Additional file

Studying and mitigating the effects of data drifts on ML model performance at the example of chemical toxicity data

Andrea Morger, Marina Garcia de Lomana, Ulf Norinder, Fredrik Svensson, Johannes Kirchmair, Miriam Mathea and Andrea Volkamer

A1 Additional information on data and methods

A1.1 Target selection for the ChEMBL datasets

Target datasets were selected following a collection of 1360 ligand sets provided by Škuta et al.[1] for similarity searching, bioactivity classification and scaffold hopping. First, the 29 target datasets, for which Škuta et al. found $\geq$ 1000 compounds with reported pIC50 values, were downloaded, including pIC50 values and publication year. The following cleaning procedure was applied to each target dataset: If there were multiple measurements per compound and endpoint, the mean and standard deviation were calculated. Only the mean measurement of those duplicates was kept if the standard deviation was lower or equal than 0.5, otherwise they were discarded. The oldest publication year (i.e. lowest number) was kept for aggregated data points. The compounds were standardised as described in the main manuscript (Section 2.1.2) and temporally split into training, update1, update2, and holdout set as explained in 2.1.4. If fewer than 50 active and 50 inactive compounds were left in the holdout set after the time-split, the target dataset was excluded from the study. Finally, 20 targets remained which match the filtering criteria. Of these, a total of twelve targets were selected that are linked to toxicity. A target was defined to be associated to toxicity if it was either assayed in ToxCast[2], or part of the list of targets that are recommended to early assess the potential hazard of a compound[3].

A1.2 Public datasets for liver toxicity and MNT

To assess drifts between data originating from different sources, public and proprietary datasets for liver toxicity and micro nucleus test (MNT) were collected. For conformal prediction (CP) model training, the same public datasets for liver toxicity (more specifically here drug-induced liver injury (DILI)) and MNT in vivo were used as described by Garcia de Lomana et al.[4]. Data for the DILI endpoint were gathered from the U.S. Food and Drug Administration (FDA)[5] and for the MNT in vivo endpoint from three sources (eChemPortal[6], the work of Benigni et al.[7] and Yoo et al.[8]). The respective datasets contain 692 (445 active and 247 inactive compounds) and 1791 compounds (316 active and 1475 inactive compounds) after the data pre-processing and deduplication steps conducted by Garcia de Lomana et al.[4].

A1.3 Inhouse datasets for liver toxicity and MNT

Two inhouse datasets for liver toxicity and MNT in vivo, with data generated by BASF SE, were used as holdout and update set to investigate data drifts between data with different origin. Liver toxicity was measured in oral assays on rats (including OECD Guidelines 407, 408 and 422, as well as range finding oral studies). Compounds showing adverse or adaptive effects in the liver in any of these studies were labelled as active. MNT in vivo was determined in mice in an assay following the OECD Guideline 474 or in (non-GLP) screening assays (with 18 animals). The liver toxicity dataset contains 140 (63 active and 77 inactive) compounds and the MNT in vivo dataset contains 366 (194 active and 172 inactive) compounds after the data pre-processing and deduplication steps (following the same procedure as Garcia de Lomana et al.[4], see "Chemical structure standardisation").

A1.4 Time-splitting procedure

Note that all compounds published (ChEMBL data) or assayed (inhouse data) in the same year were assigned to the same split.

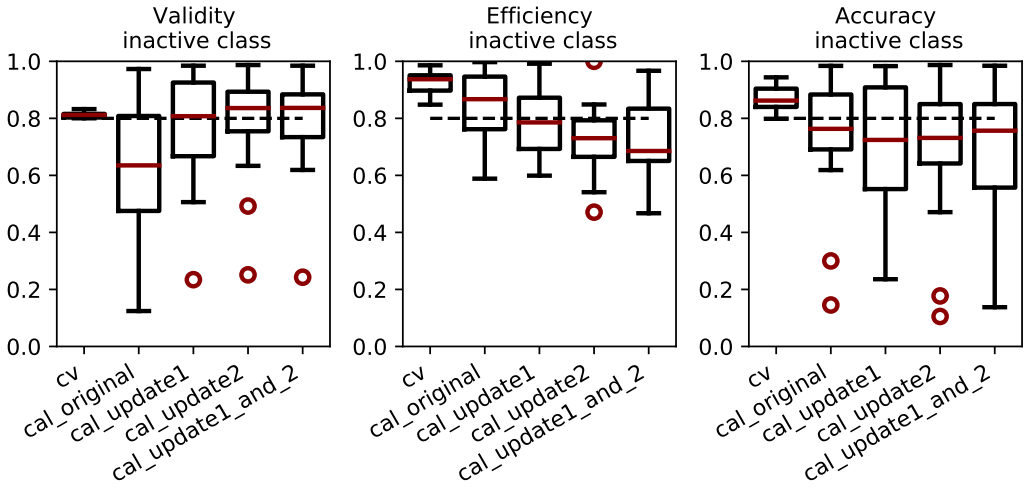
ChEMBL data After standardising the compounds (see 2.1.2), the ChEMBL data were time-split into four datasets, i.e. train, update1, update2, and holdout set based on the publication year. A minimum number of compounds per dataset was defined based on a predefined ratio, i.e. the training set must contain at least 50% of the total number of compounds, the update1 and update2 sets must contain at least 12% each. Starting from the earliest year, all compounds published in that year were assigned to the training set and the number of training compounds was assessed. Same for the next year(s) until the training set contained at least the minimum number of training compounds defined. Then, all compounds published in the following year(s) were assigned to the update1 set until the respective threshold was reached. With the same procedure, the compounds published in the subsequent year(s) were allocated to the update2 set. All remaining compounds belong to the holdout set. The number of active and inactive compounds available per subset of the twelve holdout ChEMBL target datasets, as well as the corresponding time thresholds for splitting, are provided in Table 2.

Liver toxicity and MNT data To investigate the occurrence of discrepancies between external and internal data (see A1.2), the liver toxicity and MNT datasets were investigated. The external data were used for model building as well as for the original calibration set. The internal data were time-split into update and holdout set based on the date they were measured internally. Due to the small number of available inhouse compounds, only one update set was deducted. The data was selected by year as described for the ChEMBL data until at least 50% of the compounds were assigned to the update set. The number of training, update and holdout compounds available for the liver toxicity and MNT endpoints are shown in Table 2.

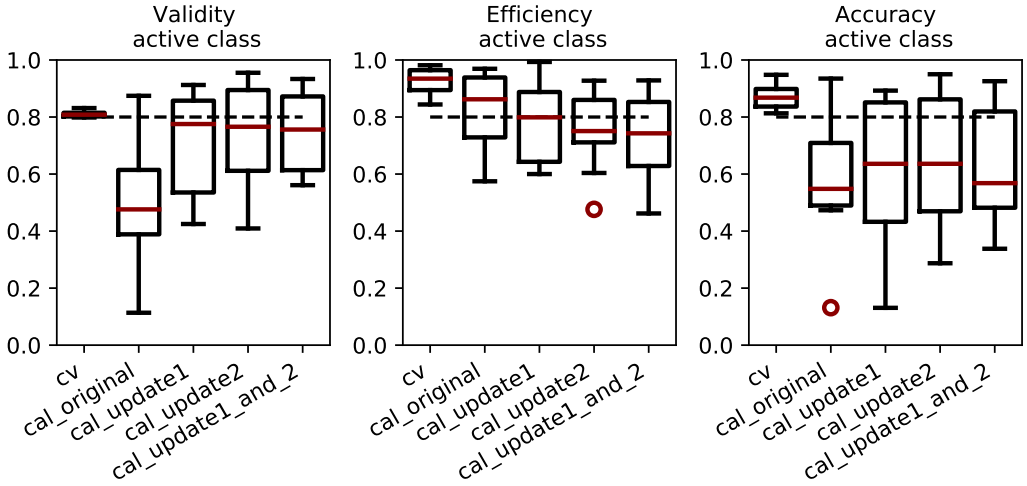
Table S1. ChEMBL datasets and their biological relevance. A selection of possible toxicological or adverse effects due to agonism (or activation) or antagonism (or inhibition) with the targets is provided.

ChEMBL ID	name	toxicological or adverse effects
CHEMBL220	Acetylcholinesterase (human)	decreased blood pressure or heart rate, increased GI motility[3,9]
CHEMBL4078	Acetylcholinesterase (fish)	decreased blood pressure or heart rate, increased GI motility[3,9]
CHEMBL5763	Cholinesterase	decreased heart rate, QT interval prolongation[10]
CHEMBL203	EGFR erbB1	skin toxicity, cardiotoxicity[11,12]
CHEMBL206	Estrogen receptor alpha	antiandrogenic effects, hormone-dependent cancers[13,14]
CHEMBL279	VEGFR 2	hypertension, disturbed wound healing, GI and skin toxicity[15]
CHEMBL230	Cyclooxygenase-2	myocardial infarction, increased blood pressure, ischaemic stroke, atherothrombosi[3,16]
CHEMBL340	Cytochrome P450 3A4	drug-drug interactions, detoxification by metabolism, activation of toxic metabolites[17]
CHEMBL240	HERG	QT interval prolongation[18]
CHEMBL2039	Monoamine oxidase B	cell death[19]
CHEMBL222	Norepinephrine transporter	increased heart rate or blood pressure, constipation[3,20]
CHEMBL228	Serotonin transporter	increased GI motility, insomnia, anxiety, sexual dysfunction[3,21]

A2 Additional information on results



(a) Evaluation for inactive compounds



(b) Evaluation for active compounds

Figure S1. Class-wise time split evaluation (validity, efficiency, accuracy) of cross-validation (CV) experiments and predictions for the holdout set using the original (cal_original), update1 (cal_update1), update2 (cal_update2) and combined update1_and_2 (cal_update1_and_2) calibration sets for twelve ChEMBL datasets.

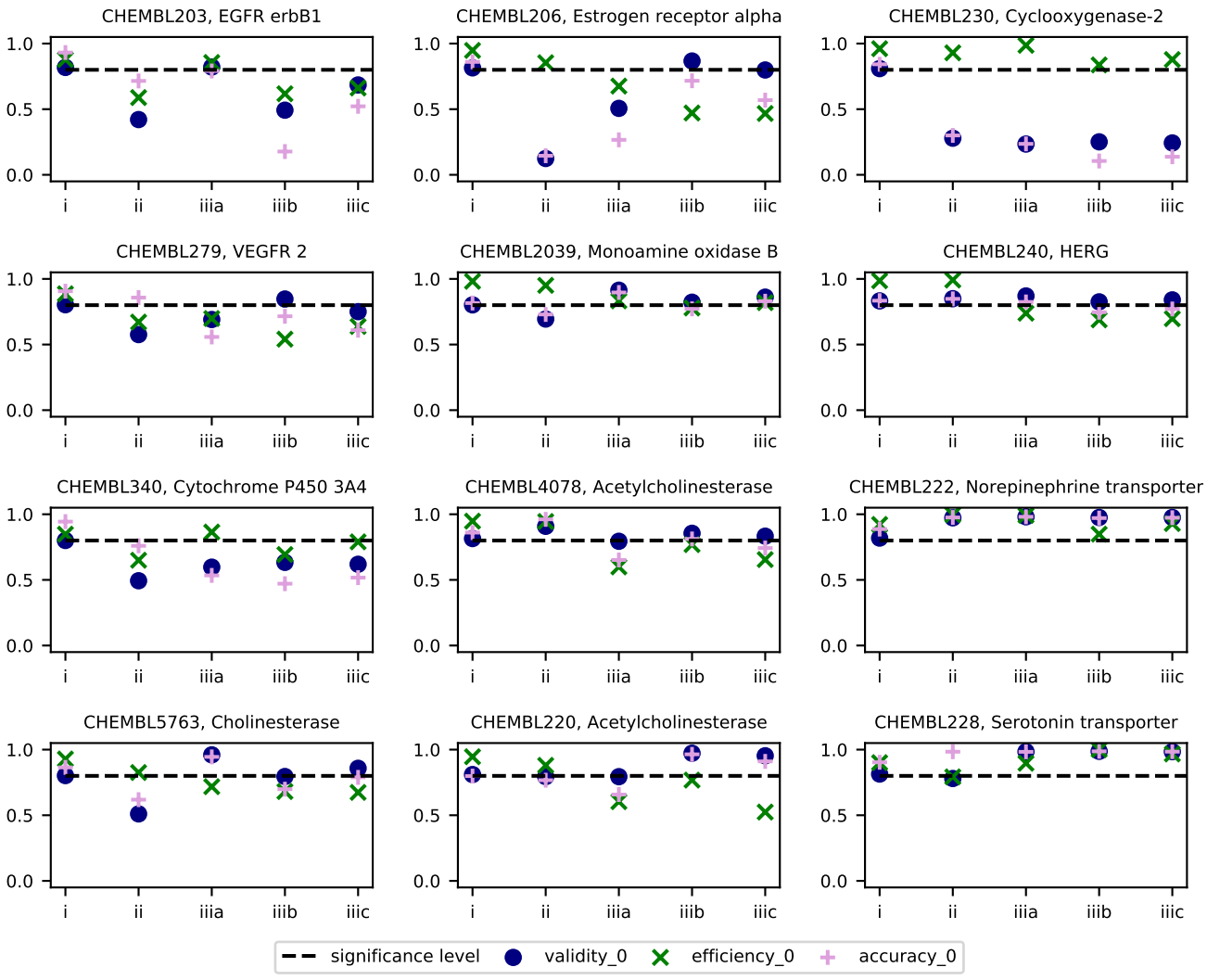


Figure S2. Inactive compounds evaluation of time-split experiments for individual ChEMBL endpoints. i) cross-validation on training data, predict holdout data using ii) original calibration set iiia) update1, iiib) update2, iiic) combined update1+2 calibration sets.

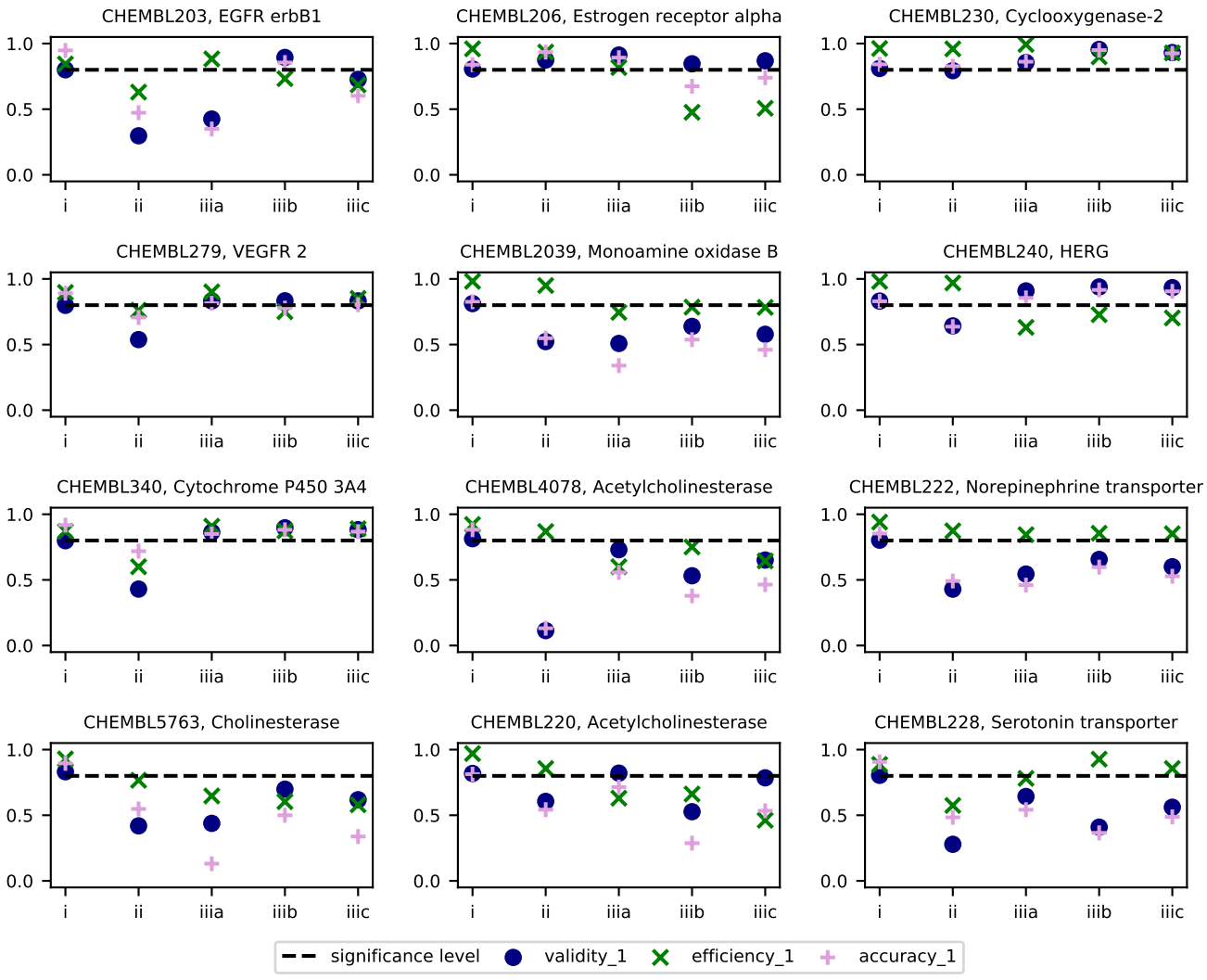


Figure S3. Active compounds evaluation of time-split experiments for individual ChEMBL endpoints. i) cross-validation on training data, predict holdout data using ii) original calibration set iiia) update1, iiib) update2, iiic) combined update1+2 calibration sets.

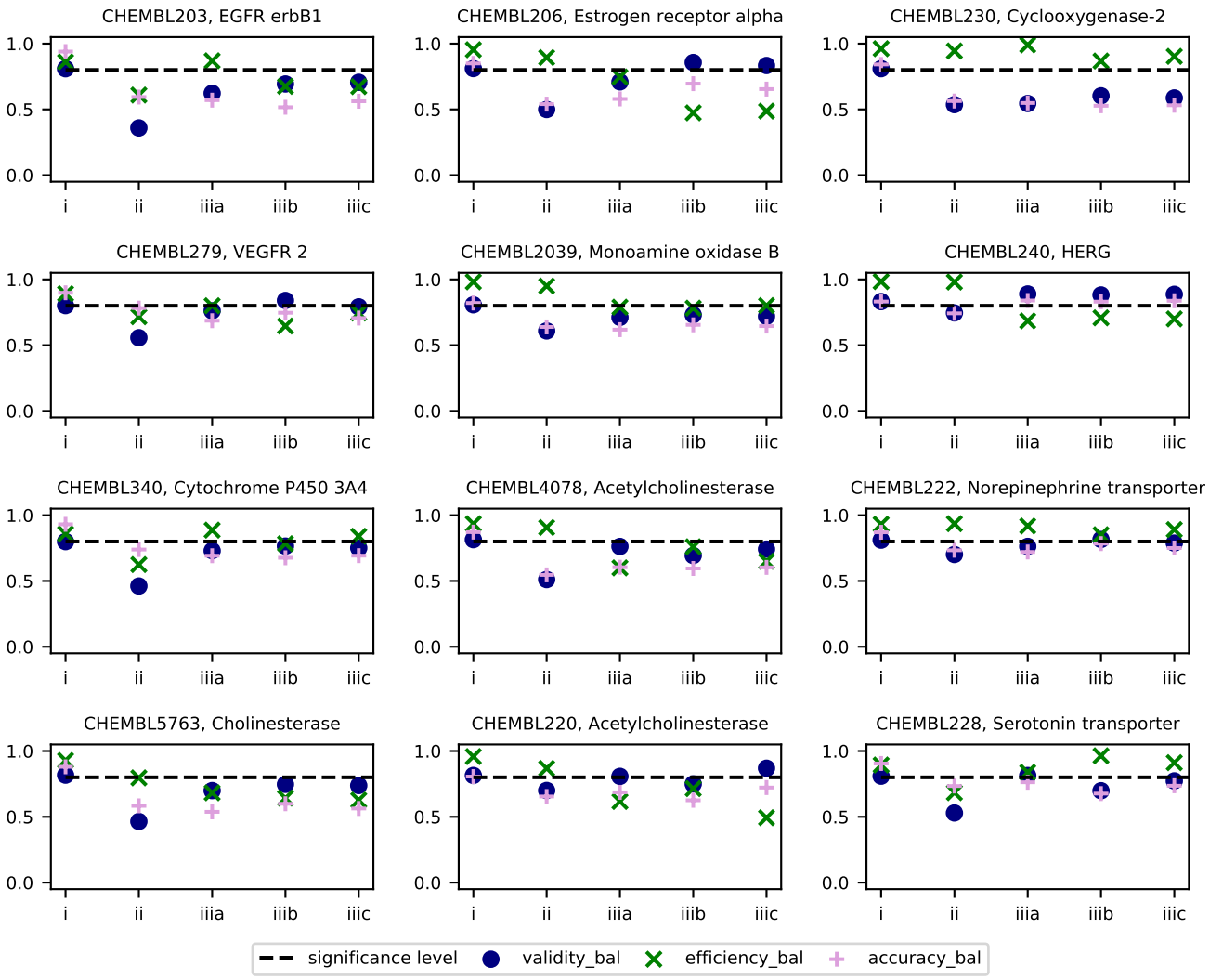


Figure S4. Balanced evaluation of time-split experiments for individual ChEMBL endpoints. i) cross-validation on training data, predict holdout data using ii) original calibration set, iii) updated calibration set, a) update1, b) update2, c) combined update1+2 sets. The doted line at 0.8 denotes the expected validity for the chosen significance level (of 0.2).

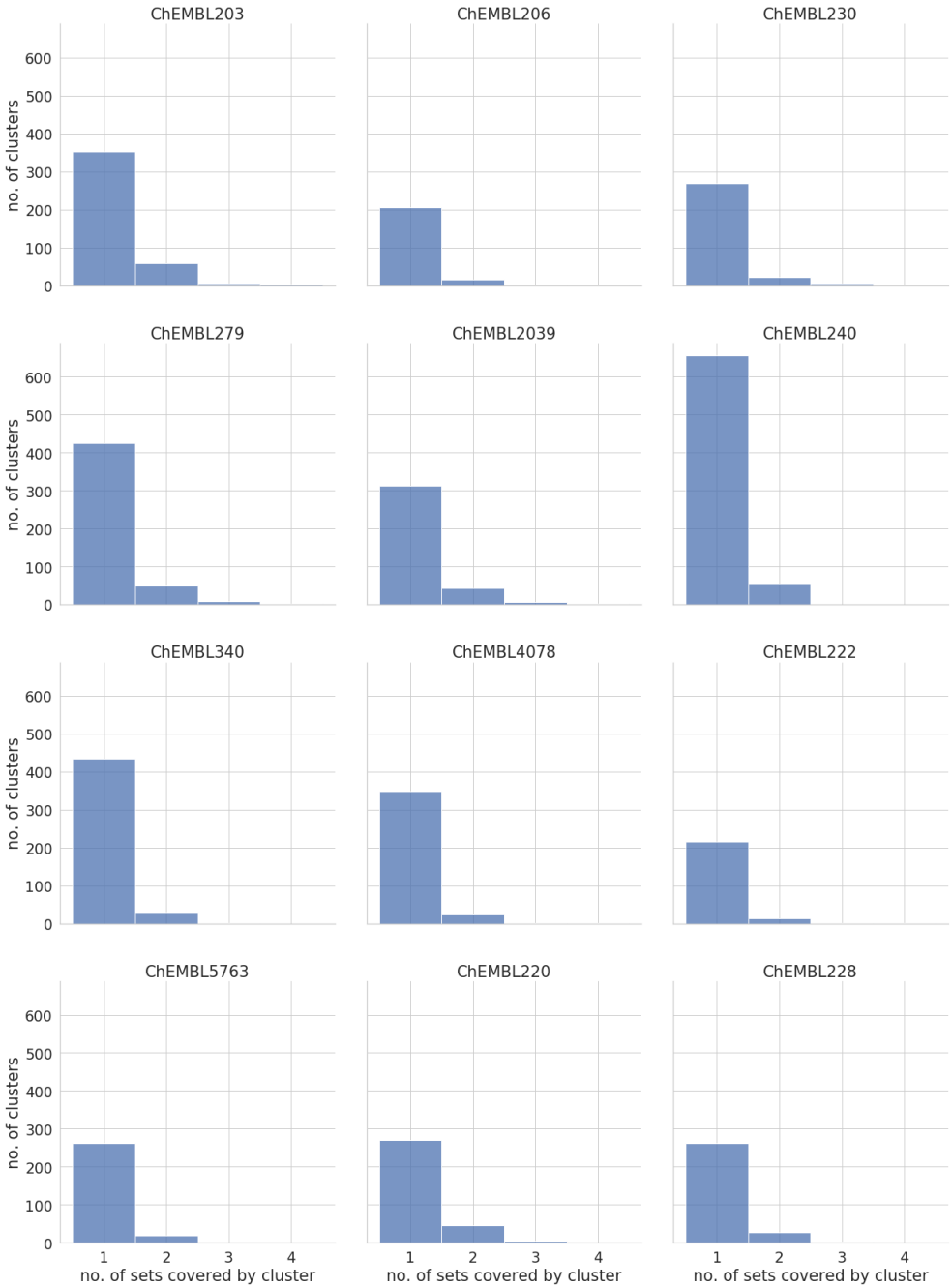
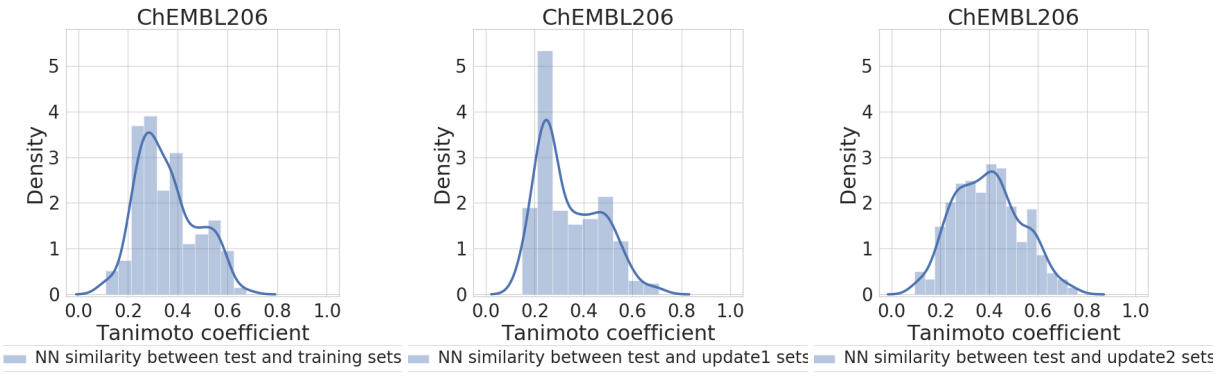


Figure S5. Spreading of clusters amongst the data subsets (i.e. splits) for the ChEMBL datasets. Most of the clusters (with at least two compounds) do not spread over more than one subset (i.e. training, update1, update2 or holdout set).



(a) training set **(b)** Update1 set **(c)** Update2 set

Figure S6. Distribution of Tanimoto coefficients between each holdout compound to its nearest neighbour in the corresponding subset (training, update1 and update2) for ChEMBL206 endpoint .

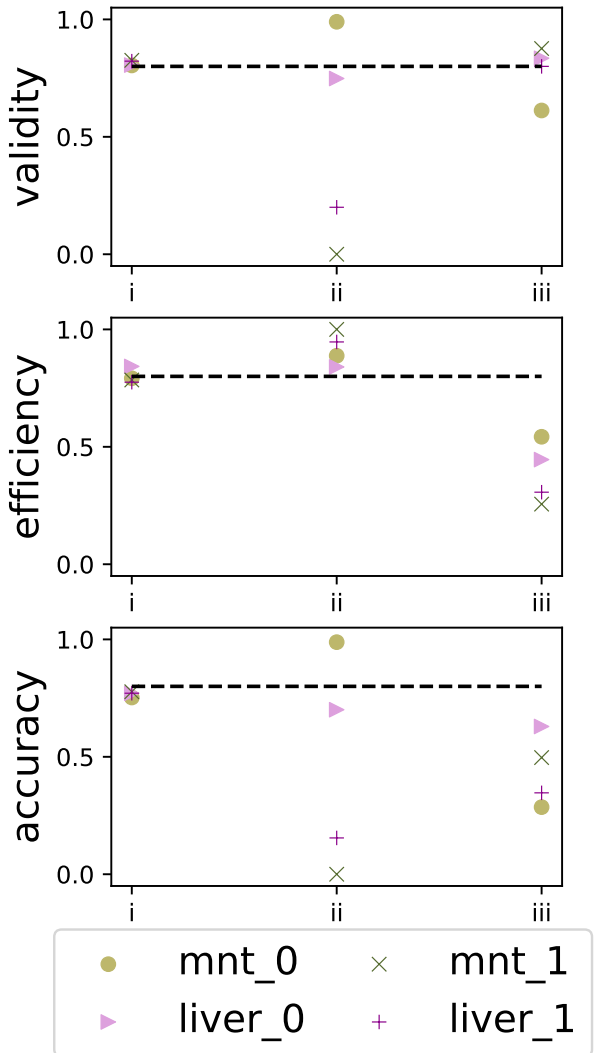


Figure S7. Time split evaluation (validity, efficiency, accuracy) of experiments i) CV, predictions using ii) original calibration set, iii) update calibration set for the liver toxicity and MNT inhouse datasets.

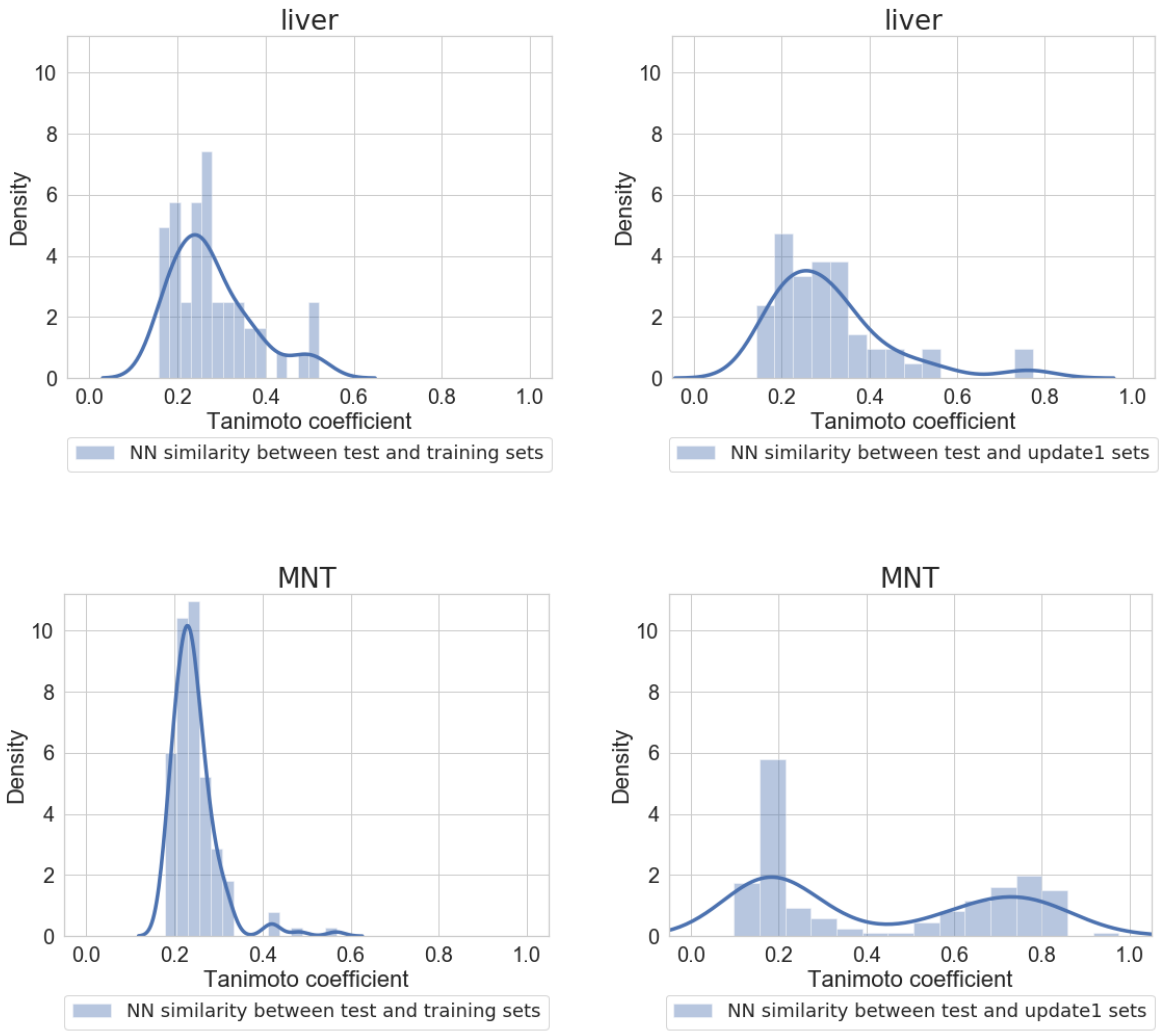


Figure S8. Distribution of Tanimoto coefficients between each holdout compound to its nearest neighbour in the training (left) and update (right) set for the liver (top) and MNT (bottom) endpoints.

References

1. Škuta, C. *et al.* QSAR-derived affinity fingerprints (part 1): Fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping. *J. Cheminformatics* **12**, 1–16, DOI: 10.1186/s13321-020-00443-6 (2020).
2. Richard, A. M. *et al.* ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **29**, 1225–1251, DOI: 10.1021/acs.chemrestox.6b00135 (2016).
3. Bowes, J. *et al.* Reducing safety-related drug attrition: The use of in vitro pharmacological profiling, DOI: 10.1038/nrd3845 (2012).
4. Garcia de Lomana, M. *et al.* ChemBioSim: Enhancing Conformal Prediction of in vivo Toxicity by Use of Predicted Bioactivities. *J. Chem. Inf. Model.* DOI: 10.1021/acs.jcim.1c00451 (2021).
5. Chen, M. *et al.* DILIrank: The largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov. Today* **21**, 648–653, DOI: 10.1016/j.drudis.2016.02.015 (2016).
6. OECD. eChemPortal (2020).
7. Benigni, R. *et al.* Evaluation of the applicability of existing (Q)SAR models for predicting the genotoxicity of pesticides and similarity analysis related with genotoxicity of pesticides for facilitating of grouping and read across. *EFSA Support. Publ.* **16**, DOI: 10.2903/sp.efsa.2019.en-1598 (2019).
8. Yoo, J. W. *et al.* Development of improved QSAR models for predicting the outcome of the in vivo micronucleus genetic toxicity assay. *Regul. Toxicol. Pharmacol.* **113**, 104620, DOI: 10.1016/j.yrtph.2020.104620 (2020).
9. Moretto, A. Experimental and clinical toxicology of anticholinesterase agents. *Toxicol. Lett.* **102-103**, 509–513, DOI: 10.1016/S0378-4274(98)00245-8 (1998).
10. Hogan, D. B. Long-term efficacy and toxicity of cholinesterase inhibitors in the treatment of Alzheimer disease. *Can. J. Psychiatry* **59**, 618–623, DOI: 10.1177/070674371405901202 (2014).
11. Bianchini, D., Jayanth, A., Yu, J. C. & Cunningham, D. Epidermal growth factor receptor inhibitor-related skin toxicity: Mechanisms, treatment, and its potential role as a predictive marker. *Clin. Color. Cancer* **7**, 33–43, DOI: 10.3816/CCC.2008.n.005 (2008).
12. Hervent, A. S. & De Keulenaer, G. W. Molecular mechanisms of cardiotoxicity induced by ErbB receptor inhibitor cancer therapeutics. *Int. J. Mol. Sci.* **13**, 12268–12286, DOI: 10.3390/ijms131012268 (2012).
13. Buluş, A. D. *et al.* The evaluation of possible role of endocrine disruptors in central and peripheral precocious puberty. *Toxicol. Mech. Methods* **26**, 493–500, DOI: 10.3109/15376516.2016.1158894 (2016).
14. La Merrill, M. A. *et al.* Consensus on the key characteristics of endocrine-disrupting chemicals as a basis for hazard identification. *Nat. Rev. Endocrinol.* **16**, 45–57, DOI: 10.1038/s41574-019-0273-8 (2020).
15. Eskens, F. A. & Verweij, J. The clinical toxicity profile of vascular endothelial growth factor (VEGF) and vascular endothelial growth factor receptor (VEGFR) targeting angiogenesis inhibitors; A review. *Eur. J. Cancer* **42**, 3127–3139, DOI: 10.1016/j.ejca.2006.09.015 (2006).
16. Grosser, T., Fries, S. & FitzGerald, G. A. Biological basis for the cardiovascular consequences of COX-2 inhibition: Therapeutic challenges and opportunities. *J. Clin. Investig.* **116**, 4–15, DOI: 10.1172/JCI27291 (2006).
17. Guengerich, F. P. Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem. Res. Toxicol.* **14**, 611–650, DOI: 10.1021/tx0002583 (2001).
18. Vandenberg, J. I., Walker, B. D. & Campbell, T. J. HERG K+ channels: Friend and foe. *Trends Pharmacol. Sci.* **22**, 240–246, DOI: 10.1016/S0165-6147(00)01662-X (2001).
19. Nicotra, A. & Parvez, S. H. Cell death induced by MPTP, a substrate for monoamine oxidase B. *Toxicology* **153**, 157–166, DOI: 10.1016/S0300-483X(00)00311-5 (2000).
20. Mayer, A. F. *et al.* Influences of norepinephrine transporter function on the distribution of sympathetic activity in humans. *Hypertension* **48**, 120–126, DOI: 10.1161/01.HYP.0000225424.13138.5d (2006).
21. Stahl, S. M. Mechanism of action of serotonin selective reuptake inhibitors. Serotonin receptors and pathways mediate therapeutic effects and side effects. *J. Affect. Disord.* **51**, 215–235, DOI: 10.1016/S0165-0327(98)00221-3 (1998).