Published in final edited form as:
J Chem Inf Model. 2016 July 25; 56(7): 1243–1252. doi:10.1021/acs.jcim.6b00129.

Trust, But Verify II: A Practical Guide to Chemogenomics Data Curation

Denis Fourches[1,*], **Eugene Muratov**[2], and **Alexander Tropsha**[2,*]

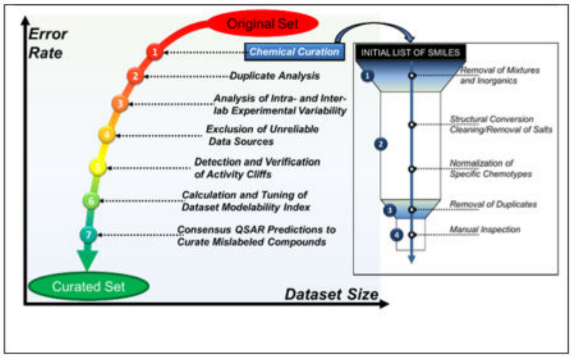
[1]Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, NC, 27695, USA

[2]Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA

Abstract

There is a growing public concern about the lack of reproducibility of experimental data published in peer-reviewed scientific literature. Herein, we review the most recent alerts regarding experimental data quality and discuss initiatives taken thus far to address this problem, especially in the area of chemical genomics. Going beyond just acknowledging the issue, we propose a chemical and biological data curation workflow that relies on existing cheminformatics approaches to flag, and, when appropriate, correct possibly erroneous entries in large chemogenomics datasets. We posit that the adherence to the best practices for data curation is important for both experimental scientists who generate primary data and deposit them in chemical genomics databases and computational researchers who rely on these data for model development.

Graphical Abstract


Error Rate
Original Set
Chemical Curation
INITIAL LIST OF SMILES
1
Removal of Mixtures and Inorganics
2
Duplicate Analysis
3
Analysis of Intra- and Inter-lab Experimental Variability
Structural Conversion Cleaning/Removal of Salts
2
4
Exclusion of Unreliable Data Sources
5
Detection and Verification of Activity Cliffs
Normalization of Specific Chemotypes
6
Calculation and Tuning of Dataset Modelability Index
Removal of Duplicates
3
7
Consensus QSAR Predictions to Curate Mislabeled Compounds
Manual Inspection
4
Curated Set
Dataset Size

[*]please address the correspondence to these authors; dfourch@ncsu.edu or alex_tropsha@unc.edu.

Competing financial interests

The authors declare no competing financial interests.

1. Introduction

Massive screening of large chemical libraries against panels of biological targets (*e.g.,* kinases, GPCRs, or cytochromes)[1] have led to the rapid expansion of publicly available chemogenomics repositories such as ChEMBL[2], PubChem[3], or PDSP[4]. These depositories fuel the initiatives such as the Big Data to Knowledge (BD2K) program at the NIH (https://datascience.nih.gov/bd2k) and enable the development of computational models of chemical bioactivity to guide chemical probe and drug discovery projects.[5,6]

The excitement concerning the growth and availability of chemogenomics data notwithstanding, many serious alerts concerning poor quality and irreproducibility of both chemical and biological records have appeared in the literature[7,8]. For example, Olah et al.[9] showed that on average there were two molecules with erroneous chemical structures per each medicinal chemistry publication with an overall error rate of 8% for compounds indexed in the WOMBAT database[10]. Similarly, Young et al.[11] found error rates for chemical structures in several public and commercial databases ranging from 0.1 to 3.4% depending on the nature of the database. Conversely, looking at the biological data accuracy, Prinz et al.[12] observed that only 20–25% of published assertions concerning purported biological functions for novel deorphanized proteins was consistent with the Bayer's in-house findings. Begley and Ellis[13] discussed a similar analysis performed at Amgen, yielding an even lower rate of reproducibility of 11%. Kramer et al.[14] analyzed the experimental uncertainty of 7,667 independent measurements for 2,540 protein-ligand systems extracted from ChEMBL12; they found a mean error of 0.44 pKi units and a standard deviation of 0.54 pKi units.

In some cases, subtle experimental details such as differences in biological screening technologies were the source of inconsistency. For instance, Ekins et al.[15] observed that the type of dispensing techniques (tip-based versus acoustic) used in HTS could significantly influence the experimental responses measured for the same compounds tested in the same assay; they also showed that these variations could dramatically affect both the prediction performances and interpretation of computational models built for that dataset. While both dispensing techniques are acceptable, this example illustrates the sensitivity of molecular modeling results to even subtle experimental variations sometimes well understood only by specialists in the respective experimental techniques.

A recent editorial in Nature Chemical Biology[16] discussed the urgent need to address the problem of data reproducibility. This problem was also given scrupulous attention by the NIH leadership in another Nature publication[17] co-authored by the NIH Director Francis Collins and the Principal Deputy Director Lawrence Tabak. A recent virtual issue of Nature summarizing all publications on this subject further highlighted the significance of the irreproducibility issue for modern research[18]. In examining various causes of this problem, these papers mainly alluded to the incorrect and inappropriate use of statistics, the clear limitations of preclinical models, and the selective data presentation combined with poor-to-mediocre study design. To start dealing with this issue, Nature recently reinforced the acceptance criteria for manuscripts by removing the space restrictions for method sections and requesting to have external statisticians to verify the correctness of statistical tests reported in the manuscripts considered for publication. This policy change has caused other journals of the Nature family to follow the suit[19]; and now, the NIH maintains the current list of journals and associations or societies publishing preclinical research that endorse NIH-supported principles and guidelines facilitating the reproduction of published experiments. NIH also started a new "*rigor and reproducibility*" web portal[20] in order "*to communicate NIH endorsed principles and guidelines […] concerning rigor and reproducibility*".

Given that even data published in the highest-ranking peer reviewed journals can suffer from poor reliability, it is non-trivial to compile, integrate, and utilize chemogenomics data from any source without at least minimum scrutiny. Data curation is especially critical for computational modelers because their success depends inherently on the accuracy of the data used for model development. Previously, we[21] and others[11] have demonstrated that the prediction performances of QSAR models can be affected by inaccurate and inconsistent representations of chemical structures. In the first paper of this series, we have proposed a workflow focusing on chemical data curation[21]. However, biological data quality also has a direct influence on model accuracy. Recently, we articulated a need to develop a comprehensive chemical-biological data curation workflow that addresses the accuracy of both chemical structures and bioactivities in chemical genomics datasets[22]. Herein, we extend our previous work on chemical data curation[21] and expand upon general principles outlined in the recent brief communication.[22] We propose an integrated chemical and biological data curation workflow incorporation specific protocols for curating both chemical structures and bioactivities in chemical genomics databases that should precede any model development. We posit that both experimental and computational researchers should consider the proposed workflow as a practical guide to the chemogenomics data curation that should be accomplished prior to or in conjunction with the data deposition into public repositories and databases. We expect that adherence to these best practices will prevent the proliferation of irreproducible data in both publications and online repositories and improve the accuracy of data models.

2. An Integrated Workflow for Chemical and Biological Data Curation

Curating both chemical and biological data, *i.e.,* verifying the accuracy, consistency, and reproducibility of the reported experimental data is critical for the success of any cheminformatics studies, but it is especially true for Quantitative Structure-Activity Relationships (QSAR) modeling[11,21]. Curation of chemical structures is a non-trivial task as was shown in our previous publication on this subject[21] but curation of biological data is even more challenging. Indeed, whereas correct canonical chemical structures based on the rules of chemistry do exist for nearly all organic molecules, there are no rules that define what the true accurate value of a biological measurement should be. Nevertheless, we posit that it is possible to flag (and in some cases even fix) suspicious entries in large chemogenomics datasets by applying a series of cheminformatics approaches. To this end, we propose an integrated chemical and biological data curation workflow (Figure 1), that complements our previous protocol for chemical data curation.[21] This workflow includes the following key steps:

(1)

Chemical curation steps include both the identification and correction of structural errors for a set of compounds. Depending on the type of analysis and/or modeling study, this curation process starts with the removal of incomplete or confusing records, such as inorganics, organometallics, counterions, biologics, and mixtures, which most programs for computing molecular descriptors are not equipped to handle. Curation elements should also include structural cleaning (*e.g.,* detection of valence violations or extreme bond lengths and angles), ring aromatization, the normalization of specific chemotypes, and the standardization of tautomeric forms. For large arrays of chemicals, most but not all of these tasks are fully automated. The treatment of tautomers[23] is particularly challenging since the ratio of tautomeric forms can be context-dependent. Sitzmann et al.[24] established empirical rules to consistently treat and represent tautomers to account for the most populated tautomers of a given chemical. Numerous software tools are available to help users conduct the aforementioned tasks such as Molecular Checker/Standardizer (available in Chemaxon JChem[25], which is free for academic organizations), RDKit program tools[26] (free software) or LigPrep (available in the Schrodinger Small Molecule Discovery Suite[27] but only for subscribers of the Schrodinger license). Importantly, one can integrate all these different functions for structural cleaning as a sharable Knime[28] workflow to streamline curation procedures. For instance, the variation of our original data curation workflow[21] was reported in a recent publication[29].

As bioactive chemicals often incorporate stereocenters, it is highly recommended to verify the correctness of stereochemistry: the more asymmetric carbons are present in a molecule, the more likely are the errors in their assignment. The comparison of that chemical entry to similar compounds in online databases may facilitate the detection of incorrect structures and erroneous stereocenters. To this end, PubChem[3] recently implemented a structural standardization workflow for ensuring that all chemicals stored in the database are processed, represented, and standardized the same way using a structured and consistent protocol. Furthermore, Chemspider[30] represents a great example of a crowd-curated database illustrating the power of community engagement and expertise for chemical structure verification. For any chemical, Chemspider indicates how many stereocenters are properly defined and confirmed. Despite these automatic curation tools, manual curation is still critical because some errors obvious for chemists are not obvious for computers.

Even in the case of large datasets in excess of thousands of compounds, we strongly recommend to check manually at least a fraction of the dataset. For instance, to reduce the amount of effort one could check only compounds with complex structures or having a large number of atoms. The most obvious advice is to generate a representative sample of the dataset and check it for the presence of potential erroneous structures.[21] In addition, identification of the subset of “suspicious” compounds with high probability of error for additional checking is described in the step 7 of the proposed workflow. However, inspection or even rejection of the entire data set, however long it may take, may become unavoidable if significant amount of errors is found.

Another solution for processing large data sets, where manual checking of structures and data is almost impossible, is an engagement of scientific community in crowd-sourced curation efforts. This approach seems to be very promising, especially given the success of similar initiatives in Chemspider or Wikipedia, where the quality of crowd-curated chemical data is comparable or higher than the quality of expert-curated data in databases such as ChemIDPlus, DrugBank, etc.

(2)

Processing of bioactivities for chemical duplicates. Often, the same compound is recorded multiple times in chemogenomics depositories[31]. For instance, identical chemicals ordered from different suppliers can be tested in the same assay, sometimes in different laboratories, resulting in different internal substance IDs, different experimental responses, and *in fine* multiple records[32]. QSAR models built with datasets containing many structural duplicates will have artificially skewed predictivity (over-optimistic if activities are similar, or of low accuracy if activities are dissimilar) due to the likely presence of the same compounds in both training and test sets[21]. Dealing with this issue requires the detection of structurally identical compounds in a dataset followed by the comparison of bioactivities reported for the retrieved duplicates. The definition of "identical" compounds depends on how the chemical similarity is computed and the type of chemical descriptors used (*e.g.,* 1D or 2D descriptors cannot differentiate between stereoisomers). Processing of structural duplicates solely based on chemical names, SMILES, and/or CAS numbers is inefficient compared to using InChI and/or 2D structures (see an example in Figure 2). We use standard molecular indices (available in both RDKit and CDK toolkits) and fragment descriptors to compute the similarity between any two compounds; if the similarity is equal to 1, then the two compounds are recognized as duplicates. Freely-accessible tools such as ISIDA-Duplicates[33] or HiT QSAR[34] can identify pairs of duplicates based on molecular descriptors or canonical numeration of chemical graphs computed for each molecule. For instance, we have identified more than 1,200 pairs of structural duplicates (with different substance IDs from different chemical vendors) in the NCGC Cytochrome P450 screening collection[32] in which over 17,000 compounds were tested against five major P450 isozymes. Importantly, as many as 874 out of 1,200 pairs had different reported CYP bioprofiles (see Supplementary Table 1) requiring further examination and curation prior to QSAR modeling.

Finding duplicates in a set of chemicals is relatively trivial with the aforementioned similarity-based programs but the automatic comparison of the bioprofiles for duplicates is not. If the two bioprofiles are identical, one structure can simply be deleted. The task becomes more difficult and time-consuming when these experimental bioprofiles are not identical. In such cases, there are several scenarios to consider:

a. The property value for one compound is clearly wrong (*e.g.,* mis-annotation in the database, labelling error, wrong unit) in which case a neighborhood analysis can help identifying which value is most likely to be correct (see an example in Figure 3);

b. The curation workflow (*e.g.,* the removal of salts and counterions, the standardization of chemical groups) has changed the representation of one substance (or both) and created duplicates. In this case, one should check the original record and verify whether the difference in activity can be explained by the fact that in one case the compound has been tested in its neutral form and in the other case, the compound was tested as a salt. As these instances are common with salts, we recommend determining early on how many chemical salts are present in the dataset and deciding whether only neutral compounds must be considered. In the situation when only a few salts have dramatically different properties than those compounds in their neutral forms, we recommend removing them from the dataset. Otherwise, they will become prediction outliers if compounds in neutral form are used for QSAR modeling.

c. Experimental variability may be at the origin of the discordance for the duplicates' bioprofiles. A specific set of actions is required for the analysis of experimental variability (see next section).

In our practice, activity values expressed in different units are among the most common sources of discordance for duplicates bioprofiles (*e.g.,* a compound recorded as a millimolar inhibitor in database A and as a micromolar inhibitor in database B). The automatic treatment of duplicates is obviously facilitated when all bioactivities are expressed in the same units[35]. As the last resort, suspicious pairs of duplicates with different bioactivities can be discarded altogether; however, the prediction of such compound bioactivity with a QSAR model built with the remaining compounds could help establishing which of the conflicting values is likely to be correct (see also step 7 for more details).

(3)

Analysis of inter- and/or intra-lab experimental variability. The analysis of bioactivities for duplicate compounds also enables the evaluation of both inter- and intra-lab experimental variability. It is a common laboratory practice to measure the bioactivity of a compound in multiple replicates in the same assay. For a given ADMET endpoint (*e.g.,* metabolic stability) with tens of thousands of measured data points available in-house, Big Pharma companies often test hundreds of chemicals in duplicates or triplicates. Datasets containing such information are particularly helpful to study the global experimental variability of an assay (Figure 4) across multiple series of compounds (*e.g.,* experimental variability for most active compounds *versus* most inactive ones) as well as to analyze the local variability within a given range of bioactivity or within a chemical series. Moreover, the bioactivity of reference compounds can even be measured hundreds of times over many months and even years, offering the ultimate assessment of the experimental variability using appropriate metrics[35]. As such type of data is typically not accessible to academic researchers, alternative approaches could be used to assess the experimental variability and its influence on models' prediction performances.

Modelers could rely on data precision as reported in a publication (*e.g.,* "± 0.2 log units"), which is often associated with one (or a few) data point(s) [35]. In practice, it means that the experimental variability should be assessed as constant (Figure 4A). As a result, the highest expected accuracy a QSAR model may not be higher than the experimental variability. For instance, if the mean absolute error for a model is reported to be 0.1 log unit despite an experimental variability close to 0.3 log units, it may be indicative of overfitting.

However, the experimental variability is not necessarily the same for all compounds in a dataset. As illustrated in Figure 4, the assay's variability can be considerably lower for compounds with extreme activities (Figure 4B), or for the more active compounds (Figure 4C), or the less active compounds (Figure 4D). Obviously, the variability can also be random (Figure 4E). We highlight these different profiles of experimental variability as they may have different effect on the prediction performances of QSAR models. For instance, the higher experimental variability for less active compounds may have negligent effect on the model accuracy if such compounds are chemically distinct from more active and more chemically similar molecules.

For modelers who may have access to large screening data collections, we recommend analyzing the baseline history of the target/endpoint of interest as well as all measurements obtained for the reference controls. The baseline history (*e.g.,* per plate, per batch, per week, per month, etc.) facilitates the identification of false-positives and false-negatives in HTS screening (cf. software tools such as HTS Navigator[36]).

It is extremely difficult to assess the inter-laboratory variability since it involves the identification of duplicate compounds having two (or more) different bioactivity values measured in two (or more) independent laboratories. Such replicate measurements are rarely done and/or rarely published, except for some very well-known active (and/or toxic) molecules and reference controls. Drawing any conclusions requires the analysis of many duplicates in addition to globally-accepted estimation of experimental reproducibility associated with a particular assay (*e.g.,* the overall reproducibility is ~85% for the Ames mutagenicity test[37]).

To summarize this step, we shall emphasize that although the duplicated records should be excluded prior to modeling, the analysis of duplicates present in the original dataset is extremely useful to estimate the experimental uncertainty of the data. Grouping of duplicates by data source allows to estimate the inter- and intra-lab variability. In case of high (>30%) discordance between duplicates no reliable models could be obtained. Thus, based on such estimates we could decide (i) whether the development of reliable models is possible for a dataset compiled from the different sources; (ii) whether we should use the data generated in different labs separately; or (iii) whether no model development is possible at all.

(4)

Exclusion of unreliable data sources: The identification and exclusion of data sources with inconsistencies, biases, systematic errors, and/or outdated records is not trivial. One can consider any data source unreliable if there are statistically significant differences in measured bioactivities for a consistent fraction of chemicals. An example could be given by a paper or a series of papers published by the same lab that is not conducting the assay of interest the same way as the majority of research groups do. In such case, the associated data points should probably not be part of the integrated modeling set for developing predictive QSAR models. Other examples of experimental discrepancies relate to chemogenomics measurements published prior to a radical change in the protocol for a given assay, or experimental data from a company site different from the other sites of the same company (*e.g.,* one site is running an assay at pH=7.0 whereas the other site is running the same assay at pH=8.0). These data points are not incorrect *per se* but they will likely be problematic and even incompatible if they are integrated in the same dataset.

It is critical to establish "*golden lists*" of carefully curated chemical databases. Our recent experience in developing "*The Children's Pharmacy Collaborative*™" database[38] shows that establishing such "*golden list*" of sources is rather tedious due to the missing and unclear information requiring expert analysis. For instance, marimastat is a drug approved for the treatment of various cancers according to DrugBank but this compound was annotated elsewhere[39] as performing poorly in clinical trials for the same indication so its development was terminated. Besides the elimination of unreliable sources as a whole, the use of strict filters[35] to identify unwanted data points is an important approach to better ensure the consistency and validity of a set of compounds prior to model development. For instance, Kalliokoski et al.[35] tested a series of filters for automatic preprocessing of IC_{50} values extracted from ChEMBL14, *e.g.,* automatic deletion of data points taken from reviews or articles with undefined authors, deletion of records with unclear units, or records with extreme values.

(5)

__Detection and verification of activity cliffs__: It is important to identify pairs of molecules sharing very high structural similarity but having drastically different bioactivities. Such "activity cliffs"[40] have been viewed as one of the major challenges for accurate bioactivity prediction using cheminformatics approaches, especially QSAR models (*e.g.,* by affecting dataset "modelability"[41]) or any other approach based on chemical similarity. There are different categories of activity cliffs[40] depending on how the similarity of compounds is measured (*e.g.,* 2D, 3D, matching molecular pairs). Prior to initiating the computational study of a dataset, all activity cliffs must be detected, verified, and treated. For each pair of compounds forming an activity cliff, there are two main questions to address: Is it a true cliff? If yes, (how) should these compounds be incorporated in the modeling dataset? The aforementioned software for duplicate searches is useful for retrieving activity cliffs. However, once identified, the activity cliff-forming pair of compounds needs to be analyzed according to the following principles (Figure 5):

a. Bioactivities associated with each compound must be carefully checked against the original data source for accuracy. Mislabeling is one of the main sources of false activity cliffs: for instance, compound A is a 10 nM inhibitor (*very active*) and its most similar molecule in the dataset is compound B annotated as 12 mM inhibitor (*inactive*). An example is given by the case discussed in Figure 3 where using a high activity value for (deemed inaccurate by the chemical similarity clustering as discussed above) Tocris-0740 would make it appear as an activity cliff as compared to any chemically similar molecule in the same table. It is also possible that merely the wrong unit (mM instead of nM) may have been reported and inserted in the database;

b. 2D structural differences between the two compounds must be analyzed and interpreted in the context of the underlying assay or target. To do so, we recommend analyzing chemical features differing between the two compounds:

what is this chemical feature (*e.g.,* a carboxylic moiety, a slightly longer carbon chain, an additional hydroxyl group, a different substituent attached to an aromatic ring)? Is it likely to be responsible for that dramatic change in activity? If no, there is likely something wrong with the record. If yes, what descriptors can reflect such an extreme structure-activity relationship?

c. 3D structural differences should also be considered. For such activity cliffs, very significant differences in experimental bioactivities can be due to minor changes in the receptor-ligand interactions. Thus, it makes sense to study 2D activity cliffs identified in a dataset by analyzing them in 3D, preferably in the context of receptor-ligand interactions. If the 3D structure of the target is known as well as the binding mode of at least one compound from the pair under consideration, it is feasible to compare the difference in receptor-ligand interactions. Tools like the structure builder in Schrodinger's Maestro are very useful in that regard. These differences in receptor-ligand interactions (e.g., steric constrains, H-bond or π-π stacking disruption) might be at the origin of the cliff formation. If no 3D structure is available for the receptor, one can still superimpose the two compounds in 3D and study how these two conformations differ from each other. We should still underline that determining the 3D conformation of a flexible chemical is very much context-dependent and even the "active conformation" of a molecule in the binding pocket of an enzyme corresponds to an ensemble of local metastable conformations. Therefore, the calculation of 3D descriptors for chemicals requires detailed justification of the choice of molecular conformations.

d. Finally, the modeler should decide whether to keep or discard any activity cliff-forming compounds in the modeling set.

Interestingly, one can enrich the set of chemical descriptors used for QSAR modeling for better discriminating activity cliffs: for example, 2D descriptors cannot differentiate 3D activity cliffs; therefore, adding chirality-encoding descriptors may help converting a 2D activity cliff into structurally more different compounds. Moreover, the researcher can choose a different modeling technique (*e.g.,* docking) to better model the cliffs. In a recent study, Hu and Bajorath[42] have compared 2D versus 3D activity cliffs identified for different targets such as β-secretase1 and factor Xa. They found that more than 60% of 2D activity cliffs could be distinguished at the 3D level. At the same time, these authors noticed that some 3D activity cliffs with no notable differences regarding their receptor-ligand interaction could easily be distinguished at the 2D level. Thus, 2D and 3D similarity assessment should complement each other.[43]

(6)

<u>Calculation and tuning of the dataset modelability index (MODI)</u>: Recently, we introduced the concept of "*dataset modelability*"[41], *i.e.,* an *a priori* estimate of the feasibility to obtain predictive QSAR models for a given set of chemicals. This concept has emerged from analyzing the impact of activity and similarity cliffs on the overall performance of QSAR models[40]. The calculation of MODI helps modelers to (i) quickly evaluate the feasibility of obtaining QSAR models with significant predictive power; (ii) identify a subset of

compounds with higher modelability (especially when the activity distribution is biased towards one activity class or an activity range); and (iii) choose which set of chemical descriptors will likely produce QSAR models of the higher prediction power. If the MODI index is low, then the development of QSAR model for the respective dataset is not recommended. One should note that different sets of 2D chemical descriptors will return different but similar MODI values for a given dataset. Switching from 2D to 3D descriptors[44] or from chemical to biological descriptors[45] can help to increase the modelability of a given dataset. This work was recently extended by Marcou et al.[46]

(7)

Consensus QSAR prediction to curate mislabeled compounds. The underlying idea of consensus predictions is that an implicit SAR for a given dataset can be formally manifested by a variety of QSAR models built with different types of molecular descriptors and diverse machine learning approaches. Rigorously built individual models form an ensemble that allows for consensus bioactivity prediction using all models at once. The development of consensus models is generally recommended as they reach higher prediction performances as well as better dataset coverage due to their larger applicability domains[47]. As a result, consensus models can be used to flag and sometimes correct the experimental measurements of mislabeled compounds in a dataset. Consensus models can also be utilized to forecast the properties of pairs of stereoisomers and potentially identify the mis-annotated ones. Briefly, a compound can be considered "*suspicious*" and selected for deeper examination if: *(i)* all models involved in the consensus ensemble failed to predict its activity accurately; and *(ii)* it belongs to a tight cluster of two or more structurally similar compounds where all the other compounds have different (but mutually similar) bioactivities. We have demonstrated the success of this approach using Ames mutagenicity[21] data. Employing both manual and automatic literature mining tools, our analysis revealed that 31 out of 140 "suspicious" compounds (~22.1%) were annotated erroneously in the original dataset. Recently, the same approach was used for the identification of mislabeled compounds in the largest set of chemical-induced skin sensitization records[48] available in the public domain. Thus, predictive QSAR models obtained for carefully curated datasets can be successfully employed as a key component in biological data curation workflows.

Concluding this section, we shall note that although each step of the developed workflow can be done using existing cheminformatics techniques and software tools, project-specific decision-making is an inherent part of almost every part of the process. In addition to the detection and elimination of erroneous, non-standardized, and duplicated chemical structures (steps 1 and 2); records associated with unreliable data sources or high experimental variability (steps 3 and 4); structural outliers and unverified activity cliffs (steps 5 and 6), some mislabeled compounds can also be identified and corrected (step 7). Moreover, global decision regarding the very feasibility of model development could become a necessity when analyzing the outcome of the steps 2, 3, and 6. For instance, if duplicate analysis (step 2) will show high (>30%) discordance of associated activity values that could not be explained by inter-lab variability (step 3) or MODI value (step 6) of the resulting dataset would be lower than cut-off value of 0.65, we would not recommend the development of QSAR models.

Sometimes, the dataset curated using the proposed workflow could become too small or unbalanced with one activity class underrepresented to ensure the development of reliable models. Thus, if after curation the dataset includes less than 40 compounds;[47] activity range is less than 3 logarithmic units or contains large gaps that exceed 10%–15% of the entire range (for continuous datasets);[47] one activity class contain less than 20 compounds (for binary models),[47] we recommend the use of similarity searching or docking (if possible) instead of QSAR modeling. At the same time, the curated dataset could be stored and later enriched with new experimental data that would make it suitable for modeling.

3. Perspectives

The data-processing steps summarized in the previous section incorporate the key procedures our research groups at NCSU and UNC currently employ to prepare, curate, and standardize any chemogenomics dataset prior to its modeling. It is important to note that the order of execution of these procedures can be modified according to the size and exact nature of the underlying dataset. In the absence of such well-defined curation workflow, researchers that in any way rely on historical data for planning their future experiments are running the risk of conducting their analysis and modeling with incomplete, incorrect, inaccurate, inconsistent, or irreproducible data points (the “five i’s” danger)[22]. In our previous study[21], we showed that chemical curation was critical to maximize the prediction abilities of QSAR models. We cannot stress enough how important data curation is to achieve reliable and useful QSAR models.

With the rapid accumulation of experimental data in publicly-available repositories, the problem of knowledge extraction from data, i.e., *data science*, becomes increasingly critical to enable further progress of research. The knowledge is summarized in the form of models that enable rational, data driven decision support. In a recent famous statement, the President of the Royal Society of Chemistry Professor Dominic Tildesley predicted that “*the advances in modeling and informatics are so dramatic that in 15 years’ time, no chemist will be doing any experiments at the bench without trying to model them first*”.[49] This expectation can be extended towards many biological and biomedical disciplines implying that experimental scientists will increasingly rely on the power of computational models to rationally direct their studies. The aforementioned ability of models to provide as accurate estimates of the experimental data as the experiment itself does not appear to be restricted to chemical biology. For instance, pharmacokinetic models often yield time-dependent drug concentration curves that are more accurate than the experimental ones.[50] Specifically, it has been acknowledged that individual time-point measurements could be off significantly whereas predicted values give accurate results that can be confirmed when the experiment is reproduced. However, computational model accuracy depends directly on the accuracy of the data used to build a model. This consideration places the issue of data irreproducibility and accuracy at the forefront of computational modeling research, emphasizing the need for data curation as the first critical step of model development.

The vast majority of synthesized compounds is reported in the literature only once.[14] Moreover, it is uncommon to find their biological assay measurements reported in replicates by multiple groups. Therefore, a full analysis of the reliability of all chemical and biological records in a dataset is extremely difficult to accomplish[14,35]; thus, the most logical way to ensure the accuracy of the reported data is to place this responsibility for ensuring data accuracy on the experimental researchers who generate data at first. Clearly, experimental scientists should know and understand their data the best. Therefore, ideally, the best approach to minimizing the risk of errors is to have experimental scientists upload their data electronically to the respective databases simultaneously with the manuscript submission. This is the standard practice in macromolecular crystallography where coordinates of protein or nucleic acid structures must be uploaded to the Protein Data Bank (PDB) prior to the manuscript submission. This should also become mandatory for newly synthesized compounds tested in bioassays prior to their publication in medicinal chemistry and chemical biology journals.

Journal editors should also consider supporting and encouraging the implementation of electronic protocols and associated file formats for chemical data storing and sharing. Indeed, the pdf format used to store and distribute journal papers is inefficient for extracting chemical data, re-plotting the graphs, or mining molecular structures, i.e., the pdf format is far from being cheminformatics-ready. Fortunately, new file formats are slowly emerging to enable the aforementioned tasks: for instance, MIABE[51] and BAO[52] were specifically designed to ensure a consistent storage, ontology and information architecture to characterize bioassays and their results. Extending the use of these formats is the only rational way to enable machine-readable descriptions of bioassays, molecular structures, experimental protocols, and NMR spectra reported in scientific publications. These formats will also facilitate the integration and comparison of experimental data from different research groups. Another plausible approach to increase data completeness and accuracy is to employ crowd-solving and crowd-checking approaches that will help in discovering and correcting erroneous entries in publicly available databases (*e.g.,* Chemspider or Wikipedia). For instance, our group had a very encouraging experience with ChEMBL, when a reported inaccuracy in β_2-adrenergic agonists binding affinity was fixed in less than 24 hours by the ChEMBL team at EBI. However, one should note that the traceability of these corrections is almost nonexistent, i.e., correcting a pKi value in an online database such as ChEMBL will not autocorrect all the instances of that particular record in other online databases, e.g., PubChem that includes data from ChEMBL, nor will it correct the same instance in a locally-stored SD file. In the absence of such connectivity between databases containing similar data, data curation workflows described in this paper should be applied to every database and dataset independently.

4. Conclusions

The exploitation of today's online repositories containing large sets of heterogeneous chemogenomics data requires the use of powerful, transparent, and robust data curation workflows. Although such procedures are required and implemented for submitting novel protein crystal structures to the PDB (*e.g.,* AutoDep Input Tool), curation is still far from being *condicio sine qua non* in chemical and biological data analysis leading to reporting erroneous or irreproducible data in published manuscripts. Since the presence of erroneous data points is known to cause computational models to fail or have low predictive power, chemical biological curation workflows can be utilized to flag (and where possible fix) those records and ultimately improve the quality of data analysis and prediction performances of modeling techniques (*e.g.,* pharmacophore, QSAR, and docking). Although this Perspective focuses predominantly on chemical biology, similar problems are common for the entire biomedical research[53] and they plague other fields as well, *e.g.,* nanotechnology[54].

Reducing the amount of erroneous or irreproducible results requires coordinated efforts between research community, funding agencies, and journal publishers. As funding agencies such as NIH or NSF are starting to establish data sharing policies, we believe the chemical biology community (and for that matter, research community of any data-rich discipline) should adopt the culture and mechanisms for data sharing established within the structural biology community. To do so, experimental researchers should be provided with computational tools to curate, organize, and submit their data to specialized repositories or databases. Importantly, these databases should be certified by the respective funding agencies and supported by peer-reviewed, competitive grants, which is how Protein Databank has been supported for many years. An agreement should be reached between funding agencies and scientific journals that no paper reporting new data could be accepted without providing a statement from the respective database or repository that they have received that data. This would be similar to rules established in structural biology where most journals will not even consider a manuscript describing a new X-ray or NMR-characterized protein structure without a confirmation from the Protein Databank that coordinates have been deposited. Such agreements are possible and can be illustrated by the practice established by the NIH several years ago that all published papers should be uploaded to PubMed Central within a year following the original publication.

The basic gold standard for reporting scientific results and ensuring their correctness will always rely on whether or not the experiments described in a study can be reproduced using the information provided by the authors. Nevertheless, curation workflows for chemical genomics data may contribute to establishing the best practices and culture of data curation as an essential component of further progress in our discipline.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully thank the financial support from NIH (GM096967 and GM66940). DF thanks the NCSU Chancellor's Faculty Excellence Program.

References

1. Elkins JM, Fedele V, Szklarz M, Abdul Azeez KR, Salah E, Mikolajczyk J, Romanov S, Sepetov N, Huang XP, Roth BL, Al Haj Zen A, Fourches D, Muratov E, Tropsha A, Morris J, Teicher BA, Kunkel M, Polley E, Lackey KE, Atkinson FL, Overington JP, Bamborough P, Müller S, Price DJ, Willson TM, Drewry DH, Knapp S, Zuercher WJ. Comprehensive Characterization of the Published Kinase Inhibitor Set. Nat Biotechnol. 2015; 34:95–103. [PubMed: 26501955]
2. [accessed Mar 1, 2016] ChEMBL Database. https://www.ebi.ac.uk/chembl/
3. PubChem. [accessed Mar 1, 2016] http://pubchem.ncbi.nlm.nih.gov/
4. PDSP. [accessed Mar 1, 2016] http://pdsp.med.unc.edu

5. Hu Y, Bajorath J. Learning from "Big Data": Compounds and Targets. Drug Discov Today. 2014; 19:357–360. [PubMed: 24561327]

6. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A. QSAR Modeling: Where Have You Been? Where Are You Going To? J Med Chem. 2014; 57:4977–5010. [PubMed: 24351051]

7. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O, Rhee SY. Big Data: The Future of Biocuration. Nature. 2008; 455:47–50. [PubMed: 18769432]

8. Frye SV, Arkin MR, Arrowsmith CH, Conn PJ, Glicksman MA, Hull-Ryde EA, Slusher BS. Tackling Reproducibility in Academic Preclinical Drug Discovery. Nat Rev Drug Discov. 2015; 14:733–734. [PubMed: 26388229]

9. Olah, M., Mracec, M., Ostopovici, L., Rad, R., Bora, A., Hadaruga, N., Olah, I., Banda, M., Simon, Z., Oprea, TI. WOMBAT: World of Molecular Bioactivity. In: Oprea, TI., editor. Chemoinformatics in Drug Discovery. Wiley-VCH; New York: 2005. p. 223-239.

10. Sunset Molecular Discovery, L. L. C. Wombat. [accessed Mar 1, 2016] http://sunsetmolecular.com/products/?id=4

11. Young D, Martin D, Venkatapathy R, Harten P. Are the Chemical Structures in Your QSAR Correct? QSAR Comb Sci. 2008; 27:1337–1345.

12. Prinz F, Schlange T, Asadullah K. Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets? Nat Rev Drug Discov. 2011; 10:712. [PubMed: 21892149]

13. Begley CG, Ellis LM. Drug Development: Raise Standards for Preclinical Cancer Research. Nature. 2012; 483:531–533. [PubMed: 22460880]

14. Kramer C, Kalliokoski T, Gedeck P, Vulpetti A. The Experimental Uncertainty of Heterogeneous Public K(i) Data. J Med Chem. 2012; 55:5165–5173. [PubMed: 22643060]

15. Ekins S, Olechno J, Williams AJ. Dispensing Processes Impact Apparent Biological Activity as Determined by Computational and Statistical Analyses. PLoS One. 2013; 8:e62325. [PubMed: 23658723]

16. Editorial. Facilitating Reproducibility. Nat Chem Biol. 2013; 9:345. [PubMed: 23689620]

17. Collins FS, Tabak LA. Policy: NIH Plans to Enhance Reproducibility. Nature. 2014; 505:612–613. [PubMed: 24482835]

18. Editorial. Repetitive Flaws. Nature. 2016; 529:256. [PubMed: 26791685]

19. Editorial. Joining the Reproducibility Initiative. Nat Nanotechnol. 2014; 9:949. [PubMed: 25466531]

20. National Institute of Health. [accessed May 25, 2016] NIH Rigor and Reproducibility. https://www.nih.gov/research-training/rigor-reproducibility

21. 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]

22. Fourches D, Muratov E, Tropsha A. Curation of Chemogenomics Data. Nat Chem Biol. 2015; 11:535. [PubMed: 26196763]

23. Martin YC. Let's Not Forget Tautomers. J Comput Aided Mol Des. 2009; 23:693–704. [PubMed: 19842045]

24. Sitzmann M, Ihlenfeldt WD, Nicklaus MC. Tautomerism in Large Databases. J Comput Aided Mol Des. 2010; 24:521–551. [PubMed: 20512400]

25. ChemAxon. [accessed Mar 1, 2016] http://www.chemaxon.com

26. RDKit. [accessed Mar 1, 2016] http://www.rdkit.org/

27. Schrodinger. [accessed Mar 1, 2016] LigPrep. http://www.schrodinger.com/LigPrep/

28. KNIME. [accessed Mar 1, 2016] http://www.knime.org/

29. Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, Zakharov A, Worth A, Richard AM, Grulke CM, Trisciuzzi D, Fourches D, Horvath D, Benfenati E, Muratov E, Wedebye EB, Grisoni F, Mangiatordi GF, Incisivo GM, Hong H, Ng HW, Tetko IV, Balabin I, Kancherla J, Shen J, Burton J, Nicklaus M, Cassotti M, Nikolov NG, Nicolotti O, Andersson PL,

Zang Q, Politi R, Beger RD, Todeschini R, Huang R, Farag S, Rosenberg SA, Slavov S, Hu X, Judson RS. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. Environ Health Perspect. 2016; Advance publication. doi: 10.1289/ehp.1510267

30. ChemSpider. [accessed Mar 1, 2016] http://www.chemspider.com

31. Baurin N, Baker R, Richardson C, Chen I, Foloppe N, Potter A, Jordan A, Roughley S, Parratt M, Greaney P, Morley D, Hubbard RE. Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds. J Chem Inf Comput Sci. 2004; 44:643–651. [PubMed: 15032546]

32. Veith H, Southall N, Huang R, James T, Fayne D, Artemenko N, Shen M, Inglese J, Austin CP, Lloyd DG, Auld DS. Comprehensive Characterization of Cytochrome P450 Isozyme Selectivity across Chemical Libraries. Nat Biotechnol. 2009; 27:1050–1055. [PubMed: 19855396]

33. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, Solov'ev V, Hoonakker F, Tetko IV, Marcou G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. Curr Comput Aided Drug Des. 2008; 4:191–198.

34. Kuz'min VE, Artemenko AG, Muratov EN. Hierarchical QSAR Technology Based on the Simplex Representation of Molecular Structure. J Comput Aided Mol Des. 2008; 22:403–421. [PubMed: 18253701]

35. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P. Comparability of Mixed IC50 Data - a Statistical Analysis. PLoS One. 2013; 8:e61007. [PubMed: 23613770]

36. Fourches D, Sassano MF, Roth BL, Tropsha A. HTS Navigator: Freely-Accessible Cheminformatics Software for Analyzing High-Throughput Screening Data. Bioinformatics. 2013; 30:588–589. [PubMed: 24376084]

37. Benigni R, Giuliani A. Computer-Assisted Analysis of Interlaboratory Ames Test Variability. J Toxicol Environ Health. 1988; 25:135–148. [PubMed: 3418743]

38. Blatt J, Farag S, Corey SJ, Sarrimanolis Z, Muratov E, Fourches D, Tropsha A, Janzen WP. Expanding the Scope of Drug Repurposing in Pediatrics: The Children's Pharmacy Collaborative^{TM}. Drug Discov Today. 2014; 19:1696–1698. [PubMed: 25149597]

39. Sparano JA, Bernardo P, Stephenson P, Gradishar WJ, Ingle JN, Zucker S, Davidson NE. Randomized Phase III Trial of Marimastat versus Placebo in Patients with Metastatic Breast Cancer Who Have Responding or Stable Disease after First-Line Chemotherapy: Eastern Cooperative Oncology Group Trial E2196. J Clin Oncol. 2004; 22:4683–4690. [PubMed: 15570070]

40. Maggiora GM. On Outliers and Activity Cliffs--Why QSAR Often Disappoints. J Chem Inf Model. 2006; 46:1535. [PubMed: 16859285]

41. Golbraikh A, Muratov EN, Fourches D, Tropsha A. Dataset Modelability by QSAR. J Chem Inf Model. 2014; 54:1–4. [PubMed: 24251851]

42. Hu Y, Bajorath J. Exploration of 3D Activity Cliffs on the Basis of Compound Binding Modes and Comparison of 2D and 3D Cliffs. J Chem Inf Model. 2012; 52:670–677. [PubMed: 22394306]

43. Fourches D, Muratov E, Ding F, Dokholyan NV, Tropsha A. Predicting Binding Affinity of CSAR Ligands Using Both Structure-Based and Ligand-Based Approaches. J Chem Inf Model. 2013; 53:1915–1922. [PubMed: 23809015]

44. Kuz'min VE, Muratov EN, Artemenko AG, Varlamova EV, Gorb L, Wang J, Leszczynski J. Consensus QSAR Modeling of Phosphor-Containing Chiral AChE Inhibitors. QSAR Comb Sci. 2009; 28:664–677.

45. 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]

46. Marcou G, Horvath D, Varnek A. Kernel Target Alignment Parameter: A New Modelability Measure for Regression Tasks. J Chem Inf Model. 2016; 56:6–11. [PubMed: 26673976]

47. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inform. 2010; 29:476–488. [PubMed: 27463326]

48. Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. Predicting Chemically-Induced Skin Reactions. Part I: QSAR Models of Skin Sensitization and

Their Application to Identify Potentially Hazardous Compounds. Toxicol Appl Pharmacol. 2015; 284:262–272. [PubMed: 25560674]

49. [accessed May 25, 2016] Royal Society of Chemistry Next RSC president predicts that in 15 years no chemist will do bench experiments without computer-modelling them first. http://www.rsc.org/AboutUs/News/PressReleases/2013/Dominic-Tildesley-Royal-Society-of-Chemistry-President-Elect.asp

50. Nielsen EI, Friberg LE. Pharmacokinetic-Pharmacodynamic Modeling of Antibacterial Drugs. Pharmacol Rev. 2013; 65:1053–1090. [PubMed: 23803529]

51. Orchard S, Al-Lazikani B, Bryant S, Clark D, Calder E, Dix I, Engkvist O, Forster M, Gaulton A, Gilson M, Glen R, Grigorov M, Hammond-Kosack K, Harland L, Hopkins A, Larminie C, Lynch N, Mann RK, Murray-Rust P, Lo Piparo E, Southan C, Steinbeck C, Wishart D, Hermjakob H, Overington J, Thornton J. Minimum Information about a Bioactive Entity (MIABE). Nat Rev Drug Discov. 2011; 10:661–669. [PubMed: 21878981]

52. Clark AM, Litterman NK, Kranz JE, Gund P, Gregory K, Bunin BA. BioAssay Templates for the Semantic Web. PeerJ Comput Sci. 2016; 2:e61.

53. Editorial. Journals Unite for Reproducibility. Nature. 2014; 515:7–7.

54. Fourches D, Pu D, Li L, Zhou H, Mu Q, Su G, Yan B, Tropsha A. Computer-Aided Design of Carbon Nanotubes with the Desired Bioactivity and Safety Profiles. Nanotoxicology. 2015:1–10.

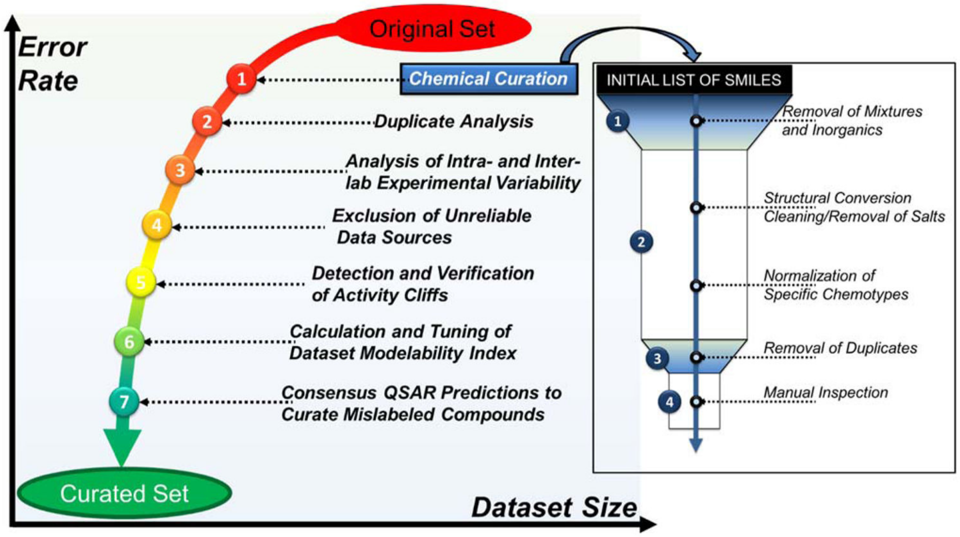
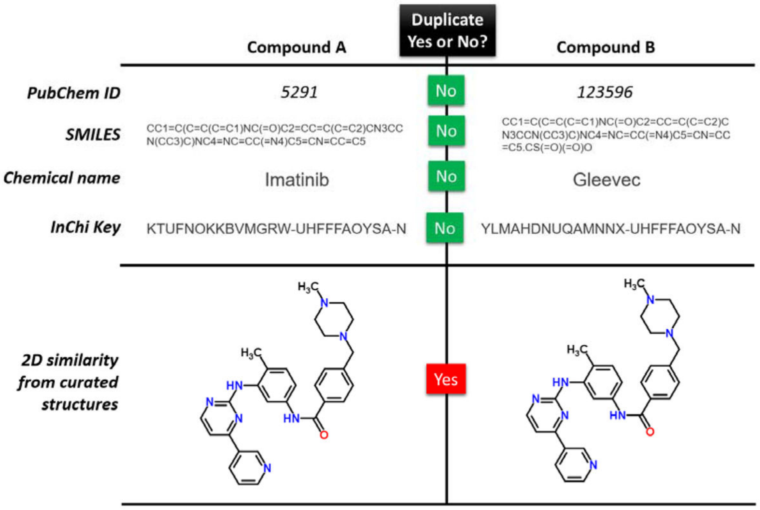

Error Rate
Original Set
Chemical Curation
1
2 Duplicate Analysis
3 Analysis of Intra- and Inter-lab Experimental Variability
4 Exclusion of Unreliable Data Sources
5 Detection and Verification of Activity Cliffs
6 Calculation and Tuning of Dataset Modelability Index
7 Consensus QSAR Predictions to Curate Mislabeled Compounds
Curated Set
Dataset Size
INITIAL LIST OF SMILES
1 Removal of Mixtures and Inorganics
2
Structural Conversion Cleaning/Removal of Salts
Normalization of Specific Chemotypes
3 Removal of Duplicates
4 Manual Inspection

Figure 1.

General workflow for comprehensive curation of chemogenomics datasets.

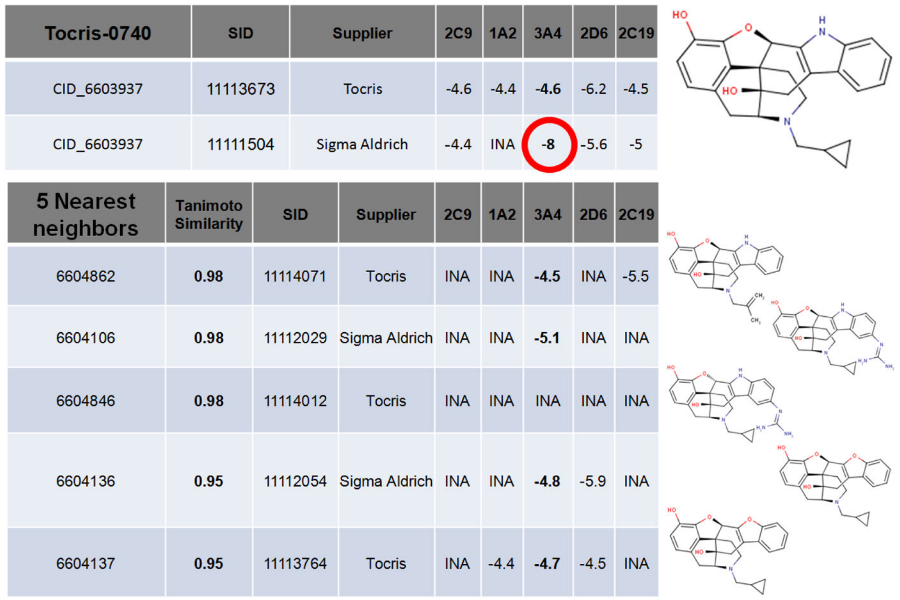


	Compound A	Duplicate Yes or No?	Compound B
PubChem ID	*5291*	No	*123596*
SMILES	CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5	No	CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5.CS(=O)(=O)O
Chemical name	Imatinib	No	Gleevec
InChi Key	KTUFNOKKBVMGRW-UHFFFAOYSA-N	No	YLMAHDNUQAMNNX-UHFFFAOYSA-N
2D similarity from curated structures		Yes	

H$_3$C
N
N
H$_3$C
HN
N
N
HN
O
N
H$_3$C
N
N
H$_3$C
HN
N
N
HN
O
N

Figure 2.

Example of duplicate retrieval using PubChem ID, smiles, chemical names, InChI, and 2D similarity. Note the 2D similarity was computed as Tanimoto coefficient using CDK descriptors and had Tc = 1 (implicating structural duplicates) for the two curated compounds (no salts, standardized functional groups and aromatization).



Tocris-0740	SID	Supplier	2C9	1A2	3A4	2D6	2C19
CID_6603937	11113673	Tocris	-4.6	-4.4	**-4.6**	-6.2	-4.5
CID_6603937	11111504	Sigma Aldrich	-4.4	INA	**-8**	-5.6	-5

5 Nearest neighbors	Tanimoto Similarity	SID	Supplier	2C9	1A2	3A4	2D6	2C19
6604862	**0.98**	11114071	Tocris	INA	INA	**-4.5**	INA	-5.5
6604106	**0.98**	11112029	Sigma Aldrich	INA	INA	**-5.1**	INA	INA
6604846	**0.98**	11114012	Tocris	INA	INA	INA	INA	INA
6604136	**0.95**	11112054	Sigma Aldrich	INA	INA	**-4.8**	-5.9	INA
6604137	**0.95**	11113764	Tocris	INA	-4.4	**-4.7**	-4.5	INA

Figure 3.

Chemical similarity analysis for a pair of substances with duplicate structures found in the NCGC Cytochrome P450 screen[32] in which 17,000 compounds were tested against five major isozymes. LogAC$_{50}$ = -8 for CID_6603937 at CYP 3A4 (highlighted by the red circle) is automatically flagged as incorrect because highly similar molecules in the same dataset have CYP 3A4 activities consistent with an alternative measurement for the same compound.

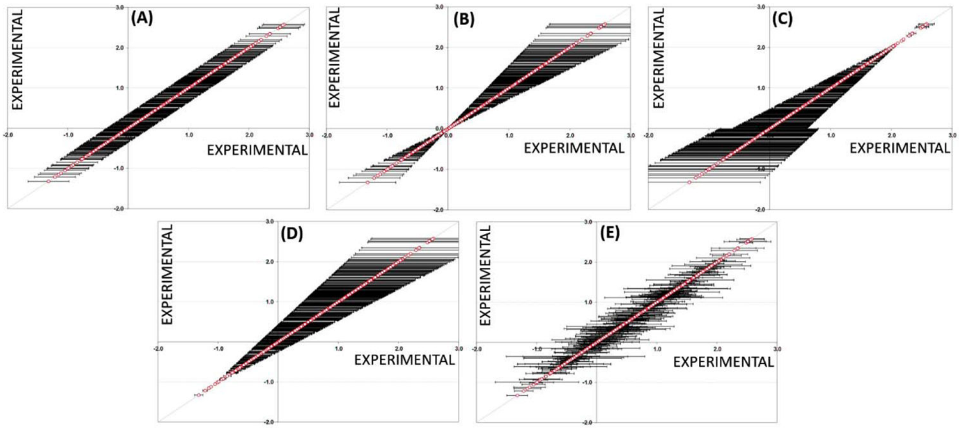


Figure 4.

Illustration of different types of experimental variability (see text for details).

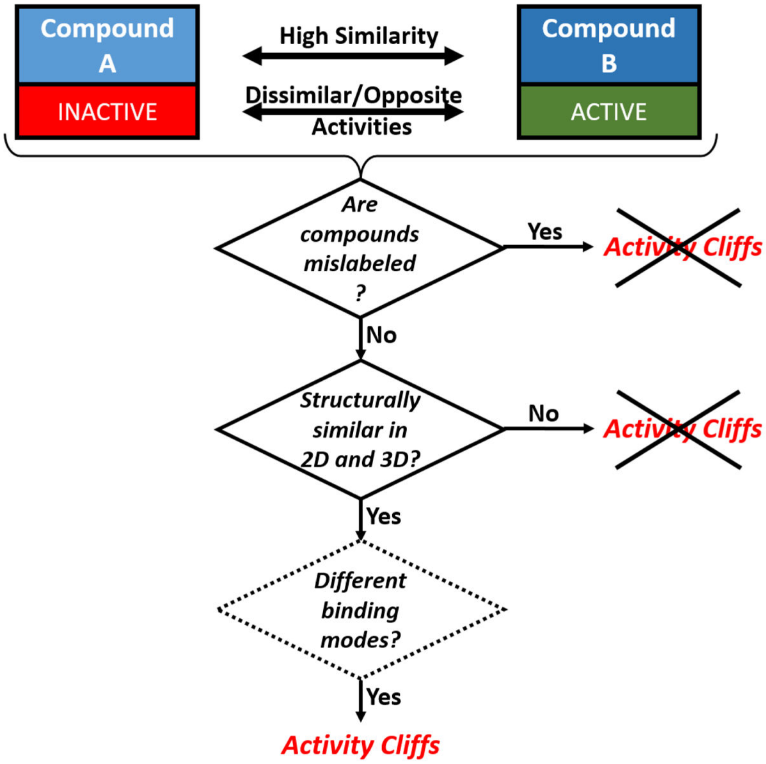


Figure 5.
Initial workflow to analyze activity cliffs.