Published in final edited form as:

Expert Opin Drug Discov. 2016 ; 11(3): 241–256. doi:10.1517/17460441.2016.1135126.

Providing data science support for systems pharmacology and its implications to drug discovery

Thomas Hart^{1,2} and Lei Xie^{3,4,*}

¹The Rockefeller University, New York, New York, United States of America

²Department of Biological Sciences, Hunter College, The City University of New York, New York, New York, United States of America

³Department of Computer Science, Hunter College, The City University of New York, New York, New York, United States of America

⁴The Graduate Center, The City University of New York, New York, New York, United States of America

Abstract

Introduction—The conventional one-drug-one-target-one-disease drug discovery process has been less successful in tracking multi-genic, multi-faceted complex diseases. Systems pharmacology has emerged as a new discipline to tackle the current challenges in drug discovery. The goal of systems pharmacology is to transform huge, heterogeneous, and dynamic biological and clinical data into interpretable and actionable mechanistic models for decision making in drug discovery and patient treatment. Thus, big data technology and data science will play an essential role in systems pharmacology.

Areas covered—This paper critically reviews the impact of three fundamental concepts of data science on systems pharmacology: similarity inference, overfitting avoidance, and disentangling causality from correlation. The authors then discuss recent advances and future directions in applying the three concepts of data science to drug discovery, with a focus on proteome-wide context-specific quantitative drug target deconvolution and personalized adverse drug reaction prediction.

Expert opinion—Data science will facilitate reducing the complexity of systems pharmacology modeling, detecting hidden correlations between complex data sets, and distinguishing causation from correlation. The power of data science can only be fully realized when integrated with mechanism-based multi-scale modeling that explicitly takes into account the hierarchical organization of biological systems from nucleic acid to proteins, to molecular interaction networks, to cells, to tissues, to patients, and to populations.

*To whom correspondence should be addressed, lei.xie@hunter.cuny.edu.

Financial and Competing Interests Disclosure

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.

1. Introduction

The conventional one-drug-one-target-one-disease drug discovery process has been less successful in tracking multi-genic, multi-faceted complex diseases. We lack fundamental knowledge of the mechanisms that drive the development, persistence, and transformation of complex diseases. Furthermore, a drug's efficacy and side effects depend on each individual's genetic and environmental backgrounds. Drug designers remain ignorant of both the causal genetic underpinning of human pathophysiology and pharmacology and a complete picture of the complex interplay of genetic, molecular, and environmental components. This lack of understanding underlies the current innovation gap in drug discovery [1].

Quantitative systems pharmacology (QSP) [2] and structural systems pharmacology (SSP) [3] have emerged as new disciplines to tackle the current challenges in drug discovery. The goal of QSP is “to understand, in a precise, predictive manner, how drugs modulate cellular networks in space and time and how they impact human pathophysiology. QSP aims to develop formal mathematical and computational models that incorporate data at several temporal and spatial scales; these models will focus on interactions among multiple elements (biomolecules, cells, tissues etc.) as a means to understand and predict therapeutic and toxic effects of drugs” [2]. SSP adds a new dimension to systems pharmacology. The aim of SSP is to understand the atomic details and conformational dynamics of molecular interactions in the context of the human genome and interactome, and to link them systematically to the human drug response under diverse genetic and environmental backgrounds [3]. Thus systems pharmacology modeling holds great potential to reduce the attrition rate of drug discovery, to enhance drug safety in the clinic, and to develop precision medicine.

The holy grail of systems pharmacology (both QSP and SSP) is to integrate biological and clinical data, and to transform them into interpretable and actionable mechanistic models for decision making in drug discovery and patient care. Biological and clinical data have the same characterizations of big data that are defined as volume, variety, velocity, and veracity. In terms of volume, advances in high-throughput techniques have generated unprecedented amounts of omics data. These data are across the hierarchical organizations of an organism (molecule, pathway, cell, tissue, organ, patient, and population), across a wide spectrum of time scales, and across multiple species. Thus they are in the form of high variety. Furthermore, the biological response to drug perturbation is dynamic. For example, cancer cells, bacteria and viruses can evolve rapidly to gain drug resistance. Systems pharmacology modeling should take the velocity of drug response into account. Finally, in terms of veracity, systems pharmacology must not only consider the signal to noise ratio of the various experimental methods and datasets, but also incorporate noise and stochasticity into its models, as they are an intrinsic property of biological processes [4].

These huge, complex, heterogeneous, dynamic, and noisy data offer great opportunities for systems pharmacology modeling, but impose great challenges in data management, data processing, data mining, and knowledge discovery. Cloud computing-based data processing technologies have significantly enhanced our capability for handling big data. With the high

availability of processed and organized data, the next challenge in systems pharmacology is how to use these big data to build interpretable and actionable computational models that are able to support decision making in the process of drug discovery and development. Data science, as an emerging discipline that supports the extraction of information and knowledge from data on top of data processing technology, will play a significant role in harnessing big data for systems pharmacology, ultimately supporting the whole drug discovery process (Figure 1). This review will focus on the application of data science to systems pharmacology. First, the three fundamental concepts of data science and their impacts on systems pharmacology will be critically reviewed. Then recent advances and future directions in applying data science to drug discovery, particularly drug target deconvolution and adverse drug reaction (ADR) prediction, will be discussed.

2. Fundamental concepts of data sciences and their impacts on systems pharmacology

The existing paradigm of systems pharmacology is centered around data-driven network-based association studies, mathematics-based modeling of biological networks, and physics-based modeling of the energetics and dynamics of molecular interactions. The primary challenge of data-driven network-based association studies is that the number of observations is often much smaller than the number of variables or parameters. It is not trivial to establish causal relationships from correlations due to a large number of confounding factors. The observed correlation often neither robustly predicts new instances (due to overfitting) nor provides biological insight into the underlying molecular and cellular mechanisms. On the other hand, current mathematics- and physics-based multi-scale models are often too complex to be supported by existing data to readily model the global behavior of the physiological system. In spite of tremendous efforts in developing specialized hardware [5] and utilizing cloud computing environments [6,7], the molecular dynamic simulation of molecular interaction on a large scale is still hindered by its high computational cost. As a result, it is yet infeasible to model the detailed kinetic process of drug-target binding/unbinding on a large scale.

Advances in big data technology and data science have provided new opportunities to address the aforementioned challenges in systems pharmacology. Although data science lacks clear boundaries as an emerging discipline, and is directly related to the broad field of statistical learning and data mining, it bears a set of distinguishing principle concepts that guide data-driven modeling and decision making. As summarized by Provost and Fawcett [8], these concepts include: 1) Similarity inference, 2) Overfitting avoidance, and 3) Correlation vs causality. In the following section, we will detail the application of each of these concepts to advancing network-based association studies and mechanism-based molecular and network modeling. Ultimately, these advances will facilitate the development of systems pharmacology, which will have major impacts on drug discovery.

2.1 Similarity as a foundation of systems pharmacology

The concept of similarity is the foundation of bioinformatics and chemoinformatics. For two biological entities (e.g. proteins), if their known attributes (e.g. sequences) are similar, then

their unknown attributes (e.g. structures) are similar. Extending the similarity concept to system level measurements (e.g. high-dimensional data, graph representation etc.) and using multi-faceted similarity metrics to integrate heterogeneous data is pivotal to the advancement of systems pharmacology.

As shown in Figure 2, the problem of detecting the association between two biological entities A and B can often be formulated first as a heterogeneous graph linking them together. The association graph commonly has two types of edges: edges representing the known positive (or negative) associations between two different entities, and edges representing similarity or interaction between those same entities. Taking advantage of advances in data science, a number of novel similarity metrics for small molecule, sequence, structure, biological network, molecular phenotype, and organism phenotype have emerged. For example, the data fusion of several chemical structure fingerprints enables the prediction of binding affinity based on chemical similarity [9]. Lhota et al. have developed a new similarity framework, Enrichment of Network Topological Similarity (ENTS), to relate similarities of different attributes of biological entities, and to assess the statistical significance of similarity measurements [10]. ENTS has demonstrated superior performance in linking protein sequence similarity to its structural similarity. Amaratunga et al. proposed a resampling-based similarity measurement for high-dimensional data [11]. This can be a powerful tool for measuring the similarity of omic data such as transcriptome profiles. Several text mining techniques such as Word2Vec has been developed to represent words as a vector [12]. The similarity between word vectors can be used to capture the semantic similarity of biological concepts extracted from electronic medical records and biomedical literature. These heterogeneous similarity metrics for different data types can be integrated through multi-kernel learning, Random Forest, and other machine learning techniques. For example, multi-kernel learning that integrates multiple profiling data achieved the best performance in the DREAM challenge for anti-cancer drug sensitivity prediction [13]. Gray et al. applied a random forest-based similarity method to combine similarities from multiple modalities including high-dimensional imaging data [14]. They demonstrated that using the joint embedding of features from multiple modalities improved the classification of Alzheimer's diseases. With the availability of a similarity metric, data mining techniques such as collaborative filtering [15,16], graph mining (e.g. random walk [17], k diverse shortest paths [18], meta-path [19] etc.), and kernel-based learning [20] can be used to predict new associations from known associations.

In the case of mechanism-based modeling, the similarity between molecular entities and their interactions can be derived from evolutionary, biophysical, spatial, and other types of attributes. Their similarity relationship can serve as a constraint to significantly reduce the search space of mathematical parameters and conformational dynamics in the mechanism-based model. For example, comparison of the electrostatic potentials of structurally similar proteins may assist in estimation of the kinetic parameters [21], as electrostatic potential in the active site determines not only the stability of transition states but also the diffusion association rate. In principle, many properties of biological systems (e.g. enzyme reaction rate, binding free energy etc.) can be accurately calculated using quantum mechanics (QM), the theory of physical processes at the atomic scale. However, solving Schrödinger's equation, and even its numerical approximations, is notoriously computationally intensive. It

is only applicable to small systems. Recently, kernel similarity has been applied to scale up QM calculations [22]. The basic idea is to only solve numerical solutions for a limited number of points in the property space of interest. Kernel-based machine learning techniques are then applied to interpolate them to the whole property space. Using kernel ridge regression, Dral et al. (2015) add corrections to computationally inexpensive approximations of QM, and produce highly accurate predictions of enthalpies, free energies, entropies, and electron correlation energies of molecules [23]. These studies open the door to the large scale application of quantum mechanics to systems pharmacology.

Several challenges remain in applying similarity metrics to systems pharmacology. The similarity measure is often entangled with other computational components. One critical issue in the similarity measure is how to represent the feature space. Context-specific feature selection or dimensional reduction may be required when handling high-dimensional data, which are common in systems pharmacology. Moreover, in many problems applied to systems pharmacology, the linear representation of the feature space (e.g. genes) may be not sufficient due to the interactions between biological components (e.g. epistasis). Thus the nonlinear information encoding of biological spaces of interests is an important topic [24]. We will discuss the problem of feature selection and representation in more detail in the following sections. With the availability of hundreds of existing similarity metrics [25] and the continuing emergence of new ones, it is not straightforward to determine which similarity metric is the most suitable for a specific data set and problem of interest. An efficient, objective evaluation is needed to automatically select the best similarity metrics (or kernel). If training and testing data are large enough (unfortunately, this is often not the case in systems pharmacology), computational tools (e.g. Spearmint [26]) that are able to run experiments automatically to optimize the parameters of the problem space for a defined objective function will be of great interest. With the rapid advance of data science, it is expected that new developments in similarity measure and the related techniques of feature representation, dimension reduction, and optimization will further enhance the power of network-based association studies to detect novel drug-target, gene-disease, chemical-phenotype, and other associations, as well as reduce the complexity of mathematics- and physics-based modeling.

2.2 Avoidance of overfitting for predictive modeling

Integration and analysis of multiple omics data is one of central tasks of systems pharmacology. The high-dimensional nature of omics data often leads to the $n \ll p$ problem, where the number of observations n is much smaller than the number of variables or parameters p . Moreover, a complex phenotype is often associated with interactions among multiple molecular components, any of which alone is not sufficient to drive phenotypic change. For example, the genetic underpinning of asthma is a set of interacting genes. The GWAS p -value of each gene against the background of random variation is modest [27]. A non-linear model can in principle be more powerful, but demands a more complex parameter space than a linear model. Thus, overfitting can be a serious concern when developing a predictive model in the systems pharmacology framework; that is, a model may fit the observed data very well but not be generalizable beyond the observed data.

Overfitting problems could be addressed with new techniques in statistical learning. In addition to using common practices such as cross-validation to overcome overfitting, a number of new methods for feature selection [28] and dimensional reduction [29] have emerged in recent years. For example, sparse coding has been proposed to reduce overfitting while at the same time improving the accuracy of support vector machines when using high-dimensional omics data [30]. Taking advantage of algorithms to avoiding overfitting, including unsupervised pre-training, regularization (L1 and L2), max norm constraints, and dropout, the machine learning field has taken a notable renewed interest in artificial neural networks (ANNs). Built from ANNs, deep learning algorithms can (1) use multiple layers or stages of ANNs for nonlinear information processing, and (2) extract abstract representations at successively higher, more abstract layers from raw data without human annotations [31]. Deep learning emulates the human brain's ability to observe, analyze, learn, and make decisions in order to solve extremely complex problems. The abstract representation produced from deep learning generates global relationships beyond immediate neighbors in the data. It can be invariant to local changes. These characteristics of deep learning may be very useful in handling biological data that are non-linear and stochastic in nature. Thus deep learning could be a powerful tool for predictive modeling in systems pharmacology. Although it is relatively new, deep learning has already demonstrated promising results in addressing biological problems. For example, Chen et al developed a bimodal deep brief network to predict human responses to stimuli from rat responses to the same stimuli [32]. More example on predicting drug-target interactions will be given in the section 3.1. In spite of these successes, the full power of deep learning remains to be seen in systems pharmacology. In addition to the high computational costs associated with the deep learning, the fundamental challenge is how to train a model from high-dimensional, sparse, and heterogeneous data. For example, new methods are needed to extract the learning patterns and relationships from combined chemical, genomic, epigenomic, transcriptional, and other omics data, and correlate them with personalized drug responses. It will be interesting to incorporate other machine learning techniques (e.g. kernel similarity) and multi-scale modeling that explicitly represents the hierarchical structure of organisms (as detailed in the next section) into the framework of deep learning.

Context-specific biological mechanisms must be embedded in statistical learning to avoid overfitting and select biologically meaningful features and rules [33]. This is directly relevant to identifying the causal factors of an event of interest, as discussed in the following section. One of the best known approaches is Gene Set Enrichment Analysis and its many variants [34,35]. The recent development of topology-based biological pathway analysis methods [36], such as PARADIGM [37], SPIA [38], and EnrichNet [39] etc., may provide a powerful set of tools for predicting drug response phenotypes and identifying therapeutic biomarkers. For example, it has been shown that PARADIGM improves the prediction of anti-cancer drug sensitivity and resistance [40]. Based on a simple mechanistic DIGRE (drug-induced genomic residual effect) model of synergistic drugs, which used cancer-related KEGG pathways as features, Yang et al. achieved the top performance of anti-cancer activity prediction of drug combinations in one of the DREAM challenges [41]. One of the fundamental problems of GSEA-related methods is that existing biological pathways are highly biased and not context-specific. Cross-talk between pathways and dynamic rewiring

of biological networks can be a major driver of phenotypic changes. Thus, GSEA may miss novel mechanisms of drug actions. Identification of functional modules from an incomplete but less biased interactome may identify novel causal genetic factors and pathways for investigation into pathophysiology and drug response [27,42]. Hofree et al. have developed a network-based stratification (NBS) method that integrates somatic tumor genomes with gene interaction networks. NBS demonstrated excellent performance in the prediction of clinical outcomes such as patient survival, response to therapy, and tumor histology [43]. Of course, the resolution of network analyses may be not high enough to distinguish the functional impact of subtle genetics alterations, environmental changes, and drug perturbations. As the physical and chemical principles of molecular interactions govern the biological processes, the functional prediction of protein structure may provide valuable information for pharmacogenomics and disease-gene associations. For example, SNPinfo uses the protein structure to identify functional variations for SNPs not in high linkage disequilibrium with any GWAS SNP [44]. Xie et al. developed a multi-scale modeling framework to integrate protein structural analysis and biological network analysis for the identification of causal mutations from GWAS data [45]. Porta-Pardo and Godzik used structural information to prioritize cancer driver mutations [46]. The forthcoming challenge is to integrate data-driven predictive modeling and mechanism-based modeling into a unified framework.

2.3 Distinguishing causation from correlation

Statistical learning is a powerful technique to identify correlations of known events X with the event of interests Y for predictive modeling. However, correlation can be confused with causation due to confounding factors, selection bias, and transportability bias. A confounding factor is an unobserved common cause that influences both X and Y . Selection bias is caused by use of a non-representative sample or bias in data acquisition of X and Y . Transportability bias arises when the population from which data are acquired is different from the one for which the inference is intended. Thus, causal conclusions cannot be simply drawn from correlations without clearly understanding the underlying assumptions and inspecting possible sources of confounding factors.

It is extremely difficult to distinguish causation from correlation when observation are derived from complex biological phenomena. The lack of understanding of causal genetic factors and disease genes may account for the current failure in drug discovery [1]. For example, the correlation of low HDL with heart disease raises a great interest in developing HDL-targeted therapy for heart disease (e.g. CETP inhibitors). However, the causal relationship between HDL and heart disease is questionable. It is not conclusive whether HDL causes healthy heart or a healthy heart produces HDL [47]. Such uncertainty casts doubts on the development of CETP inhibitors that have failed in several clinical trials. Thus, the misinterpretation of correlation as causation may waste huge amounts of resources in drug discovery.

In principle, the causal relationship can be identified using controlled experiments. In practice, in the context of systems pharmacology it is often infeasible to inspect all possible confounding factors necessary to thoroughly control the experiment. Many factors contribute to selection bias and transportability bias, both of which are quite common in systems

pharmacology. For example, a group of patients of interests may be different from the group of patients studied due to variability in genes, environments, and life style. Such selection bias could be a serious concern for anti-cancer drug development due to the high heterogeneity of cancer. One of the fundamental questions in systems pharmacology is how to link *in vitro* drug potency to *in vivo* drug activity, and how to extrapolate the drug response in animal models to that in humans. Solving these problems requires effective methods to reduce the transportability bias. The transportability bias also arises when the target event is inferred from the integration of heterogeneous data from different hierarchies, different species, and different time-scales of biological system, which is a common practice in systems pharmacology.

Recent new developments in data science and high performance computing may alleviate the challenge in disentangling causation and correlation. Mooij et al has proposed a method to distinguish cause from correlation from observational data [48]. The idea underlying the method is simple: If event X causes event Y , then the random noise in the event X will be reflected in the event Y . Bareinboim et al. introduced graphical and algebraic methods to mitigate and even eliminate selection bias [49,50]. They further propose an algorithmic framework to apply these methods to transportability bias [51], and extend them to the integration of heterogeneous data [52].

The above statistical formulizations are not magic bullets to solve the problems of disentangling causation from correlation. They only work under specific conditions. Bareinboim's framework requires conditional independences between subsets of observed variables. Mooij's method assumes that there are no confounding factors, selection bias, or feedback, although the conditional independence between observed variables is not required. In the context of systems pharmacology, the observed association of genetic and environmental factors with drug response phenotypes involves many confounding factors, selection bias, and transportability bias, as described before. The conditional independence condition does not always hold, as many dependent variables may be unknown.

One possible approach to reducing the number of confounding factors and dependent variables is to decompose a big problem into a set of small sub-problems. In each sub-problem, the number of confounding factors and dependent variables are minimal. Blois's scalar theory of biomedical information may provide a general framework to decompose the problem in systems pharmacology [53]. The fundamental concept of the scalar theory is that the complex phenotypes observed at a higher scale come from the emergent properties at a lower scale that has an intermediate phenotype (or mesophenotype), as shown in Figure 3. Based on the scalar theory, an endpoint phenotype (e.g. disease or drug response) is caused by individual-specific pathways whose observable variables could be genome-wide signatures such as gene expression profiles. The genome-wide signature results from the molecular machinery in the cell, which is characterized by molecular interactions. The molecular interaction is predominantly determined by the geometric, dynamic, and physiochemical properties of biomolecules. These properties can be changed by genetic alterations. The identification of causal genetic or environmental factors of the drug response or disease phenotype now consists of multiple steps to identify the effect of the emergent properties from the lower scale of a mesophenotype. This is different from the current

paradigm that directly links the bottom to the top of the phenotype hierarchy, but bypasses the intermediate phenotypes. Under the framework of multi-scale modeling, data-driven analysis may distinguish causal effects from confounding factors for each mesophenotype. Aligned with this concept, Gerstung et al. recently developed statistical models to untangle the effect of genetic alterations on gene expression, clinical variables, and outcomes. Combining features from mutation and gene expression improves the performance of predictive modeling [54]. Along the same lines, the causal effect of chemical structures on phenotype response could be established by modeling proteome-wide drug-target interactions [55].

In spite of the ability for multi-scale modeling to reduce the complexity of the problem, the challenge to infer causal-effect relationships still remains. In the case of mathematical modeling and simulation of molecular interaction networks, the parameters of the model must be trained from experimental data. However, because experimental data are commonly inadequate, the trained model is often non-identifiable, meaning that there is not a unique parameter set to explain the experimental observations. Increasing the amount of data for model training may not help in solving this problem, because existing fitting procedures are 'sloppy' in nature [56]. It has been proposed that approximate Bayesian computation may be employed for model selection [57,58]. Even when using efficient algorithms, the computational resources required for model selection can be massive. High performance computing and community efforts are needed to fulfill the central task in systems pharmacology: discovery of the genetic and molecular underpinnings of pathophysiology and drug action, as suggested by An [59].

3. Case studies of data science applications to drug discovery

The three fundamental principles of data science discussed previously often work together in statistical learning and data-driven decision making. They have penetrated into all aspects of the drug discovery process, from target identification to lead optimization, from ADMET (absorption, distribution, metabolism, excretion, and toxicity) evaluation to clinical trials. In the following section, we will discuss two cases that showcase the impact of data science on drug discovery, namely proteome-wide quantitative drug-target interaction prediction and personalized prediction of adverse drug reactions. In addition to briefly reviewing the state-of-the-art techniques in addressing these problems, we will analyze unsolved problems and propose future directions.

3.1. Proteome-wide context-specific quantitative drug-target interaction prediction

Phenotypic screening and target-based screening are the two major approaches in the early stage of drug discovery. Phenotypic screening is a strategy to identify molecules that cause a desirable change in phenotype in cellular or animal disease models without prior understanding of the molecular mechanism of action. Only after the molecules have been discovered, an effort is made to determine the biological target of the molecules, so that lead compounds can be optimized. Target-based screening is hypothesis-driven. Only drug candidates binding to a pre-designated protein target are assessed. A drawback of target-based screening is that the answer to the hypothesis may not be relevant to the disease

pathogenesis or provide a sufficient therapeutic ratio [2]. Unexpected drug off-target interaction often occurs, which is responsible for either therapeutic effects or side effects [60]. Therefore, fishing the molecular targets of lead compounds on a proteome scale will be critical for designing efficient and safe therapeutics using a systems pharmacology approach.

The general strategy for drug-target interaction predictions has been presented in several previous reviews [61,62]. Recent advances in data science, chemical genomics, and polypharmacology provide new opportunities for improving the performance of computational drug-target interaction predictions. One of the fundamental assumptions in applying statistical learning methods to drug-target interaction prediction is that similar chemicals bind to similar protein targets and induce similar phenotypic responses such as side effects or differentially-expressed genes, and vice versa. Based on this similarity principle, both semi-supervised and supervised machine learning techniques have been applied to predict drug-target interactions. Due to the lack of high quality negative interaction data, semi-supervised learning that uses positive interaction data only has been widely applied. The semi-supervised learning methods either build statistical models for the k nearest neighbors (k -NN) of the query compound with similar compounds in the database (e.g. Parzen-Rossette Window [63], Set Ensemble Analysis (SEA) [64], etc.) or explore a heterogeneous network of chemical and target space using graph mining, as discussed in the previous section. By using randomly generated or algorithmically derived negative cases [65], the drug-target interaction prediction can be formulated as a supervised classification problem. When the data points are sufficiently large, wide and deep Artificial Neural Network (ANN) outperformed random forests, gradient boosted decision tree ensembles, logistic regression, and several other state-of-the art methods in QSAR modeling [66,67] and compound virtual screening [68], and achieved the top performance in the compound toxicity prediction challenge [69]. The forthcoming challenge for deep learning based method is to incorporate protein relationships into the learning framework. Using individual or combined similarity measurements (or kernels) of both chemical structures and protein features, a number of techniques such as Regularized Least Squares (RLS) classifier [70,71] and Kernelized Bayesian Matrix Factorization with Twin Kernels (KBMF2K) [72] have the potential to integrate chemical and genomic space. More recently, a novel method that analyzes drug induced differential gene expressions in the context of protein-protein interaction network has been developed to predict drug-target associations [73].

In spite of these advances, the existing drug-target interaction prediction has limited coverage of the proteome, relies on unrealistic assumption of protein-ligand interactions [74], and ignores the causal and context-specific relationships between molecular interactions and systematic responses. Big data technology and data science will play critical roles in addressing the aforementioned drawbacks in the drug-target interaction prediction. Moreover, mechanism-based modeling is needed to enhance the power of data science.

The existing bioassay data are highly biased to thousands of pharmaceutically-investigated drug targets, which only cover a small portion of the human and pathogen proteomes. Many uncharacterized proteins may play key roles in mediating therapeutic or side effects of drugs. Thus, drug-target prediction should be extended to whole human and pathogen proteomes. This requires novel techniques to measure chemical and protein similarities, to

represent chemical and protein space, and to scale up data mining algorithms to millions of chemicals and proteins.

The similarity-based method is on center stage of drug-target prediction. The underlying assumption of the similarity-based method is that the chemical similarity and target similarity are continuously correlated with drug-target interaction similarity. There is a serious concern with the existing representation of protein similarity, which is usually based on global protein sequence/domain similarity or Gene Ontology (GO) similarity. Proteins with similar sequence or GO do not necessarily bind to similar chemicals, as protein-ligand interaction is governed by the spatial organization of amino acid residues in the protein structure [75]. A single amino acid mutation or post-translational modification may alter the binding of the ligand through direct modification of the ligand binding site or allosteric interaction. Furthermore, different conformations of ligand binding sites on the same protein (e.g. agonist conformation vs antagonist conformation) will allow the protein to interact with ligands with different structures (e.g. agonist vs antagonist). A protein may also consist of multiple binding sites (e.g. ATP binding site and allosteric site in kinases) that accommodate different types of ligands. On the other hand, two non-homologous proteins can bind to the same ligand if they have a similar ligand binding site [76,77]. The similarity between those residues that determine the binding specificity and promiscuity may provide a more sensible measure of protein-protein similarity for the prediction of drug-target interaction [75]. The binding residue similarity allows us to detect ligand binding promiscuity across gene families [78-83]. This is important for understanding drug actions across gene families. For example, a drug whose primary target is not an ion channel may still bind to an anti-target such as hERG, leading to serious side effects. New feature selection methods in data science are needed to identify critical binding residues that are often not continuous in the sequence [75]. In addition, new similarity measures are required to determine the similarity between binding residues, which can account for both binding promiscuity and specificity [84].

Chemical structural similarity is the most commonly used metric for similarity in chemical compounds. Although a large number of 2D and 3D fingerprints of chemical structures have been developed, the correlation of chemical structure similarity, calculated from existing structural fingerprints, with the binding similarity is not continuous. There exist activity cliffs in the chemical space, i.e. a small modification of chemical structure can lead to a dramatic change in binding activity [85], as protein-ligand interaction depends on both ligand conformation and receptor conformation. Furthermore, chemical similarity should correlate with the *proteome-wide* target binding profile instead of that of a single target interaction in a specific context. In this regard, the similarity between phenotypic signatures of chemicals, such as network perturbation profile [73], transcriptional profile [86] and cell imaging [55] etc. should be a better measure of chemical similarity, as they result from the collective effects of proteome-wide drug-target interactions. New deep learning techniques that can learn non-linear, hierarchical relationships will play increasingly important roles in representing chemical space [87].

Most semi-supervised and supervised machine learning methods consider the drug-target interaction as having only two states: an interaction occurs or it does not. This simple treatment is not sufficient to capture the quantitative nature of drug-target interactions. In the

context of systems pharmacology, both theoretical and experimental evidence suggests that multiple weak bindings may have significant influence on a drug's activity [60]. Furthermore, the impact of drug-target interactions on therapeutics and side effects depends not only on how tightly the drug binds to its targets, which is characterized by K_d , K_i or IC_{50} , but also how long the drug can remain bound to its targets, which is measured by drug-target residence time ($1/k_{off}$). A drug with low binding affinity may have a long residence time. Thus it is important to extend drug-target interaction predictions to (1) weakly interacting compounds as well as inactive compounds; and (2) predictions of combined k_{on} and k_{off} ($K_d = k_{off}/k_{on}$). As the dynamic process of protein-ligand interaction is governed by the basic principles of physics, mechanism relevant features are critical for the success of data-driven modeling. Recently, Chiu et al. has proposed a high-throughput predictive modeling framework to select key pairwise interactions between ligands and receptors, which are potential determinants of the protein binding/unbinding process, and use them to predict kinetic constants k_{on} and k_{off} [88]. One of the key components of their method is the use of conformational dynamic features derived from Normal Mode Analysis (NMA), a coarse-grained model of molecular dynamics along with energetic features calculated from force fields. These physical features are then used to train a multi-task machine learning model.

It is important to integrate drug response omics data at the system level with physical drug-target interaction data at the molecular level. Such integration will not only alleviate the problem of data bias but also is important for linking *in vitro* drug potency with *in vivo* drug activity [60]. Drug-target interaction is tissue-, context, and concentration-dependent. Proteins not only have different expression levels but also may present different isoforms in different tissues. Variations in an individual's genetics, environment, and behavior also contribute to differences in protein expression and post-translational activity. A precise identification of drug interactions with specific protein isoform(s) expressed in each tissue under diverse genetic and environmental factors is critically important for systems pharmacology modeling of drug actions. Recent effort in the reconstruction of tissue-specific protein functional networks may provide new opportunities towards this end [89]. However, the integrated analysis of omics data may only provide correlations between drug-target interactions and system level response upon drug perturbation. It may not be able to distinguish direct physical drug-target interactions from the down-stream induced effect of drug binding. In this regard, the data mining techniques that can infer causality as discussed in Section 2.3 will be an invaluable tool to disentangle confounding factors in the complex drug action. Furthermore, the binding of a chemical compound to a protein is just the start of the process of drug action. There are many types of interactions that are determined by protein conformational dynamics, from antagonism to partial agonism, inverse agonism, biased signaling, and allosteric modulation. The type of interaction will ultimately drive the biological output. In order to fully understand this process and use it as an actionable tool for drug discovery, domain expertise under the form of mechanism-based modeling needs to be formally and mathematically introduced. Data science coupled with mechanism-based modeling is thus required to address the above challenges.

3.2 Personalized adverse drug reaction prediction

One of the unprecedented opportunities that systems pharmacology presents for drug discovery is the improvement of drug safety [90]. Unexpected adverse drug reaction (ADR) is one of major factors that lead to the attrition of drugs in the late stages of drug discovery and development. It is even more challenging to predict rare ADRs that occur in 1 out of every 1000 or more cases. Statistical analysis that only detects correlations between data sets is not sufficient to forecast rare events. Systems pharmacology modeling is required to establish causal relationship between genetic and environmental factors and ADRs. The success of systems pharmacology modeling relies on the integration of heterogeneous data resources, and reliable estimation of parameter spaces to construct system-level, mechanism-based models. Data science will provide essential supports for the development of actionable models of systems pharmacology to predict ADRs *de novo* via using individual-specific omics data, for example, in the framework of constraint modeling [91].

Both on-target and off-target effects are responsible for ADRs. The identification of proteome-wide, context-specific, and quantitative drug-target interactions, as discussed in the previous section, will play a critical role in the predictive modeling of ADR using systems pharmacology. It is, however, not sufficient for systems pharmacology to model drug actions precisely. We need to know how drug perturbations on the target quantitatively influence cellular machinery through gene regulation, signal transduction, and metabolic networks. A number of studies have linked drug-target interactions with drug phenotypic responses via canonical biological pathways [92]. Nonetheless, biased, incomplete, context-independent, and isolated biological pathways cannot model the cellular behavior as a set of integrated machinery, and take into account the genetic, environmental, and behavioral differences between individuals. Genome-wide network modeling is more powerful in predicting phenotypic changes resulting from genetic and environmental perturbations [93]. Because drug actions depend on drug concentrations and gene expression levels, they are ideally simulated using a dynamic network model. However, the kinetic parameters that are required for a dynamic model are often missing. Thus the application of a dynamic model is limited to the small scale. Genome-wide tissue- and context-specific stoichiometric modeling is a powerful approach to simulate cellular behavior at the tissue and organ level [94,95]. Figure 4 shows an example framework in which multiple omics data were integrated with prior knowledge to construct a kidney stoichiometric model. Subsequently, constraint based modeling such as Flux Balance Analysis (FBA) was applied to predict the collective effects of drug treatment and genetic alteration on hypertension. An interesting finding from this study is that the serious side effect of CETP inhibitors occurs when their users bear non-deleterious mutations in several genes [96]. These genetic factors can serve as biomarkers to predict the safety of CETP inhibitors. Recently, personalized whole-cell kinetic models of metabolism have been constructed using a Mass Action Stoichiometric Simulation (MASS) approach [97]. In the MASS model, enzyme kinetics of the metabolic reactions are approximated by a pseudo-elementary rate constant (PERC), which is a combination of classic kinetic parameters and is consistent with all context-specific omic data used. Such data-driven model has the potential to identify specific ADRs for an individual.

The stoichiometric model is the most successful in modeling a metabolic network. Its steady state assumption holds in enzyme reactions, as they are fast compared with regulatory events. However, the cellular response to drug perturbation involves complex interplay between signaling transduction, gene regulation, and metabolism. It is expected that an integrated whole cell models will have significantly enhanced power for predicting precision context-specific ADRs. With the availability of a huge amount of multiple omics data from high-throughput experiments, especially with the efforts in toxicogenomics [98], the data-driven approach promises to integrate genome-scale metabolic models, transcription and translation network models, gene regulation network models, signaling transduction network models, and post-translational network models into a unified tissue- and context-specific model [99]. Even if we can precisely model drug response at the cellular, tissue, and organ levels based on genome-wide molecular signatures, the gap between predicted cellular behavior and clinical phenotype upon drug treatment, i.e. transportability bias, will still exist. Methods that can deal with transportability bias will no doubt be pivotal in linking cellular phenotype with clinical drug response. In summary, the principles of data science will play important roles in systems pharmacology modeling of ADRs, from the reconstruction of integrated cell models to the inference of precision drug responses.

4. Conclusion

Recent efforts in high throughput experiments have generated a huge amount of data across the hierarchy of the organism, across a wide range of time scales, and across multiple species. These data sets provide abundant opportunities to reduce the attrition rate of drug discovery, to enhance the drug safety in the clinic, and to develop precision medicine. Systems pharmacology has emerged as a new discipline to integrate and transform the huge, complex, heterogeneous, and dynamic omics data into interpretable and actionable mechanistic models for decision making in drug discovery and patient care. The advance of systems pharmacology requires the paralleled development and eventual merging of two disciplines: data science and mechanism-based modeling. New techniques in data science are needed to quantify the context-specific similarity between biological entities, to build robust predictive models from high-dimensional but sparse data, and to disentangle causation from correlation. Due to the complex hierarchical organization of living organisms, organism-level responses to genetic alterations and drug-target interactions encompass global changes in conformations, concentrations, modifications, and interactions between molecular components, which manifest at a multi-scale from individual molecules to biological networks. The number of confounding factors and the extent of selection and transportability bias involved in this process are too large to be handled by data science alone. On the other hand, mechanism-based modeling is hindered by a prohibitively large parameter space, which can be reduced using associations, patterns, and rules derived from data as constraints. Thus, the power of systems pharmacology modeling can be maximized by integrating data-driven approaches with mechanism-based multi-scale modeling.

5. Expert opinion

The conventional one-drug-one-target-one-disease drug discovery process has been less successful in tracking multi-genic, multi-faceted complex diseases. We lack fundamental

knowledge of the mechanisms that drive the development, persistence, and transformation of complex diseases. Furthermore, drug efficacy and side effects depend on each individual's genetic and environmental backgrounds. Ignorance of the causal genetic underpinning of human pathophysiology and pharmacology as well as the complex interplay of genetic, molecular, and environmental components underlies the current innovation gap in drug discovery [1]. Quantitative systems pharmacology (QSP) and structural systems pharmacology (SSP) have emerged as new disciplines to tackle the current challenges in drug discovery. The ultimate goal of systems pharmacology is to enable the prediction of drug responses, including both therapeutic effects and side effects, for individual patients through the use of biomarkers. Thus systems pharmacology modeling holds great potential to reduce the attrition rate of drug discovery, to enhance drug safety in the clinic, and to develop precision medicine.

In order to achieve this goal, we need to identify the genetic drivers of pathophysiology and pharmacology, to predict proteome-wide, context-specific and quantitative drug-target interactions, to develop tissue-specific whole-cell dynamic models that take individual genetic and environmental factors into account, and to translate the cellular behavior of drug actions into organismal phenotypes. Recent efforts in high throughput experiments have generated a huge amount of data across the hierarchy of the organism, across a wide range of time scales, and across multiple species. These heterogeneous, high-dimensional, and complex data sets provide abundant opportunities to build instruments for systems pharmacology, but impose great challenges in data processing, data mining, and knowledge discovery. In this regard, big data technology and data science will play an essential role in systems pharmacology. Specifically, based on the fundamental principles of similarity inference, overfitting avoidance, and causality, data science will facilitate the reduction of the complexity of systems pharmacology modeling, detection of hidden correlations between data sets, and the separation of causation from correlation. Due to the complexity of biological systems, the power of data science can only be fully realized when embedded with mechanism-based multi-scale modeling that explicitly takes into account the hierarchical organization of biological systems from nucleic acid to protein, to molecular interaction networks, to cells, to tissues, to patients, and to populations.

It is expected that significant advances will be made in high-throughput data generation, high performance computing, big data technology, data science, and mechanism-based multi-scale modeling, all of which will contribute to systems pharmacology. International and national consortiums such as the 1000 Genomes Project, ENCyclopedia Of DNA Elements (ENCODE), the Library of Integrated Network-based Cellular Signatures (LINCS), and The Cancer Genome Atlas (TCGA) continuously provide invaluable information on the molecular signatures of the global configuration of genetic states and their associations with disease phenotypes and responses to perturbations. The planned National Strategic Computing Initiative will maximize the benefits of High Performance Computing. The National Institutes of Health (NIH) have funded a number of projects to advance big data technologies and data science for biomedical data compression, provenance, visualization, wrangling, and analysis. The successful completion of these projects will significantly enhance the availability and quality of biological data, and the translation of big data to biological knowledge. Several predictive whole cell models for

organisms and multi-scale models for biological processes are under development. The forthcoming challenge is to merge these parallel developments to advance systems pharmacology for the effective prediction of protein binding/unbinding kinetics, identification of the genetic makeup that drives diseases and interferes with pharmacology, prediction of proteome-wide context-specific quantitative drug-target interactions, reconstruction of tissue-, and context specific dynamic network models, and discovery of robust biomarkers for drug safety and efficacy.

Acknowledgements

We sincerely thank the editor and the reviewers for their constructive suggestions

This work was supported by the National Library of Medicine of the National Institute of Health under award number R01LM011986, and the National Institute on Minority Health and Health Disparities of the National Institutes of Health under award number G12MD007599.

References

1. Zerhouni EA. Turning the titanic. *Sci Transl Med*. 2014; 6(221):221ed222.
- 2**. Quantitative and systems pharmacology in the post-genomic era: New approaches to discovering drugs and understanding therapeutic mechanisms. 2011. <http://www.nigms.nih.gov/NR/rdonlyres/8ECB1F7C-BE3B-431F-89E6-A43411811AB1/0/SystemsPharmaWPSorger2011.pdf> [A white paper that defines the field of quantitative systems pharmacology.]
- 3*. Xie L, Ge X, Tan H, et al. Towards structural systems pharmacology to study complex disease and personalized medicine. *PLoS Comput Biol*. 2014; 10(5):e1003554. [PubMed: 24830652] [A comprehensive review of emerging field of structural systems pharmacology.]
4. Tsimring LS. Noise in biology. *Rep Prog Phys*. 2014; 77(2):026601. [PubMed: 24444693]
5. Shaw DE, Deneroff MM, Dror RO, et al. Anton, a special-purpose machine for molecular dynamics simulation. *Communication of the ACM*. 2008; 51(7):91–97.
6. Schadt EE, Linderman MD, Sorenson J, et al. Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. *Nat Rev Genet*. 2011; 12(3):224. [PubMed: 21301474]
7. Harvey MJ, De Fabritiis G. Acecloud: Molecular dynamics simulations in the cloud. *J Chem Inf Model*. 2015; 55(5):909–914. [PubMed: 25849093]
- 8*. Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making. *Big Data*. 2013; 1(1):BD51–59. [An introduction to the principle of data science.]
9. Gregori-Puigjane E, Mestres J. A ligand-based approach to mining the chemogenomic space of drugs. *Comb Chem High Throughput Screen*. 2008; 11(8):669–676. [PubMed: 18795886]
10. Lhota J, Hauptman R, Hart T, Ng C, Xie L. A new method to improve network topological similarity search: Applied to fold recognition. *Bioinformatics*. 2015; 31(13):2106–2114. [PubMed: 25717198]
11. Amaratunga D, Cabrera J, Lee YS. Resampling-based similarity measures for high-dimensional data. *J Comput Biol*. 2015; 22(1):54–62. [PubMed: 25493697]
12. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. *NIPS*. 2013:3111–3119.
13. Costello JC, Heiser LM, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol*. 2014; 32(12):1202–1212. [PubMed: 24880487]
14. Gray KR, Aljabar P, Heckemann RA, et al. Random forest-based similarity measures for multi-modal classification of alzheimer's disease. *Neuroimage*. 2013; 65:167–175. [PubMed: 23041336]
15. Su X, Khoshgoftaar TM. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*. 2009; 2009:19.

16. Cacheda F, Carneiro C, Fern D, et al. Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Trans Web*. 2011; 5(1):1–33.
17. Lovasz L. Random walks on graphs: A survey. *Bolyai Society Mathematical Studies*. 1993; 2:1–46.
18. Shih YK, Parthasarathy S. A single source k-shortest paths algorithm to infer regulatory pathways in a gene network. *Bioinformatics*. 2012; 28(12):i49–58. [PubMed: 22689778]
19. Sun Y, Han J. Meta-path-based search and mining in heterogeneous information networks. *Tsinghua Science and Technology*. 2013; 18(4):329–338.
20. Vishwanathan SVN, Schraudolph NN, Kondor LR, et al. Graph kernels. *J Machine Learning Res*. 2010; 11:1201–1242.
21. Gabdoulline RR, Stein M, Wade RC. Qpipa: Relating enzymatic kinetic parameters and interaction fields. *BMC Bioinformatics*. 2007; 8:373. [PubMed: 17919319]
22. Rupp M. Machine learning for quantum mechanics in a nutshell. *Int J Quantum Chem*. 2015; 115:1058–1073.
23. Ramakrishnan R, Dral PO, Rupp M, et al. Big data meets quantum chemistry approximations: The δ -machine learning approach. *J Chem Theory Comput*. 2015; 11(5):2087–2096. [PubMed: 26574412]
24. Tan J, Ung M, Cheng C, et al. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pac Symp Biocomput*. 2015:132–143. [PubMed: 25592575]
25. Ashby FGaE, D. M. Similarity measures. *Scholarpedia*. 2007; 2(12):4116.
26. <https://github.com/HIPS/spearmint>: <https://github.com/HIPS/Spearmint>
27. Sharma A, Menche J, Huang CC, et al. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum Mol Genet*. 2015; 24(11):3005–3020. [PubMed: 25586491]
28. Vidyasagar M. Identifying predictive features in drug response using machine learning: Opportunities and challenges. *Annu Rev Pharmacol Toxicol*. 2015; 55:15–34. [PubMed: 25423479]
29. Sorzano, COC.; Vargas, J.; Pascual-Montano, A. A survey of dimensionality reduction techniques. 2014. <http://arxiv.org/pdf/1403.2877.pdf>
30. Han H, Jiang X. Overcome support vector machine diagnosis overfitting. *Cancer Inform*. 2014; 13(Suppl 1):145–158. [PubMed: 25574125]
31. Najafabadi M, Villanustre F, Khoshgoftaar T, et al. Deep learning applications and challenges in big data analytics. *Journal of Big Data*. 2015; 2(1):1.
32. Chen L, Cai C, Chen V, Lu X. Trans-species learning of cellular signaling systems with bimodal deep belief networks. *Bioinformatics*. 2015; 31(18):3008–3015. [PubMed: 25995230]
33. Geman D, Ochs M, Price ND, et al. An argument for mechanism-based statistical inference in cancer. *Hum Genet*. 2015; 134(5):479–495. [PubMed: 25381197]
34. Hung JH, Yang TH, Hu Z, et al. Gene set enrichment analysis: Performance evaluation and usage guidelines. *Brief Bioinform*. 2012; 13(3):281–291. [PubMed: 21900207]
35. Mutation C. Pathway Analysis working group of the International Cancer Genome C: Pathway and network analysis of cancer genomes. *Nat Methods*. 2015; 12(7):615–621. [PubMed: 26125594]
36. Mitrea C, Taghavi Z, Bokanizad B, et al. Methods and approaches in the topology-based analysis of biological pathways. *Front Physiol*. 2013; 4:278. [PubMed: 24133454]
37. Ng S, Collisson EA, Sokolov A, et al. Paradigm-shift predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics*. 2012; 28(18):i640–i646. [PubMed: 22962493]
38. Tarca AL, Draghici S, Khatri P, et al. A novel signaling pathway impact analysis. *Bioinformatics*. 2009; 25(1):75–82. [PubMed: 18990722]
39. Glaab E, Baudot A, Krasnogor N, et al. Enrichnet: Network-based gene set enrichment analysis. *Bioinformatics*. 2012; 28(18):i451–i457. [PubMed: 22962466]

40. Brubaker D, Difeo A, Chen Y, et al. Drug intervention response predictions with PARADIGM (DIRPP) identifies drug resistant cancer cell lines and pathway mechanisms of resistance. *Pac Symp Biocomput.* 2014;125–135. [PubMed: 24297540]
41. Bansal M, Yang J, Karan C, et al. A community computational challenge to predict the activity of pairs of compounds. *Nat Biotechnol.* 2014; 32(12):1213–1222. [PubMed: 25419740]
42. Sun J, Zhao M, Jia P, et al. Deciphering signaling pathway networks to understand the molecular mechanisms of metformin action. *PLoS Comput Biol.* 2015; 11(6):e1004202. [PubMed: 26083494]
43. Hofree M, Shen JP, Carter H, et al. Network-based stratification of tumor mutations. *Nat Methods.* 2013; 10(11):1108–1115. [PubMed: 24037242]
44. Xu Z, Taylor JA. SNPinfo: Integrating gwas and candidate gene information into functional snp selection for genetic association studies. *Nucleic Acids Res.* 2009; 37(Web Server issue):W600–605. [PubMed: 19417063]
45. Xie L, Ng C, Ali T, et al. Multiscale modeling of the causal functional roles of nssnps in a genome-wide association study: Application to hypoxia. *BMC Genomics.* 2013; 14(S3):S9. [PubMed: 23819581]
46. Porta-Pardo E, Godzik A. E-Driver: A novel method to identify protein regions driving cancer. *Bioinformatics.* 2014; 30(21):3109–3114. [PubMed: 25064568]
47. Rader DJ, Hovingh GK. HDL and cardiovascular disease. *Lancet.* 2014; 384(9943):618–625. [PubMed: 25131981]
- 48*. Mooij, JM.; Peters, J.; Janzing, D., et al. Distinguishing cause from effect using observational data: Methods and benchmarks. 2015. <http://arxiv.org/pdf/1412.3773v2.pdf> [A novel statistical method to distinguish causation from correlation.]
49. Bareinboim E, Pearl J. Controlling selection bias in causal inference. *JMLR Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS).* 2012; 22:100–108.
- 50*. Bareinboim, E.; Tian, J. Recovering causal effects from selection bias.. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; Menlo Park, CA.* 2015; p. 3475–3481. [An algorithm to reduce selection bias.]
- 51*. Bareinboim E, Pearl J. Transportability from multiple environments with limited experiments: Completeness results. *NIPS.* 2014:280–288. [A proposal mathematical framework to address transportability problem.]
52. Bareinboim E, Pearl J. Causal inference from big data: Theoretical foundations and the data-fusion problem. *UCLA Cognitive Systems Laboratory, Technical Report (R-450).* 2015
53. Blois, MS. *Information and medicine: The nature of medical descriptions.* University of California Press; 1984.
- 54*. Gerstung M, Pellagatti A, Malcovati L, et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat Commun.* 2015; 6:5901. [PubMed: 25574665] [An example of establishing genotype-phenotype association in the spirit of multi-scale modeling.]
55. Feng Y, Mitchison TJ, Bender A, et al. Multi-parameter phenotypic profiling: Using cellular effects to characterize small-molecule compounds. *Nat Rev Drug Discov.* 2009; 8(7):567–578. [PubMed: 19568283]
56. Gutenkunst RN, Waterfall JJ, Casey FP, et al. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol.* 2007; 3(10):1871–1878. [PubMed: 17922568]
57. Toni T, Stumpf MP. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics.* 2010; 26(1):104–110. [PubMed: 19880371]
58. Liepe J, Barnes C, Cule E, et al. Abc-sysbio--approximate bayesian computation in python with gpu support. *Bioinformatics.* 2010; 26(14):1797–1799. [PubMed: 20591907]
59. An G. Closing the scientific loop: Bridging correlation and causality in the petaflop age. *Sci Transl Med.* 2010; 2(41):41ps34.
60. Xie L, Kinnings SL, Bourne PE. Novel computational approaches to polypharmacology as a means to define responses to individual drugs. *Annu Rev Pharmacol Toxicol.* 2012; 52:361–379. [PubMed: 22017683]

61. Schirle M, Jenkins JL. Identifying compound efficacy targets in phenotypic drug discovery. *Drug Discov Today*. 2015
62. Wang L, Xie XQ. Computational target fishing: What should chemogenomics researchers expect for the future of in silico drug design and discovery? *Future Med Chem*. 2014; 6(3):247–249. [PubMed: 24575960]
63. Koutsoukas A, Lowe R, Kalantarmotamedi Y, et al. In silico target predictions: Defining a benchmarking data set and comparison of performance of the multiclass naive bayes and parzen-rosenblatt window. *J Chem Inf Model*. 2013; 53(8):1957–1966. [PubMed: 23829430]
64. Keiser MJ, Roth BL, Armbruster BN, et al. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*. 2007; 25(2):197–206. [PubMed: 17287757]
65. Liu H, Sun J, Guan J, Zheng J, et al. Improving compound-protein interaction prediction by building up highly credible negative samples. *Bioinformatics*. 2015; 31(12):i221–229. [PubMed: 26072486]
66. Ma J, Sheridan RP, Liaw A, et al. Deep neural nets as a method for quantitative structure-activity relationships. *J Chem Inf Model*. 2015; 55(2):263–274. [PubMed: 25635324]
67. Dahl GE, Jaitly N, Salakhutdinov R. Multi-task neural networks for qsar predictions. arXiv: 1406.1231 [statML]. 2014
68. Ramsundar B, Kearnes S, Riley P, et al. Massively multitask networks for drug discovery. arXiv: 1502.02072 [statML]. 2015
69. Unterthiner T, Mayr A, Klambauer G, et al. Toxicity prediction using deep learning. arXiv: 1503.01445 [statML]. 2015
70. van Laarhoven T, Marchiori E. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS One*. 2013; 8(6):e66952. [PubMed: 23840562]
71. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics*. 2011; 27(21):3036–3043. [PubMed: 21893517]
72. Gonen M. Predicting drug-target interactions from chemical and genomic kernels using bayesian matrix factorization. *Bioinformatics*. 2012; 28(18):2304–2310. [PubMed: 22730431]
- 73*. Woo JH, Shimoni Y, Yang WS, et al. Elucidating compound mechanism of action by network perturbation analysis. *Cell*. 2015; 162(2):441–451. [PubMed: 26186195] [A novel method to predict drug-target relationship using network-based differential gene expression analysis.]
74. Pahikkala T, Airola A, Pietila S, et al. Toward more realistic drug-target interaction predictions. *Brief Bioinform*. 2015; 16(2):325–337. [PubMed: 24723570]
75. Creixell P, Palmeri A, Miller CJ, et al. Unmasking determinants of specificity in the human kinome. *Cell*. 2015; 163(1):187–201. [PubMed: 26388442]
76. Lin H, Sassano MF, Roth BL, et al. A pharmacological organization of g protein-coupled receptors. *Nat Methods*. 2013; 10(2):140–146. [PubMed: 23291723]
77. Xie L, Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc Natl Acad Sci U S A*. 2008; 105(14):5441–5446. [PubMed: 18385384]
78. Xie L, Wang J, Bourne PE. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comput Biol*. 2007; 3(11):e217. [PubMed: 18052534]
79. Kinnings SL, Liu N, Buchmeier N, et al. Drug discovery using chemical systems biology: Repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput Biol*. 2009; 5(7):e1000423. [PubMed: 19578428]
80. Xie L, Li J, Xie L, Bourne PE. Drug discovery using chemical systems biology: Identification of the protein-ligand binding network to explain the side effects of cetp inhibitors. *PLoS Comput Biol*. 2009; 5(5):e1000387. [PubMed: 19436720]
81. Durrant JD, Amaro RE, Xie L, et al. A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. *PLoS Comput Biol*. 2010; 6(1):e1000648. [PubMed: 20098496]
82. Xie L, Evangelidis T, Xie L, et al. Drug discovery using chemical systems biology: Weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. *PLoS Comput Biol*. 2011; 7(4):e1002037. [PubMed: 21552547]

83. Ho Sui SJ, Lo R, Fernandes AR, et al. Raloxifene attenuates pseudomonas aeruginosa pyocyanin production and virulence. *Int J Antimicrob Agents*. 2012; 40(3):246–251. [PubMed: 22819149]
84. Xie L, Xie L, Bourne PE. Structure-based systems biology for analyzing off-target binding. *Curr Opin Struct Biol*. 2011; 21(2):189–199. [PubMed: 21292475]
85. Cruz-Monteagudo M, Medina-Franco JL, Perez-Castillo Y, et al. Activity cliffs in drug discovery: Dr jekyll or mr hyde? *Drug Discov Today*. 2014; 19(8):1069–1080. [PubMed: 24560935]
86. Lamb J, Crawford ED, Peck D, et al. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006; 313(5795):1929–1935. [PubMed: 17008526]
87. Bengio Y. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013; 35(8):1798–1828. [PubMed: 23787338]
88. Chiu SH, Xie L. Toward high-throughput predictive modeling of protein binding/unbinding kinetics. *bioRxiv*. 2015 10.1101/024513.
89. Greene CS, Krishnan A, Wong AK, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet*. 2015; 47(6):569–576. [PubMed: 25915600]
90. Bai JP, Fontana RJ, Price ND, et al. Systems pharmacology modeling: An approach to improving drug safety. *Biopharm Drug Dispos*. 2014; 35(1):1–14. [PubMed: 24136298]
91. Bordbar A, Monk JM, King ZA, et al. Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet*. 2014; 15(2):107–120. [PubMed: 24430943]
92. Urban L, Maciejewski M, Lounkine E, et al. Translation of off-target effects: Prediction of adrs by integrated experimental and computational approach. *Toxicology Research*. 2014; 3(6):433–444.
93. Joyce AR, Palsson BO. Toward whole cell modeling and simulation: Comprehensive functional genomics through the constraint-based approach. *Prog Drug Res*. 2007; 64:267–309.
94. Shlomi T, Cabili MN, Herrgard MJ, et al. Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol*. 2008; 26(9):1003–1010. [PubMed: 18711341]
95. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol*. 2008; 4(5):e1000082. [PubMed: 18483554]
- 96*. Chang RL, Xie L, Xie L, et al. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput Biol*. 2010; 6(9):e1000938. [PubMed: 20957118] [An example of combine data-driven network reconstruction and mechanism-based modeling to predict collective effect of drug perturbation and genetic alternation.]
- 97*. Bordbar A, McCloskey D, Zielinski Daniel C, et al. Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics. *Cell Systems*. 1(4):283–292. [PubMed: 27136057] [Development of personalized whole-cell kinetic models of metabolism and their potential applications to ADR.]
98. Waters MD, Fostel JM. Toxicogenomics and systems toxicology: Aims and prospects. *Nat Rev Genet*. 2004; 5(12):936–948. [PubMed: 15573125]
99. Imam S, Schauble S, Brooks AN, et al. Data-driven integration of genome-scale regulatory and metabolic network models. *Front Microbiol*. 2015; 6:409. [PubMed: 25999934]

Article highlights box

- Systems pharmacology will play increasingly important roles in accelerating drug discovery and developing precision medicine.
- Data science is the foundation of data-driven decision making. It follows the fundamental principles of similarity inference, overfitting avoidance, and causality discovery.
- Systems pharmacology modeling will benefit from exploiting the principles of data science. Data science will facilitate reducing complexity of systems pharmacology modeling, detecting hidden correlations between data sets for predictive modeling, and distinguishing causation from correlation to discover new biological knowledge.
- Data-driven reconstruction of proteome-wide, tissue- and context-specific, quantitative drug-target interaction networks and integrated whole-cell, tissue, and organ models will enable the prediction of drug responses under diverse genetic and environmental backgrounds using systems pharmacology modeling.
- The power of data science application in systems pharmacology will be enhanced when embedded with mechanism-based multi-scale modeling.

Figure 1. Relationships between big data technology, data science, systems pharmacology, and drug discovery

High throughput techniques have generated huge volumes of complex, heterogeneous, dynamic, and noisy biological and clinical data. The management and processing of such big data require big data technology based on cloud computing platforms. Data science, which includes many fields, integrates multiple data sets, recognizes patterns, infers causal relationships, makes novel predictions, and discovers new knowledge from the data. Data-driven modeling will support systems pharmacology to develop actionable and interpretable mechanism-based models. Systems pharmacology modeling will ultimately facilitate decision making in drug discovery from target identification to lead optimization, from toxicity evaluation to clinical trials.

Figure 2. A general framework to infer unknown relationships between two biological entities based on known relationships

Solid triangles and squares represent two different categories of biological entities (e.g. drugs and proteins). Solid lines represent similarity or other association-based edges between entities in the same category (e.g. chemical-chemical similarity or protein-protein interaction). Solid arrows represent known associations between two entities in different categories (e.g. known drug-target interaction). Dashed lines represent predicted new associations between two entities in different categories (e.g. predicted new drug-target interaction).

Figure 3. A multi-scale modeling strategy for the understanding of the genetic underpinning of pathophysiology and predictive modeling of drug responses for individual patients
 Instead of directly associating genetic variations with organismal phenotypes, the multi-scale modeling approach will model the emergent properties of a mesophenotype, and establish their causal relationships with an upper level mesophenotype.

Figure 4. An example framework for personalized side effect prediction by combining data-driven analysis and mechanism-based multi-scale modeling [93]

A context-specific organ model was constructed by integrating genome-scale metabolic networks, metabolomic data, transcriptomic data, and biological knowledge. Structural proteome-wide drug-target interactions were predicted using ligand binding site similarity and protein-ligand docking. The collective effect of drug-target interaction and genetic mutation on drug side effect was simulated using Flux Balance Analysis (FBA) on the organ model.