HHS Public Access

Author manuscript

Curr Top Med Chem. Author manuscript; available in PMC 2017 March 09.

Published in final edited form as:

Curr Top Med Chem. 2014 ; 14(11): 1356–1364.

Integrative Approaches for Predicting in vivo Effects of Chemicals from their Structural Descriptors and the Results of Short-term Biological Assays

Yen S. Low[1,2], **Alexander Sedykh**[1], **Ivan Rusyn**[2], and **Alexander Tropsha**[1,*]

[1]Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, University of North Carolina, Chapel Hill, USA

[2]Department of Environmental Sciences and Engineering, University of North Carolina, Chapel Hill, USA

Abstract

Cheminformatics approaches such as Quantitative Structure Activity Relationship (QSAR) modeling have been used traditionally for predicting chemical toxicity. In recent years, high throughput biological assays have been increasingly employed to elucidate mechanisms of chemical toxicity and predict toxic effects of chemicals *in vivo*. The data generated in such assays can be considered as biological descriptors of chemicals that can be combined with molecular descriptors and employed in QSAR modeling to improve the accuracy of toxicity prediction. In this review, we discuss several approaches for integrating chemical and biological data for predicting biological effects of chemicals *in vivo* and compare their performance across several data sets. We conclude that while no method consistently shows superior performance, the integrative approaches rank consistently among the best yet offer enriched interpretation of models over those built with either chemical or biological data alone. We discuss the outlook for such interdisciplinary methods and offer recommendations to further improve the accuracy and interpretability of computational models that predict chemical toxicity.

Keywords

predictive toxicology; QSAR; bioinformatics; systems pharmacology

INTRODUCTION

Predictive toxicology is often evaluated at the initial stages of regulatory assessment of environmental chemicals or drug discovery to prioritize high-risk chemicals for further testing or eliminate such chemicals from further consideration, respectively. In the current age of chemical innovation, hundreds to thousands of new chemicals are introduced each year [1] creating an urgent need to substantially optimize testing resources and reduce

*Corresponding author: 100K Beard Hall, Campus Box 7568, University of North Carolina, Chapel Hill, NC 27599-7568, Telephone: +1 (919) 966-2955, FAX: +1 (919) 966-0204, alex_tropsha@unc.edu.

The authors declare that there are no conflicts of interest.

animal use. Current toxicity evaluation protocols increasingly follow a tiered approach where chemicals are funneled through *in silico, in vitro* and *in vivo* tests in order of decreasing throughput [2–4]. Among the *in silico* methods, cheminformatics and bioinformatics have been established as integral parts of toxicity testing, especially at the initial stages.

Most of the current computational tools employed in toxicity assessment rely either on chemical or biological data. Specifically, cheminformatics approaches attempt to predict toxicity from chemical structure alone while ignoring the underlying complex biological mechanisms whereas bioinformatics approaches ignore the inherent structural features of chemical molecules that may enrich and improve modeling outcomes. In contrast, integrative chemical-biological modeling may both improve the prediction performance of models and uncover insights previously invisible to either informatics discipline alone. The realization that chemical and biological entities interact at various levels of organization in the body has spawned the emerging fields of *systems chemical biology* [5–7], *systems toxicology* [8], or *systems pharmacology* [9–12]. Several recent reviews [6,10,12–16] have focused on the current state of each individual discipline and proposed general schemes to integrate cheminformatics and bioinformatics approaches for improved understanding of chemical effects on biological systems. Few integrative studies, however, have been reported; their paucity is stemming from the lack of both suitable data and integrative methods. Nevertheless, a new data landscape for predictive toxicology has emerged due to new toxicity testing paradigms such as REACH (Registration, Evaluation and Authorization of CHemicals)[17] and *Toxicity Testing for the 21st Century* [18]. These programs have stimulated the proliferation of short-term biological assays employed for testing of growing collections of chemicals [19]. These transformative experimental programs offer new opportunities for data-driven learning beyond the traditional methods of cheminformatics or bioinformatics.

In this review, we reiterate the case for integrative chemical-biological approaches for predicting chemical effects *in vivo* with the ultimate goal of developing safer pharmaceutical or industrial chemicals. We assess the strengths and limitations of current predictive toxicology efforts based on either cheminformatics or bioinformatics and then discuss studies drawing from the two disciplines concurrently. Lastly, we put forth our vision for such interdisciplinary methods and offer recommendations to further improve the accuracy and interpretability of chemical toxicity prediction models.

CHEMINFORMATICS IN TOXICITY PREDICTION

The availability of large toxicity datasets including hundreds, even thousands of chemicals tested as part of ToxCast [20] and Tox21 [21] projects has re-established an interest in cheminformatics as a powerful computational approach for predicting chemical toxicity. In particular, Quantitative Structure-Activity Relationships (QSAR) modeling is often used as a first-line tool for toxicity prediction [22] given that it requires the knowledge of molecular structure only. The first QSAR study was published in 1962 [23]; it employed regression model correlating plant growth to molecular electronic parameters. Since then, QSAR has grown in sophistication to incorporate thousands of chemical descriptors and machine

learning methods. Despite its popularity, QSAR modeling has been criticized for poor predictivity and interpretability [24–27]. Measures to address weaknesses include OECD guidelines [28] and implementation of best practices [29] which advocate careful data curation [30] and processing [31], representative sampling of the chemical space [32], stringent validation [33] and rational descriptor selection driven by a mechanistic basis to simplify interpretation [26].

Despite the above measures, cheminformatics-based prediction of complex toxic phenomena has fallen short of expectations. In reality, the relationship between chemical structures and toxicity is far more circuitous than the models assume, involving many non-chemical factors, e.g., those dependent on complex biological mechanisms. The significance of these non-chemical factors depends on the prediction target. Generally, QSAR models are more successful at predicting direct chemical-induced outcomes (e.g., mutagenicity) than those farther downstream of chemical-initiating events (e.g., carcinogenicity) [34]. Indeed, in some cases when large datasets are available, e.g., for mutagenicity (that largely depends on molecular interactions between chemical and DNA) the QSAR model accuracy approaches that of the experimental Ames assay [34,35]. On the contrary, carcinogenicity has been notoriously difficult to predict because of its heterogeneous modes of action and the biological host's adaptive capacity for recovery [34]. One way to account for complex biological mechanisms underlying many *in vivo* effects and achieve better modeling outcomes is to integrate multiple biological characteristics of chemicals obtained in short term assays with inherent chemical properties of compounds. This emerging integrative modeling approach at the interface between bio- and cheminformatics is the main theme of this review.

BIOINFORMATICS IN TOXICITY PREDICTION

The post-genome era saw a dramatic rise of bioinformatics. While the field of bioinformatics is broad, involving the computational analysis of biological information arising from the detailed characterization of an organism at various levels (molecular, cellular, tissue, organ, system), this section focuses on applying bioinformatics approaches in toxicology where the goal is to study multiple biological perturbations in response to chemical insult.

Simultaneously studying thousands of bioassays offers several advantages: key biomarkers can be quickly identified and interactions between the biomarkers characterized, allowing for a systems toxicology approach. In drug discovery, the use of diverse bioassay panels helps to quickly identify potentially toxic properties (e.g., cytochrome P450 inhibition, transporter blockage) which may provide clues into the pathogenesis of undesired effects caused by a compound. The bioassay signatures of compounds reflecting certain toxic modes of action may be used to probe for compounds with similar mode of action. An example of broad biological characterization of drugs is provided by the Japanese Toxicogenomics Project where toxicogenomic signatures representative of various types of hepatotoxicities (e.g. phospholipidosis, glutathione depletion) have been determined [36]. These signatures can be generated for new drugs or drug candidates to predict their long-term toxicities.

Advances in assay technology have given rise to a diversity of biological measurements such as ‘omics signatures, enzymatic activity, receptor binding affinity, cytotoxicity, and histology imaging, allowing toxicologists to probe into both microscopic and macroscopic changes in the body. These bioassays may have different predictive power depending on the experimental error and biological relevance. High-dimensional ‘omics’ data, especially transcriptomics, were shown to predict long term effects such as hepatic tumorigenicity with high accuracy [37–40]. In contrast, hundreds of bioassays capturing a large diversity of biological characteristics in ToxCast Phase I [41,42], were less predictive [43]. Reasons cited by the authors include inadequate experimental fidelity, inadequate biological relevance, and poor interspecies extrapolation.

Certain successes notwithstanding, the use of biological data and bioinformatics approaches in chemical toxicity prediction is not problem-free. The ease of collecting large-scale bioassay data has encouraged fishing expeditions where assays often produce poor quality results or may be irrelevant to any toxicity leading to false discoveries. Overly sensitive ‘omics’ markers may be producing more noise than signal [44]. Countermeasures include proper statistical correction (e.g. Bonferroni, Holm) and proper application of biological context to draw meaningful conclusions from the data.

The focus on biological information has also regrettably overlooked another important dimension of toxicology: chemical information. While bioassays were previously performed for a few chemicals due to throughput limitations, it is now possible to perform high throughput screening (HTS) for large chemical libraries [45]. Consequently, *in vitro* toxicity data is rich in both biological and chemical information. The underlying chemical patterns, a traditional and rich source of data in cheminformatics, have not been capitalized upon by bioinformatics. A reasonable approach may be to combine bioinformatics and cheminformatics approaches for improved toxicity prediction.

INTEGRATIVE APPROACH COMBINING CHEMINFORMATICS AND BIOINFORMATICS

Given the lack of consideration of biological factors in cheminformatics and ignorance of chemical structures in bioinformatics, the concurrent study of both biological and chemical domains may reveal new discoveries not possible with either domain alone. Such integrated approaches recognize that *in vivo* effects, whether occurring at the cellular, or systemic level, emerge from a complex interplay between the chemical inducer and the biological host. Chemical factors govern the molecular interactions between the chemical and its protein targets. These molecular interactions then initiate a cascade of interactions within the cell, organ or organism, eventually giving rise to the observed phenotype as a response to the chemical action on the biological system.

The rise of several recent enabling trends facilitates chemical-biological integration. First, there is an increased demand and acceptance of toxicity prediction from *in silico* and *in vitro* tests instead of *in vivo* tests in efforts to boost testing throughput, improve animal welfare and deepen our understanding of the toxicological mechanisms; these new paradigms are accelerated by regulatory programs such as REACH[17] and *Toxicity Testing for the 21st*

Century [18]. Consequently, in large-scale programs such as ToxCast [20], Tox21 [21], and Molecular Libraries Initiative [46] thousands of chemicals are tested in thousands of biological assays, with the results of these HTS studies placed in publicly available repositories such as PubChem [47] or ToxNet [48]. For instance, toxicity databases now contain large amounts of chemical and biological information through data consolidation (e.g. ACToR [19], Bio2RDF [49], OpenPHACTS[50], PredPharmTox [51]; see [11] for table of databases).

The unprecedented growth of data in terms of the number and diversity of chemicals, and comprehensive biological assay characterization has afforded new research opportunities for both cheminformatics and bioinformatics. Where previously only a few chemicals were tested, the new data landscape has reinvigorated interest in cheminformatics as a means of transforming latent chemical patterns into useful chemical insights. On the other hand, the deeper biological assay characterization allows one to learn more about each chemical in terms of underlying biological mechanisms of its *in vivo* effects. Yet, sticking to the approaches of only chemical or biological modeling is unlikely to take full advantage of the richness of the modern data streams that effectively capture chemical-biological interactions.

The many parallels between bioinformatics and cheminformatics provide points of commonality to facilitate integration. Underpinning both fields are statistical techniques relating molecular features of a chemical to its biological effects. These statistical relationships rely on the similarity principle which expects chemicals similar in their molecular feature profiles to exhibit similar behavior. The key difference between bioinformatics and cheminformatics here lies in the choice of appropriate molecular features, i.e., either ‘omics’ profiles assayed by HTS or molecular structural information represented by chemical descriptors. The statistical techniques, whether as simple as read-across or as complex as machine learning, are equally applicable to both fields.

A simple means of integration is to apply existing statistical methods to both chemical and biological types of molecular features. Another way is to merge chemical models with biological models. Other approaches may be less straightforward, strategically combining chemical structures and biological assays such that the two data sources compensate for each other’s shortcomings and the complementary information between them is maximally used. Generally, modeling studies combining chemical and biological data have reported increased predictivity and interpretability (Table 1). Integrative chemical-biological approaches attempted in those studies may be broadly classified into three types: 1) data pooling (Fig. 1A), 2) model pooling (Fig. 1B), or 3) other integrative strategies that exploit the multi-domain data (e.g. hierarchical local models shown in Fig 1C).

Data pooling

In data pooling, disparate data sources are pooled to create a larger, “hybrid” data matrix for modeling by existing statistical methods. This has been aided by the growing availability of consolidated databases. Besides HTS assays, new data streams can include text annotations automatically mined from biomedical literature (ChemoText [52]), product labels (SIDER) [53–55] and clinical notes [56,57].

Table 1 lists several studies predicting toxicity from pooling various data streams. Generally, prediction performance improved after pooling, although several exceptions exist. Among the exceptions, a comprehensive modeling of 60 *in vivo* toxicities based on chemical structures and/or *in vitro* assays of ToxCast phase I data described limited if any success with data pooling models [43]. Several other studies reported that models' accuracy dropped when chemical descriptors were added on top of bioassay data such as toxicogenomics [58], hepatocyte imaging indicators [59], and protein targets [60]. Noteworthy, all four of the mentioned studies [43,58–60] employ rather small and structurally diverse sets of chemicals. Understandably, chemical models performed worse than biological ones [58–60], while in the ToxCast study [43] both were similarly poor. Thus, we caution against favoring either chemical or biological model as either performance will depend on many contributing factors such as the size and structural diversity of the chemical datasets and quality of biological data. Where biological data included considerable noise, additional data treatment may improve prediction outcomes from data pooling. For example, Sedykh et. al. [61] introduced a noise filter to transform cytotoxicity profiles into dose-response curve parameters that, when pooled with chemical structures, provided more accurate models of rat acute toxicity than the original cytotoxicity assay values.

Model pooling

Another way of integrating chemical and biological data is by meta-analysis or ensemble modeling, which pools individual predictions from several models into a final predicted value. The main benefit of ensemble modeling, i.e., increased predictivity, arises when the constituent models compensate for the errors of one another [62]. The notion that many models are better than one is best exemplified by the random forest algorithm which seeks the consensus vote of numerous constituent decision tree models within its "forest" [63]. In the case of toxicity modeling, chemical-based models and biological-based models may be pooled such that their consensus vote provides the final prediction outcome.

Such model pooling is already widely practiced in regulatory chemical risk assessment and drug discovery during where all the prediction outcomes from various toxicity models are weighted before arriving at a consensus decision [64,65]. For example, drugs must not contain structural alerts of mutagenicity and their bioassay profiles must indicate the lack of inhibitory effects on the major cytochrome P450 enzymes required for drug metabolism.

Ensemble modeling can be used in one of two ways. One approach is to require that all the constituent models for a compound point to the same prediction outcome so that the end point toxicity can be estimated with higher confidence. Alternatively, one can argue that ensemble modeling enlarges the modelable space of molecules such that compounds that cannot be predicted with confidence by one model receive their prediction from another model. In the first case, Vilar et. al. showed increased precision when a chemical similarity model was pooled with a model based on clinical notes [66,67]. In the second case, an ensemble chemical-biological model may compensate for the invalid predictions by the QSAR model outside its chemical coverage area. However, ensemble models may not always outperform their constituent chemical and biological models, as we have demonstrated recently using four different data sets [68]. Especially where a constituent

model is already highly predictive, adding another inferior model may lead to reduced predictive power of the consensus model. Thus, model pooling should be done with care paying attention to relative predictive power of each constituent model.

Other integrative methods

The mixed success of pooling data or models has led to the development of innovative approaches that leverage prior knowledge of the data structure and optimize the use of the disparate data sources. Network modeling that allows the simultaneous study of disparate entities (chemicals, targets and phenotypes) is employed increasingly [69]. In a graph representation (Figure 2), entities appear as nodes connected by edges if they are associated. Association may be defined in terms of direct physical interaction (*e.g.,* drug binds to target) or statistically (*e.g.,* disproportionately more reports of adverse effect with drug). In such way modeling, the goal is to infer new associations among pairs of entities through indirect associations. This is best illustrated by Swanson's ABC paradigm [70] in which association between entities A and C is inferred if there exist direct associations between pairs A–B and B–C (Figure 2A). Networks may be further enriched by chemical similarity [71,72], protein sequence similarity [73], or side effect similarity [74] such that novel inferences with higher confidence can be drawn (Figure 2B [75]). Associations successfully predicted in recent studies include those of phenotype-target [72], chemical-phenotype [76,77], and chemical-target [74,78,79] type. For examples of broader efforts to infer more than a single type of associations, the readers are referred to several recent studies [69,71,80].

Strategic use of biological data to stratify data sets into distinct clusters for further separate or localized modeling can be a promising direction. Zhu et. al.[81] described a two-step hierarchical approach, in which the authors first stratified compounds by their *in-vitro*/*in-vivo* correlation into two classes , i.e., a group of compounds whose in vitro/in vivo data correlated and a remaining group where no correlation was observed. The authors then built a classification model using this biologically-inferred strata and then also built stratum-specific QSAR models. It has been shown that such a hierarchical workflow where a new compound was first assigned to one of the two strata followed by the prediction using stratum-specific models afforded overall improved prediction accuracy [81,82]. Other strategic use of biological data to stratify data sets into clusters for localized modeling was also attempted by Lounkine et al. [83] who clustered compounds by chemical similarity and their bioactivity. Analogously, chemical structural data can provide useful input for biological modeling. For example, pharmacokinetics parameters, where unknown, may be estimated by QSAR models from molecular structure and then used in subsequent physiological-based models to simulate chemical toxicity in the body [84–86].

Another recently published novel integrative method, quantitative chemical-biological read-across (CBRA) [68], relied on the principles of k nearest neighbors. The CBRA approach can be viewed as an ensemble model, in which chemical- and biological-based predictions for a new chemical are weighted by similarity to known both chemical and biological analogs. This enhanced pooling of chemical and biological neighbors helps to maximize the complementarities between chemical and biological data. In particular, conflicting predictions from chemical and biological models are resolved, resulting in overall

predictivity gains. The authors compared CBRA with two other types of integrative approaches (data pooling, model pooling) on four data sets and found that none of the three approaches was markedly superior to others. We believe that this holds true in general, and no single integrative technique is likely to solve all modeling problems. Instead, this we emphasize the importance of employing set of modeling tools from which most appropriate and expedient ones can be selected and attempted for each complex dataset.

RECOMMENDATIONS AND OUTLOOK

The mixed success of using both chemical and biological features suggest the following methodological implications for predicting chemical effects. First, consider information-rich and biologically relevant assays as features. Information-rich assays such as gene expression may have more predictive value than non-descript assays measuring binary biological responses (e.g., binding/nonbinding to a target protein) [87]. The bioassays may be selected rationally according to biological pathways to reflect their relevance to the *in vivo* effect [43,88]. Third, careful variable selection [43], modeling and validation according to OECD (Q)SAR principles [28] are necessary to ensure robust and accurate models [29]. Lastly, consider the choice of modeling methods. Irrelevant variables may affect some classification methods more than others. For example, instance-based methods including CBRA are more susceptible to irrelevant variables while others such as random forest can better tolerate noisy variables [89,90].

A multidisciplinary systems approach is increasingly seen as the key solution to translating molecular and preclinical insights into desired clinical outcomes of drug use. In addition to addressing the issue of data quality, further gains through methodological innovations and cohesive integration of the various disciplines will be necessary. Scientists who develop and employ such approaches need to have profound understanding of both the data and data-analytical techniques. The ingredients for such multi-disciplinary efforts are unlikely to occur organically and will require deliberate efforts to foster a collaborative environment. As more data come online and advances in assay technologies reduce experimental variability, we expect integrative approaches to play a greater role in toxicology and drug discovery applications.

Acknowledgments

The work was supported in part by grants from NIH (GM076059, GM066940) and EPA (RD83272001, RD83382501).

REFERENCES

1. Stephenson, J. Chemical Regulation: Observations on Improving the Toxic Substances Control Act. Washington DC: 2009.
2. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. Toxicol. Sci. 2007; 95:5–12. [PubMed: 16963515]
3. Keller DA, Juberg DR, Catlin N, Farland WH, Hess FG, Wolf DC, Doerrer NG. Identification and Characterization of Adverse Effects in 21st Century Toxicology. Toxicol. Sci. 2012; 126:291–297. [PubMed: 22262567]

4. Merlot C. In Silico Methods for Early Toxicity Assessment. Curr. Opin. Drug Discov. Devel. 2008; 11:80–85.
5. Oprea TI, Tropsha A, Faulon J-LL, Rintoul MD. Systems Chemical Biology. Nat. Chem. Biol. 2007; 3:447–450. [PubMed: 17637771]
6. Wild DJ, Ding Y, Sheth AP, Harland L, Gifford EM, Lajiness MS. Systems Chemical Biology and the Semantic Web: What They Mean for the Future of Drug Discovery Research. Drug Discov. Today. 2011; 00:8–13.
7. Iskar M, Zeller G, Zhao X-M, van Noort V, Bork P. Drug Discovery in the Age of Systems Biology: The Rise of Computational Approaches for Data Integration. Curr. Opin. Biotechnol. 2011:1–8. [PubMed: 21190838]
8. Spurgeon DJ, Jones OaH, Dorne J-LCM, Svendsen C, Swain S, Stürzenbaum SR. Systems Toxicology Approaches for Understanding the Joint Effects of Environmental Chemical Mixtures. Sci. Total Environ. 2010; 408:3725–3734. [PubMed: 20231031]
9. Sorger, PK., Allerheiligen, SRB., Abernethy, DR., Altman, RB., Brouwer, KLR., Califano, A., David, Z., Argenio, D., Iyengar, R., Jusko, WJ., Lalonde, R., Lauffenburger, DA., Shoichet, B., Stevens, JL., Subramaniam, S., Van Der, Graaf P., Ward, R., Ma, B. Quantitative and Systems Pharmacology in the Post-Genomic Era : New Approaches to Discovering Drugs and Understanding Therapeutic Mechanisms An NIH White Paper by the QSP Workshop Group – October, 2011; 2011. p. 47
10. Bai JPF, Abernethy DR. Systems Pharmacology to Predict Drug Toxicity: Integration across Levels of Biological Organization. Annu. Rev. Pharmacol. Toxicol. 2013; 53:451–473. [PubMed: 23140241]
11. Lesko LJ, Zheng S, Schmidt S. Systems Approaches in Risk Assessment. Clin. Pharmacol. Ther. 2013; 93:413–424. [PubMed: 23531724]
12. Zhao S, Iyengar R. Systems Pharmacology: Network Analysis to Identify Multiscale Mechanisms of Drug Action. Annu. Rev. Pharmacol. Toxicol. 2012; 52:505–521. [PubMed: 22235860]
13. Valerio LG, Choudhuri S. Chemoinformatics and Chemical Genomics: Potential Utility of in Silico Methods. J. Appl. Toxicol. 2012
14. Rusyn I, Sedykh A, Low Y, Guyton KZ, Tropsha A. Predictive Modeling of Chemical Hazard by Integrating Numerical Descriptors of Chemical Structures and Short-Term Toxicity Assay Data. Toxicol. Sci. 2012; 127:1–9. [PubMed: 22387746]
15. Pujol A, Mosca R, Farrés J, Aloy P. Unveiling the Role of Network and Systems Biology in Drug Discovery. Trends Pharmacol. Sci. 2010; 31:115–123. [PubMed: 20117850]
16. Kienhuis AS, Bessems JGM, Pennings JLA, Driessen M, Luijten M, van Delft JHM, Peijnenburg AACM, van der Ven LTM. Application of Toxicogenomics in Hepatic Systems Toxicology for Risk Assessment: Acetaminophen as a Case Study. Toxicol. Appl. Pharmacol. 2011; 250:96–107. [PubMed: 20970440]
17. Regulation (EC) No 1907/2006 (REACH Regulation). 2006; 3:1–347.
18. National Academy of Sciences; National Research Council. Toxicity Testing in the 21st Century: A Vision and a Strategy. Washington DC: National Academies Press; 2007. p. 216
19. Judson R, Richard A, Dix DJ, Houck K, Martin M, Kavlock R, Dellarco V, Henry T, Holderman T, Sayre P, Tan S, Carpenter T, Smith E. The Toxicity Data Landscape for Environmental Chemicals. Environ. Health Perspect. 2009; 117:685–695. [PubMed: 19479008]
20. Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Mortensen HM, Reif DM, Rotroff DM, Shah I, Richard AM, Dix DJ. In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project. Environ. Health Perspect. 2010; 118:485–492. [PubMed: 20368123]
21. Collins FS, Gray GM, Bucher JR. Toxicology. Transforming Environmental Health Protection. Science. 2008; 319:906–907. [PubMed: 18276874]
22. OECD Environment Health and Safety Publications. Guidance Document for Using the OECD (Q)SAR Application Toolbox to Develop Chemical Categories according to the OECD Guidance on Grouping of Chemicals. Paris: 2009. Series on Testing and Assessment No 102.
23. Hansch C, Maloney P, Fujita T, Muir R. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. Nature. 1962; 194:178–180.

24. Gleeson MP, Modi S, Bender A, Robinson RLM, Kirchmair J, Promkatkaew M, Hannongbua S, Glen RC. The Challenges Involved in Modeling Toxicity Data in Silico: A Review. Curr. Pharm. Des. 2012; 18:1266–1291. [PubMed: 22316153]
25. Stouch TR, Kenyon JR, Johnson SR, Chen X-Q, Doweyko A, Li Y. In Silico ADME/Tox: Why Models Fail. J. Comput. Aided. Mol. Des. 2003; 17:83–92. [PubMed: 13677477]
26. Scior T, Medina-Franco JL, Do Q-T, Martínez-Mayorga K, Yunes Rojas Ja, Bernard P. How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. Curr. Med. Chem. 2009; 16:4297–4313. [PubMed: 19754417]
27. Zvinavashe E, Murk AJ, Rietjens IMCM. Promises and Pitfalls of Quantitative Structure-Activity Relationship Approaches for Predicting Metabolism and Toxicity. Chem. Res. Toxicol. 2008; 21:2229–2236. [PubMed: 19548346]
28. OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models; 69. Paris: 2007.
29. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol. Inform. 2010; 29:476–488. [PubMed: 27463326]
30. Fourches D, Muratov E, Tropsha A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. J. Chem. Inf. Model. 2010; 50:1189–1204. [PubMed: 20572635]
31. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. Environ. Health Perspect. 2003; 111:1361–1375. [PubMed: 12896860]
32. Golbraikh A, Tropsha A. Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection. Mol. Divers. 2002; 5:357–369.
33. Tropsha A, Golbraikh A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. Curr. Pharm. Des. 2007; 13:3494–3504. [PubMed: 18220786]
34. Benigni R. Structure-Activity Relationship Studies of Chemical Mutagens and Carcinogens: Mechanistic Investigations and Prediction Approaches. Chem. Rev. 2005; 105:1767–1800. [PubMed: 15884789]
35. Sushko I, Novotarskyi S, Körner R, Pandey AK, Cherkasov A, Li J, Gramatica P, Hansen K, Schroeter T, Müller K-R, Xi L, Liu H, Yao X, Öberg T, Hormozdiari F, Dao P, Sahinalp C, Todeschini R, Polishchuk P, Artemenko A, Kuz'min V, Martin TM, Young DM, Fourches D, Muratov E, Tropsha A, Baskin I, Horvath D, Marcou G, Muller C, Varnek A, Prokopenko VV, Tetko IV. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. J. Chem. Inf. Model. 2010; 50:2094–3111. [PubMed: 21033656]
36. Uehara T, Ono A, Maruyama T, Kato I, Yamada H, Ohno Y, Urushidani T. The Japanese Toxicogenomics Project: Application of Toxicogenomics. Mol. Nutr. Food Res. 2010; 54:218–227. [PubMed: 20041446]
37. Afshari CA, Hamadeh HK, Bushel PR. The Evolution of Bioinformatics in Toxicology: Advancing Toxicogenomics. Toxicol. Sci. 2011; 120(Suppl):S225–S237. [PubMed: 21177775]
38. Heijne WHM, Kienhuis AS, van Ommen B, Stierum RH, Groten JP. Systems Toxicology: Applications of Toxicogenomics, Transcriptomics, Proteomics and Metabolomics in Toxicology. Expert Rev. Proteomics. 2005; 2:767–780. [PubMed: 16209655]
39. Chen B, Ding Y, Wild DJ. Assessing Drug Target Association Using Semantic Linked Data. PLoS Comput. Biol. 2012; 8:e1002574. [PubMed: 22859915]
40. Fielden MR, Brennan R, Gollub J. A Gene Expression Biomarker Provides Early Prediction and Mechanistic Assessment of Hepatic Tumor Induction by Nongenotoxic Chemicals. Toxicol. Sci. 2007; 99:90–100. [PubMed: 17557906]
41. Sipes NS, Martin MT, Reif DM, Kleinstreuer NC, Judson RS, Singh AV, Chandler KJ, Dix DJ, Kavlock RJ, Knudsen TB. Predictive Models of Prenatal Developmental Toxicity from ToxCast High-Throughput Screening Data. Toxicol. Sci. 2011; 124:109–127. [PubMed: 21873373]
42. Martin MT, Knudsen TB, Reif DM, Houck KA, Judson RS, Kavlock RJ, Dix DJ. Predictive Model of Rat Reproductive Toxicity from ToxCast High Throughput Screening. Biol. Reprod. 2011; 85:327–339. [PubMed: 21565999]

43. Thomas RS, Black MB, Li L, Healy E, Chu T-M, Bao W, Andersen ME, Wolfinger RD. A Comprehensive Statistical Analysis of Predicting in Vivo Hazard Using High-Throughput in Vitro Screening. Toxicol. Sci. 2012; 128:398–417. [PubMed: 22543276]
44. Zhang M, Chen M, Tong W. Is Toxicogenomics a More Reliable and Sensitive Biomarker than Conventional Indicators from Rats to Predict Drug-Induced Liver Injury in Humans? Chem. Res. Toxicol. 2012; 25:122–129. [PubMed: 22122743]
45. Schmidt CW. TOX 21: New Dimensions of Toxicity Testing. Environ. Health Perspect. 2009; 117:A348–A353. [PubMed: 19672388]
46. Austin CP, Brady LS, Insel TR, Collins FS. NIH Molecular Libraries Initiative. Science. 2004; 306:1138–1139. [PubMed: 15542455]
47. Bolton EE, Wang Y, Thiessen PA, Bryant SH. Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities. Annu. Rep. Comput. Chem. 2008; 4:217–241.
48. ToxNet - Databases on toxicology, hazardous chemicals, environmental health, and toxic releases. http://toxnet.nlm.nih.gov/
49. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems. J. Biomed. Inform. 2008; 41:706–716. [PubMed: 18472304]
50. Blomberg N, Ecker GF, Kidd R, Williams-jones B. Knowledge Driven Drug Discovery Goes Semantic. In. EFMC Yearbook. 2011:39–43.
51. Galaxy. PredPharmTox.
52. Baker NC, Hemminger BM. Mining Connections between Chemicals, Proteins, and Diseases Extracted from Medline Annotations. J. Biomed. Inform. 2010; 43:510–519. [PubMed: 20348023]
53. Boyce RD, Horn JR, Hassanzadeh O, de Waard A, Schneider J, Luciano JS, Rastegar-Mojarad M, Liakata M. Dynamic Enhancement of Drug Product Labels to Support Drug Safety, Efficacy, and Effectiveness. J. Biomed. Semantics. 2013; 4:5. [PubMed: 23351881]
54. Bisgin H, Liu Z, Fang H, Xu X, Tong W. Mining FDA Drug Labels Using an Unsupervised Learning Technique--Topic Modeling. BMC Bioinformatics. 2011; 12:S11.
55. Kuhn M, Campillos M, González P, Jensen LJ, Bork P. Large-Scale Prediction of Drug-Target Relationships. FEBS Lett. 2008; 582:1283–1290. [PubMed: 18291108]
56. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, Ferris TA, Shah NH. Pharmacovigilance Using Clinical Notes. Clin. Pharmacol. Ther. 2013; 93:547–555. [PubMed: 23571773]
57. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. Clin. Pharmacol. Ther. 2012; 91:1010–1021. [PubMed: 22549283]
58. Low Y, Uehara T, Minowa Y, Yamada H, Ohno Y, Urushidani T, Sedykh A, Muratov E, Kuz'min V, Fourches D, Zhu H, Rusyn I, Tropsha A. Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. Chem. Res. Toxicol. 2011; 24:1251–1262. [PubMed: 21699217]
59. Zhu X-W, Sedykh A, Liu S-S. Hybrid in Silico Models for Drug-Induced Liver Injury Using Chemical Descriptors and in Vitro Cell-Imaging Information. J. Appl. Toxicol. 2013
60. Zhang P, Wang F, Hu J, Sorrentino R, Analytics H, Watson IBMTJ, York N. Exploring the Relationship Between Drug Side-Effects and Therapeutic Indications.
61. Sedykh A, Zhu H, Tang H, Zhang L, Richard A, Rusyn I, Tropsha A. Use of in Vitro HTS-Derived Concentration-Response Data as Biological Descriptors Improves the Accuracy of QSAR Models of in Vivo Toxicity. Environ. Health Perspect. 2011; 119:364–370. [PubMed: 20980217]
62. Dietterich, TG. Multiple Classifier Systems. Berlin, Heidelberg: Springer; 2000. Ensemble Methods in Machine Learning; p. 1-15.
63. Breiman L. Random Forests. Mach. Learn. 2001; 45:5–32.
64. Kruhlak NL, Benz RD, Zhou H, Colatsky TJ. (Q)SAR Modeling and Safety Assessment in Regulatory Review. Clin. Pharmacol. Ther. 2012; 91:529–534. [PubMed: 22258468]
65. Wang NCY, Jay Zhao Q, Wesselkamper SC, Lambert JC, Petersen D, Hess-Wilson JK. Application of Computational Toxicological Approaches in Human Health Risk Assessment. I. A Tiered Surrogate Approach. Regul. Toxicol. Pharmacol. 2012; 63:10–19. [PubMed: 22369873]

66. Vilar S, Harpaz R, Santana L, Uriarte E, Friedman C. Enhancing Adverse Drug Event Detection in Electronic Health Records Using Molecular Structure Similarity: Application to Pancreatitis. PLoS One. 2012; 7:e41471. [PubMed: 22911794]

67. Vilar S, Harpaz R, Chase HS, Costanzi S, Rabadan R, Friedman C. Facilitating Adverse Drug Event Detection in Pharmacovigilance Databases Using Molecular Structure Similarity: Application to Rhabdomyolysis. J. Am. Med. Inform. Assoc. 2011; 18(Suppl 1):i73–i80. [PubMed: 21946238]

68. Low Y, Sedykh AY, Fourches D, Golbraikh A, Whelan M, Rusyn I, Tropsha A. Integrative Chemical-Biological Read-Across Approach for Chemical Hazard Classification. Chem. Res. Toxicol. 2013; 26:1199–1208. [PubMed: 23848138]

69. Berger SI, Iyengar R. Network Analyses in Systems Pharmacology. Bioinformatics. 2009; 25:2466–2472. [PubMed: 19648136]

70. Swanson DR. Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. Perspect.Biol.Med. 1986; 30:7–18. [PubMed: 3797213]

71. Oprea TI, Nielsen SK, Ursu O, Yang JJ, Taboureau O, Mathias SL, Kouskoumvekaki L, Sklar La, Bologa CG. Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing. Mol. Inform. 2011; 30:100–111. [PubMed: 22287994]

72. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Côté S, Shoichet BK, Urban L. Large-Scale Prediction and Testing of Drug Activity on Side-Effect Targets. Nature. 2012; 486:361–367. [PubMed: 22722194]

73. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of Drug-Target Interaction Networks from the Integration of Chemical and Genomic Spaces. Bioinformatics. 2008; 24:i232–i240. [PubMed: 18586719]

74. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug Target Identification Using Side-Effect Similarity. Science (80-.). 321:263–266.

75. Tatonetti NP, Liu T, Altman RB. Predicting Drug Side-Effects by Chemical Systems Biology. Genome Biol. 2009; 10:238. [PubMed: 19723347]

76. Cheng F, Li W, Wang X, Zhou Y, Wu Z, Shen J, Tang Y. Adverse Drug Events: Database Construction and in Silico Prediction. J. Chem. Inf. Model. 2013; 53:744–752. [PubMed: 23521697]

77. Cami A, Arnold A, Manzi S, Reis B. Predicting Adverse Drug Events Using Pharmacological Network Models. Sci. Transl. Med. 2011; 3:114ra127.

78. Kimura, T., Matsushita, Y., Yang, YK., Choi, N., Park, B. Pharmacovigilance Systems and Databases in Korea, Japan, and Taiwan. 2011. p. 1237-1245.

79. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ, Schneider R, Skoblo R, Russell RB, Bourne PE, Bork P, Preissner R. SuperTarget and Matador: Resources for Exploring Drug-Target Relationships. Nucleic Acids Res. 2008; 36:D919–D922. [PubMed: 17942422]

80. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-Driven Prediction of Drug Effects and Interactions. Sci. Transl. Med. 2012; 4:125ra31.

81. Zhu H, Ye L, Richard A, Golbraikh A, Wright FA, Rusyn I, Tropsha A. A Novel Two-Step Hierarchical Quantitative Structure-Activity Relationship Modeling Work Flow for Predicting Acute Toxicity of Chemicals in Rodents. Environ. Health Perspect. 2009; 117:1257–1264. [PubMed: 19672406]

82. Zhang, L. Development and Application of Cheminformatics Approaches to Facilitate Drug Discovery and Environmental Toxicity Assessment. University of North Carolina at Chapel Hill; 2011. p. 216

83. Lounkine E, Nigsch F, Jenkins JL, Glick M. Activity-Aware Clustering of High Throughput Screening Data and Elucidation of Orthogonal Structure-Activity Relationships. J. Chem. Inf. Model. 2011; 51:3158–3168. [PubMed: 22098146]

84. Fouchécourt MO, Béliveau M, Krishnan K. Quantitative Structure-Pharmacokinetic Relationship Modelling. Sci. Total Environ. 2001; 274:125–135. [PubMed: 11453289]

85. Gombar VK, Hall SD. Quantitative Structure-Activity Relationship Models of Clinical Pharmacokinetics: Clearance and Volume of Distribution. J. Chem. Inf. Model. 2013; 53:948–957. [PubMed: 23451981]

86. Wambaugh JF, Setzer RW, Reif DM, Gangwal S, Mitchell-Blackwood J, Arnot JA, Joliet O, Frame A, Rabinowitz J, Knudsen TB, Judson RS, Egeghy P, Vallero D, Cohen Hubal EA. High-Throughput Models for Exposure-Based Chemical Prioritization in the ExpoCast Project. Environ. Sci. Technol. 2013; 47:8479–8488. [PubMed: 23758710]

87. Frueh FW, Huang S-M, Lesko LJ. Regulatory Acceptance of Toxicogenomics Data. Environ. Health Perspect. 2004; 112:A663–A664. [PubMed: 15345374]

88. Judson RS, Kavlock RJ, Setzer RW, Cohen Hubal EA, Martin MT, Knudsen TB, Houck Ka, Thomas RS, Wetmore Ba, Dix DJ. Estimating Toxicity-Related Biological Pathway Altering Doses for High-Throughput Chemical Risk Assessment. Chem. Res. Toxicol. 2011; 24:451–462. [PubMed: 21384849]

89. Segal, MR. Machine Learning Benchmarks and Random Forest Regression. 2004.

90. Breiman L. Statistical Modeling: The Two Cultures. Stat. Sci. 2001; 16:199–215.

91. Zhu H, Rusyn I, Richard A, Tropsha A. Use of Cell Viability Assay Data Improves the Prediction Accuracy of Conventional Quantitative Structure-Activity Relationship Models of Animal Carcinogenicity. Environ. Health Perspect. 2008; 116:506–513. [PubMed: 18414635]

92. Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen X-W, Matheny ME, Xu H. Large-Scale Prediction of Adverse Drug Reactions Using Chemical, Biological, and Phenotypic Properties of Drugs. J. Am. Med. Inform. Assoc. 2012; 19:e28–e35. [PubMed: 22718037]

93. Yang L, Wang K, Chen J, Jegga AG, Luo H, Shi L, Wan C, Guo X, Qin S, He G, Feng G, He L. Exploring off-Targets and off-Systems for Adverse Drug Reactions via Chemical-Protein Interactome--Clozapine-Induced Agranulocytosis as a Case Study. PLoS Comput. Biol. 2011; 7:e1002016. [PubMed: 21483481]

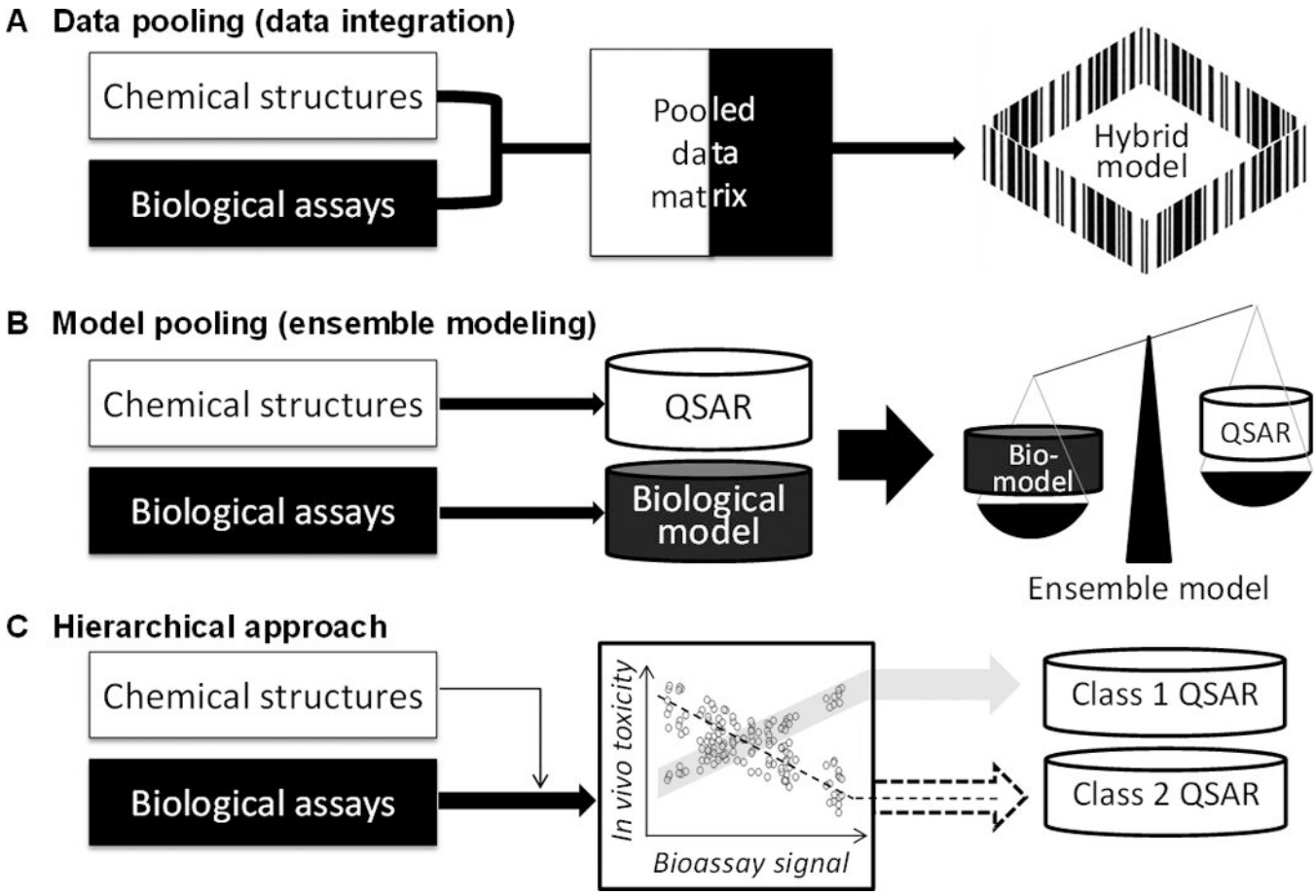

Figure 1.
Integrative chemical-biological approaches for toxicity prediction.

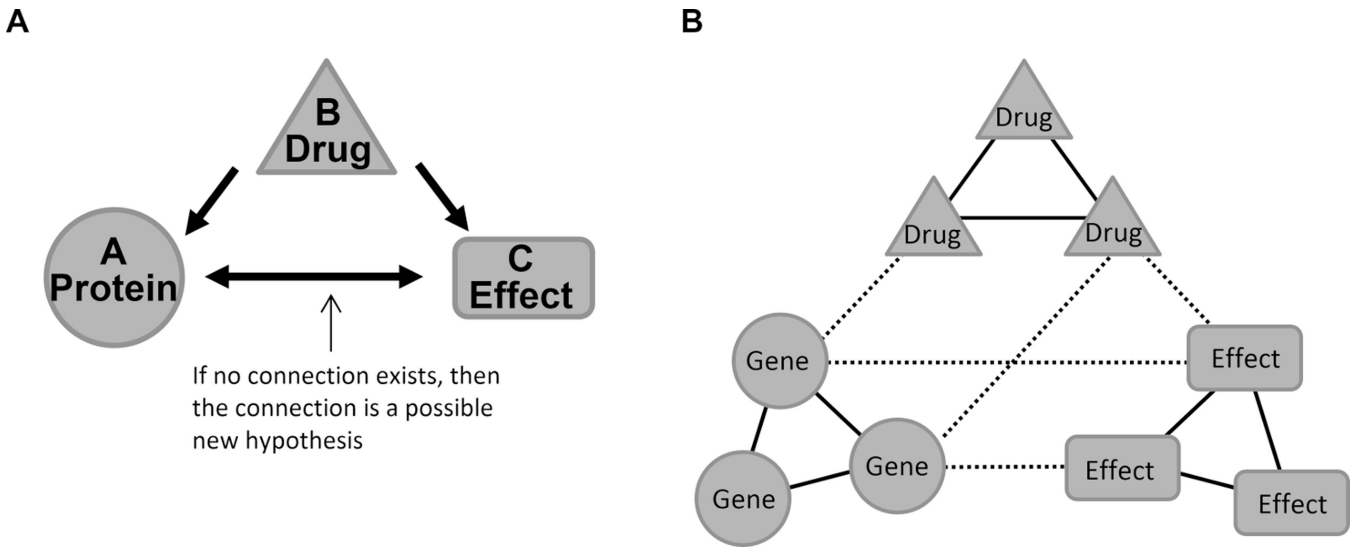

Figure 2.

Knowledge-based relationships between objects (e.g., drugs, targets, and activity phenotypes) form an object network that allows new inferences. (A) Swanson ABC paradigm, adapted from [52]; (B) Network enriched by similarity within classes of objects (solid edges) boost new inferences (dotted edges), adapted from [75].

Table 1

Integrative approaches used for toxicity prediction

Prediction target	Data sources	Type	Studies
Rat LD_{50}	Chemical structure, Cytotoxicity	Data pooling	[91]
Rat LD_{50}	Chemical structure, Dose-cytotoxicity profiles	Data pooling	[61]
Rat LD_{50}	Chemical structure, Cytotoxicity	Integrative method	[81]
Rat reproductive toxicity	Chemical structure, *In vitro* assays	Integrative method	[82]
Drug hepatotoxicity	Chemical structure, Transcriptomics	Data pooling, Model pooling, Integrative method	Chapter 2 [58] Chapter 3
Drug hepatotoxicity	Chemical structure, Hepatocyte imaging assays	Data pooling	[59]
In vivo toxicities	Chemical structure, *In vitro* assays	Data pooling	[43]
Drug properties	Chemical structure, Bioactivity	Integrative method	[83]
Adverse drug reactions	Chemical structure, Electronic health records	Model pooling	[66, 67]
Adverse drug reactions	Chemical structure, Bioactivity, Adverse drug reactions, Therapeutic indications	Data pooling	[92]
Adverse drug reactions	Chemical structure, Drug properties, Adverse drug reactions	Data pooling, Integrative method	[77]
Adverse drug reactions	Chemical structure, Drug targets, Adverse drug reactions, Clinical outcomes	Data pooling, Integrative method	[76]
Adverse drug reactions	Chemical structure, Drug targets, Adverse drug reactions, Therapeutic indications	Data pooling	[60]
Drug targets	Chemical structure, Adverse drug reactions	Data pooling, Model pooling, Integrative method	[72]
Drug targets	Chemical structure, Adverse drug reactions	Data pooling, Model pooling, Integrative method	[74]
Drug targets	Chemical structure, Adverse drug reactions	Data pooling, Model pooling, Integrative method	[71]
Drug targets	Chemical structure, Protein sequence	Data pooling, Integrative method	[73]
Drug targets	Chemical structure, Adverse drug reactions, Therapeutic indications	Data pooling, Integrative method	[39]
Drug targets associated with agranulocytosis	Protein docking profiles Transcriptomics	Integrative method	[93]