RESEARCH ARTICLE

Construction of an integrated database for hERG blocking small molecules

Tomohiro Sato, Hitomi Yuki, Keiji Ogura, Teruki Honma*

Center for Life Science Technologies, RIKEN, Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, Japan

* honma.teruki@riken.jp

Abstract

The inhibition of the hERG potassium channel is closely related to the prolonged QT interval, and thus assessing this risk could greatly facilitate the development of therapeutic compounds and the withdrawal of hazardous marketed drugs. The recent increase in SAR information about hERG inhibitors in public databases has led to many successful applications of machine learning techniques to predict hERG inhibition. However, most of these reports constructed their prediction models based on only one SAR database because the differences in the data format and ontology hindered the integration of the databases. In this study, we curated the hERG-related data in ChEMBL, PubChem, GOSTAR, and hERGCentral, and integrated them into the largest database about hERG inhibition by small molecules. Assessment of structural diversity using Murcko frameworks revealed that the integrated database contains more than twice as many chemical scaffolds for hERG inhibitors than any of the individual databases, and covers 18.2% of the Murcko framework-based chemical space occupied by the compounds in ChEMBL. The database provides the most comprehensive information about hERG inhibitors and will be useful to design safer compounds for drug discovery. The database is freely available at http://drugdesign.riken.jp/hERGdb/.

Citation: Sato T, Yuki H, Ogura K, Honma T (2018) Construction of an integrated database for hERG blocking small molecules. PLoS ONE 13(7): e0199348. https://doi.org/10.1371/journal.pone.0199348

Editor: Chunhua Song, Pennsylvania State University, UNITED STATES

Received: March 5, 2018

Accepted: June 6, 2018

Published: July 6, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information file.

Funding: This work was supported by AMED Japan Agency for Medical Research and Development (https://www.amed.go.jp/en/); Program: Development of a Drug Discovery Informatics System (https://www.amed.go.jp/en/program/list/06/03/001_02-02.html); Recipient: Teruki Honma. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Introduction

Blockade of the human ether à-go-go related gene potassium channels is associated with drug-induced QT interval prolongation, which could cause arrhythmia and more severe heart failure [1–3]. The inhibition of hERG has become the major reason for drug withdrawals in the late 1990s, as represented by the withdrawals of terfenadine [4, 5], astemizole [6], and cisapride [7]. Thus, the assessment of the risk of hERG inhibition in the early stages of drug discovery, such as the screening and hit to lead stages, could effectively decrease the cost and failure risk of drug discovery.

Various computational methods to predict hERG inhibition were recently reported [8–23]. These studies included statistical models based on the 2D or 3D structures of small compounds, and structure-based approaches employing docking simulations using a modeled 3D structure of hERG. Although the electronic microscopy structure of hERG was recently solved [24], docking simulations of hERG are still a difficult challenge due to its high flexibility. However, the rapid growth of bioactivity information about hERG in various databases, and the improvement

Competing interests: The authors have declared that no competing interests exist.

of machine learning techniques have encouraged the use of statistical methods to predict hERG inhibition as summarized by Wang *et al.* [8] and Villoutreix *et al.* [9]. Among recent studies, some prediction models have been built using more than tens of thousands of compounds (for example, 306,895 compounds from hERGCentral database [10], 58,963 compounds from patented data [11]). Despite such studies using so-called big data, most of the previous studies to predict the hERG inhibitory activity are based on fewer than 5,000 compounds, which are often derived from a single database as reported by Villoutreix *et al.* [9]. One major obstacle that limited the amount of hERG-associated data was the differences in the format and ontologies between various databases, which hampered the use of multiple data source.

This study pursued the comprehensive and careful integration of the hERG-associated bioactivity information available in various databases. The hERG-associated data entries were derived from ChEMBL (https://www.ebi.ac.uk/chembl/) [25], GOSTAR (https://www.gostardb.com/index.jsp) [26], NIH Chemical Genomics Center data set registered in PubChem (https://pubchem.ncbi.nlm.nih.gov/bioassay/588834) [27], and hERGCentral (www.hergcentral.org (currently not working)) [28], and then merged by chemical structures after standardization. The procedure included unifications of the data format, value type, units, assay types, and chemical structures, based on the text analysis assisted by visual inspections by medicinal chemistry experts. Although a dataset built from the results of a single assay protocol can make the resulting values consistent and provide high quality data for quantitative analysis, it raises the risk of systematic bias, such as false positive observations in a patch-clamp assay by membrane damaging compounds. Since the integrated dataset contained heterogeneous data entries, the deviations of the hERG inhibitory activities due to the differences in the assay protocols were analyzed, to assess the influence of the deviations on the classification of the tested compounds into hERG inhibitors and non-inhibitors.

The development of high-throughput automated patch clamp assays has increased the amount of hERG-associated data available in public databases. Thus, the time-series increase of the hERG-associated entries was also investigated, using the constructed dataset. The transitions in the number of reported compounds, the number of chemical scaffolds, and the coverage of the chemical space compared to all known biologically active compounds in ChEMBL were assessed. The analysis provides useful insights to interpret the applicability of previously reported statistical models, using different databases. Ultimately, various physicochemical properties were calculated for hERG inhibitors and inactive compounds in the dataset. The results were used to assess the chemical nature of hERG inhibitors, and to determine whether the differences in the data sources affect the statistical analysis or the prediction models built from the data.

Materials and methods

Data set

The hERG inhibitory activity information was derived from ChEMBL, GOSTAR, NIH Chemical Genomics Center (NCGC) dataset in PubChem bioassay, and hERGCentral, and was integrated as briefly summarized in Fig 1. Procedures for data extraction, formatting activity information, standardization of chemical structures to merge data entries, filtering non-drug-like compounds, and classification of hERG inhibitors and inactive compounds, considering the deviation of experimental values are presented in this section. The data set was retrieved from the respective sources in April 2017.

ChEMBL. ChEMBL is the bioactivity database maintained by the European Bioinformatics Institute, and is frequently used in various cheminformatics researches as the *de facto* standard database. The entire ChEMBL database (version 22) was downloaded from the web site,

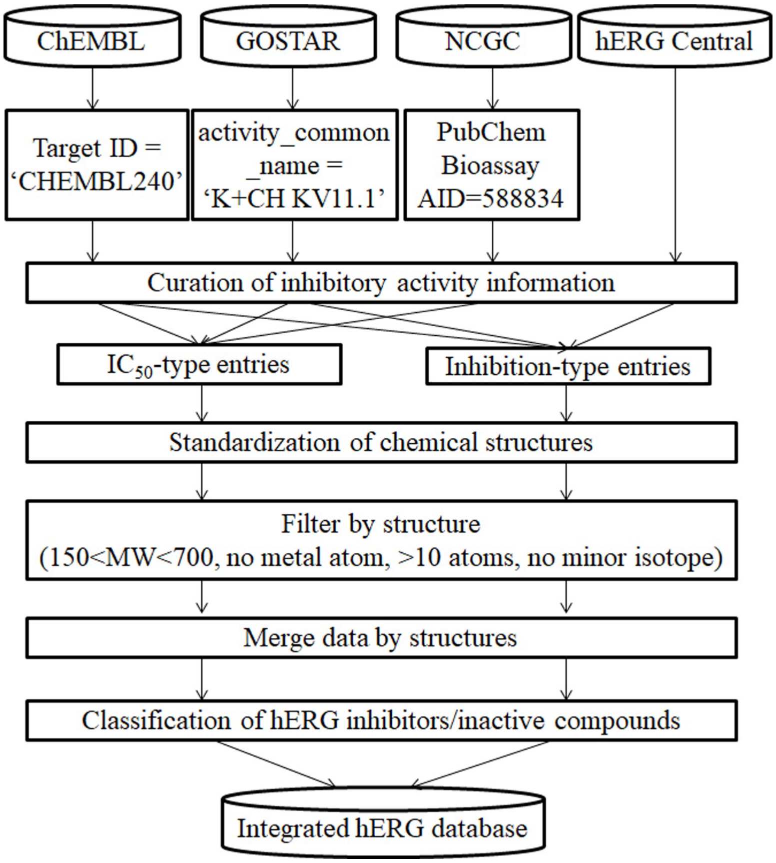

ChEMBL
GOSTAR
NCGC
hERG Central
Target ID = 'CHEMBL240'
activity_common _name = 'K+CH KV11.1'
PubChem Bioassay AID=588834
Curation of inhibitory activity information
IC50-type entries
Inhibition-type entries
Standardization of chemical structures
Filter by structure
(150<MW<700, no metal atom, >10 atoms, no minor isotope)
Merge data by structures
Classification of hERG inhibitors/inactive compounds
Integrated hERG database

Fig 1. Schematic procedure of the database integration.

https://doi.org/10.1371/journal.pone.0199348.g001

and 17,952 activity entries about 2,153 hERG-related bioassays were extracted, according to the Target ID (CHEMBL240) assigned to biological assays. To ensure the validity of the data, entries with data validity comments of "Nonstandard unit for type" and "Outside typical range", and entries tagged as potential duplicates were excluded. For example, the hERG-related assays reporting selectivity between inhibitors or raw values of tail currents, were removed by investigating their value types and assay descriptions. After manually checking assays reporting more than 20 compounds, several redundant entries that represented the same assay results were removed.

GOSTAR. GOSTAR is an online scientific database product of Excelra Knowledge Solutions, consisting of published and patented inhibitors against various biological targets and their associated SAR data. The hERG-associated data entries with activity common name of "K+CH KV11.1" were derived from the GOSTAR database. Although GOSTAR has four subset databases, Drug Database (DD) was excluded in this study, because it includes some critical curation errors during the ontology standardization. The subset will be integrated after the errors are corrected by the vendor. Excluding the 106 entries in the DD subset, GOSTAR contained 14,176 entries about hERG inhibition in the remaining three subsets. Since both ChEMBL and GOSTAR contained hERG activity entries derived from medicinal chemistry journals, the redundant entries with ChEMBL were identified by the reference field and omitted from the data set.

NIH Chemical Genomics Center. Quantitative high throughput screening data to determine in vitro hERG channel blockage by NCGC were derived from PubChem bioassays

(AID = 588834, v1.1). The data related to hERG, including about 2,688 compounds from the LOPAC1280 library (Sigma), the NTP collection and the NCGC Pharmaceutical Collection (NPC), were determined by FluxORTM thallium flux assays. The dataset contains EC_{50} values for both hERG inhibitors and activators along with some undefined data. As the EC_{50} values contained in the NCGC dataset were calculated from automated sigmoid curve-fitting to the dose responses of hERG activities by the Hill equation, the values had to be interpreted by the Hill coefficient (positive value for hERG inhibitor and negative value for activator) and the fitting quality (low correlation for inconclusive entries). In PubChem, the entries in the NCGC dataset were classified into "Inhibitor", "Activator", and "Inactive" in terms of the phenotype. For compound with low quality curve fitting, an "Inconclusive" description was included in the outcome comments. Although the EC_{50} values of the NCGC dataset were also included in the ChEMBL database, the information about the phenotype and outcome comments to interpret the assay results were omitted in ChEMBL. Thus, the corresponding entries were excluded from the ChEMBL dataset. Accordingly, the hERG inhibitors were defined as the entries specified as "Inhibitor" with sufficient inhibitory activity ($EC_{50} \leq 10 \mu M$ in this case) and without outcome comments of "Inconclusive". All compounds with EC_{50} values exceeding 10μM, were defined as negative compounds in this study.

hERGCentral. hERGCentral [28] is a database containing the hERG activity information of more than 300,000 compounds. Since the hERGCentral database (www.hergcentral.org) is currently out of order, the values of the percent inhibitory activities of 318,496 compounds at a 10 μM concentration determined by IonWorks Quattro (MDC, Sunnyvale, CA) in the population patch clamp (PPC) mode were retrieved from the supporting information of a manuscript published by Fang *et al.* [10], describing a statistical analysis of the hERGCentral dataset.

Formatting activity information

To integrate the hERG activity information collected from the databases, the derived entries were formatted as follows. At first, the entries were classified as either constant concentration values describing inhibitory activity, such as IC_{50}, EC_{50}, ED_{50}, K_i, K_d and percent inhibitory activity at a certain concentration. For convenience, the former data were categorized as IC_{50}-type, and the latter were categorized as inhibition-type in this study.

The IC_{50}, K_d, K_i, EC_{50}, and their log-unit values, such as pIC_{50}, were obtained from the ChEMBL, GOSTAR, and NCGC data sets for IC_{50}-type entries. All values were converted to nM order from their various units. In terms of inhibition-type entries in ChEMBL, GOSTAR, and hERGCentral, entries reported as remaining enzyme activity were converted to inhibition percentage format. For entries in which the assay concentration was not specified, assay descriptions were scanned to complement the compound concentration if possible. Other values, such as concentrations using thresholds other than 50% inhibition (IC_{70}, IC_{30}, etc.), raw values of measured current, prolonged QT interval ratios, etc., were discarded in this study, because they cannot be directly compared to other entries. Several value types can be used for both hERG activation and inhibitory activities, and thus the assay descriptions and/or reference manuscripts were manually checked for "Kd", "EC_{50}", "ED_{50}", and "activity" entries to remove the assay entries describing hERG activation.

The assay protocols of hERG blocking activities could roughly classified into electrostatic assays such as automated patch clamp assays that measure the change in the voltage between the cell-membrane by the presence of small molecules, and binding assays, such as radioligand replacement assays that measure the binding affinity of small molecules by the replacement ratio of radiolabeled inhibitors. The IC_{50} values and inhibitory percentages generally depend on the assay protocols, and the differences in the hERG assay protocols mentioned above could potentially affect the results, especially in quantitative analyses such as regression models of binding affinity. Using compounds with IC_{50} values determined by both binding assays and electrostatic assays, the correlation between the two methods was investigated. The deviation of the IC_{50} values of each compound was investigated for hERG inhibitors for which more than three IC_{50} values were reported to assess how the deviation of the heterogeneous data affected the characterization of the corresponding compound.

Standardization of chemical structures

To merge the databases, the chemical structures were standardized for comparison. The structures were obtained from sd files provided by ChEMBL and PubChem (NCGC and hERGCentral) or constructed from the SMILES string (GOSTAR). The obtained structures were desalted, standardized to a neutral ionization state, and then converted to the canonical tautomers using the Pipeline Pilot program. To exclude non-drug like compounds, metal containing molecules, molecules with molecular weights less than 150 or more than 700, molecules with fewer than 10 atoms, and molecules with minor isotopes were removed from the data set. After this structural filtering, the data size of each database was reduced to 14,991 activity entries for 11,993 compounds in ChEMBL, 13,861 activity entries for 8,338 compounds in GOSTAR, 1,894 activity entries for 1,733 compounds in NCGC, and 305,928 activity entries for 303,351 compounds in hERGCentral. These activity entries from the 4 databases were finally merged by the standardized structures.

Classification of hERG inhibitors and inactive compounds

Using the integrated database, the distributions of the physicochemical properties and the structural diversity of the hERG inhibitors (positive compounds) and the inactive chemicals (negative compounds) were investigated. In this study, positive compounds were defined as hERG inhibitors showing $IC_{50} \leq 10\mu M$ or $\geq 50\%$ inhibition at 10μM. Since some compounds had multiple assay results, contradictory results due to experimental errors or differences in assay methods were found. As the determination of the IC_{50} value requires measurements of inhibitory activities at multiple concentrations, the IC_{50}-type information was considered to be more reliable, and was given higher priority over the inhibition-type entries for positive/negative classification. When the deviation of the assay results was still significantly large among the IC_{50}-type entries or inhibition-type entries, the omission of outlier values and majority vote-based procedures was performed to classify hERG inhibitors and inactive compounds as follows.

At first, the compounds for which all assay results unanimously indicated either the existence or absence of sufficient inhibitory activities were assigned to hERG inhibitors or inactive compounds. When contradictory assay values were found, the assay entries that were at least 10-fold higher or lower than the mean value of the compound were defined as outliers, and were removed. Subsequently, a compound was classified as either a positive (hERG inhibitor) or negative (inactive compound), when more than 2/3 of the assay entries of the compound agreed in either category. Qualitative assay results, such as comments describing "no activity" were also included in the voting. Compounds for which comparable numbers of entries indicated opposite results were discarded as inconclusive compounds from the comparison between hERG inhibitors and inactive compounds. The classification procedure was first performed using IC_{50}-type entries, and then inhibition-type entries were considered for the remaining unclassified compounds. An investigation of the distribution of the percent inhibition data in hERGCentral revealed that the single concentration results from high-throughput screening (HTS) assays left the low confidence region around the threshold value (50% inhibition at 10μM in this study), as compared to the assay values published in peer-reviewed journals, resulting in some contrary results to IC_{50} values reported in other databases (data not shown). Thus, a more strict threshold (>70% inhibition at 10μM for hERG inhibitors and <30% inhibition at 10μM for inactive compounds) was applied to the inhibition type results from the HTS results in which more than 200 compounds were assayed in a single data source.

Assessment of structural diversity

To evaluate the usefulness of the integrated database, we assessed the structural diversity of the hERG inhibitors and inactive compounds registered in the database. As the metrics for the structural diversity of the compounds, the number of Murcko framework [29] was counted as reported by Langdon *et al.* [30] and Karawajczy *et al.* [31]. The Murcko framework of a compound was defined as the union of rings and linker atoms connecting them. After the decomposition of all compounds in the dataset as Murcko frameworks, the unique numbers of the derived frameworks were counted for both the hERG inhibitors and inactive compounds. The number of Murcko frameworks was then compared to those generated from the whole ChEMBL database. Since ChEMBL is a bioactivity database for various target proteins, their structural diversity could approximate the chemical space covered by all currently available bioactive compounds.

Physicochemical properties

Twelve physicochemical properties were computed using Pipeline Pilot, including molecular weight (MW), Ghose-Crippen-Viswanadhan neutral form's octanol-water partition coefficient (AlogP), octanol-water distribution coefficient with all forms (logD), number of hydrogen bond acceptors (HBA), number of hydrogen bond donors (HBD), number of positively charged atoms when ionized at pH7.4 (N_Cations), number of negatively charged atoms when ionized at pH7.4(N_Anions), molecular surface area (MSA), molecular polar surface area (MPSA), number of rotatable bonds (N_Rot), pKa value of most basic atom (pKa_base), and pKa value of most acidic atom (pKa_acid). This descriptor set based on physicochemical properties is widely used to compare compound data sets in drug discovery [32, 33] and some descriptors that seemed to affect hERG binding, such as pKa, were added according to advice from medicinal chemistry experts. The distributions of the physicochemical properties for both hERG inhibitors and inactive compounds were calculated and compared to assess the characteristic features of hERG inhibitors.

Results and discussion

The number of assay records for each compound

Literature-based bioactivity databases, such as ChEMBL, often contain multiple assay entries for a compound. In the integrated dataset, 329,243 assay records for 319,631 compounds were registered. Using the dataset, the assay records that specified certain values (not using NULL value, '>', or '<') for each compound were counted. The distribution of the number of assay records for each compound is shown in Fig 2. The most frequently reported compounds were cisapride (61 assay records), terfenadine (60), dofetilide (54), E-4031 (52), and astemizole (42). While 81 compounds had more than 10 assay records, only one assay record was found for 310,180 compounds mainly reported in the HTS results in the hERGCentral database, resulting in the low average frequency of 1.02 for a compound.

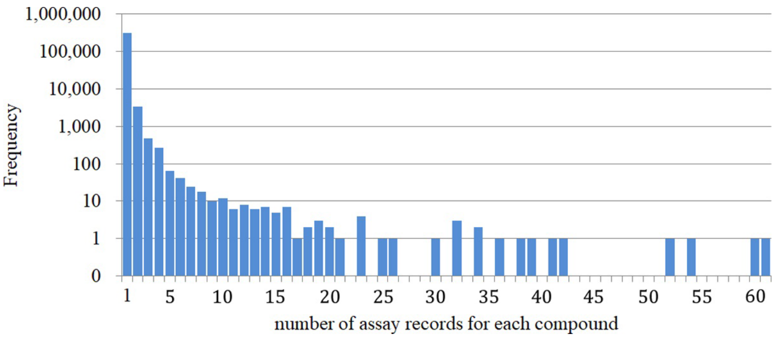


Fig 2. Histogram showing the number of assay records for each compound. The vertical axis representing the frequency is shown in logarithmic scale.

https://doi.org/10.1371/journal.pone.0199348.g002

Comparison between binding assays and electro static assays

To assess the difference between binding assays such as radio ligand replacement assay and electrostatic assays such as automated patch-clamp assays, the IC_{50} values determined by both methods were investigated using the integrated dataset. In the dataset, 15,932 IC_{50} values were reported for 11,956 compounds. To ensure the validity of the IC_{50} values, 6,594 data entries using inequality signs, NULL values, and improper value range (IC_{50}>1mM) were excluded. The remaining dataset contained 4,173 IC_{50} values for 3,449 compounds measured by binding assays, and 3,082 IC_{50} values for 2,246 compounds measured by electrostatic assays. To compare the two methods, 209 compounds for which the IC_{50} values were measured by both methods were investigated. For each compound, the mean pIC_{50} values were respectively calculated for both binding assays and electrostatic assays, and plotted in Fig 3. The averages of the mean IC_{50} values were 2.41μM in binding assays and 1.64μM in electrostatic assays. The coefficient of determination and the root mean square deviation between the pIC_{50} values measured by binding assays and electrostatic assays were 0.517 and 0.737, respectively. While the IC_{50} values determined by both methods showed moderate correlation, slightly higher potencies tended to be observed by electrostatic assays. Among the 209 compounds, 119 compounds showed higher potencies in electrostatic assays. In contrast to binding assays, which directly measure the binding affinity of a compound to the pore region of hERG, electrostatic assays, which measure the decrease of tail-current, could be affected by various additional factors, including non-specific binding to other ion channels, binding to a different region of hERG, or membrane toxicity, possibly resulting in the slightly higher potency observed in the dataset.

Deviation of IC_{50} values and classification of hERG inhibitors and inactive compounds

In some case, the resulting values had large deviations due to differences in the assay protocols or curation errors such as misinterpretation of units in the manuscript, which could confuse the classification of positive (hERG inhibitors) and negative (hERG inactive) compounds for the statistical analysis. To exemplify the deviation of the hERG inhibitory activities registered in the currently available databases, the distribution of the IC_{50} values of 263 compounds, for which more than three IC_{50} values were reported, is shown in Fig 4. To ensure the validity of the IC_{50} values and exclude extreme values due to curation errors, the IC_{50} values lower than 1pM or higher than 1mM were excluded, because such values were often caused by incorrect registration of the units or misinterpretation of the digits of assay values in the curation procedure.

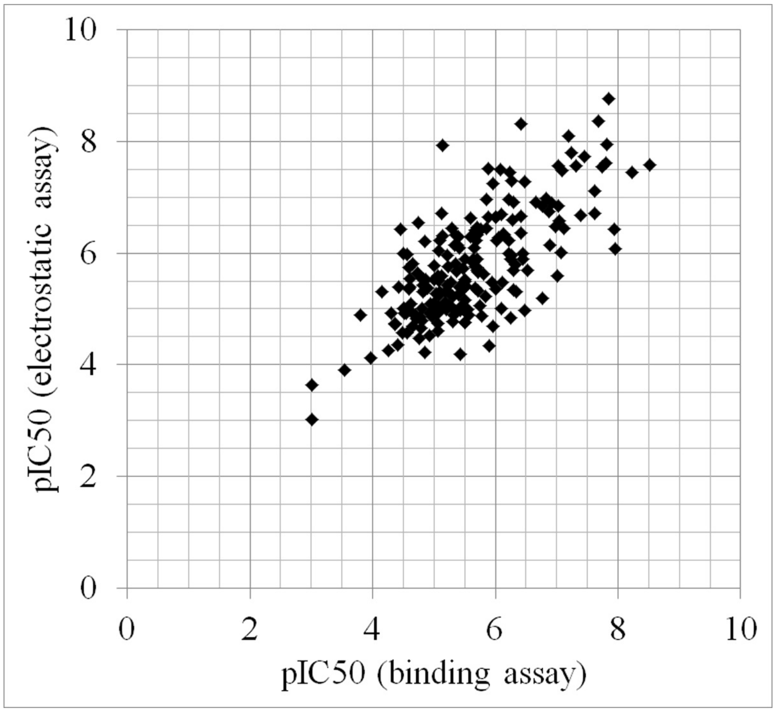


Fig 3. Comparison between mean pIC_{50} values of 209 compounds measured by binding assays and electrostatic assays.

https://doi.org/10.1371/journal.pone.0199348.g003

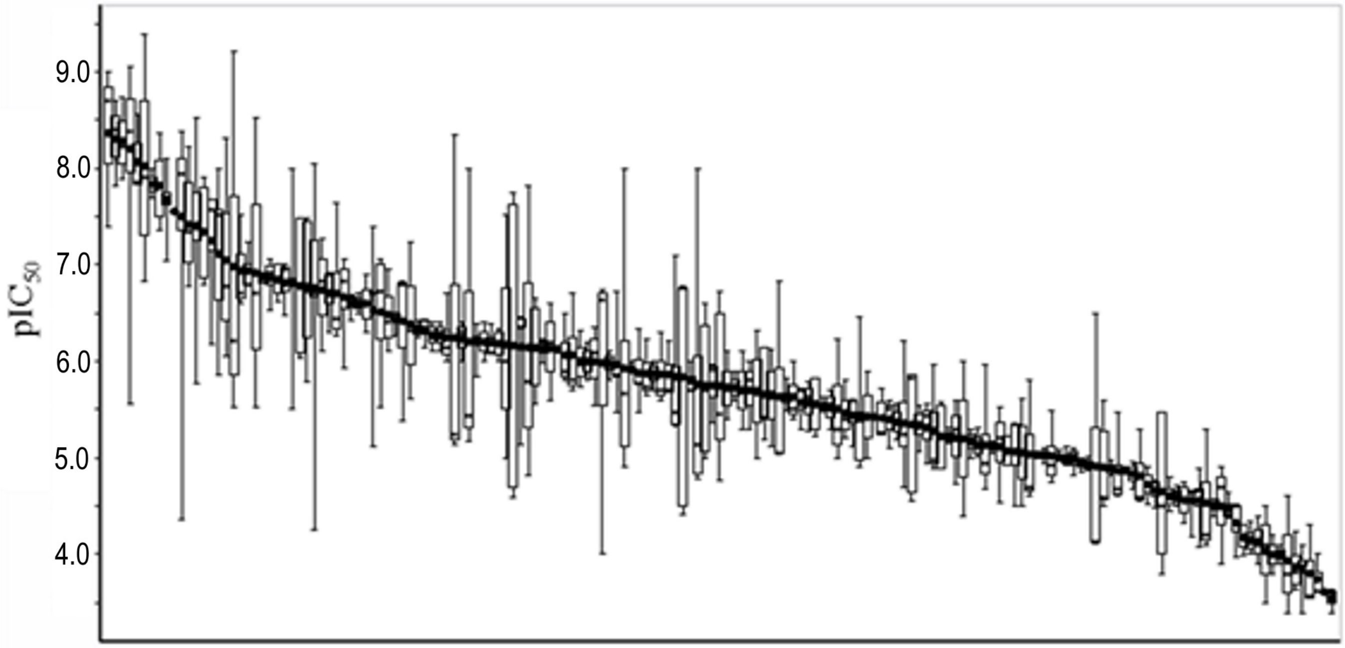


Fig 4. Box plot showing the distribution of IC_{50} values of 263 compounds with more than three reported IC_{50} values. The compounds were sorted by mean values.

https://doi.org/10.1371/journal.pone.0199348.g004

Among the 263 compounds, 144 compounds showed consistent IC_{50} values with less than one order of magnitude differences between the maximum and minimum results. However, 47 compounds recorded more than 100-fold differences between the maximum and minimum IC_{50} values. An inspection of such cases revealed that only a small number of outlier values often caused such large deviations. Therefore, the differences between the 25th percentile and 75th percentile of IC_{50} values were far less than those between the maximum and minimum values. Only two compounds had more than a 100-fold difference between the 25th percentile and 75th percentile, and those of 224 out of the 263 compounds were lower than 10-fold. These observations emphasize the importance of outlier elimination, to construct a robust dataset and avoid incorrect classification of positive and negative compounds. The integration of various databases increases the number of assay results and enables the detection of more outlier values, as compared to relying on a single database.

In this study, compounds were classified into either positive (hERG inhibitors showing $IC_{50} \leq 10\mu M$ or $\geq 50\%$ inhibition at 10μM) or negative compounds ($IC_{50} > 10\mu M$ or $< 50\%$ inhibition at 10μM). Fig 5 presents the step-wise proceeding of the classification procedure employed to minimize the influence of outlier values and achieve robust classification through unanimous agreement (Fig 5(A)),removal of outlier values (Fig 5(B)), and majority vote using criteria of two-thirds (Fig 5(C)). Considering the large deviation of the IC_{50} values, majority voting was applied to assign the labels rather than using the mean values which could be more sensitive to outlier values. The compounds without IC_{50} information were classified based on

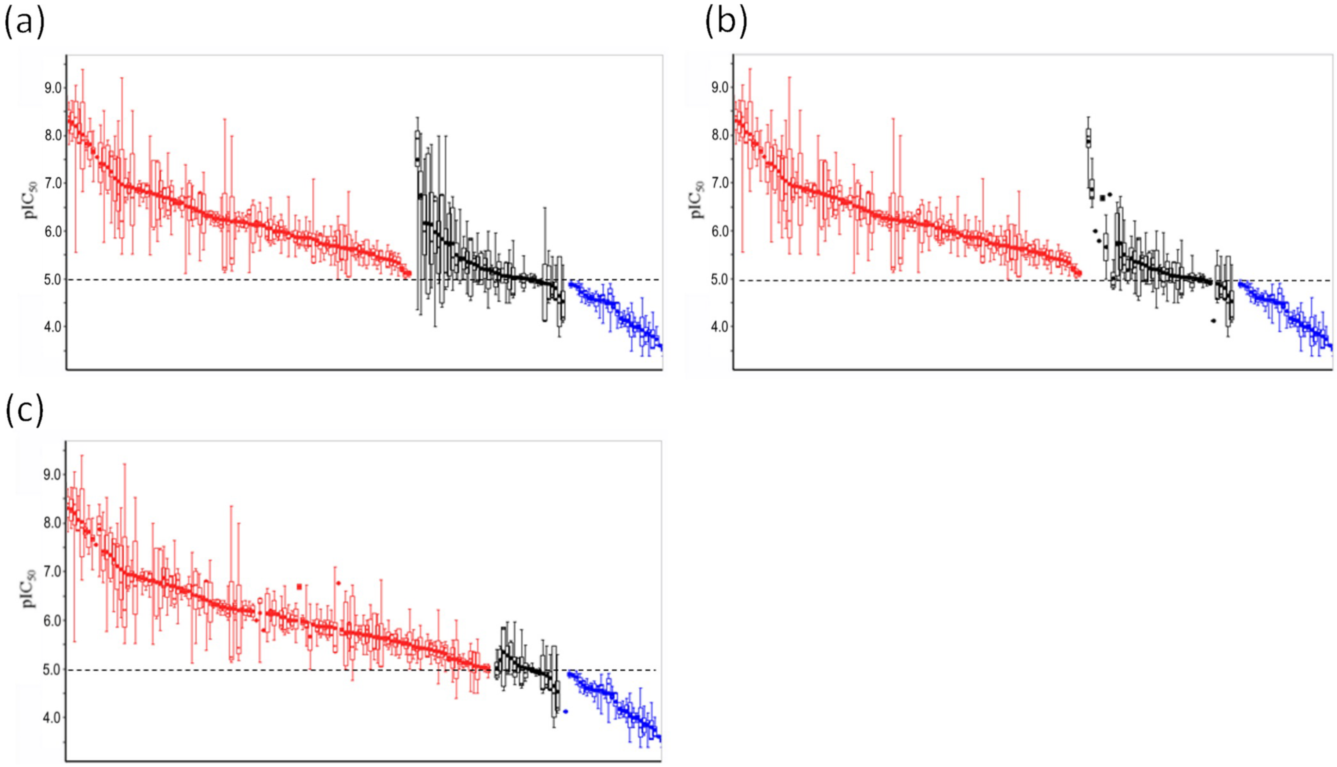


Fig 5. Step-wise procedure for compound classification. (a) Classification by unanimous results using all assay values. Positive, unclassified, and negative compounds are shown in red, black, and blue, respectively. The dashed line indicates the threshold value between positive and negative compounds (IC_{50} = 10μM) (b) Distribution of assay values after removal of outlier values differing more than 10-fold from the mean IC_{50} values from undefined compounds. (c) Classification by a two-thirds majority of filtered IC_{50} values. The remaining unclassified compounds with nearly equal numbers of positive and negative assay results were considered as inconclusive compounds, and omitted from the subsequent analysis of physicochemical properties.

https://doi.org/10.1371/journal.pone.0199348.g005

the percentage inhibition data. As in the case of the IC_{50} data, assay results showing more than 50% inhibition at a concentration lower than 10μM, and less than 50% inhibition at a concentration higher than 10μM, were counted for each compound, and then assessed according to whether more than two thirds of the data showed consistent results. When both the IC_{50} and percentage inhibition data were available, only the IC_{50} information was considered for the classification because the determination of an IC_{50} value generally requires assays with about 7 different concentrations for sigmoid curve fitting, and thus it could be more reliable than percentage inhibition data measured at a single concentration.

Structural diversity of the integrated database

Using the standardized structures, the hERG activity entries from ChEMBL, GOSTAR, NCGC, and hERGCentral were merged and classified into hERG inhibitors and inactive compounds according to the aforementioned criteria. Table 1 presents the numbers of compounds in each database and the resulting integrated database, after the removal of inconclusive compounds. The ratios of inhibitors and inactive compounds were nearly 1:1 in ChEMBL and GOSTAR, while they were about 1:6 in NCGC and 1:64 in hERGCentral, respectively. ChEMBL and GOSTAR are literature-based databases from medicinal chemistry journals and patents, and thus they often contain structure activity relationships of compounds sharing certain scaffolds, resulting in a high ratio of hERG inhibitors. The NCGC data set consisted of a pharmacologically active compound library from SIGMA (LOPAC1280 library) and NCGC (NPC). hERGCentral uses the National Institutes of Health (NIH) Molecular Library Small Molecule Repository (MLSMR) as a compound source. Since hERGCentral uses general small molecule library, which is irrelevant to hERG inhibition or other bioactivities, its inhibitor ratio would most likely approximate those in practical HTS in drug discovery project. By merging the compounds from the four databases, a database consisting of 9,890 hERG inhibitors and 281,329 inactive compounds was constructed. The size of the integrated dataset exceeded those of most previous studies. As reported by Villoutreix *et al.* [9], most of the previous studies used datasets consisting of less than 5,000 compounds from public databases, or used in house datasets. The only exception was the study employing 306,895 compounds from hERGCentral dataset by Du *et al.* [10]. The slight decrease of the integrated dataset from original hERGCentral mainly came from the

Table 1. The number of compounds and their Murcko frameworks in each database.

Database	Class	Number of compounds	Number of Murcko frameworks
ChEMBL	Inhibitors	4,793	2,474
	Inactives	5,275	3,012
	All	10,068	4,954
GOSTAR	Inhibitors	3,260	1,727
	Inactives	3,509	1,692
	All	6,769	3,098
NCGC	Inhibitors	232	173
	Inactives	1,234	504
	All	1,466	639
hERGCentral	Inhibitors	4,321	2,708
	Inactives	274,536	73,419
	All	278,857	74,687
Integrated database	Inhibitors	9,890	5,516
	Inactives	281,329	76,420
	All	291,219	79,806

https://doi.org/10.1371/journal.pone.0199348.t001

removal of the near 50% inhibition entries in HTS assays in hERGCentral. Because the strict criteria (>70% inhibition at 10uM for HTS assays) was set to remove low confidence entries, the integrated dataset was expected to achieve both the high reliability of assay results and the comprehensiveness of structural diversity.

The numbers of Murcko frameworks for hERG inhibitors and inactive compounds in each database are shown in Table 1. For hERG inhibitors, the integrated database contains more than twice as many scaffolds than any of the individual databases. Conversely, the structural diversity of the inactive compounds in the integrated database was almost equivalent to that of the hERGCentral dataset. This result revealed the unique nature of the hERGCentral dataset, which contains comprehensive HTS results against large compound libraries covering a broader chemical space than manuscripts or patents focused on specific compounds or chemical series. In total, the compounds in the integrated hERG database contained 79,806 Murcko frameworks. By estimating that the 438,551 Murcko frameworks found in the whole ChEMBL database represent the structural diversity of the all existing bioactive compounds, the integrated database successfully covered 18.2% of that chemical space.

The transition of the number of compounds/scaffolds with hERG activities registered in the databases is shown in Fig 6. Large increases in both hERG inhibitors and inactive compounds were found in 2011, according to the hERGCentral release. Apart from this event, about 200–700 hERG inhibitors and 200–1,000 inactive compounds have been reported each year since 2006. Considering that only 618 hERG inhibitors and 279 inactive compounds were reported in 2006, the recent increase in the hERG-related bioactivity data (covering 18.2% of the Murcko frameworks found in ChEMBL) enables us to construct accurate and robust hERG prediction models for diverse drug-like compounds including newly synthesized ones. The full information about the integrated database at various time points is provided in S1 Table.

SAR information of each chemical scaffold

To assess the SAR information of each chemical scaffold contained in the integrated database, the distribution of IC_{50} values among the compounds sharing the same Murcko frameworks was investigated. In the integrated dataset, the IC_{50} values were reported for 3,361 Murcko frameworks. The number of compounds with IC50 entries in each Murcko frameworks is shown in Fig 7. The distribution of IC_{50} values for 82 Murcko frameworks containing more

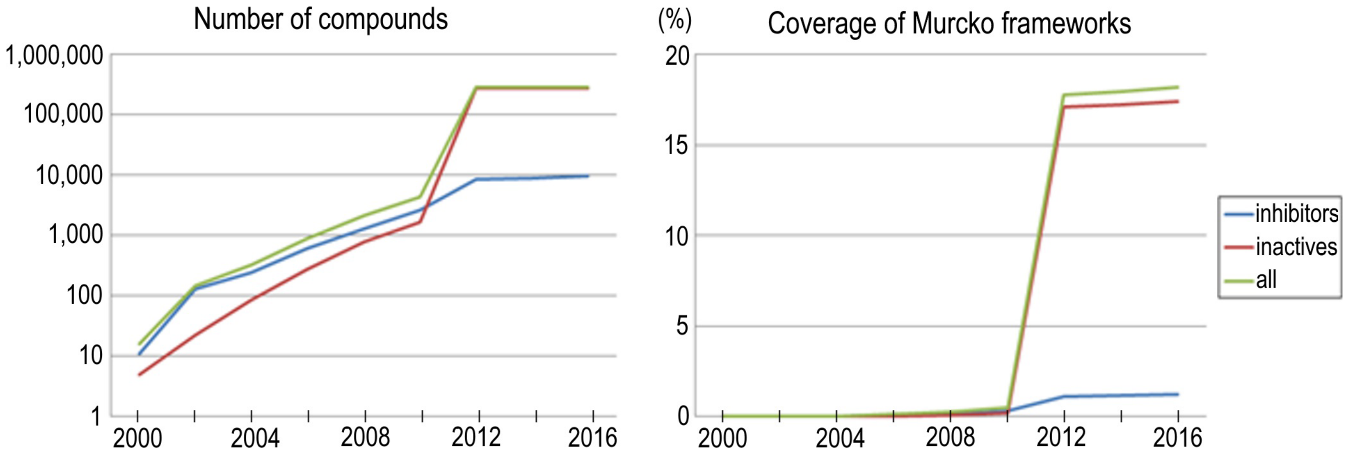


Fig 6. Transition in the number of unique compounds and coverage of Murcko frameworks for hERG inhibitors, inactive compounds, and all reported compounds. Coverage of Murcko frameworks was calculated as the ratio to those of all ChEMBL22 compounds (438,551).

https://doi.org/10.1371/journal.pone.0199348.g006

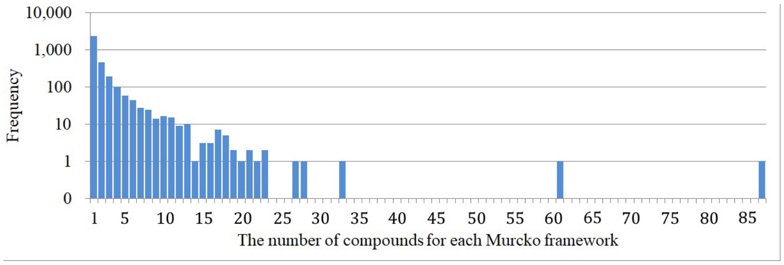


Fig 7. Histogram of the number of compounds in each Murcko frameworks.

https://doi.org/10.1371/journal.pone.0199348.g007

than 10 compounds is presented as a box plot sorted by the IC_{50} value of the most potent compound in each group in Fig 8(A). The mean pIC_{50} value was employed when multiple IC_{50} values were reported for a compound. Six scaffolds showed more than 10,000-fold differences between the IC_{50} values of the most and least potent compounds (Fig 9). All six Murcko frameworks met the requirement of common hERG pharmacophores for charged compounds, consisting of a positively charged atom and two aromatic rings reported in previous studies [34]. Thus, the compounds with various IC_{50} values could provide useful information about the substituents conferring substantial interaction energy with hERG and those decreasing the binding affinity, to design molecules with desirable properties. Such SAR information has also been hugely improved by the recent increase in hERG-related data entries. The IC_{50}

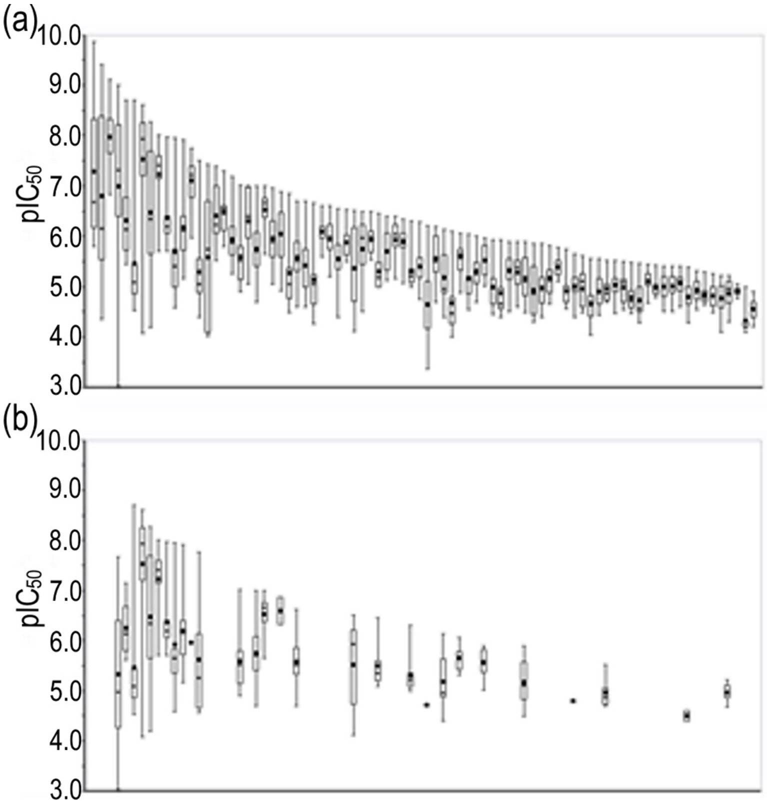
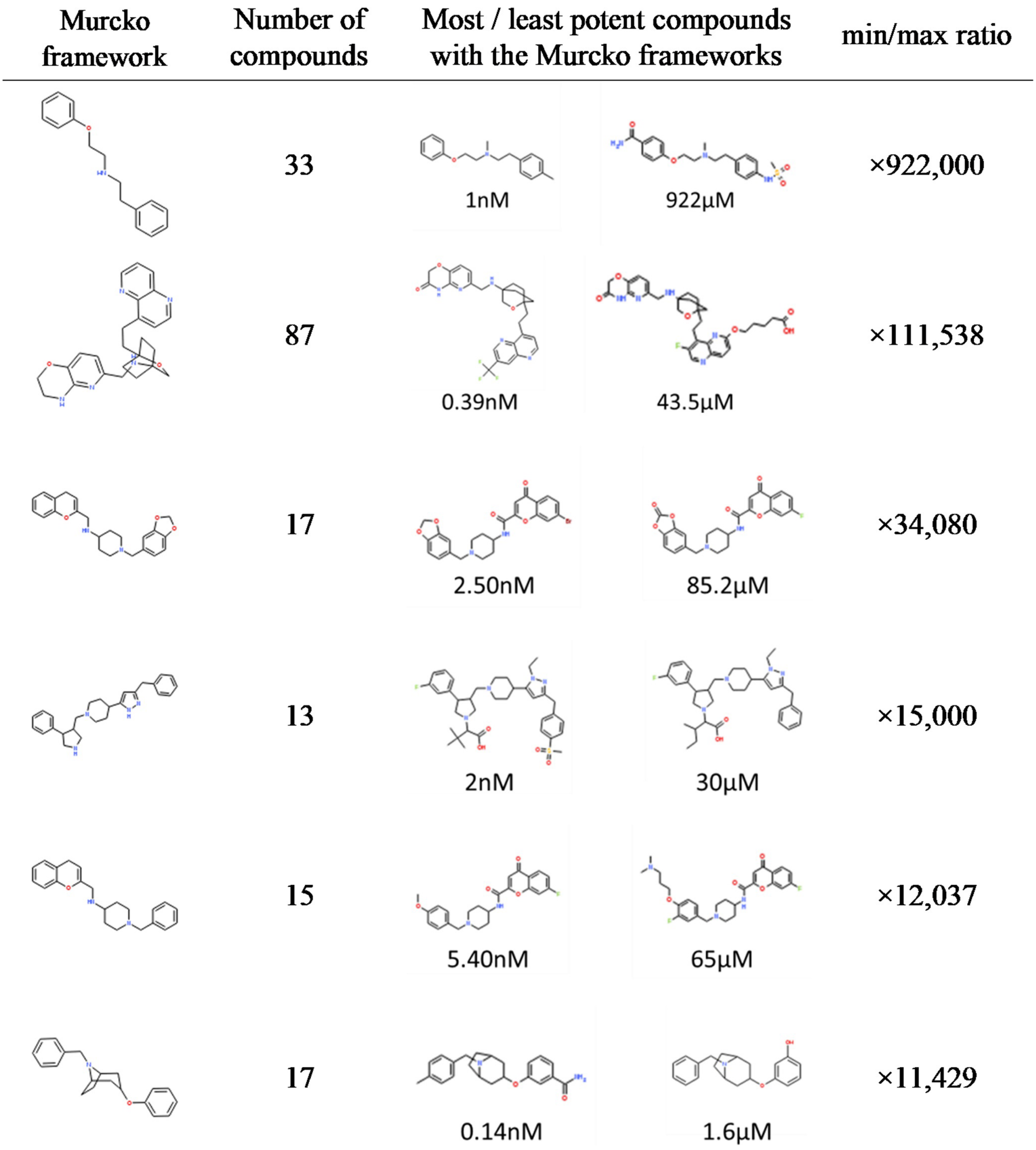


Fig 8. Box plot of IC_{50} distribution about each Murcko framework. (a) IC_{50} values for 82 Murcko frameworks containing more than 10 compounds sorted by the IC_{50} values of the most potent inhibitors. (b) Corresponding IC_{50} values reported before 2009. The horizontal axes of both plots represent each of the 82 Murcko frameworks.

https://doi.org/10.1371/journal.pone.0199348.g008



Murcko framework	Number of compounds	Most / least potent compounds with the Murcko frameworks		min/max ratio
	33	1nM	922μM	×922,000
	87	0.39nM	43.5μM	×111,538
	17	2.50nM	85.2μM	×34,080
	13	2nM	30μM	×15,000
	15	5.40nM	65μM	×12,037
	17	0.14nM	1.6μM	×11,429

Fig 9. Six Murcko frameworks showing more than 10,000-fold potency differences.

https://doi.org/10.1371/journal.pone.0199348.g009

distributions of the same chemical scaffolds, reported before 2009, are shown in Fig 8(B). Among the 82 chemical scaffolds, only 28 scaffolds were associated with hERG inhibition before 2009, and several scaffolds lacked highly potent compounds at the time, which could lead to severe bias in the risk assessment of the scaffolds.

Physicochemical properties

To assess the contribution of the physicochemical properties to the hERG inhibitory activity, the distributions of 12 physicochemical properties for hERG inhibitors, inactive compounds, and all compounds in the integrated database were compared. The mean and standard deviation of 12 physicochemical properties for both hERG inhibitors and inactive compounds in each database are shown in Table 2. Among the 12 physicochemical properties, the distributions of MW, AlogP, N_Cations, MPSA, N_Rot, and pKa_base, which show significant differences between hERG inhibitors and inactive compounds, are highlighted in Fig 10. The results indicated that hERG inhibitors tend to have a larger molecular weight (418.7, as compared to 355.9 for inactive compounds), higher hydrophobicity according to AlogP and logD (3.80 and 3.22, as compared to 2.77 and 2.58), have more cations (0.68, as compared to 0.22), and less basic substituents (pKa value of most basic substituents was 8.18 as compared to 6.33). While about 80% of the inactive compounds had no positively charged atoms, more than half of the hERG inhibitors contained at least one positively charged atom. These differences were clearly consistent with the previously reported pharmacophores [34] and the expected binding modes of the known hERG inhibitors, because Tyr652 and Phe656 were identified as the important residues frequently forming cation-π and π-π stacking interactions with various hERG inhibitors according to site-directed mutagenesis analyses [35, 36].

Although the aforementioned general trends between hERG inhibitors and inactive compounds were commonly observed in each database, slightly different property distributions were found across the databases (S1 Table). ChEMBL and GOSTAR, which are based on scientific journals and patents, often contain hERG inhibitors and structurally similar derivatives synthesized from an initial hit compound. The mean MWs were 438.7 for hERG inhibitors and 417.0 for inactive compounds in ChEMBL, and 432.7 for hERG inhibitors and 417.8 for inactive compounds in GOSTAR. In contrast, the NCGC dataset, consisting of the LOPAC1280 library, contained much smaller and more hydrophilic compounds as compared to ChEMBL and GOSTAR. The mean MWs were 345.1 for inhibitors and 273.1 for inactives; and for AlogP they were 3.75 for inhibitors

Table 2. Mean values (standard deviations) of 12 physicochemical properties for hERG inhibitors, inactive compounds, and all compounds in the integrated database.

Property	Inhibitors	Inactives	All
MW	418.7(77.1)	355.9(76.6)	358.0(77.5)
AlogP	3.80(1.39)	2.77(1.34)	2.80(1.35)
logD	3.22(1.47)	2.58(1.41)	2.60(1.42)
HBA	4.47(1.67)	4.41(1.64)	4.41(1.64)
HBD	1.07(0.94)	1.25(0.91)	1.24(0.91)
N_Cations	0.68(0.63)	0.22(0.46)	0.23(0.47)
N_Anions	0.03(0.17)	0.11(0.33)	0.11(0.33)
MSA	424.7(75.7)	356.8(75.0)	359.1(76.0)
MPSA	71.0(30.6)	86.1(31.6)	85.6(31.7)
N_Rot	6.24(2.43)	5.22(2.32)	5.25(2.33)
pKa_base	8.18(2.74)	6.33(3.31)	6.41(3.31)
pKa_acid	3.33(4.15)	2.67(3.71)	2.69(3.73)

https://doi.org/10.1371/journal.pone.0199348.t002

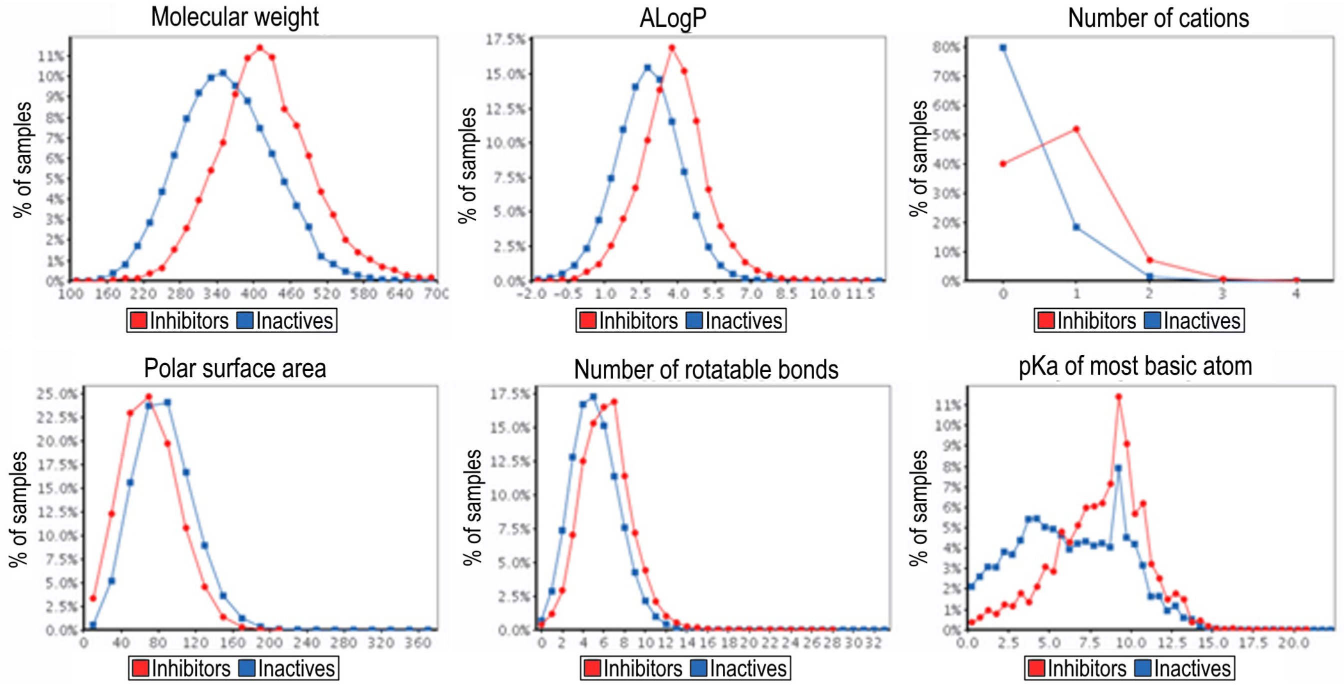


Fig 10. Distribution of 6 physicochemical properties showing significant differences. The distributions of hERG inhibitors and inactive compounds are shown as red and blue lines, respectively.

https://doi.org/10.1371/journal.pone.0199348.g010

and 1.87 for inactives on average. The size of the compounds in the hERGCentral dataset fell between them, with mean MWs of 399.2 for inhibitors and 354.5 for inactive compounds. The difference in the molecular sizes could be due to the fact that the compounds published in medicinal chemistry journals or patents were often synthetically optimized to achieve high potency or desirable properties, and this process generally increased their size from the initial hit compounds in chemical libraries for HTS. Thus, the construction of hERG prediction models based on only a single database could be severely affected by the choice of the database. These results emphasize the importance of the collection and integration of SAR information from various databases, to cover a wider chemical space and enable a more robust analysis. The full description of the physicochemical properties for each database is available in S1 Table.

Conclusion

By integrating the ChEMBL, GOSTAR, NCGC, and hERGCentral datasets, the largest dataset for the hERG blockade activities of small compounds was constructed in this study. The assay deviation of each compound revealed that large deviations up to more than 100-fold could occur by incorrect curation, emphasizing the importance of removing the outlier values. After the consideration of multiple assay entries, the dataset consisting 9,890 hERG inhibitors and 281,329 inactive compounds was built. The database covered 18.2% of all of the Murcko framework-based chemical space occupied by ChEMBL compounds. The amount of hERG activity information has dramatically increased over the past decade. The number of reported hERG inhibitors increased almost 10-fold in the past 10 years, along with about a 100-fold increase of inactive compounds. The variety of chemical scaffolds commonly (more than 10 compounds) found in hERG inhibitors also increased nearly 3-fold (only 28 scaffolds were reported before 2009, among the 82 scaffolds currently reported).

Investigations of the physicochemical properties of hERG inhibitors in the integrated database reproduced their well-known characteristics. hERG inhibitors tend to exhibit larger molecular weight, more hydrophobicity, more cationic atoms, and less polar surface area. A notable observation was found in the difference of the property distribution among the individual databases. The molecular weight distribution in each database reflected the origin of the compounds where the database collected the SAR information. These differences in the data sources could affect the interpretation of the physicochemical properties in the construction of prediction models for hERG inhibition when the prediction model is built using only one data source. The integration of hERG-associated information from various databases would decrease the bias of the data sources and provide a robust data set for statistical analysis.

The integrated database is available at our home page (http://drugdesign.riken.jp/hERGdb/), with the exception of GOSTAR, which is a commercial database that is not publicly accessible. The current interface allows structural searches and filtering by assay types or data sources. More features such as keyword search, ftp downloads, etc. are planned for future development. The authors are currently constructing a discrimination model of hERG inhibitors and inactive compounds based on the integrated database, which has already showed promising prediction performance exceeding those by commercial software to predict hERG inhibition. The prediction model will be released publicly with the integrated database itself on our homepage. Along with the functional additions, updates of the integrated database to accommodate to latest versions of the ChEMBL and NCGC datasets are ongoing. The corresponding changes in the basic statistics of the updated database and the distribution of molecular properties will be presented on the home page.

Supporting information

S1 Table. Statistics of the physicochemical properties for both hERG inhibitors and inactive compounds in each individual database and the integrated database at various time points.
(XLS)

Acknowledgments

This research is funded by the AMED Drug Discovery Informatics project. The authors thank Dr. Konagaya, Mr. Mungkalaton, and the members of the AMED Drug Discovery Informatics project for useful discussion about data format.

Author Contributions

Data curation: Tomohiro Sato, Hitomi Yuki, Keiji Ogura.

Methodology: Tomohiro Sato.

Project administration: Teruki Honma.

Supervision: Teruki Honma.

Visualization: Tomohiro Sato.

Writing – original draft: Tomohiro Sato.

References

1. Pearlstein RA, Vaz RJ, Kang J, Chen XL, Preobrazhenskaya M, Shchekotikhin AE, et al. Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. Bioorg Med Chem Lett. 2003; 13: 1829–1835 PMID: 12729675

2. De Ponti F, Poluzzi E, Montanaro 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 PMID: 11497335

3. Czodrowski P. hERG Me Out. J Chem Inf Model. 2013; 53(9): 2240–2251. https://doi.org/10.1021/ci400308z PMID: 23944269

4. Roy M, Dumaine R, Brown AM. HERG, A primary human ventricular target of the nonsedating antihistamine terfenadine. Circulation. 1996; 94: 817–823 PMID: 8772706

5. Suessbrich H, Waldegger S, Lang F, Busch AE. Blockade of HERG channels expressed in *Xenopus* oocytes by the histamine receptor antagonists terfenadine and astemizole. FEBS Lett. 1996; 385: 77–80 PMID: 8641472

6. Zhou Z, Vorperian VR, Gong Q, Zhang S, January CT. Block of HERG potassium channels by the antihistamine astemizole and its metabolites desmethylastemizole and norastemizole. J Cardiovasc Electrophysiol. 1996; 10: 836–843

7. Rampe D, Roy ML, Dennis A, Brown AM. 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 PMID: 9395068

8. Wang S, Li Y, Xu L, Li D, Hou T. Recent Developments in Computational Prediction of hERG Blockage. Curr Top Med Chem. 2013; 13(11): 1317–1326. PMID: 23675938

9. Villoutreix BO, Taboureau O. Computational Investigations of hERG Channel Blockers: New Insights and Current Predictive Models. Adv Drug Deliv Rev. 2015; 86: 72–82. https://doi.org/10.1016/j.addr.2015.03.003 PMID: 25770776

10. Du F, Babcock JJ, Yu H, Zou B, Li M. Global Analysis Reveals Families of Chemical Motifs Enriched for hERG Inhibitors. PLoS One 2015; 10(2): 1–21.

11. O'Brien SE, De Groot MJ. Greater than the Sum of Its Parts: Combining Models for Useful ADMET Prediction. J Med Chem. 2005; 48(4): 1287–1291. https://doi.org/10.1021/jm049254b PMID: 15715500

12. Song M, Clark M. Development and Evaluation of an in Silico Model for bERG Binding. J Chem Inf Model. 2006; 46(1): 392–400. https://doi.org/10.1021/ci050308f PMID: 16426073

13. Tobita M, Nishikawa T, Nagashima RA. Discriminant Model Constructed by the Support Vector Machine Method for HERG Potassium Channel Inhibitors. Bioorganic Med Chem Lett. 2005; 15(11): 2886–2890.

14. Xu C, Cheng F, Chen L, Du Z, Li W, Liu G, et al. In Silico Prediction of Chemical Ames Mutagenicity. J Chem Inf Model. 2012; 52(11): 2840–2847. https://doi.org/10.1021/ci300400a PMID: 23030379

15. Yu P, Wild DJ. Fast Rule-Based Bioactivity Prediction Using Associative Classification Mining. J Cheminform. 2012; 4(11): 1–10.

16. Su BH, Slien MY, Esposito EX, Hopnnger AJ, Tseng YJ. In Silico Binary Classification QSAR Models Based on 4D-Fingerprints and MOE Descriptors for Prediction of hERG Blockage. J Chem Inf Model. 2010; 50(7): 1304–1318. https://doi.org/10.1021/ci100081j PMID: 20565102

17. Jia L, Sun H. Support Vector Machines Classification of hERG Liabilities Based on Atom Types. Bioorganic Med Chem. 2008; 16(11): 6252–6260.

18. Braga RC, Alves VM, Silva MFB, Muratov E, Fourches D, Lião LM, et al. Pred-hERG: A Novel Web-Accessible Computational Tool for Predicting Cardiac Toxicity Mol Inform. 2015; 34(10): 698–701. https://doi.org/10.1002/minf.201500040 PMID: 27490970

19. Schyman P, Liu R, Wallqvist A. General Purpose 2D and 3D Similarity Approach to Identify hERG Blockers. J Chem Inf Model. 2016; 56(1): 213–222. https://doi.org/10.1021/acs.jcim.5b00616 PMID: 26718126

20. Wang S, Sun H, Liu H, Li D, Li Y, Hou T. ADMET Evaluation in Drug Discovery. 16. Predicting hERG Blockers by Combining Multiple Pharmacophores and Machine Learning Approaches. Mol. Pharm. 2016; 13 (8):285–2866.

21. Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. Mol Pharm. 2017; 14 (12):4462–4475. https://doi.org/10.1021/acs.molpharmaceut.7b00578 PMID: 29096442

22. Sun H, Huang R, Xia M, Shahane S, Southall N, Wang Y. Prediction of hERG Liability: Using SVM Classification, Bootstrapping and Jackknifing. Mol Inform. 2017; 36 (4).

23. Chemi G, Gemma S, Campiani G, Brogi S, Butini S, Brindisi M. Computational Tool for Fast in Silico Evaluation of hERG K^+ Channel Affinity. Front. Chem. 2017; 5 (7).

24. Wang W, MacKinnon R. Cryo-EM Structure of the Open Human Ether-À-Go-Go-Related K^+ Channel hERG. Cell. 2017; 169(3): 422–430.e10. https://doi.org/10.1016/j.cell.2017.03.048 PMID: 28431243

25. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012; 40(D1): 1100–1107.

26. GOSTAR (GVK BIO Online Structure Activity Relationship Database) GVK Biosciences Private Limited, Plot No. 28 A, IDA Nacharam, Hyderabad, India, 2010. https://gostardb.com/gostar/

27. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, et al. PubChem Substance and Compound Databases. Nucleic Acids Res. 2016; 44(D1): D1202–D1213. https://doi.org/10.1093/nar/gkv951 PMID: 26400175

28. Du F, Yu H, Zou B, Babcock J, Long S, Li M. hERGCentral: A Large Database to Store, Retrieve, and Analyze Compound-Human Ether-À-Go-Go Related Gene Channel Interactions to Facilitate Cardiotoxicity Assessment in Drug Development. Assay Drug Dev Technol. 2011; 9(6): 580–588. https://doi.org/10.1089/adt.2011.0425 PMID: 22149888

29. Bemis GW, Murcko MA. The Properties of Known Drugs. 1. Molecular Frameworks. J Med Chem. 1996; 39(15): 2887–2893. https://doi.org/10.1021/jm9602928 PMID: 8709122

30. Langdon SR, Blagg J, Brown N, Scaffold Diversity in Medicinal Chemistry Space. Scaffold Hopping. Med. Chem. 2013; 58:39–60.

31. Karawajczyk A, Giordanetto F, Benningshof J, Hamza D, Kalliokoski T, Pouwer K, Morgentin R, Nelson A, Müller G, Piechot A, Tzalis D. Expansion of Chemical Space for Collaborative Lead Generation and Drug Discovery: The European Lead Factory Perspective. Drug Discov. Today 2015, 20 (11), 1310–1316. https://doi.org/10.1016/j.drudis.2015.09.009 PMID: 26429298

32. Fernandez-de Gortari E, Medina-Franco JL. Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases. RSC Adv. 2015; 5(106): 87465–87476.

33. Prieto-Martínez FD, Gortari EF, Méndez-Lucio O, Medina-Franco JL. A Chemical Space Odyssey of Inhibitors of Histone Deacetylases and Bromodomains. RSC Adv. 2016; 6(61): 56225–56239.

34. Kratz JM, Schuster D, Edtbauer M, Saxena P, Mair CE, Kirchebner J, et al. Experimentally Validated hERG Pharmacophore Models as Cardiotoxicity Prediction Tools. J Chem Inf Model. 2014; 54(10): 2887–2901. https://doi.org/10.1021/ci5001955 PMID: 25148533

35. Kamiya K, Niwa R, Mitcheson JS, Sanguinetti MC. Molecular Determinants of HERG Channel Block. Mol Pharmacol. 2006; 69(5): 1709–1716. https://doi.org/10.1124/mol.105.020990 PMID: 16474003

36. Kamiya K, Niwa R, Morishima M, Honjo H, Sanguinetti MC. Molecular Determinants of hERG Channel Block by Terfenadine and Cisapride. J Pharmacol Sci. 2008; 108(3): 301–307. PMID: 18987434