]
20. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature.* 2012; 489:109–113. [PubMed: 22955621]
21. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. *Nature.* 2012; 489:101–108. [PubMed: 22955620]
22. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature.* 2007; 448:561–566. [PubMed: 17625570]
23. Solomon BJ, Mok T, Kim DW, Wu YL, Nakagawa K, Mekhail T, et al. First-line crizotinib versus chemotherapy in ALK-positive lung cancer. *The New England journal of medicine.* 2014; 371:2167–2177. [PubMed: 25470694]
24. Shaw AT, Kim DW, Nakagawa K, Seto T, Crino L, Ahn MJ, et al. Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *N Engl J Med.* 2013; 368:2385–2394. [PubMed: 23724913]
25. Gerber DE, Minna JD. ALK inhibition for non-small cell lung cancer: from discovery to therapy in record time. *Cancer Cell.* 2010; 18:548–551. [PubMed: 21156280]
26. Killock D. Lung cancer: a new generation of EGFR inhibition. *Nature reviews Clinical oncology.* 2015; 12:373.
27. Shaw AT, Friboulet L, Leshchiner I, Gainor JF, Bergqvist S, Brooun A, et al. Resensitization to Crizotinib by the Lorlatinib ALK Resistance Mutation L1198F. *The New England journal of medicine.* 2016; 374:54–61. [PubMed: 26698910]
28. Renfro LA, Mallick H, An MW, Sargent DJ, Mandrekars SJ. Clinical trial designs incorporating predictive biomarkers. *Cancer Treat Rev.* 2016; 43:74–82. [PubMed: 26827695]
29. Zang Y, Guo B. Optimal two-stage enrichment design correcting for biomarker misclassification. *Statistical methods in medical research.* 2015
30. Mehta C, Schafer H, Daniel H, Irle S. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Stat Med.* 2014; 33:4515–4531. [PubMed: 25130879]
31. Redig AJ, Janne PA. Basket trials and the evolution of clinical trial design in an era of genomic medicine. *J Clin Oncol.* 2015; 33:975–977. [PubMed: 25667288]
32. McNeil C. NCI-MATCH launch highlights new trial design in precision-medicine era. *J Natl Cancer Inst.* 2015; 107
33. Gannon HS, Kaplan N, Tsherniak A, Vazquez F, Weir BA, Hahn WC, et al. Identification of an "Exceptional Responder" Cell Line to MEK1 Inhibition: Clinical Implications for MEK-Targeted Therapy. *Mol Cancer Res.* 2015
34. Chen Z, Cheng K, Walton Z, Wang Y, Ebi H, Shimamura T, et al. A murine lung cancer co-clinical trial identifies genetic modifiers of therapeutic response. *Nature.* 2012; 483:613–617. [PubMed: 22425996]
35. Schork NJ. Personalized medicine: Time for one-person trials. *Nature.* 2015; 520:609–611. [PubMed: 25925459]
36. Mills JJ, Falls JG, De Souza AT, Jirtle RL. Imprinted M6p/Igf2 receptor is mutated in rat liver tumors. *Oncogene.* 1998; 16:2797–2802. [PubMed: 9652747]

37. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015; 526:75–81. [PubMed: 26432246]
38. Robinson D, Van Allen EM, Wu YM, Schultz N, Lonigro RJ, Mosquera JM, et al. Integrative clinical genomics of advanced prostate cancer. *Cell*. 2015; 161:1215–1228. [PubMed: 26000489]
39. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med*. 2015; 7:311ra174.
40. Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer biomarker discovery and validation. *Translational cancer research*. 2015; 4:256–269. [PubMed: 26213686]
41. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*. 2012; 148:1293–1307. [PubMed: 22424236]
42. Iwayanagi T, Miyamoto S, Konno T, Mizutani H, Hirai T, Shigemoto Y, et al. TP Atlas: integration and dissemination of advances in Targeted Proteins Research Program (TPRP)-structural biology project phase II in Japan. *Journal of structural and functional genomics*. 2012; 13:145–154. [PubMed: 22644393]
43. Watkins AM, Arora PS. Structure-based inhibition of protein-protein interactions. *European journal of medicinal chemistry*. 2015; 94:480–488. [PubMed: 25253637]
44. Caffrey M. A comprehensive review of the lipid cubic phase or in meso method for crystallizing membrane and soluble proteins and complexes. *Acta crystallographica Section F, Structural biology communications*. 2015; 71:3–18. [PubMed: 25615961]
45. Carroni M, Saibil HR. Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods*. 2016; 95:78–85. [PubMed: 26638773]
46. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, et al. Predicting new molecular targets for known drugs. *Nature*. 2009; 462:175–181. [PubMed: 19881490]
47. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*. 2012; 486:361–367. [PubMed: 22722194]
48. Lamb J. The Connectivity Map: a new tool for biomedical research. *Nat Rev Cancer*. 2007; 7:54–60. [PubMed: 17186018]
49. Roti G, Stegmaier K. Genetic and proteomic approaches to identify cancer drug targets. *British journal of cancer*. 2012; 106:254–261. [PubMed: 22166799]
50. Stegmaier K, Ross KN, Colavito SA, O'Malley S, Stockwell BR, Golub TR. Gene expression-based high-throughput screening(GE-HTS) and application to leukemia differentiation. *Nat Genet*. 2004; 36:257–263. [PubMed: 14770183]
51. Dudley JT, Sirota M, Shenoy M, Pai RK, Roedder S, Chiang AP, et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci Transl Med*. 2011; 3:96ra76.
52. Hieronymus H, Lamb J, Ross KN, Peng XP, Clement C, Rodina A, et al. Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell*. 2006; 10:321–330. [PubMed: 17010675]
53. Hahn CK, Ross KN, Warrington IM, Mazitschek R, Kanegai CM, Wright RD, et al. Expression-based screening identifies the combination of histone deacetylase inhibitors and retinoids for neuroblastoma differentiation. *Proc Natl Acad Sci U S A*. 2008; 105:9751–9756. [PubMed: 18607002]
54. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. *Briefings in bioinformatics*. 2016; 17:2–12. [PubMed: 25832646]
55. Power A, Berger AC, Ginsburg GS. Genomics-enabled drug repositioning and repurposing: insights from an IOM Roundtable activity. *JAMA*. 2014; 311:2063–2064. [PubMed: 24867009]
56. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006; 313:1929–1935. [PubMed: 17008526]

57. Comer E, Beaudoin JA, Kato N, Fitzgerald ME, Heidebrecht RW, Lee Mdt, et al. Diversity-oriented synthesis-facilitated medicinal chemistry: toward the development of novel antimalarial agents. *Journal of medicinal chemistry*. 2014; 57:8496–8502. [PubMed: 25211597]
58. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. *Science*. 2008; 321:263–266. [PubMed: 18621671]
59. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012; 483:603–607. [PubMed: 22460905]
60. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*. 2012; 483:570–575. [PubMed: 22460902]
61. Bachovchin DA, Koblan LW, Wu W, Liu Y, Li Y, Zhao P, et al. A high-throughput, multiplexed assay for superfamily-wide profiling of enzyme activity. *Nature chemical biology*. 2014; 10:656–663. [PubMed: 24997602]
62. Boehm JS, Golub TR. An ecosystem of cancer cell line factories to support a cancer dependency map. *Nature reviews Genetics*. 2015; 16:373–374.
63. Devillers J. Methods for building QSARs. *Computational Toxicology: Volume II*. 2013:3–27.
64. Roy, K.; Das. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. 1st. Academic Press; 2015.
65. Woo Y, Lai DY. OncoLogic: a mechanism-based expert system for predicting the carcinogenic potential of chemicals. *Predictive Toxicology*. 2005:385–413.
66. Marchant CA, Briggs KA, Long A. In silico tools for sharing data and knowledge on toxicity and metabolism: derek for windows, meteor, and vitic. *Toxicology mechanisms and methods*. 2008; 18:177–187. [PubMed: 20020913]
67. Patlewicz G, Jeliaskova N, Safford R, Worth A, Aleksiev B. An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR and QSAR in Environmental Research*. 2008; 19:495–524. [PubMed: 18853299]
68. Kirchmair J, Goller AH, Lang D, Kunze J, Testa B, Wilson ID, et al. Predicting drug metabolism: experiment and/or computation? *Nature reviews Drug discovery*. 2015; 14:387–404. [PubMed: 25907346]
69. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British journal of pharmacology*. 2007; 152:9–20. [PubMed: 17549047]
70. Smithing MP, Darvas F. HazardExpert: an expert system for predicting chemical toxicity. *ACS symposium series American chemical society*. 1992; 1992
71. Sung JH, Srinivasan B, Esch MB, McLamb WT, Bernabini C, Shuler ML, et al. Using physiologically-based pharmacokinetic-guided “body-on-a-chip” systems to predict mammalian response to drug and chemical exposure. *Experimental Biology and Medicine*. 2014 1535370214529397.
72. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, et al. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*. 2012; 486:361–367. [PubMed: 22722194]
73. Roncaglioni A, Toropov AA, Toropova AP, Benfenati E. In silico methods to predict drug toxicity. *Current opinion in pharmacology*. 2013; 13:802–806. [PubMed: 23797035]
74. Kuster A, Adler N. Pharmaceuticals in the environment: scientific evidence of risks and its regulation. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2014; 369
75. Jobling S, Williams R, Johnson A, Taylor A, Gross-Sorokin M, Nolan M, et al. Predicted exposures to steroid estrogens in U.K. rivers correlate with widespread sexual disruption in wild fish populations. *Environmental health perspectives*. 2006; 114(Suppl 1):32–39. [PubMed: 16818244]
76. Brodin T, Fick J, Jonsson M, Klaminder J. Dilute concentrations of a psychiatric drug alter behavior of fish from natural populations. *Science*. 2013; 339:814–815. [PubMed: 23413353]

77. Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Predicting multiple ecotoxicological profiles in agrochemical fungicides: a multi-species chemoinformatic approach. *Ecotoxicology and environmental safety*. 2012; 80:308–313. [PubMed: 22521812]
78. Kar S, Roy K. QSAR modeling of toxicity of diverse organic chemicals to *Daphnia magna* using 2D and 3D descriptors. *Journal of hazardous materials*. 2010; 177:344–351. [PubMed: 20045248]
79. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicological sciences : an official journal of the Society of Toxicology*. 2007; 95:5–12. [PubMed: 16963515]
80. Judson RS, Magpantay FM, Chickarmane V, Haskell C, Tania N, Taylor J, et al. Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor. *Toxicological sciences : an official journal of the Society of Toxicology*. 2015; 148:137–154. [PubMed: 26272952]
81. Law LW. Effects of combinations of antileukemic agents on an acute lymphocytic leukemia of mice. *Cancer research*. 1952; 12:871–878. [PubMed: 13009674]
82. Law LW. Origin of the resistance of leukaemic cells to folic acid antagonists. *Nature*. 1952; 169:628–629. [PubMed: 14929259]
83. Sun X, Vilar S, Tatonetti NP. High-throughput methods for combinatorial drug discovery. *Science translational medicine*. 2013; 5:205rv201.
84. Pritchard JR, Bruno PM, Gilbert LA, Capron KL, Lauffenburger DA, Hemann MT. Defining principles of combination drug mechanisms of action. *Proceedings of the National Academy of Sciences of the United States of America*. 2013; 110:E170–E179. [PubMed: 23251029]
85. Postow MA, Chesney J, Pavlick AC, Robert C, Grossmann K, McDermott D, et al. Nivolumab and ipilimumab versus ipilimumab in untreated melanoma. *The New England journal of medicine*. 2015; 372:2006–2017. [PubMed: 25891304]
86. Swain SM, Baselga J, Kim SB, Ro J, Semiglazov V, Campone M, et al. Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer. *The New England journal of medicine*. 2015; 372:724–734. [PubMed: 25693012]
87. Robert C, Karaszewska B, Schachter J, Rutkowski P, Mackiewicz A, Stroiakovski D, et al. Improved overall survival in melanoma with combined dabrafenib and trametinib. *New England Journal of Medicine*. 2015; 372:30–39. [PubMed: 25399551]
88. Skirton H, Goldsmith L, Jackson L, O'Connor A. Direct to consumer genetic testing: a systematic review of position statements, policies and recommendations. *Clin Genet*. 2012; 82:210–218. [PubMed: 22372614]
89. Eisenstein M. GSK collaborates with Apple on ResearchKit. *Nature biotechnology*. 2015; 33:1013–1014.
90. Egbring M, Kullak-Ublick GA, Russmann S. Phynx: an open source software solution supporting data management and web-based patient-level data review for drug safety studies in the general practice research database and other health care databases. *Pharmacoepidemiology and drug safety*. 2010; 19:38–44. [PubMed: 19777533]
91. Byrd JC, Devi GR, de Souza AT, Jirtle RL, MacDonald RG. Disruption of ligand binding to the insulin-like growth factor II/mannose 6-phosphate receptor by cancer-associated missense mutations. *J Biol Chem*. 1999; 274:24408–24416. [PubMed: 10446221]
92. Devi GR, De Souza AT, Byrd JC, Jirtle RL, MacDonald RG. Altered ligand binding by insulin-like growth factor II/mannose 6-phosphate receptors bearing missense mutations in human cancers. *Cancer Res*. 1999; 59:4314–4319. [PubMed: 10485478]

Key issues

- Accumulated drug knowledge bases, experimental data, clinical data, and exponentially expanding IT infrastructure now enable big data analysis to systematically explore new therapeutics.
- Omics data accumulation covering a variety of biological systems/ contexts and variations across human populations and individuals, is becoming a resource to demonstrate the proof-of-concept of Precision Medicine.
- Experimentally or computationally determined chemical, genetic, and proteomic interactions connect cross-domain knowledge, and facilitate drug discovery.
- Not only chemical/protein structures and interactions, but also transcriptome, enzymatic activity, ligands, other functional readouts, and even clinical toxicities are now available as tools for high-throughput identification of therapeutic targets and strategies.
- Drug toxicity prediction may bypass less-reliable rodent models, through the leveraging of big data-format clinical phenotypic information.
- Rapidly increasing involvement of direct-to-consumer genetic testing as well as information technology industries has been drastically accelerating the big data-based approach, further promoting individualized precision medicine and improved health.

Table 1

Resources for big data-driven therapeutic target identification

Type	Focus	Resource	Assay/datatype	Species	URL
Projects/initiatives	Genome, phenome, microbiome	Personal Genome Project (PGP)	DNA-seq, clinical phenotypes	Human	www.personalgenomes.org/Harvard
	Genomic DNA variations across populations	1000 Genomes	DNA-seq	Human	www.1000genomes.org/data
	Genomic DNA variations across populations	HapMap	SNP array	Human	hapmap.ncbi.nlm.nih.gov/downloads/index.html.en
	Regulatory elements in genome	The Encyclopedia of DNA Elements (ENCODE)	ChIP/RNA/DNA-seq, DNA microarray	Various	www.encodeproject.org
	Multi-omic data in cancer	The Cancer Genome Atlas (TCGA)	DNA-seq, RNA-seq, methylation array, clinical phenotypes	Human	cga-data.nci.nih.gov/tcga
	Multi-omic data in cancer	The International Cancer Genome Consortium (ICGC)	DNA-seq, RNA-seq, methylation array, clinical phenotypes	Human	dcc.icgc.org
	Molecular signatures of gain/loss of genes	Library of Integrated Network-Based Cellular Signatures (LINCS)	Bead-array with informatic inference	Human	www.lincsproject.org/data/tools-and-databases
	Expression/DNA variants across organs	Genotype-Tissue Expression (GTEx)	RNA-seq, SNP array	Human	www.gtexportal.org/home
	Expression/DNA variants in cancer cell lines	Cancer Cell Line Encyclopedia (CCLE)	Expression array, targeted DNA-seq	Human	www.broadinstitute.org/ccle
	Presence/variants of microorganisms in human	Human Microbiome Project (HMP)	DNA-seq, reference genome, clinical meta-data	Microorganisms	hmpdacc.org/resources/tools_protocols.php
Data repository	Somatic DNA mutations in cancer	Catalogue of Somatic Mutations in Cancer (COSMIC)	Various omic types from literature/databases	Human	cancer.sanger.ac.uk/cosmic
	Various	Gene Expression Omnibus (GEO)	Various omic types with or without clinical phenotypes	Various	www.ncbi.nlm.nih.gov/geo
	Various	ArrayExpress	Various omic types with or without clinical phenotypes	Various	www.ebi.ac.uk/arrayexpress
	Expression/DNA variations, phenotypes	European Genome-phenome Archive	Various omic types, clinical phenotypes	Human	www.ebi.ac.uk/ega/home
	Annotated gene sets	Molecular Signature Database (MSigDB)	Gene expression, knowledgebase, genomic/genetic structural	Various	www.broadinstitute.org/msigdb
	Expression/DNA variations, phenotypes in cancer	Oncomine	Various omic types with or without clinical phenotypes	Human	www.oncomine.org/resource/login.html
	Multi-omic data in mouse models	Mouse Genome Informatics (MGI)	Various omic types from literature/databases	Mouse	www.informatics.jax.org
	Clinically relevant DNA variations	Clinical Genome Resource (ClinGen)	Various DNA variant assays, clinical phenotypes	Human	www.clinicalgenome.org

Table 2

Resources for big data-driven drug identification.

Type	Focus	Resource	Assay/data type	Species	URL
Data repository	Chemical property/structure/ biological function	ChemBank	Chemical properties, phenotypic readouts	Various	chembank.broadinstitute.org
		PubChem	Chemical properties, phenotypic readouts	Various	pubchem.ncbi.nlm.nih.gov
		ChEMBL	Chemical properties, phenotypic readouts	Various	www.ebi.ac.uk/chembl
		Pharmacogenomics Knowledgebase (PharmGKB)	Drug annotations, drug-gene association	Human	www.pharmgkb.org
		KEGG DRUG	Drug annotations, drug-gene association		www.genome.jp/kegg/drug
		DrugBank	Drug and drug target information	Various	www.drugbank.ca
		FDA Orange Book	List of FDA-approved drugs	Human	www.accessdata.fda.gov/scripts/cder/ob
	Protein properties	UniProt/Swiss-Prot	Protein sequence, structure, function, ontology	Various	www.uniprot.org
		Protein Data Bank (PDB)	Protein sequence, structure, function, ontology	Various	www.pdb.org
		Max-Planck Unified Proteome Database (MAPU)	Protein sequence, structure, function, ontology	Human, mouse	mapuproteome.com
		ExPASy	Protein sequence, structure, function, ontology	Various	www.expasy.org
		Protein Information Resource (PIR)	Protein sequence, structure, function, ontology	Various	pir.georgetown.edu
		CATH	Classification of protein structures	Various	www.cathdb.info
		Plasma Proteome Database (PPD)	List of plasma/serum proteins	Human	www.plasmaproteomedatabase.org
	Protein-protein/chemical/genetic interaction	Database of Interacting proteins (DIP)	Experimentally determined protein-protein interaction	Various	dip.mbi.ucla.edu/dip
		STITCH	Chemical-protein interaction	Various	stitch.embl.de
		BioGRID	Protein, genetic, chemical interactions/associations	Various	thebiogrid.org
		Therapeutic Target Database (TTD)	Protein, genetic, chemical interactions/associations	Human	database.idrbc.cqu.edu.cn/TTD
		Potential Drug Target Database (PDTD)	Protein, genetic, chemical interactions/associations	Human	www.dddc.ac.cn/pdtd

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Type	Focus	Resource	Assay/data type	Species	URL
Tools		NIH Small Molecule Repository	Compound libraries for screening	-	nslsmr.evotec.com/evotec
		ViCi	<i>in silico</i> ligand-based drug design	-	www.embl-hamburg.de/vici

Table 3

Resources for computational prediction of drug toxicity

Focus	Resource	Assay/Data type	URL
Computational prediction of drug toxicity	Derek Nexus	Toxicological profile based on structure	www.hasalimited.org/products/derek-nexus.htm
	ToxTree	Toxicological profile based on structure	toxtree.sourceforge.net
	HazardExpert	Toxicity profile based on toxic fragments	www.computdrug.com/hazardexpertpro
	TOPKAT	Toxicological profile based on structure. Predicted ADMET properties.	accelrys.com/products/collaborative-science/biovia-discovery-studio/qsar-admet-and-predictive-toxicology.html
Database of bioactive molecules and target	CASE Ultra	Toxicological profile based on structure	www.multicase.com/case-ultra
	OECD QSAR	Toxicological profile based on structure	www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm
	ChEMBL	Chemical properties, phenotypic readouts	www.ebi.ac.uk/chembl
	DrugBank	Drug and drug target information	www.drugbank.ca
Toxicology databases	Drugs@FDA Database	FDA-approved drugs	www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm
	PubChem	Chemical properties, phenotypic readouts	pubchem.ncbi.nlm.nih.gov
	SWEETLEAD	Chemical structure database	sintk.org/home/sweetlead
	The NCGC Pharmaceutical Collection (NPC)	Chemical structure database	tripod.nih.gov/npc
	Chemical Entities of Biological Interest (ChEBI)	Small compound database	www.ebi.ac.uk/chebi
	SIDER	Adverse drug reaction database	sideeffects.embl.de
	Comparative Toxicogenomics Database	Chemical-protein and chemical-phenotype interactions	ctdbase.org
	Environmental Protection Agency Aggregated Computational Toxicology Resource (ACToR)	Chemical toxicity database	actor.epa.gov
	FDA Adverse Event	Adverse drug reaction database	www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects
	OpenTox	Multiple toxicological resources (data, computer models, validation and reporting)	www.opentox.org
Chemical property/structure	Pharmacogenomics Knowledgebase (PharmGKB)	Drug annotations, drug-gene association	www.pharmgkb.org
	T3DB	Toxin-target database	www.t3db.ca
	TOXNET	Toxicology resources	toxnet.nlm.nih.gov
	ChEMBL	Chemical properties, phenotypic readouts.	chembank.broadinstitute.org