

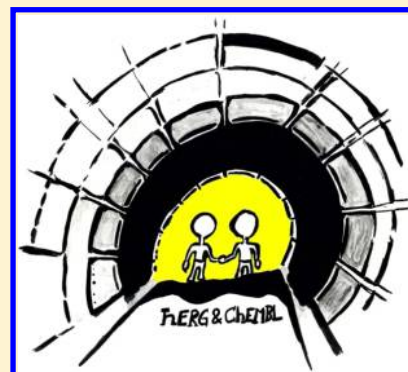
hERG Me Out

Paul Czodrowski

Merck KGaA, Small Molecule Platform, Global Computational Chemistry, Frankfurter Strasse 250, 64293 Darmstadt, Germany

S Supporting Information

ABSTRACT: A detailed analysis of the hERG content inside the ChEMBL database is performed. The correlation between the outcome from binding assays and functional assays is probed. On the basis of descriptor distributions, design paradigms with respect to structural and physicochemical properties of hERG active and hERG inactive compounds are challenged. Finally, classification models with different data sets are trained. All source code is provided, which is based on the Python open source packages RDKit and scikit-learn to enable the community to rerun the experiments. The code is stored on github (https://github.com/pzc/herg_chembl_jcim).



INTRODUCTION

Inhibition of the human ether a-go-go related gene product (aka *hERG*) is associated with QT interval prolongation and can lead to heart arrhythmia^{1,2} and death. Several drugs (e.g., terfenadine,^{3,4} astemizole,⁵ cisapride⁶) have been taken from the market due to hERG liabilities. Therefore, it is of utmost importance that potential drugs are free of this risk, and consequently, to understand and unravel the interaction with the hERG channel during the design phase of the compounds.

Many hERG-related reports have been published, ranging from QSAR-based models^{7–27} to matched-pair analysis^{28,29} to pharmacophore-based approaches^{30,31} to structure-based studies employing homology models^{32,33} of Kv11.1 (hERG: human ether a-go go related gene). Most of these publications make use of commercial or proprietary software tools, which complicates reproducibility. In addition, the original data sets are not always provided.

Recently, a few publications concerning the usage of public data extracted from ChEMBL³⁴ appeared in the literature.^{35–37} These reports encouraged us to investigate ChEMBL's content with respect to hERG. The use of public data comes at an appropriate time because the usage of open data and source software tools is becoming more prominent in the pharmaceutical industry due in part to financial constraints. Open Source tools such as RDKit³⁸ have gained a level of maturity that convince a greater proportion of computational chemists to incorporate this toolkit into their workflows. When such open source tools are employed, error-checking and free support by a large user community becomes possible (aka *crowd sourcing*). Furthermore, the exchange of source code becomes more feasible. Lastly, knowledge transfer is made easier due to public access to the source code and documentation via mailing lists and blogs.

In this publication, data from a public source (ChEMBL) in combination with open source tools (RDKit,³⁸ scikit-learn,³⁹

matplotlib,⁴⁰ SciPy,⁴¹ KNIME⁴²) are used. All data and source are freely available (see Supporting Information). This enables anyone to perform the described analysis and model building, repeat the experimentation, and also to adapt it according to his/her own needs.

Beyond the technical details and the ethical questions concerning reproducibility, the following questions will be addressed in this communication: (1) Which sources and activities does ChEMBL provide in terms of hERG data? (2) Is there a correlation between ChEMBL's functional hERG assays and the hERG binding assays? (3) Are there generally applicable rules that differentiate between hERG active and hERG inactive compounds? (4) Is it possible to train classifier models for ChEMBL's hERG data?

MATERIALS AND METHODS

The extraction from the ChEMBL database was done based on a KNIME workflow (accessible via the public Web server). In addition, ChEMBL provides a Python RESTful interface (<https://www.ebi.ac.uk/chembl/db/index.php/ws#pythonClient>). The target_ID was set to the uniprot code Q12809 (target_ID = 240) corresponding to the human form of the Kv 11.1 ion channel synonymously used for the hERG channel. The date of the extraction was February 28, 2013, and ChEMBL version 15 was employed.

This query results in a total of 11,958 compounds with any biochemical affinity (percent effect, IC₅₀, pIC₅₀, K_i) for the hERG channel. This data set is split into compounds that were analyzed by employing a functional assay or by using a binding assay. Only compounds with IC₅₀/pIC₅₀/K_i values without any operator (> or <) were retained. Compounds labeled as

Received: May 22, 2013

Published: August 14, 2013



Table 1. Overview of Data Sets

data set	number of compounds	threshold at 1 μ M		threshold at 10 μ M	
		hERG inactive	hERG active	hERG inactive	hERG active
binding assay	3721	2748 (73%)	973 (26%)	966 (26%)	2755 (74%)
functional assay	694	623 (90%)	71 (10%)	404 (58%)	290 (42%)

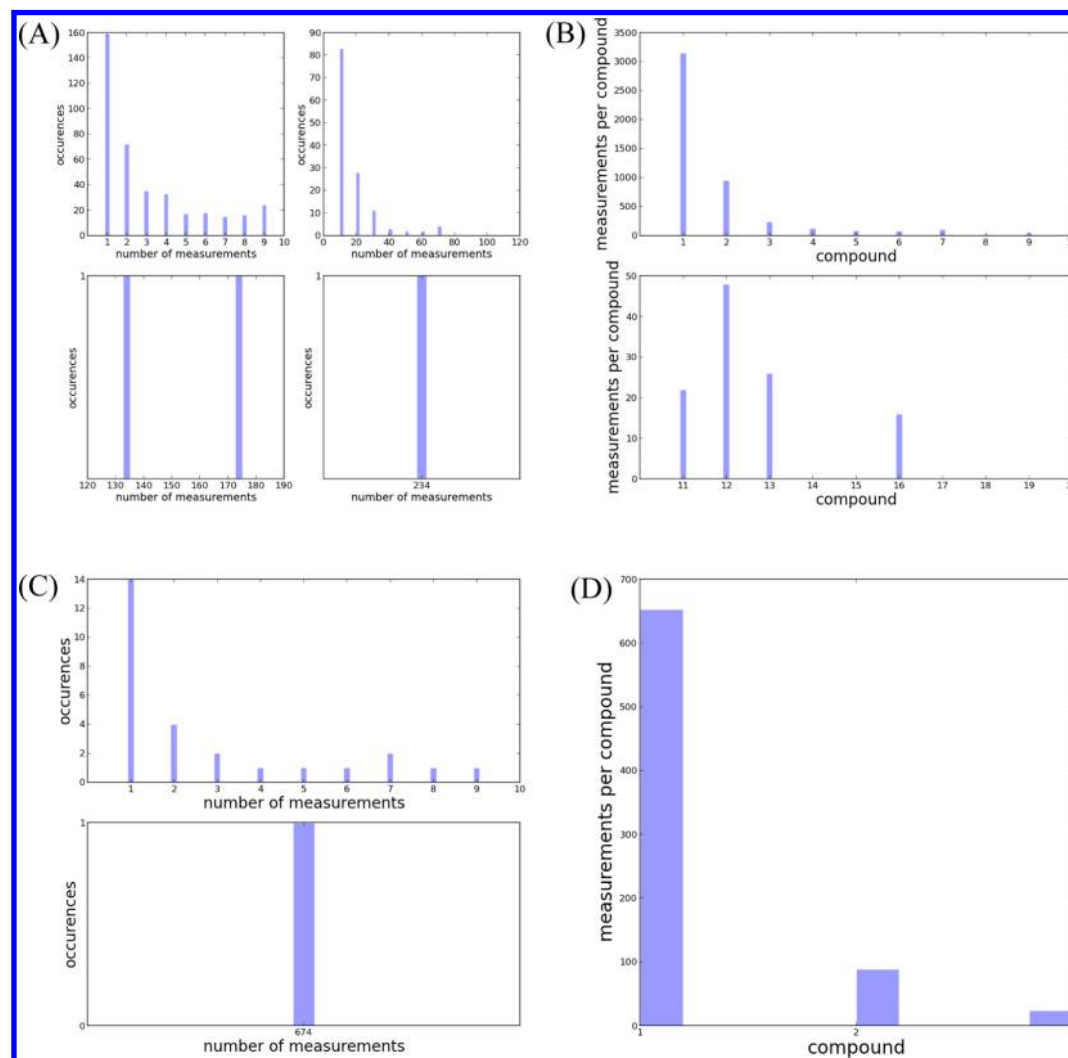


Figure 1. Measurements per assay or per compound. (A) Binding assay subset, measurements per ChEMBL assay ID. (B) Binding assay subset, measurements per compound. (C) Functional assay subset, measurements per ChEMBL assay ID. (D) Functional assay subset, measurements per compound.

inactive were discarded. This resulted in 4892 compounds that were measured in a binding assay and 765 compounds that were measured in a functional assay. Only the largest fragment of the reported compound was kept for both data sets. For duplicate entries, based on the canonical SMILES calculated by RDKit, the mean value of the affinity and the standard deviation were calculated.

Compounds with multiple hERG measurements for which the standard deviation is larger than the mean IC_{50} or those possessing a single IC_{50} value larger than thrice the standard deviation were not further considered. This resulted in a final data set consisting of 3721 compounds from the initial binding assay data set and 694 compounds from the initial functional assay data set.

The pre-processed data sets can be found as text files, in SMILES format, within the Supporting Information. SD Files can be requested from the author.

An initial binary classification was done based on the IC_{50} value; hERG_TL (TL stands for *traffic light*) was set to 1 for active compounds and 0 for inactive compounds. As threshold values for the active/inactive classification either 1 or 10 μ M was chosen (Table 1).

All available RDKit descriptors were calculated (see Supporting Information for an overview). Because of the fact that the Random Forest classifier implicitly performs a feature selection, no reduction of the descriptor space was necessary.

For the MCSS (maximum common substructure) search, the threshold was set to 0.1, i.e., the MCSS moiety needs to be contained in at least 10% of the complete structure. The

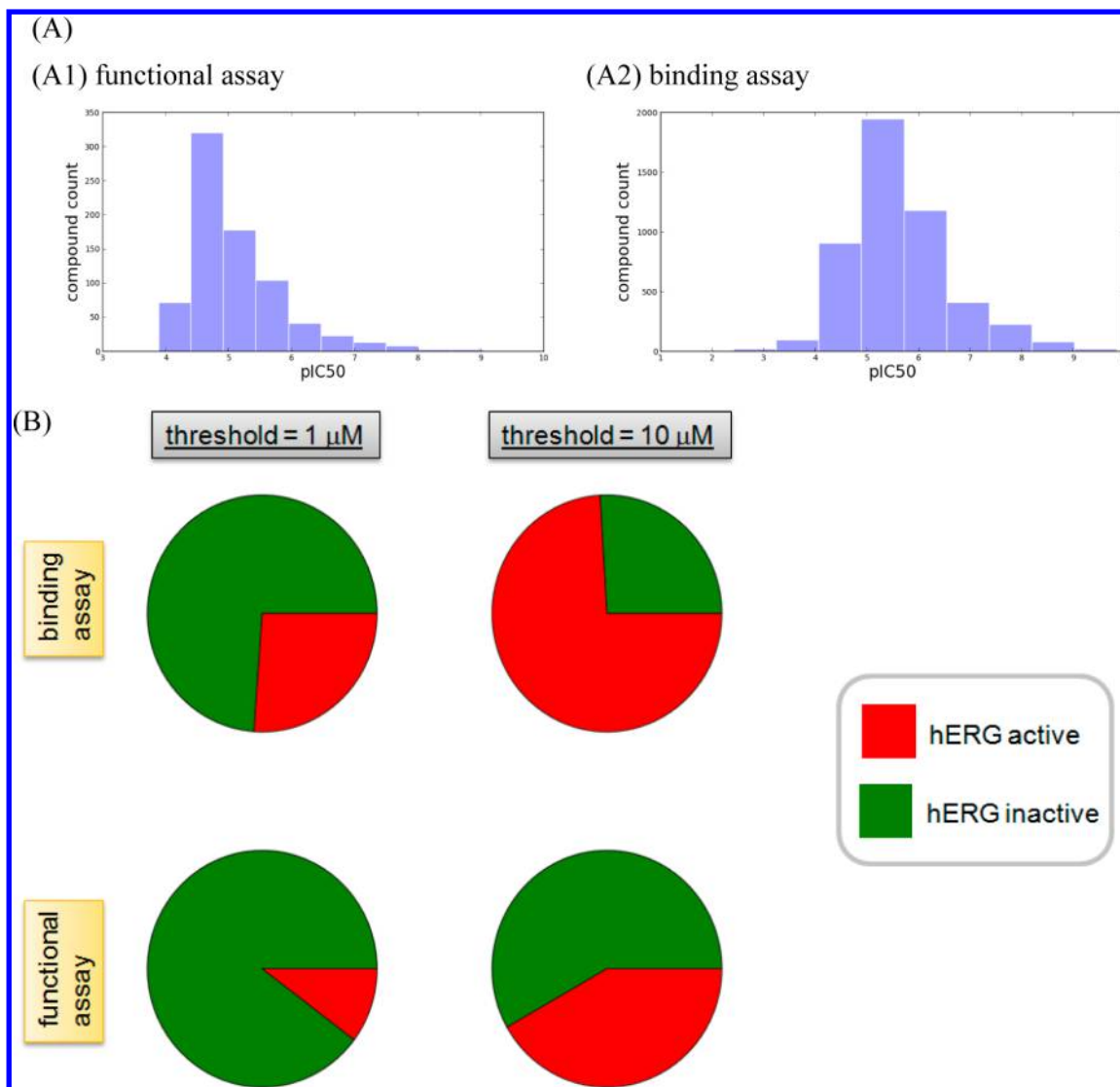


Figure 2. (A) pIC₅₀ distribution for the (A1) functional assay. (A2) Binding assay data. (B) Class distribution for the functional assay and binding assay given the two used threshold values.

minimum number of bonds is set to 5, and only complete rings were considered as part of a MCSS.

Random Forest was chosen as classifier with the following settings: (1) Number of trees was set to 100 ($n_{\text{estimators}} = 100$). (2) The minimum number of samples to split an internal node was set to 1 ($\text{min_samples_split} = 1$, default setting). (3) The minimum number of samples in newly created leaves was set to 1 ($\text{min_samples_leaf} = 1$, default setting). (4) The number of features to consider when looking for the best split was set to the square root of the number of descriptors ($\text{max_features} = \text{auto}$, default setting). (5) The maximum depth of the tree was expanded until all leaves are pure or until leaves contain less than min_samples_split samples ($\text{max_depth} = \text{none}$, default setting). (6) Bootstrap samples were used ($\text{bootstrap} = \text{true}$, default setting).

For further documentation on the Random Forest implementation in scikit-learn, the interested reader is referred to the Web site (<http://scikit-learn.org>).

A random split into training and test set was done 10 times; the ratio between training and test set is 60% and 40%. For each split, a 5-fold cross validation was performed.

The confusion matrix of the binary classification model is defined as follows

$$\text{confusion matrix} = \begin{matrix} & \text{prediction} \\ \begin{matrix} \text{experiment} \\ \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{matrix} \end{matrix}$$

On the basis of the confusion matrix, these figures of merit were used for the evaluation of the models

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

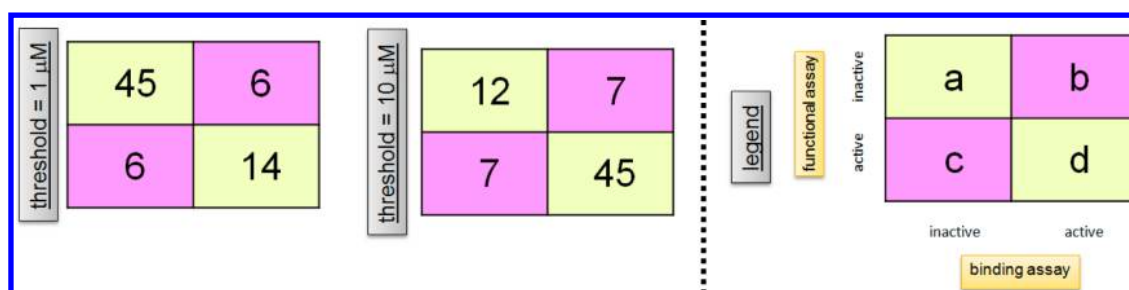


Figure 3. Compounds that have been analyzed in the binding assay as well as in the functional assay. For the matrix elements a and d, binding assay and functional assay agree in the hERG_{TL} classification scheme. In the case of matrix element c, the functional assay is more sensitive than the binding assay. The contrary is found for matrix element d.

Table 2. TPSA/MolLogP Box Plots for the ChEMBL Data Sets

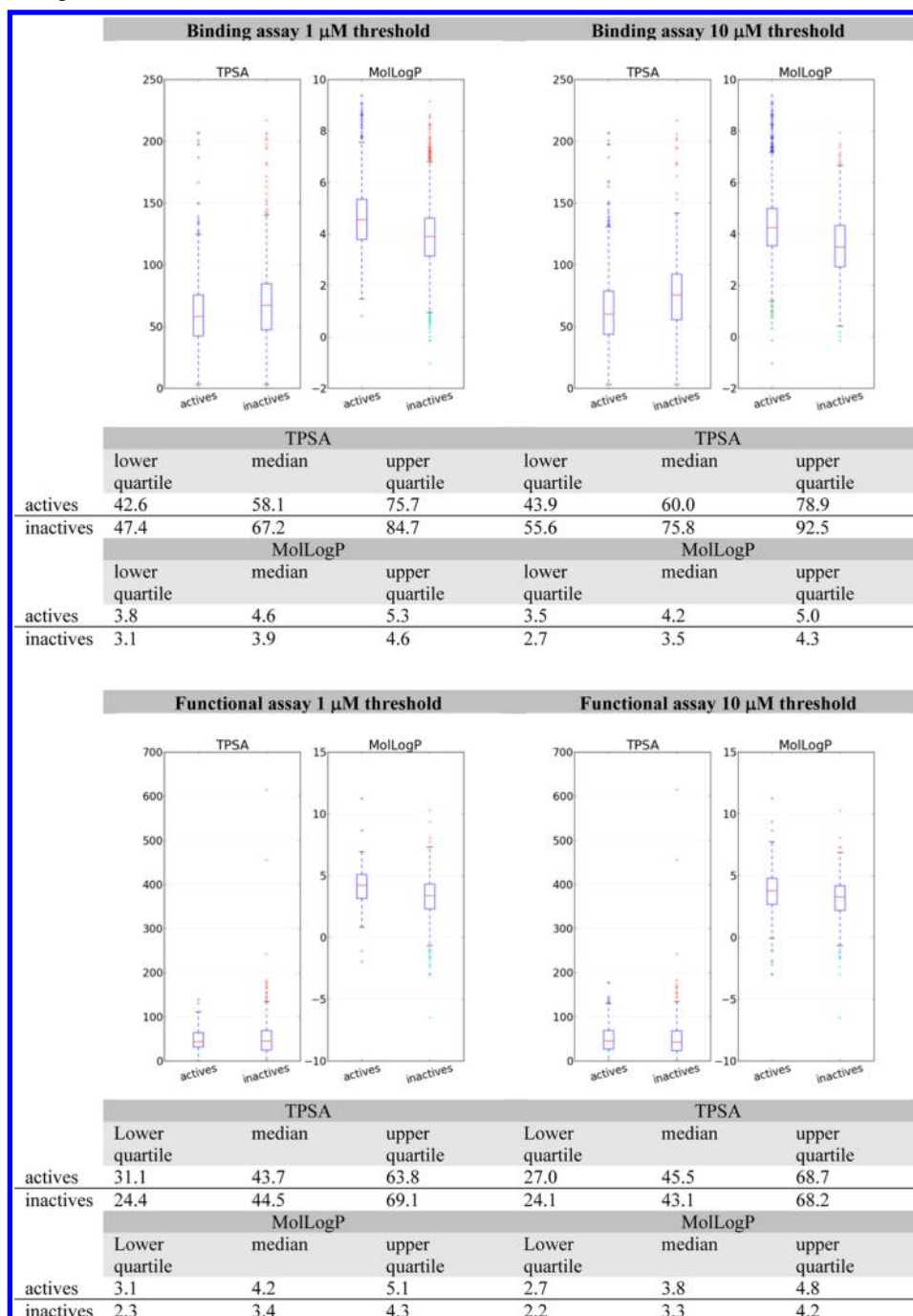
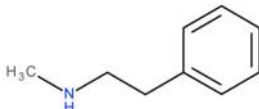

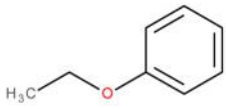

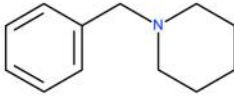
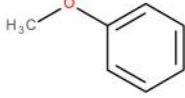
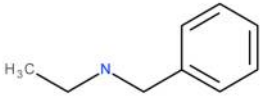
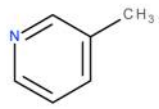
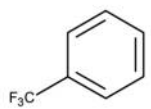
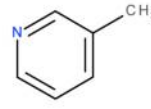
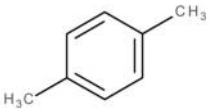
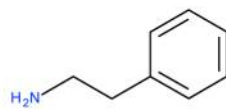
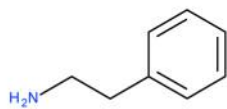
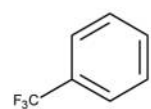
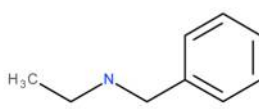
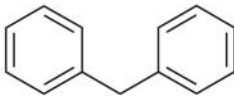
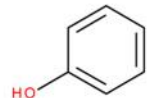
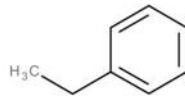
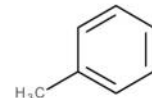
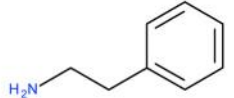
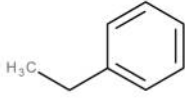
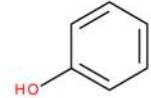
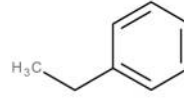
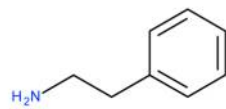


Table 3. MCSS (maximum common substructures) Found for the hERG Active and hERG Inactive Compounds of the Binding Assay and the Functional Assay^a

binding assay, threshold at 1 μ M		binding assay, threshold at 10 μ M	
active	inactive	active	inactive
 174/973	 890/2748	 316/2755	 372/966
 109/973	 677/2748	 300/2755	 168/966
 98/973	 362/2748		 119/966
	 335/2748		 117/966
	 278/2748		 117/966
functional assay, threshold at 1 μ M		functional assay, threshold at 10 μ M	
active	inactive	active	inactive
 16/71	 192/623	 97/290	 198/404
 9/71	 161/623	 84/290	 101/404
		 30/290	

^aThe numbers below the structural element refer to the frequency of the occurrence of the MCSS. To be regarded as MCSS, the motif needs to bear at least five heavy atoms, and only complete rings are considered.

Matthew's Correlation Coefficient (MCC)

$$= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The following program versions were used: KNIME 2.7.0,⁴² RDKit 2012_09,³⁸ scikit-learn 0.12,³⁹ SciPy 0.10.1,⁴¹ matplotlib 1.1.1.⁴⁰

RESULTS

Which sources and activities does ChEMBL provide in terms of hERG data? ChEMBL mainly contains fractions of

hERG data with small to medium-sized data sets. The largest homologous data set, i.e., with the same assay ID, has 234 entries. Under the ChEMBL assay ID 1827362, a collection of literature data from various sources is available.⁴³ For the functional assay subset, the largest data chunk stems from the ChEMBL assay ID 1794573. This assay is a PubChem Assay (PubChem AssayID: 588834) containing 674 data points. It employed a cell-based assay using the U2OS cell line and measured the activity using a thallium dye (FluxORMT).

With respect to measurements per compound, the majority of the assembled data points consist of only a single

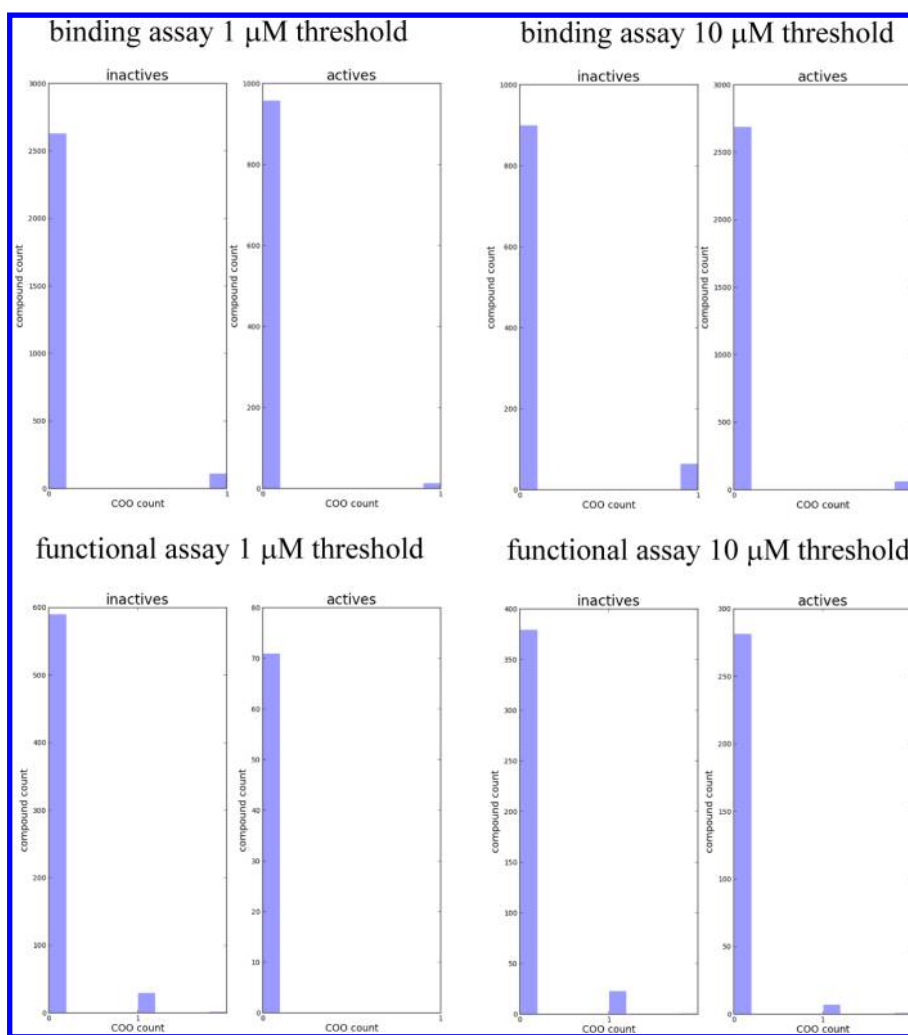


Figure 4. Frequency count for the COOH group apparent in the ChEMBL data sets.

measurement per compound. A total of 63% of the compounds in the binding data subset were measured only once, while 94% of the compounds in the functional datasubset have been measured only a single time.

The details of this analysis can be found in Figure 1(A–D). For the binding assay, terfenadine is the compound with the most measurements (16 times) over all hERG binding assays. A closer look at the reported activities reveals that not all terfenadine data points refer to measurements from a primary experiment. Roughly half of the reported data points stem from literature citing the primary experiment.

In Figure 2, the activity distributions are given. Figure 2(A) gives the pIC_{50} ($pIC_{50} = -\log_{10}(IC_{50})$) – distribution, whereas Figure 2(B) shows the class distribution after the assignment of the hERG traffic light (hERG_TL) given the two different thresholds of 1 and 10 μ M. For a more complete description, see below.

The functional assay subset shows less potent compounds compared to the binding assay subset (Figure 2(A)). For the functional assay, the peak of the distribution is between 4 and 5, which corresponds to double and triple digit μ M inhibition. In contrast, most of the compounds from the binding assay show a single to double digit μ M inhibition.

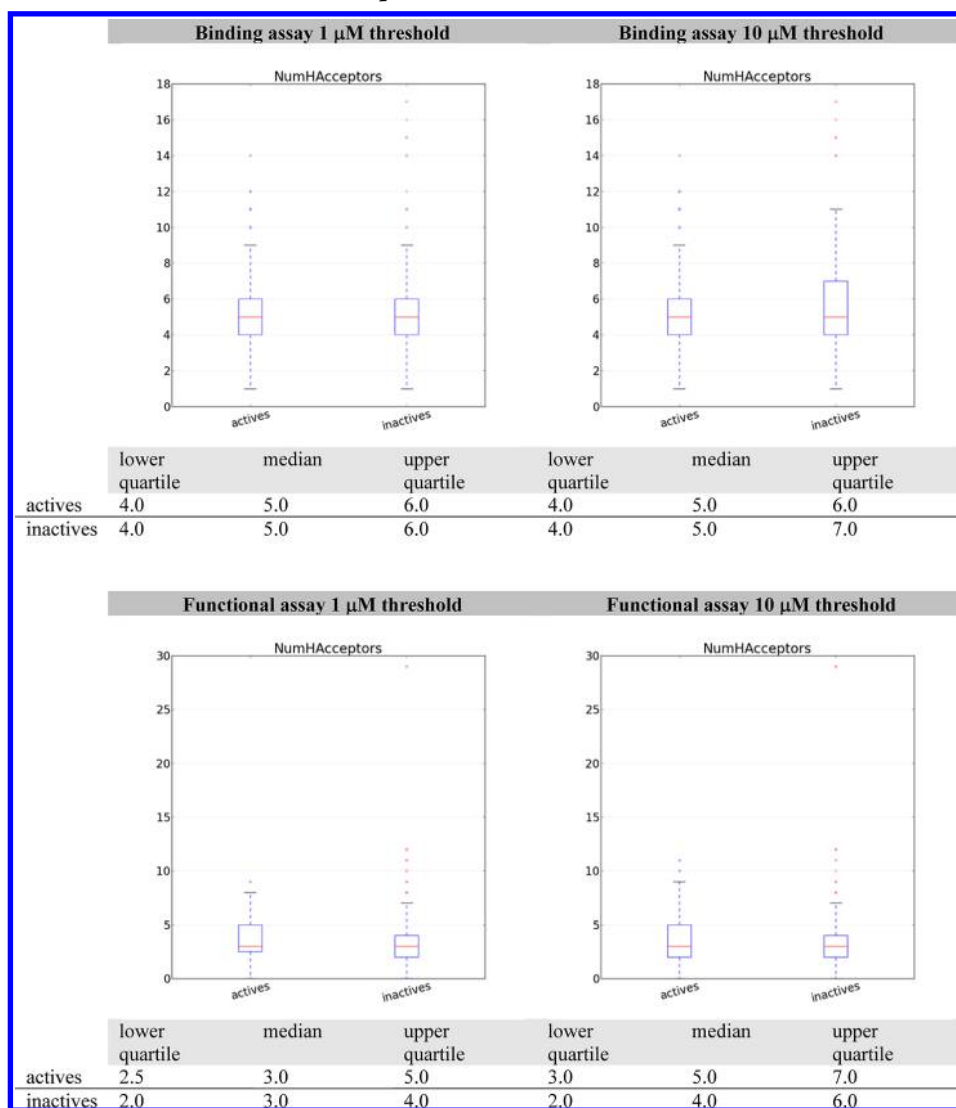
Is there a correlation between ChEMBL’s functional hERG assays and the hERG binding assays? The two assay

principles, binding assay and functional assay, sample different effects on the Kv 11.1 channel. Whereas the binding assay samples the occupation of one particular binding site, the functional assay represents a more holistic view of the effect of the compound on the Kv 11.1 channel. What makes the binding assay more attractive to the researcher are the lower costs in combination with a higher throughput.

When checking for a correlation between the two assay systems, there is an overlap of measurements with both assays for 71 compounds (Figure 3). For the two different thresholds used throughout this publication, the following is observed: (1) For the majority of the measurements, the functional assay and the binding assay are in agreement. This is the case for 59 of 71 (83%) compounds at the 1 μ M threshold and 57 of 71 (80%) of compounds at the 10 μ M threshold. (2) For the remaining 12 compounds at the 1 μ M threshold and 14 compounds at the 10 μ M threshold, there is an equal split between cases where the binding assay and functional assay disagree in the hERG_TL classification scheme.

Are there generally applicable rules that differentiate between hERG active and hERG inactive compounds? In text books, some general rules exist how to trigger hERG activity:⁴⁴ (1) reduce the lipophilicity, (2) Reduce the pK_a (basicity) of the amine, (3) add an acid moiety, (4) add an oxygen H-bond acceptor, and (5) rigidify the linkers.

Table 4. Box Plots for the Number of H-Bond Acceptors in the ChEMBL Data Sets



The TPSA/MolLogP box plot (Table 2) for the binding assay subsets indeed agree with the first general rule. The TPSA value is larger for the hERG inactive compounds, whereas MolLogP goes down when comparing hERG inactive with hERG active compounds. In the case of the functional assay subsets, the TPSA values are in a similar range for the hERG active and hERG inactive compounds, but the MolLogP trend is the same as for the binding assay.

Because RDKit does not calculate a pK_a value, a MCSS analysis was performed. This analysis aims to detect the preponderance of basic centers for hERG active compounds. In the case of the binding assay subsets, there are MCSS moieties found for the hERG inactives that bear a basic center. This is not the case for the functional assay subsets. All MCSS moieties can be found in Table 3.

In order to decipher the acid moiety rule, a MMP analysis was performed in KNIME. For the functional assay data, no transformation to an acid was found, neither at a 1 μM nor at a 10 μM threshold. In the case of the binding assay data, in total seven transformations from H to COOH can be observed. Given a 1 μM threshold, none of the H \rightarrow COOH transformations have a negative influence on the transition of the hERG traffic light, two transformations have a neutral

influence, and five transformations show a transition from hERG active to hERG inactive. Given a 1 μM threshold, five transformations have a neutral influence, and two transformations show a transition from hERG active to hERG inactive.

This shows that for the binding assay, the introduction of an acid moiety has a positive effect on preventing a potential hERG liability (for a small number of observations).

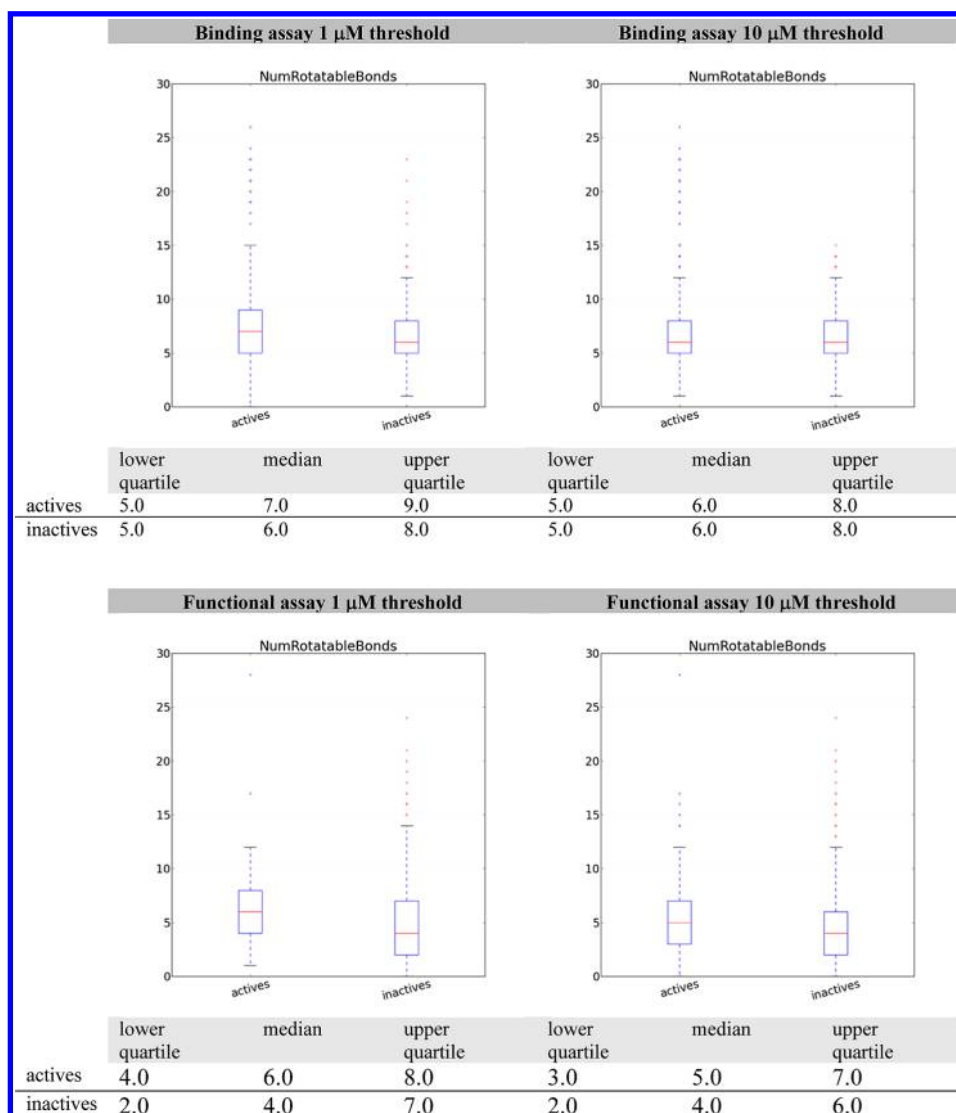
The pure COOH count is given in Figure 4. It shows that only a minority of the compounds bears an acid moiety.

In RDKit, there is no descriptor for the oxygen H-Bond acceptor count. However, it is possible to calculate the overall H-Bond acceptor count (Table 4). No clear distinction can be made on the basis of this descriptor.

The last of the accepted beliefs is challenged in Table 5. Here, the box plot shows the spread of the number of rotatable bonds. No clear distinction can be made on the basis of this descriptor.

Is it possible to train classifier models for ChEMBL's hERG data? The results from an average of all random train/test set splits are shown in Table 6(A); all detailed results can be found in the Supporting Information. The results for a

Table 5. Box Plots for the Number of Rotatable Bonds in the ChEMBL Data Sets

Table 6. (A) Results for Random Forest-Based Classification Models^a (B) Results for Random Forest-Based Classification Models Based on Y-Scrambled Data^a

(A)							
accuracy	MCC	precision	recall	f1	auc	assaytype	threshold
0.907 ± 0.014	0.24 ± 0.188	0.546 ± 0.397	0.136 ± 0.107	0.209 ± 0.156	0.564 ± 0.054	functional	1 μM
0.692 ± 0.037	0.354 ± 0.081	0.667 ± 0.061	0.527 ± 0.066	0.586 ± 0.055	0.668 ± 0.038	functional	10 μM
0.801 ± 0.012	0.423 ± 0.041	0.726 ± 0.042	0.385 ± 0.035	0.502 ± 0.037	0.666 ± 0.019	binding	1 μM
0.797 ± 0.015	0.405 ± 0.051	0.815 ± 0.011	0.94 ± 0.013	0.873 ± 0.009	0.662 ± 0.023	binding	10 μM
(B)							
accuracy	MCC	precision	recall	f1	auc	assaytype	threshold
0.481 ± 0.042	−0.04 ± 0.085	0.476 ± 0.046	0.455 ± 0.073	0.463 ± 0.055	0.48 ± 0.042	functional	1 μM
0.494 ± 0.036	−0.014 ± 0.074	0.508 ± 0.036	0.521 ± 0.063	0.513 ± 0.043	0.493 ± 0.037	functional	10 μM
0.491 ± 0.016	−0.019 ± 0.033	0.482 ± 0.018	0.457 ± 0.03	0.469 ± 0.023	0.49 ± 0.016	binding	1 μM
0.508 ± 0.016	0.016 ± 0.032	0.505 ± 0.017	0.482 ± 0.03	0.493 ± 0.022	0.508 ± 0.016	binding	10 μM

^aThe presented results are based on an average over all 10 random train/test splits.

prediction based on Y-scrambled data can be found in Table 6(B).

A comparison of these result sets indicates that model building for the functional assay given a 1 μM threshold is challenging. MCC fluctuates significantly, and in two out of ten

train/test splits, the standard deviation (based on the 5-fold cross validation) exceeds the mean value. In addition, precision, recall and f1 fluctuate to a large extent. This is also reflected in the mean value of all train/test splits and the performed cross validations.

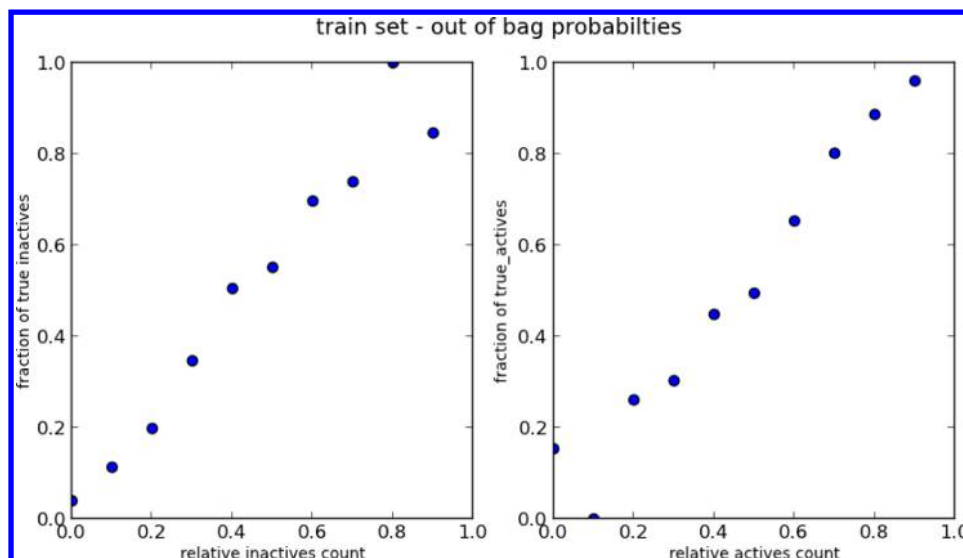


Figure 5. Plot of the out-of-classification probabilities per hERG_{TL} class. On the *x* axis, the predicted probabilities for the inactive compounds (left) or active compounds (right) are plotted. On the *y*-axis, the fraction of true inactive compounds (left) or fraction of true active compounds (right) are shown.

Given a 10 μM threshold/functional assay data, MCC fluctuates to a lesser extent, and the other parameters show a smaller standard deviation.

With respect to confidence of the models, the out-of-bag probability scores for the training sets (binding assay, 10 μM activity threshold) is computed. Here, it becomes clear that the models show a high confidence, i.e., given a probability between 90 and 100% to be hERG active, 95% of those compounds are truly hERG active. This also means that given a probability score of 70%, there are 70% true active compounds (Figure 5).

DISCUSSION

Which sources and activities does ChEMBL provide in terms of hERG data? There is large variety of different data sources available in ChEMBL (Figure 1). Most of the measurements are single measurements meaning that only for a minority multiple measurements for the same compound are available. For the binding assay, the amount of more multiple measurements is due to the fact the data basis is larger. Furthermore, the functional assay consists to a large extent of one assay only (ChEMBL assay ID 1794573).

In the binding assay subsets, different probe molecules have been used. The different probe molecules may or may not sample different binding sites. A distinction of the probe molecules is not possible given the current ChEMBL version. In contrast, the cell line used in the functional assay is reported only for a small number of compounds. Furthermore, this information cannot be automatically queried because it is not stored as separate property. It is rather noted as free text in the assay description, and no uniform format is given for the assay description. If such distinctions were made, it would be possible to further differentiate these data. However, in the case of the functional assay, this may result in quite small subsubsets, for which model building could be difficult. In the case of the binding assay, such a distinction might be more helpful.

Because the pIC_{50} spread is far from being uniform, classification models instead of regression models were trained. Furthermore, a large variety of different IC_{50} values would have been mixed. Because the assay formats cannot be compared, the combination of such IC_{50} values is not scientifically sound.

The terfenadine example shows that ChEMBL does not appropriately check for the source of the measurement. It mixes up experimentally determined IC_{50} values that are cited in, for example, a review article. Although a correct value is reported, the number of actual measurements is wrong.

Is there a correlation between ChEMBL's functional hERG assays and the hERG binding assays? Only for a small subset, affinity values from measurement in both assay types are available. In total, the full data matrix exists for 71 compounds, which comprises 10% of the functional assay data but only 2% of the binding assay data. This shows that the deposited values in ChEMBL represent the situation that researchers usually publish one measured affinity data point from one assay type.

Although the available data set is rather small, first hints can be deduced with respect to the correlation between the binding assay data and the functional assay data. For the majority of the compounds (83% at a 1 μM threshold, 80% at a 1 μM threshold), both assay types classify a compound with the same hERG traffic light. The remaining compounds do not show a uniform trend, e.g., that the binding assay is less predictive than the functional assay. This is surprising due to the fact that the binding assay per se is not capable of sampling all binding sites. Therefore, one might have assumed that there are more compounds inactive in the binding assay but active in the functional assay (matrix element c in Figure 3) compared the opposite (matrix element b in Figure 3). Because the binding assay data in ChEMBL combines all binding assays that sample different binding sites, this effect may have been diminished.

Are there generally applicable rules that differentiate between hERG active and hERG inactive compounds?

The accepted generalities of how to lower hERG activity were challenged by the descriptor distributions and analyses of the structural motifs. (1) The box plots for TPSA and MolLogP suggest that the lipophilicity rule holds true (i.e., lowering the lipophilicity makes compounds less prone to hERG risk). In the case of the functional assay, the TPSA values lie in a similar range for hERG actives and hERG inactives, but the MolLogP trend is the same as for the binding assay. (2) The effect of the pK_a value was indirectly measured. Because RDKit does not

provide a means to calculate a pK_a value, the MCSS analysis shall help to explain the importance of basic centers. Indeed, the MCSS analysis reveals that basic centers are present in hERG active compounds, but this is also true for the hERG inactive compounds, at least for the binding assay subset. Here, it is interesting to note that hERG inactives bear basic centers, which is contradictory to the common understanding in drug design. In the case of the functional assay subset, basic centers are exclusively found for the hERG actives. (3) A typical approach to lower the hERG activity of compound is to introduce an acidic center. The matched pair analysis reveals that only for the binding assay subsets are such transformations reported. None of these transformations has the effect to result in a hERG active compound, i.e., this suggests that this rule holds true. (4) For the last two rules (oxygen H-Bond acceptor rule and rigidity rule), the descriptor analysis does not support the accepted generality.

The ChEMBL data sets support the common design paradigms in only two out of the five rules. It may be interesting to follow this trend over the up-coming ChEMBL versions. Furthermore, given a different view, for example, if one restricts the data set to particular chemical classes, the conclusions might be different. Because the aim of this publication was a rather holistic analysis of the ChEMBL database, such a restriction is outside its scope.

Is it possible to train classifier models for ChEMBL's hERG data? The models based on the binding assay data show a much higher predictability compared to the models based on functional assay data. This can be due to the fact that the data basis of the binding assay data set is larger. Nonetheless, for both assay worlds, the distribution of inactive to active compounds is not in perfect balance. Only for the functional assay given at a threshold of 10 μ M, one might argue that the distribution between active and inactive compounds is almost equal. However, the obtained model is not significantly better than the models based on more imbalanced data.

The comparison with other published studies is not straightforward. A recent study making use of ChEMBL data combines data from binding and functional assay.¹¹ Furthermore, the authors report on MCSS moieties for hERG active compounds, but they do not perform such an analysis for the hERG inactive compounds. Our results indicate that no privileged hERG active moieties are found in the ChEMBL database. The work by Li et al.¹⁶ publishes models at 1 and 10 μ M. The accuracies are between 0.87 and 0.72, whereas our accuracies lie between 0.69 and 0.91. Their sensitivities are between 0.16 and 0.45, whereas our recall values fall between 0.14 and 0.94.

CONCLUSIONS

To the best of the author's knowledge, this study comprises the first detailed analysis of the hERG content inside ChEMBL. ChEMBL represents a rich data source for hERG-related data, and the database is growing due to the fact that more and more hERG data becomes published.

It becomes apparent that most of ChEMBL's data stems from small to medium-sized data chunks that were published in MedChem-related papers. There is only one large data chunk that comes from a qHTS run available via PubChem. This data chunk makes up the largest part of the functional assay subset. In the case of the binding assay subset, only small data sets make up the final data set.

Because the description of the assay formats is not standardized, a comparison can only be made on the basis of the actual activity values. Here, measurements show a large deviation between the different assays. The data basis is still not good enough to train regression models, but classification models are feasible and reasonable given the ChEMBL data basis.

The correlation between functional assay and binding assay is moderate, on average. Almost 80% of the compounds that were measured in both assay systems reveal a correlation. This is encouraging because a binding assay is usually more cost-effective and faster than a functional assay. However, ChEMBL does not provide any insight on which probe molecule was used for a particular binding assay. This would be very valuable because this would enable a more detailed analysis of the correlation between the different assay types. It would also allow to compare different binding assays with each other.

The reality check of the urban legends that are present in drug design text books was challenged by the performed analyses. For only two out of the five ascribed rules, the descriptors underline the correctness of these rules. In the case of the three other rules, some hints are given that these rules could be true. The most striking finding is the fact that the introduction of acids was only very rarely reported. In addition, it was confounding that in the binding assay subset many of the hERG inactives also bear a basic center. This is not in line with the currently used drug design paradigms on how to remove a hERG liability.

When it comes to model building, the huge difference in terms of the data basis between the functional assay and the binding assay becomes apparent. In addition, the data sets are rather imbalanced, which is a challenging scenario for machine learning algorithms. Whereas the models based on binding assay data give reasonable results, the models based on functional assay show a lower prediction quality. The explanation is that the functional assay contains a too small data basis. However, this will improve over the years when more and more functional data becomes published. Given the provided source code (see Supporting Information and browse to https://github.com/pzc/herg_chembl_jcim), everybody will then be enabled to train new models on this new data.

ASSOCIATED CONTENT

Supporting Information

Data as mentioned in the text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

E-mail: paul.czodrowski@merckgroup.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

To my colleagues at Merck KGaA, I thank Gerhard Barnickel and especially Daniel Kuhn for their proof reading and fruitful discussions. Roger Draheim (Goethe University, Frankfurt) is highly appreciated for his text modifications with respect to semantics. Christian Kramer (Innsbruck University) is acknowledged for the discussions on machine learning and ChEMBL. For his comments in an early phase of the manuscript, I thank George Papadatos (EBI) for his comments on ChEMBL and

model building. The great coding skills by Andrew Dalke (Dalke Scientific) are appreciated because he extended his original MCSS RDKit implementation based on my request. For the artwork, Steffen Schellenberger is acknowledged.

REFERENCES

- (1) Pearlstein, R. A.; Vaz, R. J.; Kang, J.; Preobrazhenskaya, M.; E. S., A.; Korolev, A. M.; Lysenkova, L. N.; Miroshnikova, O. V.; Hendrix, J.; Rampe, D. Characterization of HERG potassium channel inhibition using CoMSIA 3D QSAR and homology modeling approaches. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1829–1835.
- (2) De Ponti, F.; Poluzzi, E.; M., N. Organising evidence on QT prolongation and occurrence of Torsades de Pointes with non-antiarrhythmic drugs: A call for consensus. *Eur. J. Clin. Pharmacol* **2001**, *57*, 185–209.
- (3) Roy, M. L.; Dumaine, R.; M. B., A. HERG, A primary human ventricular target of the non-sedating antihistamine terfenadine. *Circulation* **1996**, *94*, 817–823.
- (4) Suessbrich, H.; Waldegger, S.; Lang, F.; E. B., A. Blockade of HERG channels expressed in *Xenopus oocytes* by the histamine receptor antagonists terfenadine and astemizole. *FEBS Lett.* **1996**, *385*, 77–80.
- (5) Zhou, Z.; Vorperian, V. R.; Gong, Q.; Zhang, S.; T. J., C. Block of HERG potassium channels by the antihistamine astemizole and its metabolites desmethyastemizole and norastemizole. *J. Cardiovasc. Electrophysiol.* **1996**, *10*, 836–843.
- (6) Roy, M. L.; Dennis, A.; M. B., A. A mechanism for the proarrhythmic effects of cisapride (Propulsid): High affinity blockade of the human cardiac potassium channel HERG. *FEBS Lett.* **1997**, *417*, 28–32.
- (7) Shen, M.-Y.; Su, B.-H.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. A comprehensive SVM binary hERG classification model based on extensive but biased endpoint hERG data sets. *Chem. Res. Toxicol.* **2011**, *24*, 934–49.
- (8) Su, B.; Tu, Y.; Esposito, E. X.; Tseng, Y. J. Predictive toxicology modeling: protocols for exploring hERG classification and Tetrahymena pyriformis end point predictions. *J. Chem. Inf. Model.* **2012**, *52*, 1660–73.
- (9) Su, B.-H.; Shen, M.-Y.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. In silico binary classification QSAR models based on 4D-fingerprints and MOE descriptors for prediction of hERG blockage. *J. Chem. Inf. Model.* **2010**, *50*, 1304–1318.
- (10) Polak, S.; Wiśniowska, B.; Ahmadi, M.; Mendyk, A. Prediction of the hERG potassium channel inhibition potential with use of artificial neural networks. *Appl. Soft Comput.* **2011**, *11*, 2611–2617.
- (11) Doddareddy, M. R.; Klaasse, E. C.; Shaguffa; Ijzerman, A. P.; Bender, A. Prospective validation of a comprehensive in silico hERG model and its applications to commercial compound and drug databases. *ChemMedChem* **2010**, *5*, 716–29.
- (12) Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Mol. Pharmaceutics* **2012**, *9*, 996–1010.
- (13) Marchese Robinson, R. L.; Glen, R. C.; Mitchell, J. B. O. Development and comparison of hERG blocker classifiers: Assessment on different datasets yields markedly different results. *Mol. Inf.* **2011**, *30*, 443–458.
- (14) Thai, K.-M.; Ecker, G. F. Similarity-based SIBAR descriptors for classification of chemically diverse hERG blockers. *Mol. Diversity* **2009**, *13*, 321–36.
- (15) Nisius, B.; Göller, A. H. Similarity-based classifier using topomers to provide a knowledge base for hERG channel inhibition. *J. Chem. Inf. Model.* **2009**, *49*, 247–256.
- (16) Li, Q.; Jørgensen, F. S.; Oprea, T.; Brunak, S.; Taboureau, O. hERG classification model based on a combination of support vector machine method and GRIND descriptors. *Mol. Pharmaceutics* **2008**, *5*, 117–27.
- (17) Keserü, G. M. Prediction of hERG potassium channel affinity by traditional and hologram qSAR methods. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2773–2775.
- (18) Tobita, M.; Nishikawa, T.; Nagashima, R. A discriminant model constructed by the support vector machine method for HERG potassium channel inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2886–90.
- (19) Song, M.; Clark, M. Development and evaluation of an in silico model for hERG binding. *J. Chem. Inf. Model.* **2006**, *46*, 392–400.
- (20) Leong, M. K. A novel approach using pharmacophore ensemble/support vector machine (PhE/SVM) for prediction of hERG liability. *Chem. Res. Toxicol.* **2007**, *20*, 217–26.
- (21) Jia, L.; Sun, H. Support vector machines classification of hERG liabilities based on atom types. *Bioorg. Med. Chem.* **2008**, *16*, 6252–60.
- (22) Gepp, M. M.; Hutter, M. C. Determination of hERG channel blockers using a decision tree. *Bioorg. Med. Chem.* **2006**, *14*, 5325–32.
- (23) Sun, H. An accurate and interpretable Bayesian classification model for prediction of HERG liability. *ChemMedChem* **2006**, *1*, 315–22.
- (24) Roche, O.; Trube, G.; Zuegge, J.; Pfimlin, P.; Alanine, A.; Schneider, G. A virtual screening method for prediction of the HERG potassium channel liability of compound libraries. *ChemBioChem* **2002**, *3*, 455–9.
- (25) Coi, A.; Massarelli, I.; Murgia, L.; Saraceno, M.; Calderone, V.; Bianucci, A. M. Prediction of hERG potassium channel affinity by the CODESSA approach. *Bioorg. Med. Chem.* **2006**, *14*, 3153–9.
- (26) Cianchetta, G.; Li, Y.; Kang, J.; Rampe, D.; Fravolini, A.; Cruciani, G.; Vaz, R. J. Predictive models for hERG potassium channel blockers. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3637–42.
- (27) Aronov, A. M.; Goldman, B. B. A model for identifying HERG K⁺ channel blockers. *Bioorg. Med. Chem. Lett.* **2004**, *12*, 2307–15.
- (28) Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadiramanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; Pritchard, J. M.; Cooper, A. W. J.; Macdonald, S. J. F. Lead optimization using matched molecular pairs: inclusion of contextual information for enhanced prediction of HERG inhibition, solubility, and lipophilicity. *J. Chem. Inf. Model.* **2010**, *50*, 1872–86.
- (29) Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. Lead optimization using matched molecular pairs: inclusion of contextual information for enhanced prediction of HERG inhibition, solubility, and lipophilicity. *J. Chem. Inf. Model.* **2010**, *50*, 1350–7.
- (30) Kramer, C.; Beck, B.; Kriegl, J. M.; Clark, T. A composite model for HERG blockage. *ChemMedChem* **2008**, *3*, 254–65.
- (31) Aronov, A. M. Common pharmacophores for uncharged human ether-a-go-go-related gene (hERG) blockers. *J. Med. Chem.* **2006**, *49*, 6917–21.
- (32) Zhang, S.; Du-Cuny, L.; Chen, L. A critical assessment of combined ligand- and structure-based approaches to HERG channel blocker modeling. *J. Chem. Inf. Model.* **2011**, *51*, 2948–2960.
- (33) Thai, K.-M.; Windisch, A.; Stork, D.; Weininger, A.; Schiesaro, A.; Guy, R. H.; Timin, E. N.; Hering, S.; Ecker, G. F. The HERG potassium channel and drug trapping: insight from docking studies with propafenone derivatives. *ChemMedChem* **2010**, *5*, 436–42.
- (34) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–7.
- (35) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The experimental uncertainty of heterogeneous public K(i) data. *J. Med. Chem.* **2012**, *55*, 5165–73.
- (36) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of mixed IC₅₀ data: A statistical analysis. *PLoS One* **2013**, *8*, e61007.
- (37) Zdrazil, B.; Pinto, M.; Vasanathan, P.; Williams, A. J.; Balderud, L. Z.; Engkvist, O.; Chichester, C.; Hersey, A.; Overington, J. P.; Ecker, G. F. Annotating human P-glycoprotein bioassay data. *Mol. Inf.* **2012**, *31*, 599–609.
- (38) RDKit, Open-Source Cheminformatics. <http://www.rdkit.org>.

- (39) Pedregosa, F.; Weiss, R.; Brucher, M. Scikit-learn: Machine learning in Python. *Int. J. Mach. Learn. Cybern.* **2011**, *12*, 2825–2830.
- (40) matplotlib. <http://matplotlib.org/>.
- (41) SciPy. <http://www.scipy.org/>.
- (42) KNIME. <http://www.knime.org/>.
- (43) Levoine, N.; Labeeuw, O.; Calmels, T.; Poupardin-Olivier, O.; Berrebi-Bertrand, I.; Lecomte, J.-M.; Schwartz, J.-C.; Capet, M. Novel and highly potent histamine H3 receptor ligands. Part 1: Withdrawing of hERG activity. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 5378–83.
- (44) Kerns, E.; Di, L. *Drug-Like Properties: Concepts, Structure Design and Methods: From ADME to Toxicity Optimization*; Elsevier Science: Amsterdam, The Netherlands, 2008.