JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

QSAR Modeling Using Large-Scale Databases: Case Study for HIV-1 Reverse Transcriptase Inhibitors

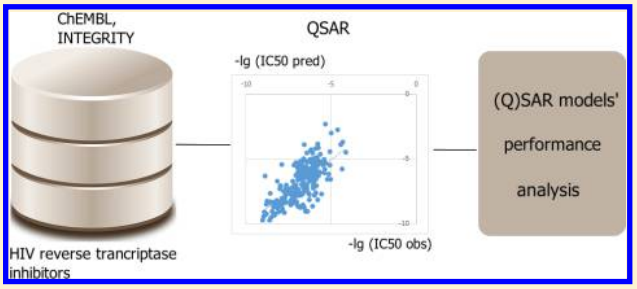
Olga A. Tarasova,*[,†] Aleksandra F. Urusova,[†] Dmitry A. Filimonov,[†] Marc C. Nicklaus,[‡] Alexey V. Zakharov,[‡] and Vladimir V. Poroikov[†]

[†]Institute of Biochemical Chemistry, 10-8, Pogodinskaya St., 119121, Moscow, Russia

[‡]CADD Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, DHHS, NCI-Frederick, 376 Boyles St., Frederick, Maryland 21702, United States

Supporting Information

ABSTRACT: Large-scale databases are important sources of training sets for various QSAR modeling approaches. Generally, these databases contain information extracted from different sources. This variety of sources can produce inconsistency in the data, defined as sometimes widely diverging activity results for the same compound against the same target. Because such inconsistency can reduce the accuracy of predictive models built from these data, we are addressing the question of how best to use data from publicly and commercially accessible databases to create accurate and predictive QSAR models. We investigate the suitability of commercially and publicly available databases to QSAR modeling of antiviral activity (HIV-1 reverse transcriptase (RT) inhibition). We present several methods for the creation of modeling (i.e., training and test) sets from two, either commercially or freely available, databases: Thomson Reuters Integrity and ChEMBL. We found that the typical predictivities of QSAR models obtained using these different modeling set compilation methods differ significantly from each other. The best results were obtained using training sets compiled for compounds tested using only one method and material (i.e., a specific type of biological assay). Compound sets aggregated by target only typically yielded poorly predictive models. We discuss the possibility of "mix-and-matching" assay data across aggregating databases such as ChEMBL and Integrity and their current severe limitations for this purpose. One of them is the general lack of complete and semantic/computer-parsable descriptions of assay methodology carried by these databases that would allow one to determine mix-and-matchability of result sets at the assay level.



INTRODUCTION

In the past decade, a great number of publicly and commercially accessible databases have become available containing information regarding the chemical structure and biological activity of drug-like organic compounds.[1] These data have become an important source of training sets for various ligand-based drug design approaches. It has been stated that the quality of publicly available data, in general, requires significant improvement.[2] Sometimes, large variability in the measured activity values for the same compound is observed for different experiments run at different times, by different technicians, and/or by different laboratories.[1,3] Apart from overt differences in protocols, many factors affecting biological activity values are poorly understood and even more poorly quantified. Several methods have been suggested to reduce this inconsistency in publicly available bioactivity databases.[1,4,5] Typically, these approaches are based on selecting only compounds investigated by a single team of authors to reduce the impact of different experimental conditions on the assay result. While this approach can certainly help with filtering out noisy data and errors, it would be of much greater practical value if the databases themselves would carry sufficient information about the assay protocols and conditions under which the compounds were tested to fully assess the comparability of, if not mutually calibrate, the various result sets. Unfortunately, ontological data about the assays is not typically present in the publicly available databases such as BindingDB,[6] ChEMBL,[7] and PubChem.[8] According to Kalliokoski et al.,[1] "the assay descriptions available within ChEMBL are too terse to permit analyzing this any further."[1] The same authors conclude that it is not possible to systematically analyze the comparability of the activity data for the same assay, or various assay types under the same conditions, due to the scarcity of details about the experimental assay setup in both large public activity databases and the original publications. Notwithstanding that IC_{50} values measured under different assay conditions cannot in general be compared, Kalliokoski and co-workers found the data quality in ChEMBL to be good enough to build large-scale computational tools, where errors partially neutralized each other.[1]

Because the inconsistency of the data sets taken from these large-scale databases for a "mix-and-match" approach is so prevalent, one important question we are trying to answer is

Received: January 13, 2015
Published: June 5, 2015

DOI: 10.1021/acs.jcim.5b00019
J. Chem. Inf. Model. 2015, 55, 1388−1399

how one should use the data from publicly and commercially accessible databases to compile QSAR modeling sets that yield the most predictive models. To answer this issue, we propose several methods for the creation of modeling sets from such databases and investigate the accuracy of the QSAR models obtained using these sets. We used the program GUSAR for building the (Q)SAR models in this study. We have shown that the combination of radial basis function interpolation and self-consistent regression (RBF-SCR) recently implemented in GUSAR produces high-accuracy models.[9] First making sure that we thoroughly test the accuracy of the obtained QSAR models with leave-30%-out cross-validation (LMO), *y*-randomization (rand), and 5-fold cross validation, we then discuss the obtained results as to the compatibility, and the possibility of "mix-and-matching," of data from ChEMBL and Integrity.[10]

The overall goal of this study is thus 2-fold and to some extent complementary: First, to test and describe selection and preparation methods and strategies for assay sets, including those that combine results from several different original sources, with the aim to obtain high-quality predictive QSAR models, but to also delineate where all such attempts failed and no responsible filtering/curation approach delivered acceptable models for combined data sets, in order to provide at least some semiquantitative indication where the limits for mix-and-match lie in the data themselves. This latter question is a multifaceted and difficult one, and we do not claim to have provided a comprehensive answer in this paper by any means. More efforts in this direction are planned by this group and hopefully others.

We selected HIV-1 reverse transcriptase inhibitors for this study because this target provides a useful case study and is important in the ongoing anti-HIV drug development work.

METHODS

Preparation of Data Sets. First, we divided the whole data set coming from each database into two general subclasses—assaying done for (1) the wild type of the target and (2) the mutant form of the target. For each of these two general subclasses, we tested several different ways to compile data sets for creating QSAR models:

1. Selection of all compounds tested against a specific end point (IC_{50}).
2. Selection of the compounds tested using one method and material (biological assay).
3. Selection of the compounds derived from one specific scientific publication.

For every set obtained using selection strategies 1−3, we removed structural duplicates and used the median IC_{50} value for each group of duplicate structures. Log-transformed IC_{50} values were used for the creation of the QSAR models.

We considered a number of freely available and commercially accessible databases of biologically active compounds (including Thomson Reuters' Integrity,[10] ChEMBL,[7] PubChem,[8] DrugBank,[11] BindingDB,[6] HMDB,[12] and TTD[13]). Based on the number of unique compounds and the quantitative data on biological activity that are available in each of these databases,[14] we selected two of them for the extraction of data sets: Integrity, because it (1) offered maximal coverage of chemical space (including data from patents) and (2) is based on manual annotation of the data, presumably providing high quality and adequate representation of the data, and ChEMBL, because it is the largest publicly available database offering multiple annotations for chemical structures such as compound, data source, activity data, and data on mutations.

The structures of all HIV reverse transcriptase inhibitors available from ChEMBL and Integrity were collected, including compounds assayed against both wild type and mutants of RT. Integrity yielded a data set of 1327 records representing 564 unique compounds tested in approximately 1300 different biological assays using approximately 200 different biological materials, such as C8166 (human T-lymphoblastoid cells), mononuclear cells (blood), human (phytohemagglutinin-stimulated), and HEK293 (human embryonic kidney cells). The structures obtained from Integrity came from more than 50 scientific publications (published in more than 10 different journals), including scientific journals and patent data. From ChEMBL, we extracted 3787 compounds representing 2297 unique compounds, stemming from more than 50 publications (however, they all derived from only two different journal titles).

Data sources used in Integrity include *Bioorg. Med. Chem. Lett.* (48 different scientific publications), *J. Med. Chem.* (73 publications), *Antimicrob. Agents Chemother.* (31 publications), *Bioorg. Med. Chem.* (33 publications), *Antivir. Chem. Chemother.* (13 publications), *Eur. J. Med. Chem.* (11 publications), ACS meeting abstracts (11 publications), and patents (35 publications). Data sources used in ChEMBL include only two, *Bioorg. Med. Chem. Lett.* (51 publications) and *J. Med. Chem.* (104 publications). See also Table 1.

Table 1. Distribution of the Journal Titles Evaluated in ChEMBL and Integrity for Reverse Transcriptase Inhibitors

journal title	ChEMBL	Integrity
Bioorg. Med. Chem. Lett.	51	48
J. Med. Chem.	104	73
Antimicrob. Agents Chemother.	0	31
Bioorg. Med. Chem.	0	33
Antivir. Chem. Chemother.	0	13
Eur. J. Med. Chem.	0	11
ACS meeting abstracts	0	11
other abstracts	0	28
other papers	0	70
patents	0	35

Modeling Set Preparation from Integrity Data. To clean up data that can introduce noise into the data set, we developed a semiautomated workflow (Figure 1) of data mining based on a set of Python scripts.

First, using a script, we collected the complete list of the values for each field, which is essential for grouping compounds into more homogeneous data sets (the description of the Integrity field values is given in Table S1 of the Supporting Information, and a full list of the field values is given in List 1 of Supporting Information). Then, we determined the combinations of fields corresponding to the target on which the particular compound acts (in this case, HIV-1 reverse transcriptase).

We then identified the fields that contained the data on the biological activity. Such fields are "Pharmacological_activity," "Experimental_activity," "Target/condition/toxicity," and "Mechanism of action." For the extraction of HIV RT inhibitors, we selected records with the mechanism of action "Reverse Transcriptase Inhibitors." We also found that molecules for which "Reverse Transcriptase Inhibitors" was

DOI: 10.1021/acs.jcim.5b00019
J. Chem. Inf. Model. 2015, 55, 1388−1399

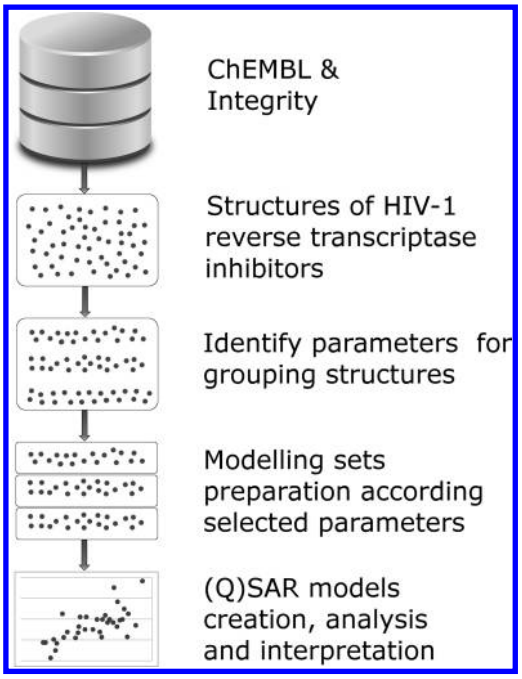


Figure 1. Preparation of data sets of RT inhibitors from ChEMBL and Integrity.

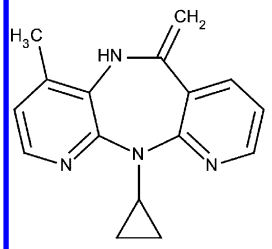
specified as the molecular mechanism of action are associated with two alternative values of the field "Target/condition/toxicity": (1) "Reverse transcriptase" and (2) "Infection, HIV." A more detailed analysis allowed us to categorize the assays into two broad classes: "directed" assays reflecting reverse transcriptase inhibition measured directly, which we term here "PCR-based" assays because they were all just different variants of PCR analysis, and "cell-based", in which cell lines were used for the indirect identification of reverse transcriptase inhibitors.

We analyzed several specific examples of the original publications that were used as the sources for the assay descriptions in Integrity. These assay descriptions are given in List 2 of the Supporting Information. All multicomponent compounds and salts were excluded, as is usual in QSAR modeling.[15]

The procedure to recognize structurally identical compounds within the data sets was done by pairwise comparison of the "uuuuu" NCI/CADD chemical identifier[16] calculated for each compound. Every data set associated with a particular assay that we subjected to this duplication analysis contained between 2 and 15 entries that were actually identical structures but with differing data for the biological activity (Figure 2). For each group of such identical compounds, we calculated the median IC_{50} value and used that value for the generation of the QSAR models. The NCI/CADD identifiers including the "uuuuu" version were obtained via the URL API we have implemented on our public CADD Group Chemoinformatics Tools and User Services web server (http://cactus.nci.nih.gov). The overlap analysis and calculation of the median activity value was done by Python scripts operating on the SD files for which the "uuuuu" identifiers had been calculated.

We calculated the overlap between the modeling sets corresponding to each one combination of the method and material. The number of matching compounds is given in the Table S2 of the Supporting Information.

Three types of modeling sets were created. The first method ("Selection of all compounds tested against a specific end-



Poly(rA)-oligo(dT)	Poly(rC)-oligo(dG)
6.74	7.42
6.72	7.42
6.55	7.08
6.40	7.06
6.40	6.82
6.22	6.70
5.77	6.70
5.56	6.70
5.65	6.70
5.54	6.70
5.39	6.52
5.36	5.74
5.19	5.56
5.16	5.35
4.98	
4.98	
4.91	
Min: 4.91	5.35
Max: 6.74	7.42
Mean: 5.73	6.61
Median: 5.55	6.73

Figure 2. Chemical structure of Nevirapine with a series of $-\log_{10}(IC_{50})$ values from different primary sources obtained for two PCR-based methods: "poly(rA)-oligo (dT) as template primer" and "poly(rC)-oligo (dG) as template primer."

point") yielded one set for wild type RT, which included a total of 564 compounds, and one set for RT mutants resistant to non-nucleoside resistant reverse transcriptase inhibitors (NNRTIs), yielding a total of 78 compounds. The second compilation method ("Selection of the compounds tested using one method and material (biological assay)") produced sets for specific assays (i.e., combinations of methods and materials constituting essentially the same assay). See Table 2 for detailed classification of these sets. Only those sets that contained 20 or more structures were admitted as suitable for our QSAR modeling.

The minimum, maximum, mean, and median values of each column of the $-\log(IC_{50})$ values are given in the bottom rows. These data are from Integrity. Note the approximately 2 orders of magnitude spread in IC_{50} data for either PCR-based method.

None of the data sets from Integrity associated with the third compilation method ("Selection of the compounds derived from one specific scientific publication") possessed at least 10 molecules, which we had set as the minimum size for a training set. Therefore, it was not possible to compile modeling sets from Integrity associated with specific scientific publications for our model building.

All these compilation methods together allowed us to compile a total of 18 modeling sets from Integrity.

DOI: 10.1021/acs.jcim.5b00019
J. Chem. Inf. Model. 2015, 55, 1388−1399

Table 2. Parameters of QSAR Models Obtained for Integrity Data Sets Compiled from (a) Compounds Tested against One End-Point (First Compilation Method) and (b) Compounds Tested in One Specific Biological Assay (Second Compilation Method)

(a)							
	N	RMSE	R^2	Q^2	R_{LMO}^2	R_k^2	R_{rand}^2
overall modeling set/HIV-1	564	1.24	0.54	0.24	0.28	N/D	0.22
Overall modeling set/HIV-1, NNRTIs resistant	78	0.99	0.70	0.65	0.39	0.10	0.01
(b)							
PCR-based methods							
method/material of testing[a]	N	RMSE	R^2	Q^2	R_{LMO}^2	R_k^2	R_{rand}^2
combinations of methods including unclassified compounds/HIV-1	319	0.98	0.64	0.54	0.35	0.29	0.12
unclassified compounds	115	0.97	0.64	0.53	0.34	N/D	0.22
combinations of methods excluding unclassified compounds/HIV-1	**215**	**0.87**	**0.99**	**0.67**	**0.60**	**0.50**	**0.29**
radioactivity assay/HIV-1	48	0.74	0.96	0.49	0.71	0.60	0.29
poly(rA)-oligo(dT) as template primer/HIV-1	**93**	**0.82**	**0.99**	**0.74**	**0.60**	**0.58**	**0.20**
poly(rA)-oligo(dT) as template primer/HIV-1, NNRTIs resistant	**32**	**0.31**	**0.97**	**0.92**	**0.90**	**0.62**	**−0.14**
poly(rC)-oligo(dG) as template primer/HIV-1	64	0.93	0.99	0.52	0.55	0.46	−0.22
poly(rC)-oligo(dG) as template primer/HIV-1, NNRTIs resistant	22	0.24	0.92	0.85	−0.01	0.59	0.73
cell culture-based methods							
method/material of testing	N	RMSE	R^2	Q^2	R_{LMO}^2	R_k^2	R_{rand}^2
combinations of methods based on cell cultures	309	1.17	0.69	0.50	0.43	0.38	−0.08
combinations of methods based on cell cultures (excluding unclassified compounds)	219	0.67	0.66	0.52	0.52	0.41	−0.12
antigen assay/C8166 (human T-lymphoblastoid cells)	22	0.92	0.90	0.83	0.89	0.66	0.19
antigen assay/mononuclear cells (blood), human (phytohemagglutinin-stimulated)	25	0.95	0.97	0.93	0.91	0.66	0.10
luciferin-luciferase assay/HEK293 (human embryonic kidney cells)	28	0.30	0.90	0.85	0.77	0.64	−0.05
antigen assay/mononuclear cells (blood) human	**52**	**0.91**	**0.85**	**0.76**	**0.75**	**0.64**	**0.04**
viral replication assay/mononuclear cells (blood) human	**82**	**0.92**	**0.76**	**0.63**	**0.71**	**0.61**	**0.01**
cytopathicity assay/MT2 human T-lymphoblastoid cells	34	0.99	0.99	0.43	0.43	0.29	−0.6
cytopathicity assay/MT4 human T-lymphoblastoid cells	270	0.91	0.86	0.61	0.56	0.49	−0.2
syncytium formation assay/MT4 human T-lymphoblastoid cells	29	0.63	0.99	0.55	0.64	0.60	−0.58
dye assay/MT4 human T-lymphoblastoid cells	115	1.17	0.69	0.50	0.41	0.33	−0.41

[a]Material and method of testing as per the classification used in Integrity. N: number of compounds in the data set. R^2: determination coefficient between measured and predicted values. Q^2: determination coefficient between measured and predicted values in leave-one-out cross-validation (LOO CV). RMSE: root mean-square error between measured and predicted values in LOO CV. R_{LMO}^2: determination coefficient between measured and predicted values obtained in leave-many-out cross-validation. R_{rand}^2: determination coefficient between measured and predicted values after randomization of y (IC_{50}) values. R_k^2: determination coefficient between measured and predicted values obtained in k-fold cross validation. Rows with bold font: the models with the highest performance.

Modeling Set Preparation from ChEMBL Data. The set of compounds and data obtained from ChEMBL was prefiltered using the following criteria directly obtained, or derived, from the ChEMBL data: (1) confidence score (quantitative characteristic reflecting the quality of the data in ChEMBL) is larger than 8; (2) expert-based curation; (3) data source (PubMed_ID) is indicated; (4) IC_{50} is a parameter of activity measurement; (5) IC_{50} is precisely defined (there is, e.g., no ">" or "<" sign before the IC_{50}); (6) the structure is not a multicomponent compound or salt, which we excluded.

We applied three compilation methods to the ChEMBL data to construct the modeling sets. ChEMBL provides labels indicating the activity against several mutant forms of RT; however, these data are not available in a form allowing its completely automated filtering. We manually deleted compounds tested against mutant forms of HIV from the general data set collected from ChEMBL. Thus, the overall set of the HIV-1 RT inhibitors extracted from ChEMBL included the compounds tested against the wild type of RT only.

We compiled one overall modeling set that included all extracted compounds from ChEMBL according to the first compilation method.

To apply the second compilation method, we divided the overall modeling set into subsets by ChEMBL assay identifier (CAID). We obtained about 40 modeling sets based on this compilation method. We should emphasize that, typically, the description of the assay type is found as part, or in the vicinity, of the original phrase or section in the abstract or body of the paper that provides the assay description. This corresponds to the earlier observations by Kaliokoski et al.[1] about the limitations of ChEMBL data. In this case, although we can divide the general data set into subsets by unique CAID, it is impossible to interpret the description of each assay to derive any assumptions about assay details.

According to the third compilation method, the compounds belonging to a single scientific publication were assigned to the same modeling set, which in total yielded 15 modeling sets coming from ChEMBL.

(Q)SAR Modeling Algorithm. To create (Q)SAR models, we used the program GUSAR.[17,18] A combination of three types of descriptors is used in GUSAR: (1) QNA (Quantitative Neighborhoods of Atoms) descriptors, (2) descriptors that are both topological (length and volume) and physicochemical parameters of the whole molecule, and (3) descriptors based on the prediction of the biological activity spectra by the program

DOI: 10.1021/acs.jcim.5b00019
J. Chem. Inf. Model. 2015, 55, 1388−1399

PASS.[19,20] The robustness of the GUSAR algorithm vis-à-vis the utilization of data sets that are nonhomogeneous (in terms of chemical similarity) has been shown previously.[21] Applicability domain calculation is integrated into GUSAR and is output together with the results of the prediction. The current version of GUSAR is based on a new machine learning approach that combines self-consistent regression with the radial basis function interpolation method (RBF-SCR).[17,18] Having shown previously that the RBF-SCR method provides higher accuracy of prediction than other modern QSAR approaches,[18] we decided to use GUSAR for the QSAR modeling of the data sets obtained from Integrity and ChEMBL.

Validation. To estimate the models' predictivity, we used the leave many out cross-validation procedure, 5-fold cross validation, as well as *y*-randomization (shuffling of the dependent values), which are considered the most powerful validation procedures.[22]

To achieve the leave-30%-out cross-validation (L30%CV), each data set was randomly divided 20 times into fractions of 70% and 30% for training and test sets, respectively. This procedure is similar to n-fold external validation, however with one difference: During n-fold validation, each compound may be used as a test compound only one time, whereas with the multiple splitting validation procedure, each compound may be used as a test compound several times, depending on where the random splitting put it. The average R^2 value (determination coefficient) obtained from this procedure was used as one predictivity measure of the obtained model.

During *y*-randomization, GUSAR calculates a determination coefficient R_{rand}^2, too, according to

$$R_{rand}^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

where y_i is the experimental value, $\hat{y}_i$ is the predicted value, and $\bar{y}_i$ is the average of the training set values.

The sum of squares of the residuals can be much higher than the total sum of squares if the prediction results are really poor. In this case, the R^2 value would be negative, which means that obtained predictions do not make any sense and should be discarded.

The sum of squares of the residuals can be much higher than the total sum of squares if the prediction results are really poor. In this case, the R^2 value would be negative, which means that obtained predictions do not make any sense and should be discarded.

High values of R^2 in L30%CV as well as low absolute values of R^2 after *y*-randomization are indications of high model quality.

We also performed 5-fold cross validation for each modeling set. To calculate the 5-fold cross validation R^2 values, we randomly divided each data set into five subsets and subsequently used a single subset for testing the model and the remaining four sets for training. This procedure was performed five times, and the R^2 values obtained from each instance were averaged.

The reason why we did not use any external test set was the dependence of the experimental IC_{50} on the parameters of a specific assay; in this case, we would *expect* low performance of the models due to the high level of discrepancy we already observed between the experimental IC_{50} values from one assay to another. In other words, we tried to use the databases of bioactive compounds along the "best practice" principle of QSAR modeling; i.e. that the most reliable models can be expected for data obtained using the same experimental approach. As a consequence, we estimated the performance of each QSAR model only intrinsically for its own data set but not across training data sets obtained with different materials and methods.

RESULTS

Results Based on Modeling Sets from Integrity. *Wild-type RT.* The results presented in Table 2 show that the quality of QSAR models was very low when they were created using the set obtained for the first compilation method (N = 564; RMSE = 1.24; R^2 = 0.54; Q^2 = 0.24; R_{LMO}^2 = 0.28; R_{rand}^2 = 0.22).

We tried to improve the model performance for this data set by dividing it into two subsets according to the basic type of assay, cell-based vs PCR-based assays. We then built the QSAR models for these two subsets separately. However, the performance of the models did not increase significantly after this 2-fold split (subset for PCR-based assays: N = 319; RMSE = 0.97; R^2 = 0.64; Q^2 = 0.54; R_{LMO}^2 =0.35; R_k^2 = 0.29; R_{rand}^2 = 0.12; subset for cell-based assays N = 309; RMSE = 1.17; R^2 = 0.69; Q^2 = 0.50; R_{LMO}^2 = 0.43; R_k^2 = 0.38; R_{rand}^2 = −0.08).

Next, we tried to improve the model performance by excluding compounds for which the method and/or material of testing were unknown. For the PCR-based assays, a total of 115 compounds had no information about the method and/or material. The model rebuilt using only the remaining 319 compounds showed a significant increase in model accuracy (N = 319; RMSE = 0.97; R^2 = 0.64; Q^2 = 0.54; R_{LMO}^2 = 0.35; R_k^2 = 0.29; R_{rand}^2 = 0.12).

Similar results were obtained for the modeling set collected for the cell-based methods; after the exclusion of 67 unclassified compounds, the characteristics of the model improved to N = 215; RMSE = 0.87; R^2 = 0.99; Q^2 = 0.67; R_{LMO}^2 = 0.60; R_k^2 = 0.50; and R_{rand}^2 = 0.29.

It thus appears that selecting compounds for which the material and method of testing are known leads to QSAR models that have a higher accuracy in comparison to using all the compounds retrieved from Integrity without any filtering, at least for this target. The results obtained for modeling sets compiled using the second method are given in Table 2. For all modeling sets associated with PCR-based assays, low values of R^2 after *y*-randomization were observed. In general, low values of *y*-randomization R^2 combined with high values of R^2 were obtained for almost all the modeling sets put together for cell-based assays (except sets associated with antigen assays tested using C8166 (human T-lymphoblastoid) cells and human (phytohemagglutinin-stimulated) mononuclear blood cells). We therefore can conclude that the Integrity wild-type RT data sets stemming from both the PCR-based assays and most of the cell-based assays allowed us to build several models that had good performance; the models obtained for (1) data associated with "Poly(rA)-oligo(dT) as template primer/HIV-1" (PCR-based model): N = 93, RMSE = 0.82, R^2 = 0.99, Q^2 = 0.74, R_{LMO}^2 = 0.60, R_k^2 = 0.58, R_{rand}^2 = 0.20; (2) antigen assay data (cell-based method, human blood mononuclear cells): N = 52, RMSE = 0.91, R^2 = 0.85, Q^2 = 0.76, R_{LMO}^2 = 0.75, R_k^2 = 0.64, R_{rand}^2 = −0.04); and (3) viral replication assay data (human mononuclear blood cells): N = 82, RMSE = 0.92, R^2 = 0.76, Q^2 = 0.63, R_{LMO}^2 = 0.71, R_k^2 = 0.61, R_{rand}^2 = 0.01. The

DOI: 10.1021/acs.jcim.5b00019
J. Chem. Inf. Model. 2015, 55, 1388−1399

Table 3. Quality of the Models Built for ChEMBL Data Sets Containing (a) Compounds Tested against One End-Point (First Compilation Method), (b) Compounds Tested in One Specific Biological Assay (Second Compilation Method), (c) Compounds Derived from Individual Scientific Publications (Third Compilation Method)[a]

(a)

	N	RMSE	R^2	Q^2	R_{LMO}^2	R_k^2	R_{rand}^2
overall modeling set of RT inhibitors	1847	1.23	0.75	0.56	0.54	0.44	−0.20

(b)

wild-type RT

assay ID[b]	description of the assay[c]	N	RMSE	R^2	Q^2	R_{LMO}^2	R_k^2	R_{rand}^2
800743	inhibitory activity against HIV-1 reverse transcriptase (HIV-RT)	73	0.77	0.92	0.44	0.40	0.32	−0.40
800589	inhibitory concentration against HIV-1 wild type reverse transcriptase (RT)	65	0.48	0.98	0.21	0.20	0.16	−0.62
797532	in vitro HIV-1 RT inhibitory activity using rC.dG as template-primer	67	0.63	0.99	0.48	0.24	0.16	0.25
802411	inhibitory activity against HIV-1 wild-type reverse transcriptase.	50	0.74	0.99	0.01	0.21	0.12	0.62
802759	in vitro inhibitory activity against human immunodeficiency virus type 1 (HIV-1) reverse transcriptase (RT)	59	0.59	0.92	0.40	0.40	0.40	−0.75
800744	tested for inhibition against HIV-1 RT (used (poly)rC-(oligo)dG as the template)	54	0.78	0.98	0.50	0.47	0.40	−0.44
800582	inhibitory activity against HIV-1 reverse transcriptase	55	0.72	0.99	0.50	0.59	0.44	−0.39
802412	inhibitory activity against HIV-1 wild-type reverse transcriptase	37	0.60	0.99	0.01	0.42	0.39	−0.78
797524	**inhibition of HIV-1 reverse transcriptase using rCdG as template and dGTP as substrate**	**54**	**0.53**	**0.96**	**0.66**	**0.59**	**0.56**	**−0.34**
798389	**concentration required to inhibit the HIV-1 reverse transcriptase activity by 50%**	**49**	**0.53**	**0.96**	**0.62**	**0.64**	**0.60**	**−0.44**
801332	in vitro inhibitory activity against HIV-1 reverse transcriptase	42	0.83	0.96	0.16	0.32	0.12	−0.86
802761	in vitro inhibitory concentration against HIV-1 reverse transcriptase using rC-dG template primer	39	0.61	0.63	0.53	0.51	0.46	0.10

mutant forms of RT

assay ID[b]	description of the assay[c]	N	RMSE	R^2	Q^2	R_{LMO}^2	R_k^2	R_{rand}^2
802599	inhibitory activity against HIV-1 Y181C reverse transcriptase (RT)	49	0.40	0.95	0.56	0.34	0.22	−0.30
800600	tested for the inhibition of mutant type (Cys181) HIV-1 (IIIB) RT	40	1.29	0.38	0.02	0.38	0.08	−0.74
800601	tested for the inhibition of mutant type (Ile100) HIV-1 (IIIB) RT	29	1.06	0.98	0.03	0.42	0.12	−0.85

(c)

Pubmed ID	N	RMSE	R^2	Q^2	R_{LMO}^2	R_k^2	R_{rand}^2
8523406	**63**	**0.52**	**0.94**	**0.56**	**0.56**	**0.48**	**−0.09**
7683054	**56**	**0.73**	**0.99**	**0.61**	**0.64**	**0.58**	**−0.24**
8523406	**63**	**0.49**	**0.74**	**0.64**	**0.64**	**0.62**	**0.05**
9685236	38	0.30	0.73	0.58	0.55	0.52	0.05
9685235	53	0.41	0.57	0.38	0.37	0.38	0.08
11708913	39	0.78	0.64	0.49	0.56	0.54	−0.01
11384233	70	0.70	0.97	0.59	0.60	0.60	−0.45
15634005	54	0.49	0.99	0.41	0.45	0.40	−0.67
14643337	54	2.16	0.62	0.44	0.35	0.41	0.02
16220981	28	0.56	0.74	0.58	0.66	0.70	0.13
1279173	70	0.54	0.98	0.67	0.51	0.48	−0.20
1712395	64	1.13	0.99	0.05	0.21	0.12	−0.44
8809165	33	0.41	0.69	0.69	0.42	0.29	−0.23
1375293	42	0.99	0.98	0.41	0.40	0.24	−0.21
7490732	61	0.57	0.97	0.10	0.14	0.09	−0.27

[a]Rows with bold font: the models with the highest performance. [b]ChEMBL assay identifier (CAID) is presented. [c]Extracted verbatim from ChEMBL.

cells of Table 2 for these top-performing models (in terms of performance) have been underlayed with bold font.

Summarizing the results of this part of the QSAR modeling, we can posit that the compilation of modeling sets according to their assay data (i.e., associated with just one material and method of testing) leads to higher consistency in the sets and thus to QSAR models with higher performance.

NNRTI-Resistant Forms of RT. Compounds tested against NNRTI-resistant forms of RT could be selected only for the PCR-based assays. The characteristics of the model associated with the first compilation method were N = 78; RMSE = 0.99; R^2 = 0.70; Q^2 = 0.65; R_{LMO}^2 = 0.39; R_k^2 = 0.10; R_{rand}^2 = 0.01. Similar to the wild type RT results, the second compilation method yielded high-accuracy models for NNRTI-resistant forms of RT. The best model obtained for the second compilation method had N = 32; RMSE = 0.31; R^2 = 0.97; Q^2 = 0.92; R_{LMO}^2 = 0.90; R_k^2 = 0.62; R_{rand}^2 = −0.14 (for assay method "poly(rA)-oligo(dT) as a template primer").

Typically, the number of compounds in the overlap set of any two assays from Integrity is 2−3 (for details, see Table S2 of the Supporting Information), with eight or more compounds found only for five pairs of assays. We therefore calculated the correlation coefficient only for these five overlap sets. Of course, it is difficult to estimate general inconsistency between

DOI: 10.1021/acs.jcim.5b00019
J. Chem. Inf. Model. 2015, 55, 1388−1399

assays from Integrity based on just these five overlap sets of (the total number of possible RT assays combinations in Integrity exceeds 60). What one can do of course as a surrogate of sorts for the above data consistency analysis, with all the necessary caveats of course, is to calculate the pairwise correlation coefficients for the *predicted* activity values generated by the models based on each of the training sets in the pair. Such pairwise correlation coefficients are given in the Supporting Information.

Results Based on Modeling Sets from ChEMBL. The quality parameters of the QSAR models for the data sets extracted from ChEMBL are given in Table 3, arranged according to the compilation method. For the second compilation method, we considered only the top 15 models built for the data sets containing the largest number of molecules.

The descriptions of assays extracted from ChEMBL "as it is" are given in Table 3 together with the results of (Q)SAR modeling. The descriptions of assays are not well-structured as was already found by others.[1] Based on these descriptions, it is impossible (1) to deduce what were the reasons to group the compounds into one set labeled by unique CAID or (2) to even venture any guess about the relationship between the (Q)SAR models' performances and the assays details. In many cases, the descriptions are essentially just synonyms of each other (for example CAID 800589, "Inhibitory concentration against HIV-1 wild type Reverse transcriptase (RT)" and CAID 802411, "Inhibitory activity against HIV-1 wild-type reverse transcriptase"). Therefore, it is practically impossible to further use the models, built using the second compilation method applied for ChEMBL. The two best models (built using the data sets for CAIDs 797524 and 798389) had higher performance compared to those obtained for the overall data set (the result of the first compilation method) by both R_{LMO}^2 and R^2 for 5-fold validation (CAID 797524: N = 54; RMSE = 0.53; R^2 = 0.96; Q^2 = 0.66; R_{LMO}^2 = 0.59; R_k^2 = 0.56; R_{rand}^2 = −0.34; CAID 798389: N = 49; RMSE = 0.53; R^2 = 0.96; Q^2 = 0.62; R_{LMO}^2 = 0.64; R_k^2 = 0.60; R_{rand}^2 = −0.44). However, in general, the performance of the models built using the data sets of the second compilation method is not higher compared to the models created for the overall data set (created by the first compilation method). This finding was different from that obtained for the modeling sets from Integrity. We propose the main reason for this observation is the limitations of the annotation procedure used to classify the assays in ChEMBL. Summarizing our observations, we can conclude that attempts to divide an overall data set (for RT) into subsets by grouping CAIDs by assay descriptions provided by ChEMBL are futile. We can also conclude that model performance is strongly dependent on the quality of the description of materials and methods (assays) used in databases of biologically active compounds.

The third compilation method led to results that are more consistent and yielded higher quality QSAR models. For example, the best model, which was obtained with the data from Hoffman et al. (PubMed_ID 7683054),[23] had the following characteristics: N = 56; RMSE = 0.63; R^2 = 0.99; Q^2 = 0.61; R_{LMO}^2 = 0.64; R_k^2 = 0.58; R_{rand}^2 = −0.24. On the other hand, several QSAR models showed very low performance.

For example, the model associated with PubMed_ID 1712395 (Hargrave et al.[24]) had very poor accuracy: N = 64; RMSE = 1.13; R^2 = 0.99; Q^2 = 0.05; R_{LMO}^2 = 0.21; R_k^2 = 0.12; R_{rand}^2 = −0.44. We suggest that the very low quality observed for some models might be a result of inconsistencies in the methods used in those studies. For example, in the above-mentioned paper of Hargrave et al.,[24] there are data for 125 structures tested in the RT assay, but only 64 of them have a precise value of IC_{50} in the range from 35 nM to 15.3 μM. The IC_{50} values of the other 61 compounds are characterized by means of approximate values, such as "> 1 μM" or "≫ 1 μM." This observation gave us reason to suspect that the assay used in this study had low robustness, which might be one reason for the extremely low accuracy and predictivity of the model. Because we cannot verify any of these suggestions (either low assay robustness or low model predictivity due to the peculiarities of the training set or of the statistical approach used in GUSAR), we can at least propose several possible reasons for the low predictivity of the model, which can be a consequence of the inconsistent (in other words, low robustness) initial data values, which we can only use for the QSAR model creation.

We would like to point out that having only one IC_{50} value does not per se guarantee high robustness of the assay if measurements were not repeated multiple times.[25] At the same time, we would like to note that although the general trend in QSAR model preparation is the creation of models using data values obtained in one laboratory using one biological material, for several activities, such as HIV reverse transcriptase (RT) inhibition, it is practically impossible to create a representative set of the compounds with multiple activity data measurements to the moment of our study, which makes it impossible to estimate the robustness of a particular assay for HIV RT activity measurement to date.

Another possible reason leading to inconsistency of the experimental results, which may then result in low accuracy of the models, is the sometimes difficult interpretation of the signal for low activity levels, i.e. IC_{50} values in the higher micromolar ranges.

Other possible reasons for the low quality of QSAR models are errors in the annotation (Kaliokoski et al.[1]) and utilization of several drug-resistant mutant forms together with the wild strain of HIV. Examples of studies that support this assumption are papers by Proudfoot et al. (PubMed_ID 7490732)[26] and Romero et al. (PubMed_ID 8809165),[27] where several drug resistant mutant strains were used together. The QSAR models built using the data sets from these sources have poor accuracy and predictivity. We should emphasize that performing automated filtering of such data from ChEMBL is problematic because this database does not provide assay classification (in the way that Integrity does) and classification of mutant vs wild type of the target in a machine-readable form.

Mix-and-Match Compatibility of ChEMBL and Integrity Data Sets. We investigated whether it is possible to mix the RT inhibition data from ChEMBL with those from Integrity for the purpose of model building. We analyze to what extent we can match up the data from these databases because if there are no matching compounds between data sets, then an important basis upon which to do true mix-and-matching is missing.

Along the lines of our compilation methods 1−3 presented above for the creation of more-consistent data sets, we applied the following corresponding analyses A−C to determine the overlap between the data sets from ChEMBL and Integrity.

A. Overlap between the modeling sets obtained with compilation method 1 (all compounds tested against a specific

DOI: 10.1021/acs.jcim.5b00019
J. Chem. Inf. Model. 2015, 55, 1388−1399

end-point): The overlap between the two overall data sets extracted from Integrity and ChEMBL, respectively (Integrity, 564 structures; ChEMBL, 1847 structures) comprised 87 structures. The IC_{50} values extracted from Integrity and ChEMBL were log-transformed. To avoid the "tyranny of averages," we removed from this overlap set those structures that had duplicates in the initial data set with IC_{50} values differing by more than 0.5 on a logarithmic scale, which left us with 57 structures. The Pearson correlation coefficient r for the two sets of IC_{50} values of these 57 compounds was −0.03. We can thus conclude that there is no concordance between the IC_{50} values for the compounds in the overlap set extracted from the two databases.

B. Overlap between the modeling sets obtained with compilation method 2 (compounds tested using one method and material (biological assay)): As already mentioned, data about assay materials and methods is available from Integrity in a well-structured form, allowing one to extract in automated manner information about the combinations of the materials and methods. While data sets extracted from ChEMBL can be divided by ChEMBL AID(s), the assay descriptions in ChEMBL are not suitable for automated parsing out of the essential assay details s that would allow matching the data between ChEMBL and Integrity. Also, we cannot combine in an automated manner data from ChEMBL and Integrity in the absence of any common terminology that would allow one to determine conditions of the experimental testing that are common for two assays. Therefore, it was not possible to determine the overlap between the modeling sets obtained with this second compilation method.

C. Overlap between data sets obtained with compilation method 3 (compounds derived from specific scientific publications): Integrity provides the data source details including the name of the scientific publication, which allows automated parsing of the data and grouping it according to its data source. ChEMBL provides information about the scientific publication including the PubMed identifier in a form that allows automated parsing.

We estimated the number of chemical structures extracted from one scientific publication in ChEMBL and Integrity. Table 4 contains the number of unique structures extracted from one scientific publication and found both in ChEMBL and Integrity. The number of unique structures deriving from one paper is about 2−3 in Integrity vs 10 and more structures in ChEMBL. In this case, it is completely impossible to build the interdatabase overlapping set of compounds for the creation of (Q)SAR models and even for the calculation of the concordance between the two data sets created according to the third compilation method.

In Table 5, we list all records associated with each specific publication found in Integrity only but not in ChEMBL. The numbers of unique compounds are given in parentheses. We used PubMed nomenclature to identify the individual publications in Tables 4 and 5.

Table 4. Number of Structures Collected from Individual Data Sources for ChEMBL and Integrity, Respectively (Publications identified by PubMed ID)

	number of unique structures	
PubMed_ID	ChEMBL	Integrity
1712395	119	1
7490733	37	3
11708913	38	3
11384233	70	3
16884295	23	3
16913724	14	5
16220981	28	2
8523406	63	4

Table 5. Number of Structures Collected from Several Individual Data Sources for Integrity with the Largest Number of Compounds Tested (Publications Identified by PubMed ID (See Text))

PubMed_ID	number of structures[a]
17910429	29(5)
7541618	24(9)
10548537	12(1)
7537029	10(1)
7475321	18(6)
8799326	12(3)
1384425	16(2)
8930167	15(4)
7694070	14(5)
15974590	12(4)
15686922	5(2)
11714616	8(4)
15189038	8(4)
10386942	11(11)

[a]Numbers in parentheses indicate unique compounds.

The data sets associated with individual scientific publications (which can be considered as data coming from one laboratory) are very small in size for Integrity (Tables 4, 5). We hypothesize that the reasons for this are (1) selection by a database annotator of only a few compounds that are typical representatives of chemical structures, (2) focus on the collection of many compounds tested in the same or similar assays/conditions, and (3) selection of only a few compounds with a high therapeutic index from one scientific publication.

The small sizes of the Integrity sets coming from single publications (Table 5) limits the number of data sets that could be collected from Integrity for QSAR modeling according to the third compilation method (modeling sets using structures coming from only one scientific publication). This may, in fact, be another implicit factor leading to the comparatively low quality of the models from the Integrity data sets in general, because the number of compounds tested by one scientific team (and reported in one publication) is restricted typically to 2−10 compounds, which is not enough for the creation of (Q)SAR models with high performance. Although Integrity contains the classification of the materials and methods of the experimental testing, obviously, this classification does not include specific details of the assay that may be associated with the peculiarities of experimental testing in a certain laboratory, which, in our opinion, makes it difficult to create the (Q)SAR models with the highest performance.

DISCUSSION

Biological Interpretation of Modeling Results. Taking into account previous argumentation about limitation of the assay classification of ChEMBL, we consider below the interpretation of the modeling results corresponding to the data sets from Integrity only.

DOI: 10.1021/acs.jcim.5b00019
J. Chem. Inf. Model. 2015, 55, 1388−1399

The best QSAR models were obtained using modeling sets based on assay data put together with the second compilation method. However, for some of them, we found significant differences in the accuracy of the predictions, even if the models were based on supposedly similar assays. To elucidate this issue, we analyzed the differences between two very similar PCR-based methods, "poly(rA)-oligo(dT) as a template primer" and "poly(rC)-oligo(dG) as template primer," from a biological point of view. It is known from the literature that "the enzymes that catalyze the polymerization or terminal addition of [H′]TMP with a poly(rA)-oligo (dT) template primer are not exclusively viral, and DNA synthesis is more specific to viral enzyme."[28] The authors also suggest that the activity of RNA-directed DNA polymerase (an indirect label of reverse transcriptase activity) is greater with poly(rA)-oligo (dT) as a template primer than with poly(rC)-oligo(dG) as a template primer. Moreover, there are also data to the effect that poly(rA)-oligo(dT) does not in itself indicate the activity of reverse transcriptase and that only one of the most frequently used primer-template complexes, namely, poly(rC)-oligo(dG), is utilized to any significant degree by reverse transcriptase exclusively. However, the sensitivity and specificity of this template complex are limited due to several issues, (1) the strong dependence on the supplier of the primers and that (2) the activity of reverse transcriptase measured based on poly(rC)-oligo(dG) as a template primer is much more variable than for the other primer templates.[28] Taking into account the strong dependence on the supplier of the primer, we can a priori predict a particularly low quality of the models based on data sets associated with poly(rC)-oligo(dG). These points allow us to propose an explanation of the results of our *t* test, according to which the IC_{50} values predicted using the poly(rC)-oligo(dG) template primer data set are significantly different from those associated with any other method of testing. The higher accuracy of predictions calculated based on HIV-1 resistant strains as the material (Table 1; and RNA-based DNA polymerase activity assay with poly(rA)-oligo(dT) or poly(rC)-oligo(dG) as the template primer as the method) can be explained by the smaller number of observations, which leads to a more consistent data set in comparison with the numerous tests done for the wild type of HIV-1. One of the main conclusions we can draw is that the data sets created for compounds tested by the "poly(rC)-oligo(dG) as a template primer" assay confirms the ambiguity of the activity observed using this type of primer.

Critical Assessment of Databases' Usefulness for QSAR Modeling. We have presented three methods to automatically arrange data derived from (large) databases of biologically active compounds for QSAR modeling. We showed that the division of the whole data set into subsets by material and method of testing or by specific publication led to an increase in model performance. However, our analysis of the results of QSAR modeling also identified several issues that may restrict the use of such databases for QSAR modeling.

We suppose that if detailed information about the correlation of the activity measurements between each pair of assays is absent, then the data set associated with one biological assay should not be used as the test set for a QSAR model built using a data set associated with another biological assay.

It has been shown that the assays used in high-throughput screening may be different in terms of their reproducibility/robustness.[25,29] We believe that nonrobustness (poor reproducibility) of assays themselves can be another factor leading to the low predictivity of the QSAR model.

Unfortunately, it is at this time impossible to quantify this type of inconsistency using statistical approaches because there are currently not enough data in the databases of biologically active compounds for this type of analysis, i.e., at least 10 compounds in each type of assay for which activity measurements were repeated multiple times.

Findings and Recommendations for Integration of Data From Different Data Sources. Our experiences have led us to a number of general guidelines we have found useful to adhere to, and which may be of more general applicability, for mix-and-match integration of assay data from different databases. Obviously, compound sets should have been measured against the same target and have the same type of end point data (IC_{50}, K_i, etc.). We then recommend to compare the structures for searching for identical molecules in two sets or more by modern tautomer-invariant identifiers such as InChI or uuuuu. If there are at least seven common compounds, the correlation coefficient between the assay result sets can be calculated. If this coefficient exceeds 0.6, then this pair of assays can be used for mix-and-match (Q)SAR model building. Subdividing the initial whole data set for one target into subsets by individual assay types (materials and method) is generally necessary to allow for fruitful (Q)SAR modeling.

While subdividing the initial set into subsets by assay type, or even by individual scientific publication, in general yielded training sets amenable for mix-and-match QSAR model building within one database, i.e. ChEMBL or Integrity, this approach did not work *across* databases. The main problem is that, at this time, the assay descriptions are too different between ChEMBL and Integrity, in terms of both the amount of detail given (assay methods, materials, and conditions) and the classifications used. This lack of unified and standardized assay descriptions across databases is one of the major obstacles for mix-and-matching public assay data.

Combining data from different large databases—or, for that matter, from any separate data sources that are not under stringent, e.g., corporate, control—remains an open problem. If one wants to combine assay result sets for the same protein target for two or more compound sets with some structural overlap and significant result discrepancy in the overlap set, one has to make the decision whether this discrepancy is (a) trivial (i.e., can be dealt with by, e.g., averaging), (b) significant but fixable (e.g., by mutual calibration), (c) serious but solvable by omission (i.e., it is justified to treat them as outliers to be removed from the training set and henceforth forgotten), or (d) fatal (i.e., one should take it as a warning flag not to build a model at all). We would argue that the state of the field is such that this situation has never been even (algorithmically or otherwise) quantified and that the choice among possibilities (a) through d is made based on habits, vague estimates, desires, and other rather nonscientific criteria. What one would like to have is a way of quantifying this issue algorithmically from the input data to be able to make an objective decision whether to mix-and-match or not, or, even better, a way of prospectively determining the possible mix-and-matchability of, e.g., a future assay with existing assay results in large databases, based on detailed assay methodology descriptions.

We have shown on examples of the data sets division into the subsets according to the materials and methods (assays) for Integrity and ChEMBL. Based on our results, we can conclude that the (Q)SAR modelers are *absolutely* dependent on the

DOI: 10.1021/acs.jcim.5b00019
J. Chem. Inf. Model. 2015, 55, 1388−1399

quality of the annotation procedure when using the large-scale databases of bioactive compounds. The arguably most important issue here is the absence of a unified terminology for the conditions of assaying small molecules against biological targets (be they enzyme/PCR-based or cell-based). When we do not have detailed information about the assay, we cannot even begin to analyze whether the experimental and predicted data are comparable. We thus see it as extremely important to develop, and bring into general use, semantically useful description protocols in which the assay procedure is defined down to all the details that are necessary to make assay/activity data from scientific literature mutually comparable and to allow fact-based decisions on which data sets allow mix-and-matching for (Q)SAR model building and which not.

We would argue that a strong biological assay ontology classification will be useful, if not essential, for better integration of data sets associated with different assays. Widespread adaptation of bioassay ontology and its usage in scientific publications as well as in databases of biologically active compounds would, in our view, allow the field to improve overall data quality and, as a consequence, the predictivity of models built from these data.

Those efforts that do exist of creating semantically useful assay description ontologies, such as the BioAssay Ontology project[30] and the format of assay result documentation proposed by Orchard et al.[31] called MIABE (Minimum Information About a Bioactive Entity), do not seem to be widely used by biologists running assays, if they are known in that community at all. For example, the number of citations of Orchard et al.[31] by Google Scholar at the time of the writing of this manuscript was 39—a rather small number—and none of the citing papers actually used the assay documentation proposed in the Orchard paper to describe their bioassays but just mentioned that paper in the discussion part.

Quantification of the correlation between assay result sets can help with the mix-and-match questions. For example, if the overlap set (by structure) between assay result sets is of sufficient size and the correlation within this subset found to be acceptable, then a mutual calibration of the different data sets can be attempted. On the other hand, it has to be clear that poor correlation between assay result sets—expressed as R^2 in the simplest case—and poor model quality are related but not the same. After all, one can always build a bad model from good data, and conversely, even a poorly mix-and-matchable combined data set (with limited overlap) can lead to an apparently good model, but it may just be overfitting. Nevertheless, we believe that we have provided evidence that the latter is, to some extent, a signature of the former, i.e., that there is some correlation between good model quality and good mix-and-matchability.

Comparison with Other Approaches from Literature. Several studies have recently been published that estimate the inconsistency of data sets from public databases.[1,4] In these studies, several filters were suggested to filter out noisy data and errors of annotation in data from ChEMBL. Although these authors proposed an efficient algorithm for excluding records with errors of annotation, they did not explicitly investigate ways to compile data sets from databases of biologically active compounds specifically for the creation of predictive QSAR models.[1,4]

Gao et al.[5] reported a study of two sets of kinase inhibitors tested in two different assays. In contrast to our approach, the authors focused on the calculation of the concordance of kinase selectivity data for overlapping sets of compounds generated from four independent profiling sources but not on the analysis of the database contents for application to QSAR modeling.

CONCLUSIONS

We have investigated possible ways to compile combined training sets from generally (commercially and publicly) available databases of bioactive compounds for the purpose of building high performance QSAR models. While we focused on two specific databases—Integrity and ChEMBL—and one molecular target—HIV-1 reverse transcriptase—we believe that the conclusions we were able to draw from the results apply in a similar manner to other large-scale databases and many other biological targets.

Apart from the creation of useful RT QSAR models, a more general goal of this study was to investigate how data from large-scale databases can be used in a multidata set combination mode, how we can improve the accuracy and predictivity of QSAR models developed based on these data, and for which combined data sets the models were so poor that one must conclude that mix-and-matching was not possible. We clearly observed from the model quality parameters (Q^2 etc.) that the combined data sets straddled the limit of mix-and-matchability. While we were able to create filtered sets that yielded good models, other combined data sets (and even a significant number of primary data sets) yielded models devoid of predictivity. Dividing training sets along the lines of assay materials and methods generally provided the best performance QSAR models, supporting the assertion that details of the assay procedure are important, even though there are often poorly understood and even more poorly quantified if not entirely unrecognized factors affecting assay reproducibility and inter-assay compatibility. Another issue that emerged is the need to have data on mutations affecting target protein function that are formatted in such a manner that they can be parsed by computer in a way that is useful for (Q)SAR modeling.

One of the central issues for a retrospective, and even more so prospective, mix-and-matchability assessment is the near-complete lack of usage of universal and strong ontologies for biological assay data both in primary scientific publications and in the aggregating databases of biologically active substances. Consequently, there is currently no way to, e.g., aggregate data from ChEMBL and Integrity automatically.

In conclusion, we believe that there are currently significant problems with interassay data set compatibility within databases and even more so across different databases. Even the quantification of these issues appears to be in its early stages and fragmentary. We believe it will require a community effort in this field to achieve better integration of data from different sources using a unified terminology. We intend to present broader studies toward this goal in the future.

ASSOCIATED CONTENT

Supporting Information

Table S1: Description of Integrity database fields. List 1: Complete list of all fields' values of Integrity database. List 2: Descriptions of the assays used in the Integrity classification. Table S2: An overlap of structures between the data sets corresponding to different material-methods of the assay used in Integrity. Table S3: Pairwise correlation coefficients between the experimental values of the test set and results of prediction of the test set by the models created using the modeling sets, corresponding to the different materials and methods. The

DOI: 10.1021/acs.jcim.5b00019
J. Chem. Inf. Model. 2015, 55, 1388−1399

Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00019.

AUTHOR INFORMATION

Corresponding Author

*E-mail: olga.a.tarasova@gmail.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation of Basic Research (grant No. 13-04-91455_NIH-a). Part of this work was supported by the Intramural Program of the U.S. National Institutes of Health. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

REFERENCES

(1) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC_{50} Data - a Statistical Analysis. *PloS One* **2013**, *8*, e61007.

(2) Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M. J.; Kogej, T.; Tyrchan, C.; Varkonyi, P.; Xie, P. H. Making Every SAR Point Count: The Development of Chemistry Connect for the Large-Scale Integration of Structure and Bioactivity Data. *Drug Discovery Today* **2011**, *16*, 1019−1030.

(3) Williams, A. J.; Ekins, S.; Tkachenko, V. Towards a Gold Standard: Regarding Quality in Public Domain Chemistry Databases and Approaches to Improving the Situation. *Drug Discovery Today* **2012**, *17*, 685−701.

(4) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K(i) Data. *J. Med. Chem.* **2012**, *55*, 5165−5173.

(5) Gao, C.; Cahya, S.; Nicolaou, C. A.; Wang, J.; Watson, I. A.; Cummins, D. J.; Iversen, P. W.; Vieth, M. Selectivity Data: Assessment, Predictions, Concordance, and Implications. *J. Med. Chem.* **2013**, *56*, 6991−7002.

(6) Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. The Binding Database: Data Management and Interface Design. *Bioinf. Oxf. Engl.* **2002**, *18*, 130−139.

(7) Papadatos, G.; Overington, J. P. The ChEMBL Database: A Taster for Medicinal Chemists. *Future Med. Chem.* **2014**, *6*, 361−364.

(8) NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2014**, *42*, D7−D17.

(9) Filimonov, D. A.; Zakharov, A. V.; Lagunin, A. A.; Poroikov, V. V. QNA-Based "Star Track" QSAR Approach. *SAR QSAR Environ. Res.* **2009**, *20*, 679−709.

(10) Grados, O. B. The laboratory in programs for enteric infection control. *Bol. Oficina Sanit. Panam. Pan Am. Sanit. Bur.* **1975**, *78*, 318−322.

(11) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091−D1097.

(12) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Res.* **2013**, *41*, D801−D807.

(13) Qin, C.; Zhang, C.; Zhu, F.; Xu, F.; Chen, S. Y.; Zhang, P.; Li, Y. H.; Yang, S. Y.; Wei, Y. Q.; Tao, L.; Chen, Y. Z. Therapeutic Target Database Update 2014: A Resource for Targeted Therapeutics. *Nucleic Acids Res.* **2014**, *42*, D1118−D1123.

(14) Southan, C.; Sitzmann, M.; Muresan, S. Comparing the Chemical Structure and Protein Content of ChEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target Database. *Mol. Inform.* **2013**, *32*, 881−897.

(15) 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.

(16) Sitzmann, M.; Filippov, I. V.; Nicklaus, M. C. Internet Resources Integrating Many Small-Molecule Databases. *SAR QSAR Environ. Res.* **2008**, *19*, 1−9.

(17) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C. A New Approach to Radial Basis Function Approximation and Its Application to QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 713−719.

(18) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C. QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem. *J. Chem. Inf. Model.* **2014**, *54*, 705−712.

(19) Poroikov, V.; Filimonov, D.; Lagunin, A.; Gloriozova, T.; Zakharov, A. PASS: Identification of Probable Targets and Mechanisms of Toxicity. *SAR QSAR Environ. Res.* **2007**, *18*, 101−110.

(20) Filimonov, D. A.; Lagunin, A. A.; Gloriozova, T. A.; Rudik, A. V.; Druzhilovskii, D. S.; Pogodin, P. V.; Poroikov, V. V. Prediction of the Biological Activity Spectra of Organic Compounds Using the Pass Online Web Resource. *Chem. Heterocycl. Compd.* **2014**, *50*, 444−457.

(21) Zakharov, A. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Quantitative Prediction of Antitarget Interaction Profiles for Chemical Compounds. *Chem. Res. Toxicol.* **2012**, *25*, 2378−2385.

(22) Kubinyi, H. QSAR in Drug Design. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-vch Verlag GmhH & Co.: Weinheim, Germany, 2003; Vol. *4*, pp 1532−1553.

(23) Hoffman, J. M.; Smith, A. M.; Rooney, C. S.; Fisher, T. E.; Wai, J. S.; Thomas, C. M.; Bamberger, D. L.; Barnes, J. L.; Williams, T. M.; Jones, J. H. Synthesis and Evaluation of 2-Pyridinone Derivatives as HIV-1-Specific Reverse Transcriptase Inhibitors. 4. 3-[2-(Benzoxazol-2-Yl)ethyl]-5-Ethyl-6-Methylpyridin-2(1H)-One and Analogues. *J. Med. Chem.* **1993**, *36*, 953−966.

(24) Hargrave, K. D.; Proudfoot, J. R.; Grozinger, K. G.; Cullen, E.; Kapadia, S. R.; Patel, U. R.; Fuchs, V. U.; Mauldin, S. C.; Vitous, J.; Behnke, M. L. Novel Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase. 1. Tricyclic Pyridobenzo- and Dipyridodiazepinones. *J. Med. Chem.* **1991**, *34*, 2231−2241.

(25) Williams, M. Qualitative Pharmacology in a Quantitative World: Diminishing Value in the Drug Discovery Process. *Curr. Opin. Pharmacol.* **2011**, *11*, 496−500.

(26) Proudfoot, J. R.; Hargrave, K. D.; Kapadia, S. R.; Patel, U. R.; Grozinger, K. G.; McNeil, D. W.; Cullen, E.; Cardozo, M.; Tong, L.; Kelly, T. A. Novel Non-Nucleoside Inhibitors of Human Immunodeficiency Virus Type 1 (HIV-1) Reverse Transcriptase. 4. 2-Substituted Dipyridodiazepinones as Potent Inhibitors of Both Wild-Type and Cysteine-181 HIV-1 Reverse Transcriptase Enzymes. *J. Med. Chem.* **1995**, *38*, 4830−4838.

(27) Romero, D. L.; Olmsted, R. A.; Poel, T. J.; Morge, R. A.; Biles, C.; Keiser, B. J.; Kopta, L. A.; Friis, J. M.; Hosley, J. D.; Stefanski, K. J.; Wishka, D. G.; Evans, D. B.; Morris, J.; Stehle, R. G.; Sharma, S. K.; Yagi, Y.; Voorman, R. L.; Adams, W. J.; Tarpley, W. G.; Thomas, R. C. Targeting Delavirdine/atevirdine Resistant HIV-1: Identification of (alkylamino)piperidine-Containing Bis(heteroaryl)piperazines as Broad Spectrum HIV-1 Reverse Transcriptase Inhibitors. *J. Med. Chem.* **1996**, *39*, 3769−3789.

(28) Hay, R. J.; Iconomi, P. Detection of Microbial and Viral Contaminants in Cell Lines. In *Cell Biology*, 3rd ed.; Celis, J. E., Ed.; Elsevier: San Diego, CA, 2006; Vol. *1*, pp 49−65.

(29) Kool, J.; Lingeman, H.; Niessen, W.; Irth, H. High Throughput Screening Methodologies Classified for Major Drug Target Classes according to Target Signaling Pathways. *Comb. Chem. High Throughput Screen.* **2010**, *13*, 548−561.

(30) Abeyruwan, S.; Vempati, U. D.; Küçük-McGinty, H.; Visser, U.; Koleti, A.; Mir, A.; Sakurai, K.; Chung, C.; Bittker, J. A.; Clemons, P. A.; Brudz, S.; Siripala, A.; Morales, A. J.; Romacker, M.; Twomey, D.; Bureeva, S.; Lemmon, V.; Schürer, S. C. Evolving BioAssay Ontology (BAO): modularization, integration and applications. *J. Biomed Semantics* **2014**, S1−S5.

(31) 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, R. K.; 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 Discovery* **2011**, *10*, 661−669.

DOI: 10.1021/acs.jcim.5b00019
J. Chem. Inf. Model. 2015, 55, 1388−1399