Nonadditivity in Public and Inhouse Data – Implications for Drug Design

Dea Gogishvili
AstraZeneca Sweden: AstraZeneca AB https://orcid.org/0000-0001-8809-0861
Eva Nittinger (✉ **eva.nittinger@astrazeneca.com**)
AstraZeneca https://orcid.org/0000-0001-7231-7996
Christian Margreitter
AstraZeneca https://orcid.org/0000-0002-5473-6318
Christian Tyrchan
AstraZeneca https://orcid.org/0000-0002-6470-984X

Research article

Keywords: Nonadditivity analysis, structure-activity relationship, matched molecular pair analysis, experimental uncertainty, machine learning, support vector machine, random forest.

Posted Date: December 16th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-127217/v1

Nonadditivity in Public and Inhouse Data – Implications for Drug Design

D. Gogishvili[1,#,¤], E.Nittinger[1,#,*], C. Margreitter[2], C. Tyrchan[1]

[1] Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

[2] Computational Chemistry, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden

[#] Shared first authors

[*] Correspondence: eva.nittinger@astrazeneca.com

ABSTRACT

Numerous ligand-based drug discovery projects are based on structure-activity relationship (SAR) analysis, such as Free-Wilson (FW) or matched molecular pair (MMP) analysis. Intrinsically they assume linearity and additivity of substituent contributions. These techniques are challenged by nonadditivity (NA) in protein-ligand binding where the change of two functional groups in one molecule results in much higher or lower activity than expected from the respective single changes. Identifying nonlinear cases and possible underlying explanations is crucial for a drug design project since it might influence which lead to follow. By systematically analyzing all AstraZeneca (AZ) inhouse compound data and publicly available ChEMBL25 bioactivity data, we show significant NA events in almost every second assay among the inhouse and once in every third assay in public data sets. Furthermore, 9.4% of all compounds of the AZ database and 5.1% from public sources display significant additivity shifts indicating important SAR features or fundamental measurement errors. Using NA data in combination with machine learning showed that nonadditive data is challenging to predict and even the addition of nonadditive data into training did not result in an increase in predictivity. Overall, NA analysis should be applied on a regular basis in many areas of computational chemistry and can further improve rational drug design.

KEYWORDS

Nonadditivity analysis; structure-activity relationship; matched molecular pair analysis; experimental uncertainty; machine learning; support vector machine; random forest.

INTRODUCTION

The similarity and additivity principles represent the basis of various well-established areas in computer-aided drug design (CADD) such as Free-Wilson (FW)[1] analysis, 2D/3D quantitative structure-activity relationship (QSAR),[2] matched molecular pair (MMP)[3] analysis and computational scoring functions.[4, 5] Similarity and additivity are often implicitly assumed in CADD approaches in order to identify favorable molecular descriptors and predict the activity of new molecules. Otherwise chemists would have to synthesize and biologically evaluate every single molecule.[6]

Yet, all the principles are subjects of frequent disruptions. The exceptions to the similarity principle often complicate SAR analysis. So-called ‘activity cliffs’ refer to structurally very similar compound pairs with large alterations in potency.[7–14] Exceptions to linearity and additivity occur when the combination of substituents boosts or significantly decreases the biological activity of a ligand.[15–19] Nonadditivity (NA) may have several underlying reasons, including inconsistency in the binding pose of the central scaffold inside the pocket[20] and steric clashes.[21] Conformational changes in the binding pocked such as complete reorientation of the ligands alter the free energy of binding.[15] Furthermore, many nonadditive ‘magic methyl’ cases[13, 14, 22], i.e. attaching a simple alkyl fragment to a ligand that greatly increases the biological activity, can be explained by conformational changes as the so-called ‘ortho-effect’.

Additivity and NA of ligand binding have been studied for many years[23, 24] and can be perceived as a specific kind of interaction between functional groups.[25, 26] By analyzing public SAR data sets for strong NA (ΔΔpActivity > 2.0 log units) and respective X-ray structures, Kramer et al. showed that the cases of strong NA are underlined by changes in binding mode.[15] Babaoglu and Schoichet applied an inverse, deconstructive logic to structure-based drug design (SBDD) and by studying β-lactamase inhibitors demonstrated that fragments often do not recapitulate the binding affinity of the parent molecule.[27] The study of Miller and Wolfenden about substrate recognition demonstrated that the combination of distinct functional groups shows strong nonadditive behavior.[28] The work of Hajduk et al.[29] on stromelysin inhibitors and Congrive et al.[30] on CDK inhibitors showed that molecular affinity after the combinations of a certain amount of functional groups is much higher than expected. Patel et al. examined various combinatorial libraries assayed on several different biological responses and concluded that only half of the data is additive.[4] McClure and colleagues developed a method to determine FW additivity in a combinatorial matrix of compounds (when multiple R groups are altered simultaneously; combinatorial analoging) and they intuitively explained the occurring NA by changes in binding mode without any structural validation.[18, 19] Water molecules are a major player in ligand−protein interactions by participating in extended hydrogen-bond networks.[31] Baum, Muley, and co-workers thoroughly analyzed the structural data and the reasons behind NA at the molecular level, [17, 32] showing that NA can be the result of entropy and enthalpy profile changes, caused by hydrophobic interactions, hydrogen bonding, loss of residual mobility of the bound ligands. In another study, Kuhn et al. proposed that internal hydrogen bonding to be the reason for NA during compound optimization.[33] Gomez et al. explained NA caused by protein structural changes upon ligand binding.[16] According to these studies, instead of seeing NA as a problem, it should be interpreted as a hint towards key SAR features and variations in the binding modes. Identifying NA and understanding the reasons behind it is crucial for rational drug design since it provides valuable information about ligand-protein contacts and molecular recognition. NA analysis helps us to identify potential SAR outliers in a data set, ultimately suggesting interesting structural properties that might change the course of a small molecule optimization. Importantly, NA might be caused by experimental noise.

Despite the clear need for NA analysis it is generally not incorporated in classical QSAR applications and publications. NA clearly creates difficulties for linear SAR analysis approaches, such as standard MMP and FW analysis. These classical QSAR models will not work if the effect of introducing group R1 in the molecule is influenced by R2 or R3.[4] NA is calculated from double-mutant/double-transformation cycles consisting of four compounds linked by two identical transformations.[15] Assuming that each measurement among these double mutants contains experimental uncertainty, the experimental noise might add up and result in false nonadditive cases. Therefore, it is critical to distinguish real NA from assay noise. Extensive work on NA has been carried out by Kramer et al. In their publications they created the statistical framework to systematically analyze NA.[6, 15] Kramer first developed a general metric and afterwards created an open-source python code to quantify NA, available on GitHub.[6]

Apart from classical CADD approaches, many machine learning (ML) and deep learning (DL) techniques became very popular and are applied to a diverse range of questions – from generation of new molecules[34–37], to predicting binding affinities[38–46] and retrosynthesis predictions[47–50]. Thus, the question arises: How much are those methods influenced by NA? When activity data is used for model training, NA might cause problems that are currently not considered adequately.

In this work we show a systematics analysis of AZ inhouse and public ChEMBL physicochemical and biological data with the aim to quantify and compare NA in assays and compounds in public and inhouse data. Nonlinear events occur in 57.8% of all the tests in AZ inhouse and in 30.3% of all public data sets, indicating the need for constantly integrating NA analysis in drug discovery projects and understanding the structural reasons behind it. Additionally, we trained ML models to evaluate the predictability of nonadditive data and could show their poor performance in both support vector machine and random forest models.

METHODS

NA analysis code

The open-source NA analysis code provided by Christian Kramer was used in this study (available on GitHub: https://github.com/KramerChristian/NonadditivityAnalysis).[6] The code is written in python making use of the cheminformatics libraries RDKit[51] as well as Pandas and NumPy. NA calculations are based on MMP analysis (upon the assembly of double-transformation cycles (DTC)), using an open-source code developed by Dalke *et al.*,[52] which is an implementation of the MMPA algorithm by Hussain and Rea.[3]

Data sets

In this study both public and inhouse data are analyzed in order to compare the occurrence of NA. By understanding both types of data valuable information can be concluded for CADD projects.

ChEMBL data set

Assay data was downloaded from ChEMBL version 25 (accessed Feb. 6, 2020).[53] A ChEMBL target confidence score of at least 4 (confidence range from 0 to 9 based on available target information) was set as a threshold, resulting in 15,504,603 values.

AstraZeneca inhouse data set

All assays with an existing target gene ID were extracted from the internal AZ screening and test database (38,356 IT tests run from 2005 until 2020 across all AZ sites, accessed September 13, 2020).

Data curation

Molecules were standardized with PipelinePilot including standardization of stereoisomers, neutralization of charges, and clearing of unknown stereoisomers. This step was followed by the enumeration of tautomeric forms and selecting the canonical tautomer with PipelinePilot. The same subsequent filtering steps were employed for both datasets using a Python script to make inhouse and public data comparable (Figure 1). The filtering steps were the following: (1) All endpoints, suitable for NA analysis, were selected based on assay description. (2) Measurements without values as well as uncertain and negative values were removed. (3) Only measurements with a defined unit (M, mM, µM, nM, pM, or fM) were kept. (4) The activity values were converted to the negative logarithm of the activity - pActivity (pAct) and unrealistic values, i.e. lower than 10 pM or higher than 10 mM, were discarded. Cases were the measurement was given as pActivity (*e.g.* pIC_{50}) but had an indicated unit were discarded. (5) All compounds with multiple measurements in one test, where the difference between the minimum and the maximum measurement was larger than 2.5 log units, were removed. For those kept, the median of the logged activity values was calculated. Only compounds with large measurement differences were removed, the assay itself was kept. (6) All compounds with different IDs and the same simplified molecular-input line-entry system (SMILES) strings were filtered out and only the compound with the highest activity value was kept. (7) The molecular size was restricted to 70 heavy atoms (atomic number > 1). (8) Last, small tests with less than 25 compounds were removed.

Data selection for QSAR models

The data sets for ML study were extracted from ChEMBL (Table 1). Public assays were chosen from the NA analysis of the ChEMBL tests that had (1) NA output, (2) >200 compounds, (3) >25 double-transformation cycles (DTC) in the test in order to observe the effect of NA on ML model performance.

Table 1. Description of ChEMBL tests selected for QSAR models.

ChEMBL data	# Cpds	# Cpds with significant NA (%)	# DTC	# DTC with significant NA (%)	ChEMBL Version (access date)
1613797	6,236	73 (1.2)	6,245	694 (11.1)	27 (08/26/2020)
1614027	2,892	69 (2.4)	4,691	582 (12.4)	27 (08/26/2020)
1613777	3,512	122 (3.5)	8,600	1606 (18.7)	26 (06/20/2020)

Data curation was conducted with the Jupyter notebook (SI 1). Molecules were standardized with the PipelinePilot protocol mentioned above.

Each assay file contains: Compound IDs, SMILES, pActivity values, number of occurrences in double-transformation cycles, and an absolute NA value per compound. An NA value above 1.0 is considered to be significant.

OPTUNA

In order to build ML models, an automatic extensive hyper-parameter optimization tool, Optuna[54], was applied. Herein the optimization strategy is based on surrogate models, which is supposed to be superior to random or grid search. In order to analyze the effect of NA on ML performance, support vector machine (SVM) and Random Forest (RF) models from the scikit-learn framework[55] were trained. The latter is often considered as a base-line algorithm, being robust against over-fitting, while SVMs often push performance a bit further than RF.

While we will analyze the binary classification problem in detail (threshold chosen: pIC50 = 5), the underlying problem is a regression problem. Thus, it seemed more appropriate to model the fit as a regression and binarize afterwards. For both SVM and RF 500 trial runs were performed using a 5-fold cross-validation to avoid overfitting. We used ECFP6 counts (as implemented in REINVENT[36]). The reported values are R^2 and RMSE from scikit learn.

Model training protocol

The following protocol was applied to ChEMBL data for training SVM and RF models. Herein, additive data refers to those compounds that had NA below the experimental uncertainty cut-off and were thus not significant.

1.1) Optimization of hyper-parameters based on the training set (80% additive observations) with 5-fold cross-validation (i.e. mean performance of 5 models trained on 80% of the training set).

1.2) Train final model on all of the training set using the best hyper-parameters from 1.1) and predict both the non-significant test (20%), i.e. additive data only and the significant hold-out sets (all significant observations), i.e. nonadditive data only.

1.3) Use R^2 and RMSE (scikit learn's function) to quantify performance.

Binary classification

2.1) The predictions from 1.2) were dichotomized (threshold based on pActivity: 0 if pActivity < 5, 1 if pActivity > 5) and then compared to the true class (same threshold).

2.2) Matthews correlation coefficient (MCC from scikit learn) is used to quantify performance. MCC is used due to several advantages for binary classification problems.[56] For binary classification problems, the MCC score is guaranteed to be between -1 (anti-correlation) and 1 (perfect correlation), with 0 being the worst possible score, i.e. random. It takes into account the complete confusion matrix and thus provides a better balance between the different categories.

"Mixin" models

The effect of NA data during training and on the model performance on the test data was analyzed by adding increasing fractions of NA observations in the respective training sets (see Results). For those, we have trained models as described above and investigated whether the model performance changes by analyzing MCC values and confusion matrices. We used the hyper-parameters established earlier for the respective datasets.

RESULTS

The curated ChEMBL dataset contains 13,620 unique tests, 799,860 unique compounds and in total 3,625,044 measurements (Figure 1), while AZ inhouse data set consists of 6,277 unique tests, 1,232,555 unique compounds and in total 5,801,969 measurements.

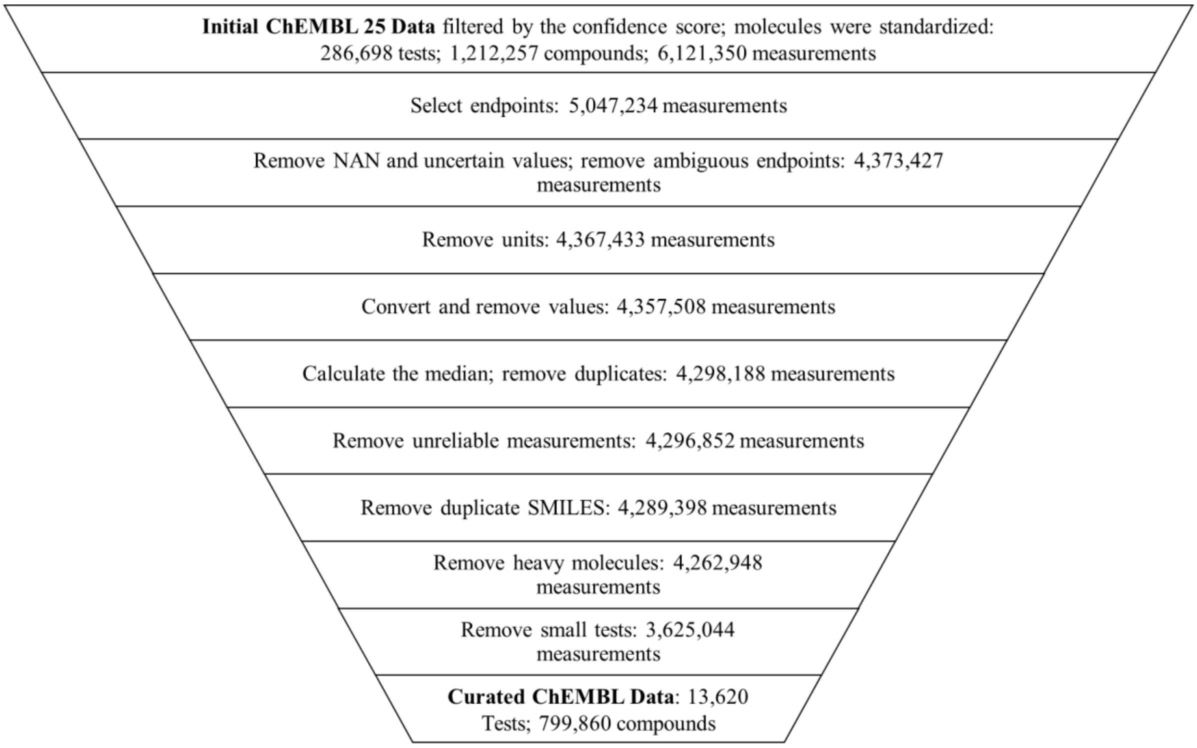

Initial ChEMBL 25 Data filtered by the confidence score; molecules were standardized: 286,698 tests; 1,212,257 compounds; 6,121,350 measurements
Select endpoints: 5,047,234 measurements
Remove NAN and uncertain values; remove ambiguous endpoints: 4,373,427 measurements
Remove units: 4,367,433 measurements
Convert and remove values: 4,357,508 measurements
Calculate the median; remove duplicates: 4,298,188 measurements
Remove unreliable measurements: 4,296,852 measurements
Remove duplicate SMILES: 4,289,398 measurements
Remove heavy molecules: 4,262,948 measurements
Remove small tests: 3,625,044 measurements
Curated ChEMBL Data: 13,620 Tests; 799,860 compounds

Figure 1. The data curation process of public ChEMBL25 data representing number of measurements after each cleaning step.

Most compounds (85%) in AZ tests have been measured more than once (Table 2), which is not the case for ChEMBL data (5%). This must be considered, during the differentiation of true NA from experimental noise. It is, indeed, easy to detect strong NA, although weak NA can be easily confused with the experimental uncertainty. On the other hand, if the experimental noise is overestimated, potentially significant cases will be ignored and not considered for compound optimization. Therefore, it is critical, to set the right threshold for experimental noise, since as mentioned before, it impacts the NA value twice as much as an individual biological measurement. Considering our data and the studies carried out by Kramer *et al.* regarding experimental uncertainty of public and inhouse data sets,[57–59] 0.3 and 0.5 log units were used as thresholds for AZ and ChEMBL data respectively. Consequently, the NA values above 0.6 (AZ) and 1.0 log (ChEMBL) units were considered significant.

NONADDITIVITY ANALYSIS

Figure 2 shows all observed NA of both AZ inhouse and ChEMBL data sets. The sign of the NA value depends on the order of the molecules within the double-transformation cycles (DTCs). Consequently, the raw data obtained after running the NA analysis contains both positive and negative values (Figure 2). Negative values have afterwards been converted to absolute values. Most of the NA cases can be explained with the experimental noise (Figure 2). Especially the major peak in the AZ and ChEMBL data are fully covered by the normal distribution expected from 0.3 and 0.5 log units of the experimental uncertainty respectively. A significant amount of DTCs not explainable by experimental uncertainty can be identified from the tail distributions.

Table 2. The numbers describing both curated AZ inhouse and ChEMBL datasets along with the output of NA analysis.

Nof	**AZ**	**ChEMBL**
Measurements	5,801,969	3,625,044
Cpds measured more than once (%)	85.8%	5.1%
Curated tests	6,277	13,620
Unique cpds	1,232,555	799,860
Tests with NA	4,030	7,534
Tests with significant NA	3,628 (57.8%)	4,128 (30.3%)
Tests with NA*	3,081 (49%)	-----
Tests with strong NA#	1,509 (24%)	1,237 (9.1%)
Unique cpds showing significant NA*	114,862 (9.4%)	40,798 (5.1%)
Unique cpds showing strong NA#	5,767 (0.5%)	8,572 (1.1%)
Median nof	**AZ**	**ChEMBL**
Unique cpds per test	233	35
Unique cpds per test with NA output	490	39
DTC per test with NA output	63	13

Unique cpds per test with significant NA*	562.5	43
DTC per test with significant NA*	88.5	23
Unique cpds per test with NA*	662	----
DTC per test with NA*	133	----
Unique cpds per test with strong NA#	1093	52
DTC per test with strong NA#	423	43

Nof = number of, DTC = double-transformation cycles

* Significant NA: 0.6 log units for AZ inhouse data, 1.0 log units for ChEMBL data

$^\#$ Strong NA: > 2.0 log units

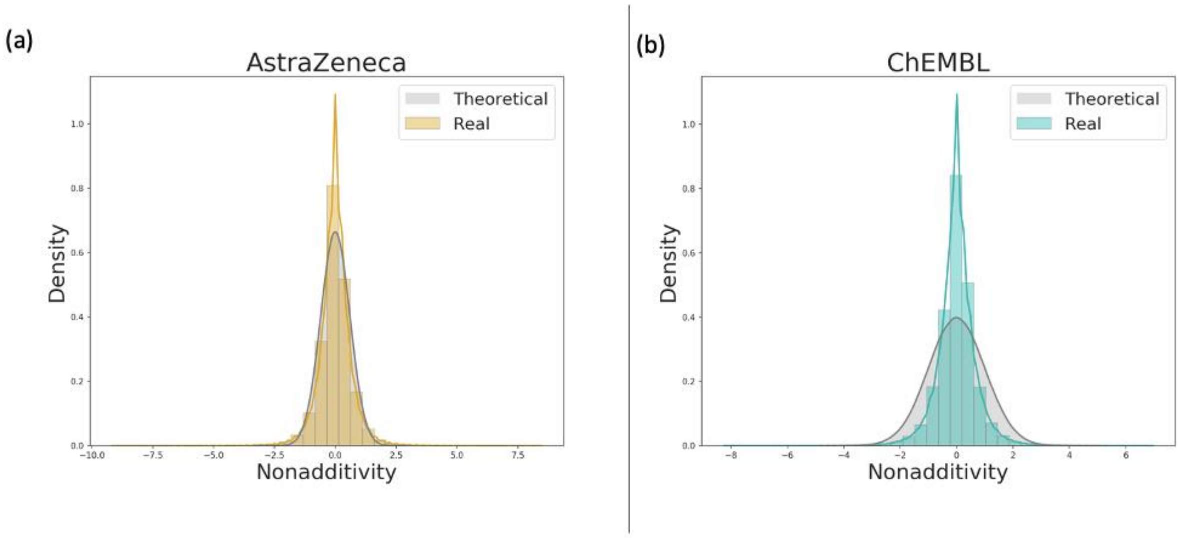


Figure 2. Theoretical NA distribution expected from an experimental uncertainty of (a) 0.3 and (b) 0.5 log units (grey lines), and observed NA distribution for all (a) AZ (yellow) tests and (b) ChEMBL (blue) tests.

According to the Figure 2 both AZ and ChEMBL NA distributions seem normal. However, the kurtosis, which is a measure of 'tailedness', is significantly large in both datasets (Table 3) and both fail the Kolmogorov-Smirnov[60, 61] tests for normality. Both AZ inhouse and public output of NA analysis is similar, yet undersampled in case of ChEMBL. Importantly, NA events occur less often in public data, based on which one might assume that nonlinear events are rare and can be disregarded. However, the pattern of nonlinear observations in AZ data sets suggests that it must be considered more carefully and structural reasons must be thoroughly investigated since they might be hinting towards important structural features.

Table 3. Descriptive statistics of NA distribution in AZ inhouse and ChEMBL data sets. Note that all NA values have not been converted to absolute values prior to these calculations.

	Observations	Mean	Variance	Std	Skewness	Kurtosis
AstraZeneca	3,053,055	0	0.42	0.65	0	3.13
ChEMBL	1,246,975	0	0.46	0.68	0.01	4.52

In order to compare the distribution of NA in two groups, two tests have been performed: (1) Kruskal-Wallis H Test[62], that does not have the assumption of normality, testing the null hypothesis that the population median of both of the groups is equal; (2) Mann-Whitney U tests[63] have been employed to test the null hypothesis that it is equally likely that a randomly selected measurement from one group of observations will be less than or greater than a randomly selected measurement from the second group of observations. According to the obtained results from both tests, the NA value distribution in AZ and ChEMBL data sets are not different from a given level of confidence (p-value = 0.07).

Importantly, public data has a larger number of tests with fewer measurements and unique compounds (Table 2). The number of tests showing significant NA in ChEMBL data is lower (30.3%, higher than 1 log unit) than in AZ inhouse data (57.8%, higher than 0.6 log units). However, ChEMBL tests, in general, contain fewer compounds, therefore the number of DTCs and hence, the chance of a strong NA occurring is lower.

Less than half of the tests (41.7%) in AZ screening and test database are either additive or no DTCs were assembled (Figure 3a). This number is higher in public bioactivity data (69.7%, Figure 3b), which can be explained by the higher threshold of experimental noise and smaller test sizes. Remarkably, in 24% of all AZ inhouse tests show strong NA (above 2 log units), whereas in ChEMBL bioactivity data strong NA is observed in 9.1% of all tests. Yet, various virtual screening studies depend on public datasets and it is crucial to take NA into account whilst judging the performance of predictive models since 1 out of 10 tests might not be additive.

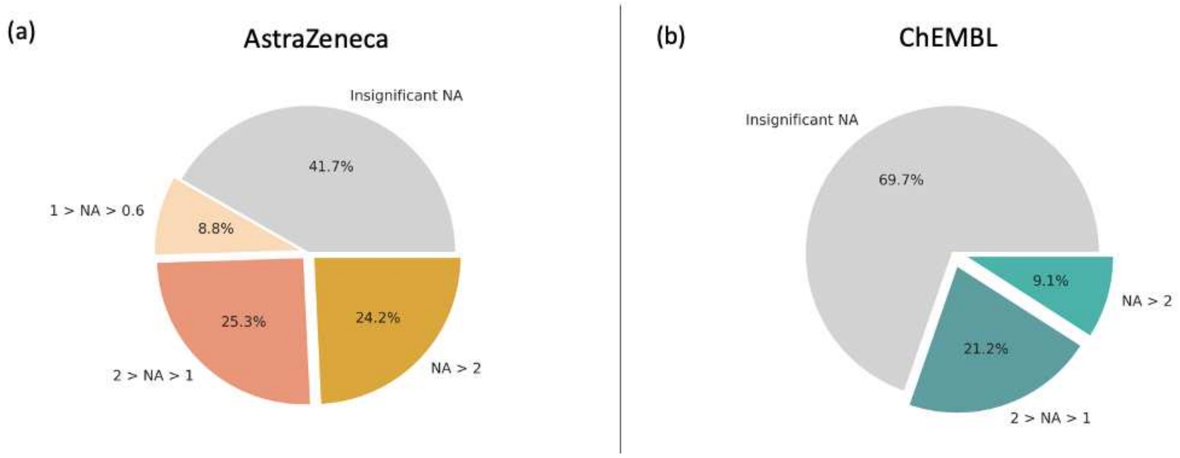


Figure 3. NA distribution among all curated tests from AZ inhouse (a) and public ChEMBL (b) data sets.

Beside the number of tests, NA can also be analyzed for DTCs. On average one out of four and one out of ten DTCs is not additive for AZ inhouse and ChEMBL data respectively (Figure 4a and b). The distribution of NA among DTCs shows significant NA up to 2 log units indicating a gradual decrease in the number of cycles with the increasing NA value (Figure 4c and d).

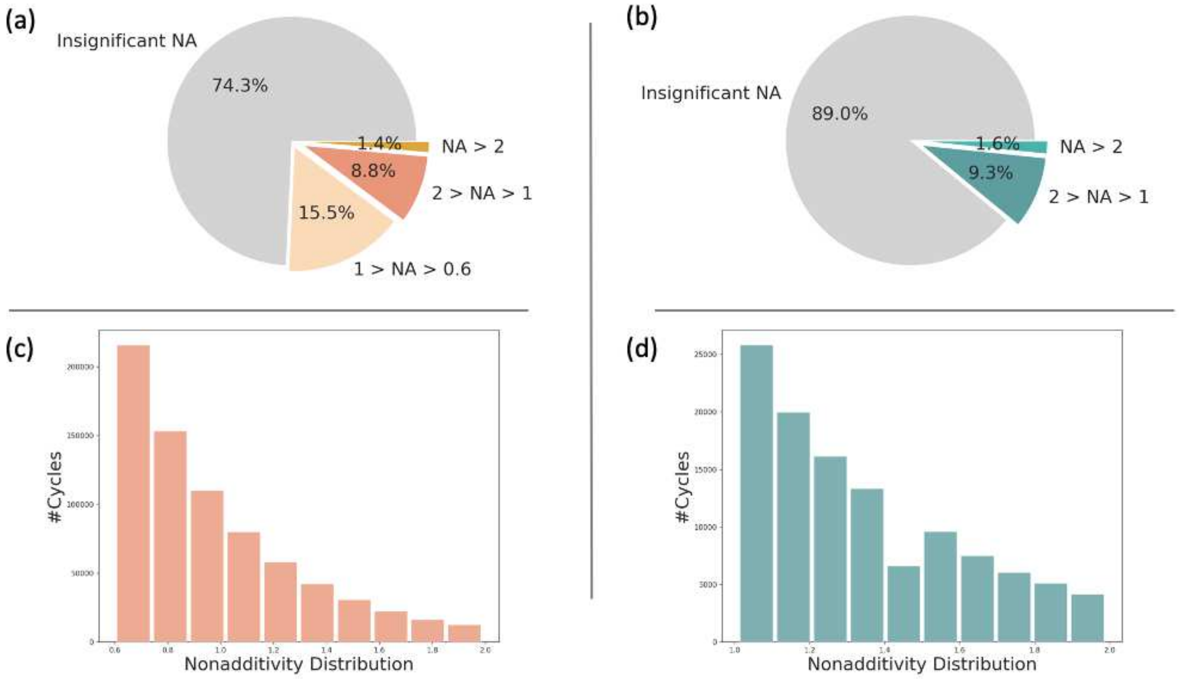


Figure 4. NA distribution for all DTCs among curated tests from AZ (a) and ChEMBL (b) data sets. (c) NA distribution of DTCs showing significant NA score (from 0.6 - up to 2 log units) in AZ (c) and (from 1 - up to 2 log units) ChEMBL (b) bioactivity data.

Out of all compounds 9.4% from AZ and 5.1% from ChEMBL data sets show a significant NA shift (Figure 5). As mentioned before, test sizes and different thresholds for the experimental uncertainty influence these numbers.

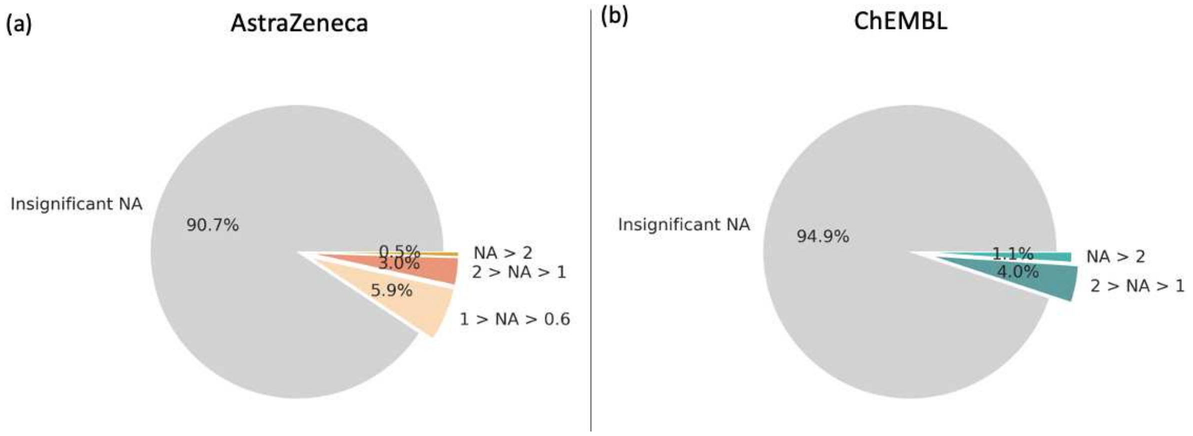


Figure 5. NA distribution among all unique compounds from AZ (a) and ChEMBL (b) data sets.

Bioactivity tests from ChEMBL have a smaller number of compounds and a lower number of DTCs per test. Yet, Figure 6a and b show the shifted distribution of the compounds occurring in double-transformation cycles per test. Surprisingly, there are more than a hundred tests in public data sets in which almost all compounds participate in the assembly of DTCs. This might be due to very small structural variations of tested molecules. AZ inhouse tests tend to be more diverse. Ultimately, testing more compounds results in a lower percentage of unique molecules showing NA. Even though the median number of DTCs is higher in AZ tests, the number of compounds tested in these data sets is also larger, resulting in a relatively lower ratio.

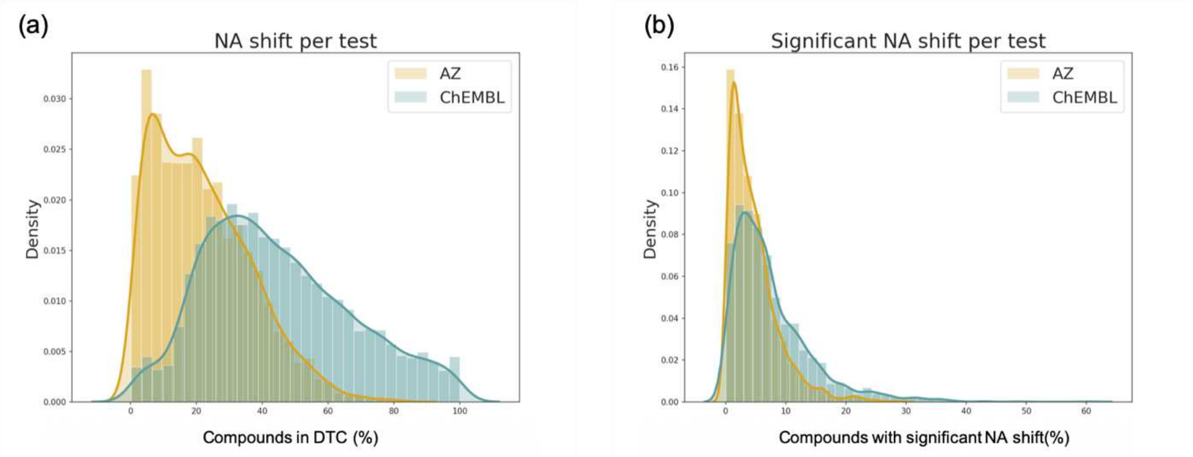


Figure 6. (a) Distribution of the compounds in DTC. (b) Distribution of the compounds showing a significant NA shift per test.

NA distribution according to the number of compounds in tests (Figure 7) indicates that most of the tests in the AZ database contain up to 20,000 compounds and generally smaller tests show higher NA. On average, ChEMBL tests are smaller (Table 2), although several large tests vary in sizes resulting in a more spread out pattern (Figure 7). Herein, highest NA values occur in both small as well as large tests (Figure S1). Furthermore, the density distribution of all tests shows the assembly around the experimental uncertainty.

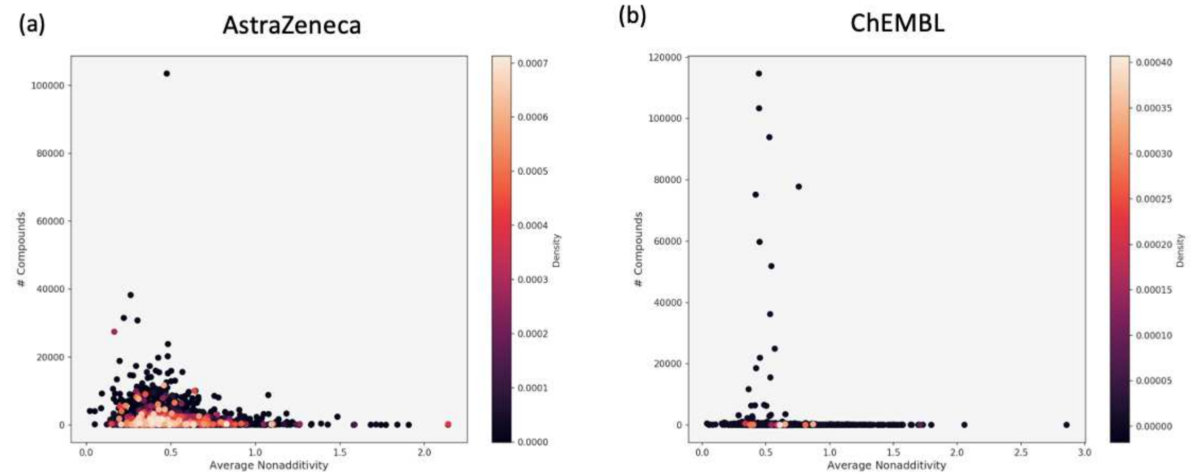


Figure 7. Density distribution of the tests showing significant NA from AZ (a) and ChEMBL (b) based on the average NA and the number of compounds in each test.

CHEMBL1794483 is the largest bioassay obtained from CHEMBL25 (Figure S 1). Initial data of the quantitative high throughput screening for the inhibitors of polymerase Iota contains 115,311 measurements, 33,777 DTCs have been assembled with an average NA score of 0.44. The NA distribution is almost entirely covered by the theoretical normal distribution expected from the experimental noise of 0.5 log units (Figure 8a). The assembled DTCs contain 24,238 compounds and the average additivity shift for each compound is depicted in Figure 8b. In general, it is impossible to point out which molecule causes the NA in a given DTC without further structural information. If the compound occurs in many DTCs with high average NA shift (always with significantly low or high potency), it indicates either a plain error, i.e. a wrong measurement, or structural properties that drastically increase or decrease the compound's biological activity.

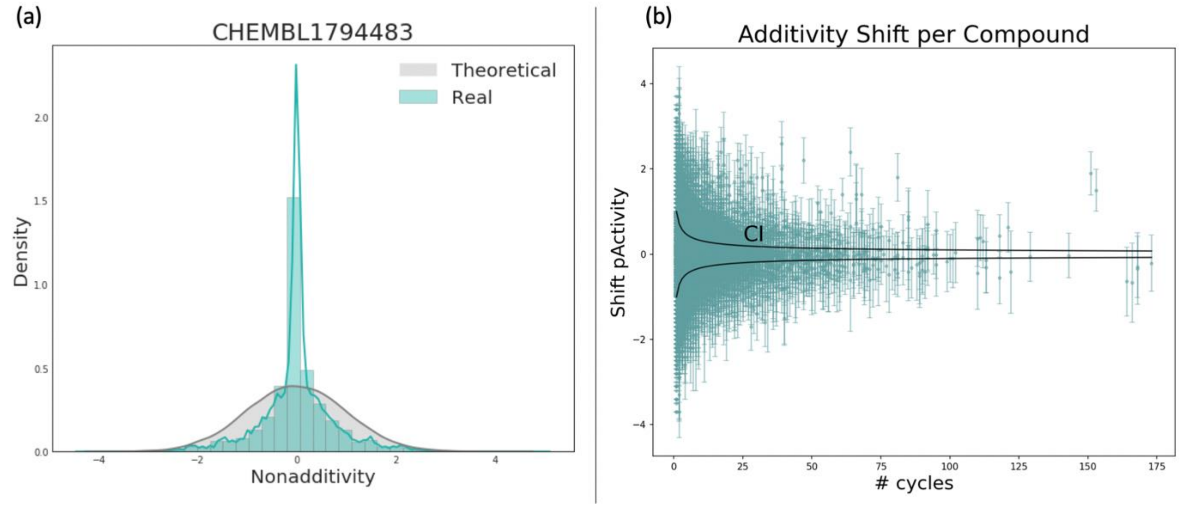


Figure 8. (a) Theoretical NA distribution expected from an experimental uncertainty of 0.5 log units (grey line), and an actual NA distribution for CHEMBL1794483 test (Blue). (b) The average additivity shifts per compound and the standard deviation of the shift for the CHEMBL1794483 data set. Black lines show the confidence interval (CI = 95%) indicating the area where the compounds should appear in case of additivity given the selected threshold of experimental uncertainty (0.5 log units in this case).

Figure 9 shows the DTC from CHEMBL1794483 test with one of the highest NA scores. If the SAR was perfectly additive then the removal of isopropyl group and attaching the benzyl group should have resulted in a significant increase of the potency, yielding pActivity of 8.35. Instead, the activity of the fourth compound even decreased and is lower than compound 1.

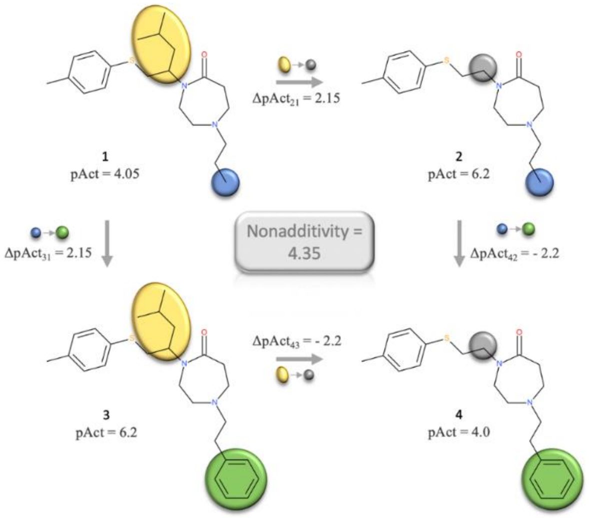


Figure 9. The DTC from CHEMBL1794483 data set with one of the highest NA score (4.35).

OPTUNA

In the second part of the results, the influence of NA on ML performance will be analyzed. Herein, three different ChEMBL assays (Table 4, Figure S 2) were used to analyze the following aspects: (1) Can NA compounds be correctly predicted from a model based on additive data? (2) Does the integration of NA data into training increase model performance?

Table 4. Selected ChEMBL data sets including NA statistics.

ChEMBL data	# Cpds	# Cpds with significant NA (%)	# Cycles	# Cycles with significant NA (%)
1613797	6,236	73 (1.2)	6,245	694 (11.1)
1614027	2,892	69 (2.4)	4,691	582 (12.4)
1613777	3,512	122 (3.5)	8,600	1606 (18.7)

The data sets for the second question was constructed based on the median number of compounds with NA observations (Figure 10). Thus, three sets were constructed for each ChEMBL test containing Q1 (0.6%), median (1.3%) and Q3 (2.6%) of NA compounds. The NA compounds were selected using a stratified split. The NA hold-out set was constructed form the Q3 (2.6%) split, i.e. all models were evaluated on the same subset of observations to ensure comparability.

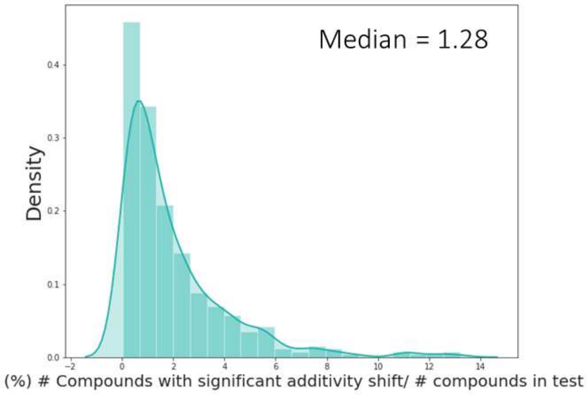


Figure 10. Distribution of NA compounds (%) and the number of DTCs (%) in ChEMBL tests that show NA.

In order to check that any difference in performance is not purely due to a different biological/chemical space, two aspects were checked: (1) the coverage of pIC50 values between training and both test sets and (2) the similarity between the compounds (Figure S 2**Error! Reference source not found.**, Figure 11).

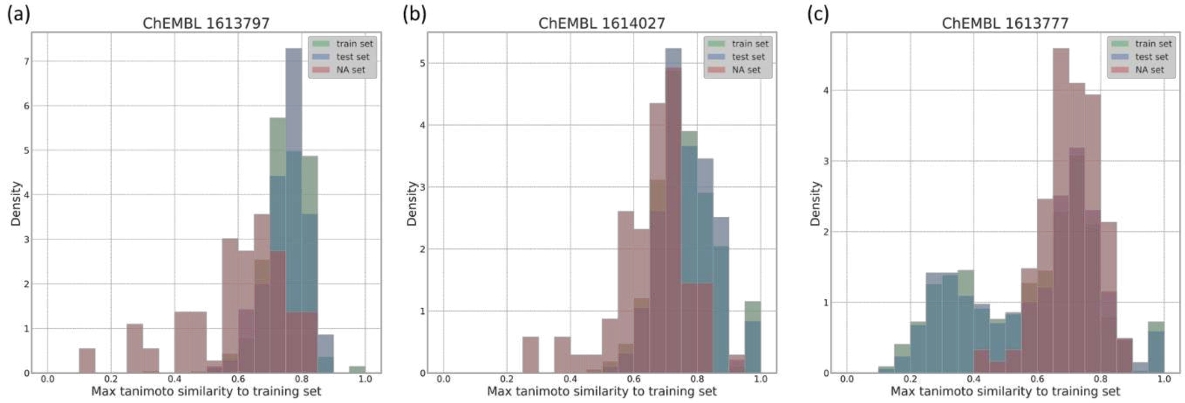


Figure 11. Overlay of tanimoto similarity distributions for training (green), and both test data sets, i.e. additive (blue) and NA (red). Tanimoto similarity was calculated using ECFP6. For training set similarity the identity for the molecule was excluded for its similarity calculation.

Based on the automatic hyper-parameter training using Optuna, RF and SVM models were generated for all three ChEMBL data sets. Both RF and SVM show similar performances for each ChEMBL data set, while SVM performance was more volatile to the actual choice of hyper-parameters. While the RF model for ChEMBL1614027 performed well on training and additive test data (Table 5, Figure 12), the models for the other two data sets performed less good with R^2=0.63/0.06 and R^2=0.72/0.24 for CHEMBL1613797 and ChEMBL1613777 training/test data respectively (Table 5, Figure S 3, and Figure S 4). Importantly, for both models (RF and SVM), as well as all three data sets the performance on NA test data consistently dropped, with the RFs typically performing slightly better than the SVM model (Table 5, Figure S 5-7). In addition to the correlation between experimental and predicted data the predicted error (RMSE) increases for all NA data sets.

Table 5. RF and SVM model performance measures.

ChEMBL data (#measures)	RF					SVM				
	Train r2 (RMSE)	Test r2 (RMSE)		Test MCC		Train r2 (RMSE)	Test r2 (RMSE)		Test MCC	
		A*	NA#	A*	NA#		A*	NA#	A*	NA#
1613797 (772)	0.63 (0.22)	**0.06** **(0.33)**	-0.27 (1.19)	0.06	**0.22**	0.51 (0.26)	**0.05** **(0.33)**	-0.35 (1.22)	**0.14**	0.07
1614027 (1024)	0.94 (0.15)	**0.68** **(0.34)**	-0.29 (1.26)	**0.53**	0.20	0.94 (0.15)	**0.68** **(0.34)**	-0.29 (1.26)	**0.54**	0.08
1613777 (3511)	0.72 (0.42)	**0.24** **(0.69)**	-0.37 (1.29)	**0.40**	-0.01	0.84 (0.32)	**0.24** **(0.69)**	-0.47 (1.33)	**0.49**	0.00

Testdata with (*) additive and (#) NA data only

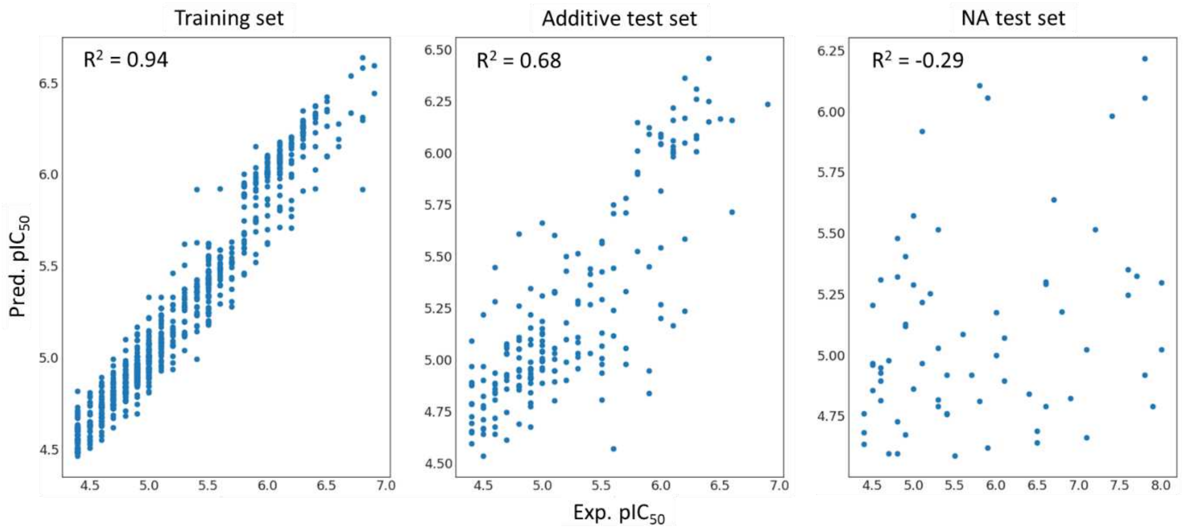


Figure 12. Correlation plots with RF predictions for ChEMBL1614027.

Furthermore, a binary classification of the predicted values was done and the MCC was calculated as well as confusion matrices generated. Both show that it is much harder to accurately predict the NA test sets (Figure S 8, Figure S 9).

In a subsequent test, NA data was added to the training data to evaluate whether this could improve the prediction for NA data (Table 6, Figure S 10). For these "mixin" trials, it appears that for all ratios and all datasets there is no significant difference in performance. This might be either because it has a hard time learning from those examples or because they are too few in number.

Table 6. Performance measures for binary classification.

ChEMBL	RF (MCC for test)			
data	Q0 (0.0%)*	Q1 (0.6%)*	Median (1.3%)*	Q3 (2.6%)*
1613797	0.22	0.16	0.16	0.16
1614027	0.20	0.20	0.12	0.10
1613777	-0.01	0.11	-0.03	-0.05

* Test set size for Q0 differs from Q1/Median/Q3.

DISCUSSION

The project aimed to analyze the occurrence of NA in public and inhouse data and its influence on machine learning performance.

One of the biggest challenges during this process is the data pre-processing to make both sets comparable. Thus, additional cleaning steps were applied to ChEMBL bioactivity data, such as filtering by the target confidence score to increase the data reliability. The final ‘cleaned’ dataset depends on the experience and decision-making of the researcher to correctly choose which tests are compatible with the analysis.

The size restriction of the molecules was based on the structural transformations and similarities, the upper limit of the molecular size included and exchanged during the transformations must be set carefully. 70 heavy atoms and the transformation of a maximum 1/3rd of the molecule were established. Without having these limitations, the following issues may arise: (1) large molecules, such as peptides are not compatible with the NA analysis since it is impossible to track small functional groups; (2) performing calculations on large molecules is computationally expensive; (3) in cases where the functional group represents 60% of the molecule will most likely result in NA since almost the whole compound is transformed and the corresponding binding mode is likely to change.

In addition to the molecules size restrictions, also test size after all the data-cleaning steps is crucial. On one hand, small tests should be discarded, because there is a lower probability of DTCs assembling. In this research project, 25 was set as the lowest number of unique compounds per test. Since most of the tests are small (half of the measurements in both inhouse and public data sets were concentrated in a few hundred assays only), it also influences the general statistics resulting in no NA output. One might argue that the majority of the tests are additive, however, most of them are too small to draw any meaningful conclusions regarding their NA.

According to the results, significant nonlinearity occurs once in every second test in AZ inhouse and once in every third biological and physicochemical tests in ChEMBL databases. Importantly, significant nonadditive events are less frequent in public data sets. The reasons for it can be: (1) potential bias in reporting single series or positive SAR results; (2) the smaller size of public bioactivity tests, resulting in less DTCs; (3) a higher threshold of the experimental uncertainty for the entire data, as some tests have significantly higher experimental noise. An additional influence is the reliability of the compounds measurements. Since in the inhouse database a majority of compounds is measured several times in each test the measurements are more reliable. This is not the case in the public data sets, where only 5% of the compounds are measured more than once in each test.

Prior to the analysis, it is crucial to carefully set the thresholds for the experimental noise to point out true NA cases. Strong NA stands out from the rest of the data and it is easy to spot, while weaker NA is usually blended with the experimental noise. As described by Kramer *et* *al.*[6] NA analysis can estimate the upper limit of an experimental uncertainty for specific biochemical assays, which is crucial in differentiating true NA from the assay artefacts. However, it is less straight forward to select the threshold for large data. While experimental noise among most of the inhouse tests might be 0.2 log units, there are still some assays with larger errors. The problem with the higher limits of the experimental noise is the higher amount of insignificant NA cases. By choosing 0.5 log units for public data, we potentially cover all the assay artefacts, still, we might have ignored potentially true NA cases.

NA can be a problem for linear SAR techniques. Yet, if used intentionally, it can be an important tool for drug discovery. This study provides a detailed picture of the NA pattern amongst the inhouse and public databases, providing the global distribution of nonlinear events amongst tests and unique compounds. A careful understanding of the data is the key to successful decision-making. By conducting NA analysis one can easily identify outliers, detect potential assay artefacts, or key conformational changes. It is crucial to understand the possible experimental noise, that can be underlying most of NA cases because of assay noise. Therefore, one must always keep in mind the origin of a given assay, the reliability of the measurements, and a possible upper limit of experimental uncertainty.

By systematically incorporating the NA analysis into the drug discovery projects, detection of interesting interactions and key SAR features will be easier and will eventually provide more structural insights for rational drug design.

CONCLUSIONS

Identifying NA in the SAR data sets can be crucial by suggesting important structural features for the compound optimization. However, nonadditive events can be caused by the random addition of experimental uncertainty, which is important to consider during the interpretation of results. The impact of the experimental noise increases with the size of the test, as more double-transformation cycles can be assembled. NA analysis in the AZ compound database suggests that significant nonlinear events are more frequent in AZ inhouse data than public ChEMBL data. By considering only public data one might assume that a NA is a rare event and important cases can be neglected. AZ data points out the fact that this is not true and the statistical framework of the NA analysis should be systematically implemented in SAR projects and discussed in publications for rational drug design.

Despite that we retrospectively cannot figure out if the optimization lead to a general increase or decrease in the activity, from MMP studies we know that 100-fold improvements are very rare events of about 1%.[64] Our numbers 1-3% point to the fact that electrostatic or steric problems occur more frequently than expected from SAR data because of the undersampling of negative data. This undersampling might be a reason why QSAR models have problems with describing activity cliffs despite being often based on non-linear algorithms. This would also be useful for setting a baseline of performance to be expected from such models.

Currently, the sign of a NA value does not provide valuable information since the order of compounds does not indicate the effect of a given transformations. In other words, one cannot guess which feature leads to the gain or loss of the activity from a specific double-transformation cycle. It would add another level of information to see the pattern of NA distribution in terms of boosting or decreasing the biological effect, whether the cases are equal or mostly lead to the loss of biological activity.

REFERENCES

1. Free SM, Wilson JW (1964) A mathematical contribution to structure-activity studies. J Med Chem 7:395–399

2. Cramer RD, Wendt B (2014) Template CoMFA: The 3D-QSAR Grail? J Chem Inf Model 54:660–671

3. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs

(MMPs) in large data sets. J Chem Inf Model 50:339–348

4. Patel Y, Gillet VJ, Howe T, et al (2008) Assessment of additive/nonadditive effects in structure− activity relationships: implications for iterative drug design. J Med Chem 51:7552–7562

5. Wang L, Wu Y, Deng Y, et al (2015) Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. J Am Chem Soc 137:2695–2703. https://doi.org/10.1021/ja512751q

6. Kramer C (2019) Nonadditivity Analysis. J Chem Inf Model 59:4034–4042. https://doi.org/10.1021/acs.jcim.9b00631

7. Dimova D, Bajorath J (2016) Advances in activity cliff research. Mol Inform 35:181–191

8. Dimova D, Heikamp K, Stumpfe D, Bajorath J (2013) Do medicinal chemists learn from activity cliffs? A systematic evaluation of cliff progression in evolving compound data sets. J Med Chem 56:3339–3345

9. Hu Y, Stumpfe D, Bajorath J (2013) Advancing the activity cliff concept. F1000Research 2:

10. Mobley DL, Gilson MK (2017) Predicting binding free energies: frontiers and benchmarks. Annu Rev Biophys 46:531–558

11. Hu H, Bajorath J (2020) Introducing a new category of activity cliffs combining different compound similarity criteria. RSC Med Chem

12. Abramyan TM, An Y, Kireev D (2019) Off-Pocket Activity Cliffs: A Puzzling Facet of Molecular Recognition. J Chem Inf Model

13. Andrews SP, Mason JS, Hurrell E, Congreve M (2014) Structure-based drug design of chromone antagonists of the adenosine A2A receptor. Medchemcomm 5:571–575. https://doi.org/10.1039/C3MD00338H

14. Schönherr H, Cernak T (2013) Profound Methyl Effects in Drug Discovery and a Call for New C-H Methylation Reactions. Angew Chemie Int Ed 52:12256–12267

15. Kramer C, Fuchs JE, Liedl KR (2015) Strong nonadditivity as a key structure-activity relationship feature: Distinguishing structural changes from assay artifacts. J Chem Inf Model 55:483–494. https://doi.org/10.1021/acs.jcim.5b00018

16. Gomez L, Xu R, Sinko W, et al (2018) Mathematical and Structural Characterization of Strong Nonadditive Structure–Activity Relationship Caused by Protein Conformational Changes. J Med Chem 61:7754–7766

17. Baum B, Muley L, Smolinski M, et al (2010) Non-additivity of functional group contributions in protein–ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry. J Mol Biol 397:1042–1054

18. McClure K, Hack M, Huang L, et al (2006) Pyrazole CCK1 receptor antagonists. Part 1: Solution-phase library synthesis and determination of Free–Wilson additivity. Bioorg Med Chem Lett 16:72–76

19. Sehon C, McClure K, Hack M, et al (2006) Pyrazole CCK1 receptor antagonists. Part 2: SAR studies by solid-phase library synthesis and determination of Free–Wilson additivity. Bioorg Med Chem Lett 16:77–80

20. Hilpert K, Ackermann J, Banner DW, et al (2002) Design and synthesis of potent and highly selective thrombin inhibitors. J Med Chem 37:3889–3901

21. Lübbers T, Böhringer M, Gobbi L, et al (2007) 1, 3-disubstituted 4-aminopiperidines as useful tools in the optimization of the 2-aminobenzo [a] quinolizine dipeptidyl peptidase IV inhibitors. Bioorg Med Chem Lett 17:2966–2970

22. Leung CS, Leung SSF, Tirado-Rives J, Jorgensen WL (2012) Methyl effects on protein–ligand binding. J Med Chem 55:4489–4500

23. Abeliovich H (2005) An empirical extremum principle for the hill coefficient in ligand-protein interactions showing negative cooperativity. Biophys J 89:76–79

24. Dill KA (1997) Additivity principles in biochemistry. J Biol Chem 272:701–704

25. Camara-Campos A, Musumeci D, Hunter CA, Turega S (2009) Chemical double mutant cycles for the quantification of cooperativity in H-bonded complexes. J Am Chem Soc 131:18518–18524

26. Cockroft SL, Hunter CA (2007) Chemical double-mutant cycles: dissecting non-covalent interactions. Chem Soc Rev 36:172–188

27. Babaoglu K, Shoichet BK (2006) Deconstructing fragment-based inhibitor discovery. Nat Chem Biol 2:720–723

28. Miller BG, Wolfenden R (2002) Catalytic proficiency: the unusual case of OMP decarboxylase. Annu Rev Biochem 71:847–885

29. Hajduk PJ, Sheppard G, Nettesheim DG, et al (1997) Discovery of potent nonpeptide inhibitors of stromelysin using SAR by NMR. J Am Chem Soc 119:5818–5827

30. Congreve MS, Davis DJ, Devine L, et al (2003) Detection of ligands from a dynamic combinatorial library by X-ray crystallography. Angew Chemie Int Ed 42:4479–4482

31. Sharrow SD, Edmonds KA, Goodman MA, et al (2005) Thermodynamic consequences of disrupting a water-mediated hydrogen bond network in a protein: pheromone complex. Protein Sci 14:249–256

32. Muley L, Baum B, Smolinski M, et al (2010) Enhancement of hydrophobic interactions and hydrogen bond strength by cooperativity: synthesis, modeling, and molecular dynamics simulations of a congeneric series of thrombin inhibitors. J Med Chem 53:2126–2135

33. Kuhn B, Mohr P, Stahl M (2010) Intramolecular hydrogen bonding in medicinal chemistry. J Med Chem 53:2601–2611. https://doi.org/10.1021/jm100087s

34. Segler MHS, Kogej T, Tyrchan C, Waller MP (2018) Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci 4:120–131. https://doi.org/10.1021/acscentsci.7b00512

35. Arús-Pous J, Blaschke T, Ulander S, et al (2019) Exploring the GDB-13 chemical space using deep generative models. J Cheminform 11:20. https://doi.org/10.1186/s13321-019-0341-z

36. Blaschke T, Arús-Pous J, Chen H, et al (2020) REINVENT 2.0 – an AI Tool for De Novo Drug Design. https://doi.org/10.26434/CHEMRXIV.12058026.V2

37. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. J Cheminform 9:48. https://doi.org/10.1186/s13321-017-0235-x

38. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P (2018) Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. Bioinformatics 34:3666–3674. https://doi.org/10.1093/bioinformatics/bty374

39. Gomes J, Ramsundar B, Feinberg EN, Pande VS (2017) Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity

40. Feinberg EN, Sur D, Wu Z, et al (2018) PotentialNet for Molecular Property Prediction. ACS Cent Sci 4:1520–1530. https://doi.org/10.1021/acscentsci.8b00507

41. Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G (2018) KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. J Chem Inf Model 58:287–296. https://doi.org/10.1021/acs.jcim.7b00650

42. Wójcikowski M, Ballester PJ, Siedlecki P (2017) Performance of machine-learning scoring functions in structure-based virtual screening. Sci Rep 7:1–10. https://doi.org/10.1038/srep46710

43. Ragoza M, Hochuli J, Idrobo E, et al (2017) Protein-Ligand Scoring with Convolutional Neural Networks. J Chem Inf Model 57:942–957. https://doi.org/10.1021/acs.jcim.6b00740

44. Pereira JC, Caffarena ER, Dos Santos CN (2016) Boosting Docking-Based Virtual Screening with Deep Learning. J Chem Inf Model 56:2495–2506. https://doi.org/10.1021/acs.jcim.6b00355

45. Wallach I, Dzamba M, Heifets A (2015) AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery

46. Ballester PJ, Mitchell JBO (2010) A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics 26:1169–1175. https://doi.org/10.1093/bioinformatics/btq112

47. Kayala MA, Baldi P (2012) ReactionPredictor: Prediction of complex chemical reactions at the mechanistic level using machine learning. J Chem Inf Model 52:2526–2540. https://doi.org/10.1021/ci3003039

48. Struble TJ, Alvarez JC, Brown SP, et al (2020) Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. J Med Chem. https://doi.org/10.1021/acs.jmedchem.9b02120

49. Segler MHS, Waller MP (2017) Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. Chem - A Eur J 23:5966–5971. https://doi.org/10.1002/chem.201605499

50. Schwaller P, Gaudin T, Lányi D, et al (2018) “Found in Translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chem Sci 9:6091–6098. https://doi.org/10.1039/c8sc02339e

51. Landrum G (2006) RDKit: Open-source cheminformatics

52. Dalke A, Hert J, Kramer C (2018) mmpdb: An Open-Source Matched Molecular Pair Platform for Large Multiproperty Data Sets. J Chem Inf Model 58:902–910. https://doi.org/10.1021/acs.jcim.8b00173

53. Gaulton A, Hersey A, Nowotka M, et al (2017) The ChEMBL database in 2017. Nucleic Acids Res 45:D945–D954

54. Akiba T, Sano S, Yanase T, et al (2019) Optuna: A Next-generation Hyperparameter Optimization Framework. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, pp 2623–2631

55. Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830

56. Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21:6. https://doi.org/10.1186/s12864-019-6413-7

57. Kramer C, Kalliokoski T, Gedeck P, Vulpetti A (2012) The experimental uncertainty of heterogeneous public K i data. J Med Chem 55:5165–5173. https://doi.org/10.1021/jm300131x

58. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P (2013) Comparability of mixed IC50 data–a statistical analysis. PLoS One 8:

59. Kramer C, Dahl G, Tyrchan C, Ulander J (2016) A comprehensive company database analysis of biological assay variability. Drug Discov. Today 21:1213–1221

60. Kolmogorov-Smirnov AN, Kolmogorov A, Kolmogorov M (1933) Sulla determinazione empírica di uma legge di distribuzione

61. Smirnov N (1948) Table for estimating the goodness of fit of empirical distributions. Ann Math Stat 19:279–281

62. Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47:583–621

63. Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 50–60

64. Hajduk PJ, Sauer DR (2008) Statistical analysis of the effects of common chemical substituents on ligand potency. J Med Chem 51:553–564. https://doi.org/10.1021/jm070838y

List of abbreviations

NAA: NA analysis; AZ: AstraZeneca; SAR: Structure-activity relationship; QSAR: Quantitative structure-activity relationship; AI: Artificial intelligence; ML: Machine learning; FW: Free-Wilson; MMPA: Matched molecular pair analysis; FBDD: Fragment-based drug discovery; CADD: Computer-aided drug design; SMILES: Simplified molecular-input line-entry system; SBDD: Structure-based drug design; RF: Random forest; SVM: Support vector machine.

Acknowledgements

Gratitudes towards Uppsala University, Dr. Lena Åslund and the colleagues from IMIM program for supporting the master thesis of DG.

Authors' contributions

DG performed data curation, NA analysis and wrote the paper. CM realized the ML study and wrote the paper. EN and CT supervised the study and wrote the paper. All the authors read and approved the final manuscript.

Funding

DG is supported financially by Erasmus Mundus Joint Master Degree scholarship 2018-2020 and AstraZeneca Master student program.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files.

- S1: Additional figures.
- The Jupyter notebook for data preparation and NA analysis is available on GitHub (https://github.com/MolecularAI/NonadditivityAnalysis).
- ChEMBL data sets (ChEMBL1613777/1613797/1614027) with obtained NA values for ML approach

are available as csv files.

Nonadditivity analysis code was made available by Christian Kramer on GitHub (https://github.com/KramerChristian/NonadditivityAnalysis).

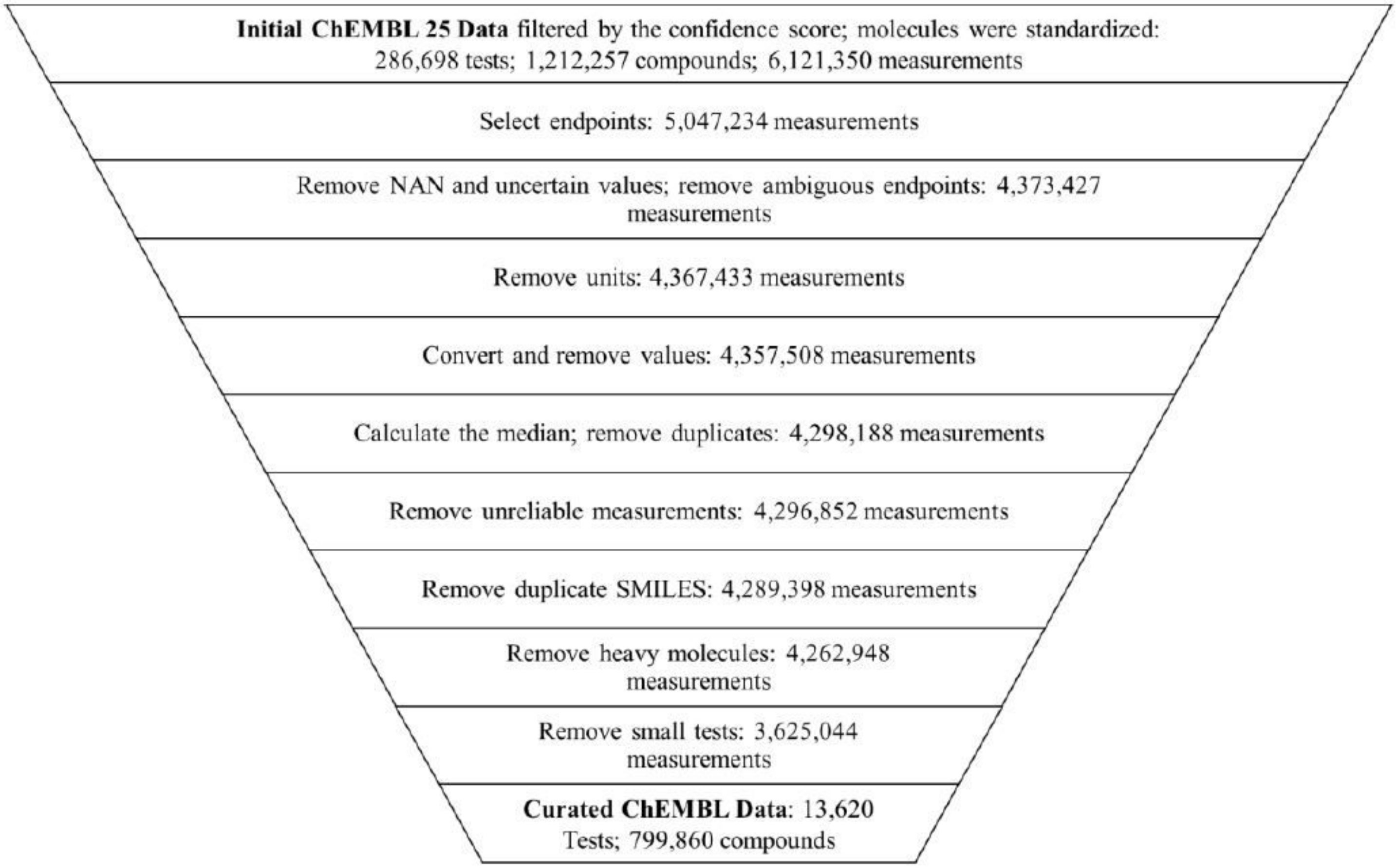
Competing interests

The authors declare that they have no competing interests. CM, CT and EN are employees of AstraZeneca and own stock options.

Current address

¤ Dea Gogishvili, Department of Computer Science, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.

Figures



Initial ChEMBL 25 Data filtered by the confidence score; molecules were standardized: 286,698 tests; 1,212,257 compounds; 6,121,350 measurements

Select endpoints: 5,047,234 measurements

Remove NAN and uncertain values; remove ambiguous endpoints: 4,373,427 measurements

Remove units: 4,367,433 measurements

Convert and remove values: 4,357,508 measurements

Calculate the median; remove duplicates: 4,298,188 measurements

Remove unreliable measurements: 4,296,852 measurements

Remove duplicate SMILES: 4,289,398 measurements

Remove heavy molecules: 4,262,948 measurements

Remove small tests: 3,625,044 measurements

Curated ChEMBL Data: 13,620 Tests; 799,860 compounds

Figure 1

The data curation process of public ChEMBL25 data representing number of measurements after each cleaning step.

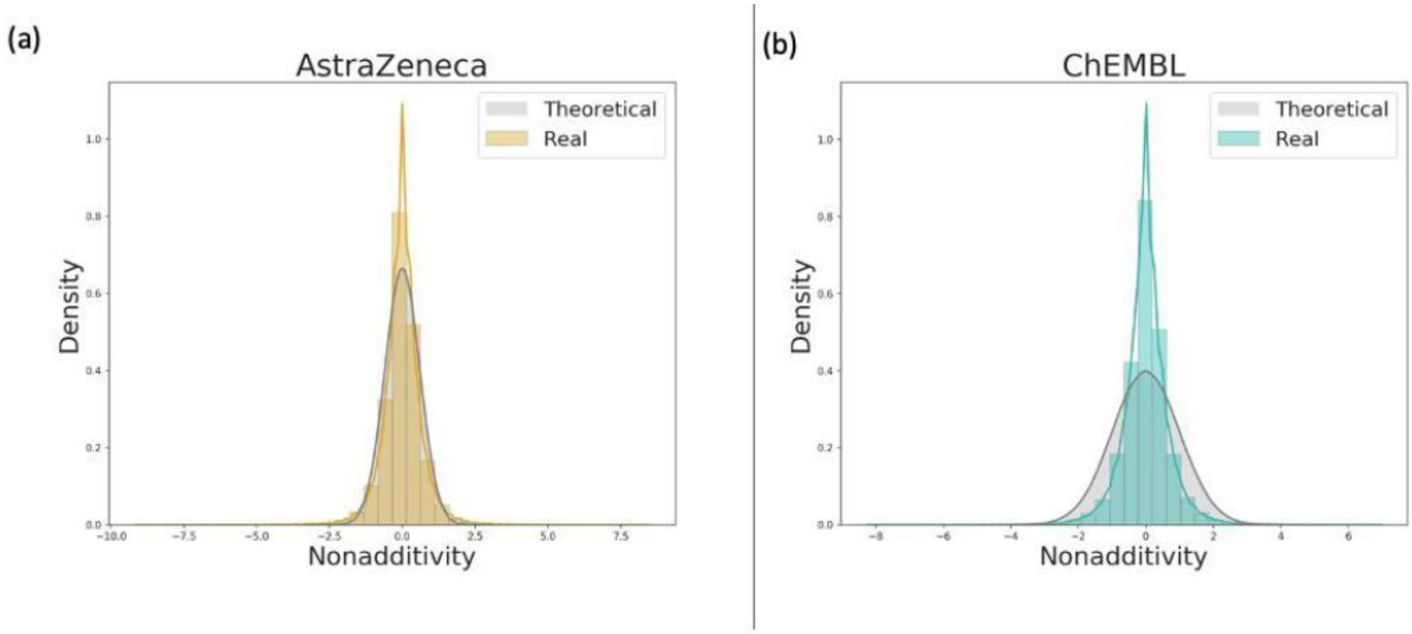


Figure 2

Theoretical NA distribution expected from an experimental uncertainty of (a) 0.3 and (b) 0.5 log units (grey lines), and observed NA distribution for all (a) AZ (yellow) tests and (b) ChEMBL (blue) tests.

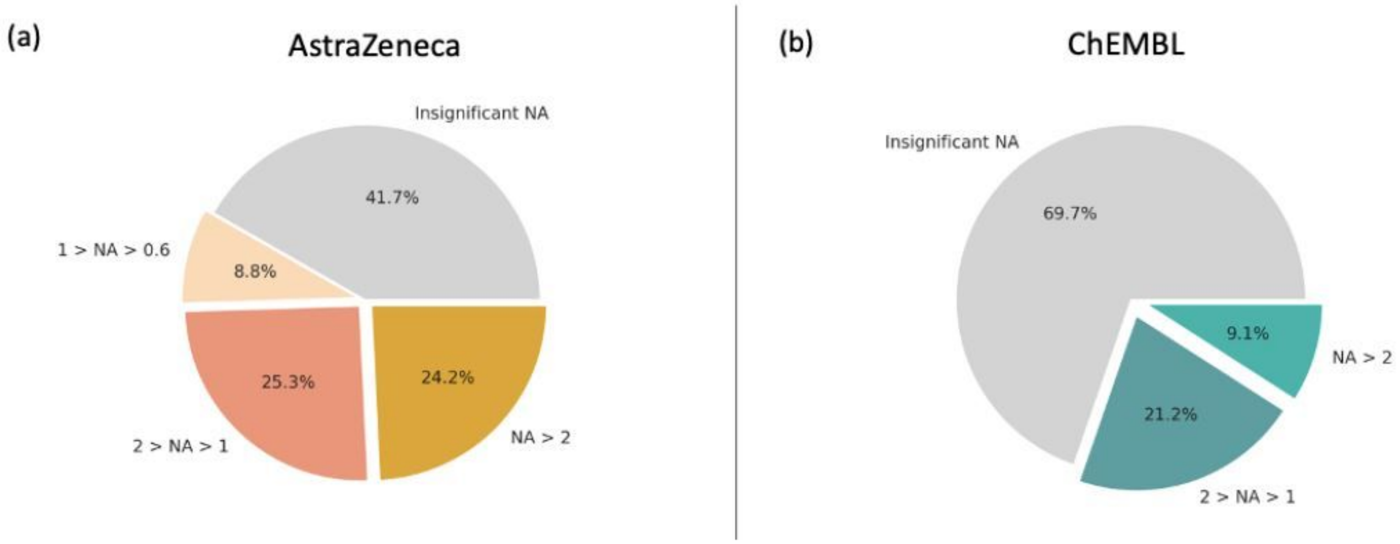


Figure 3

NA distribution among all curated tests from AZ inhouse (a) and public ChEMBL (b) data sets.

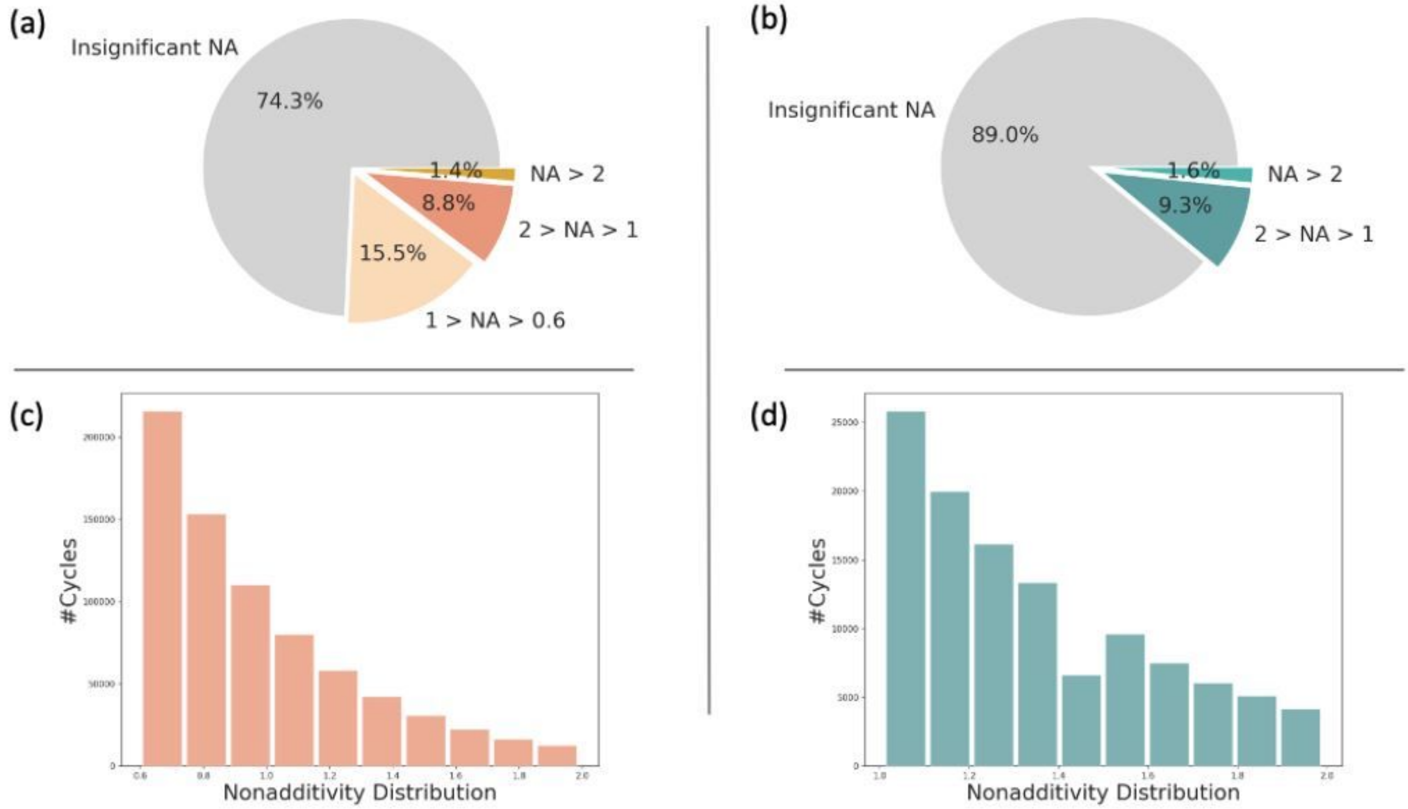


Figure 4

NA distribution for all DTCs among curated tests from AZ (a) and ChEMBL (b) data sets. (c) NA distribution of DTCs showing significant NA score (from 0.6 - up to 2 log units) in AZ (c) and (from 1 - up to 2 log units) ChEMBL (b) bioactivity data.

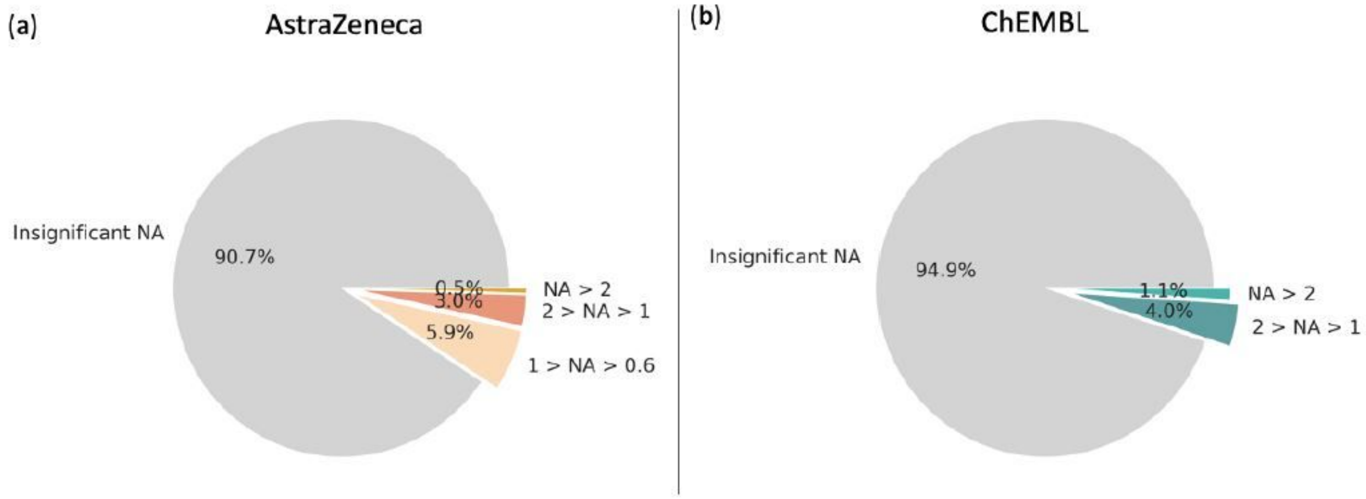


Figure 5

NA distribution among all unique compounds from AZ (a) and ChEMBL (b) data sets.

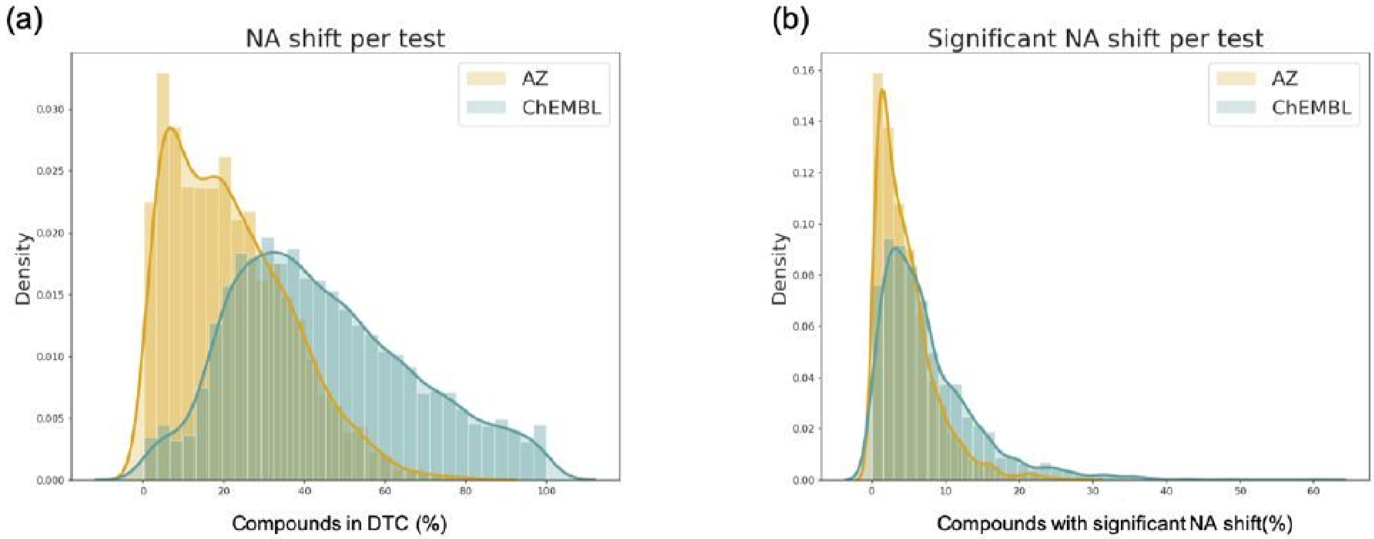


Figure 6

(a) Distribution of the compounds in DTC. (b) Distribution of the compounds showing a significant NA shift per test.

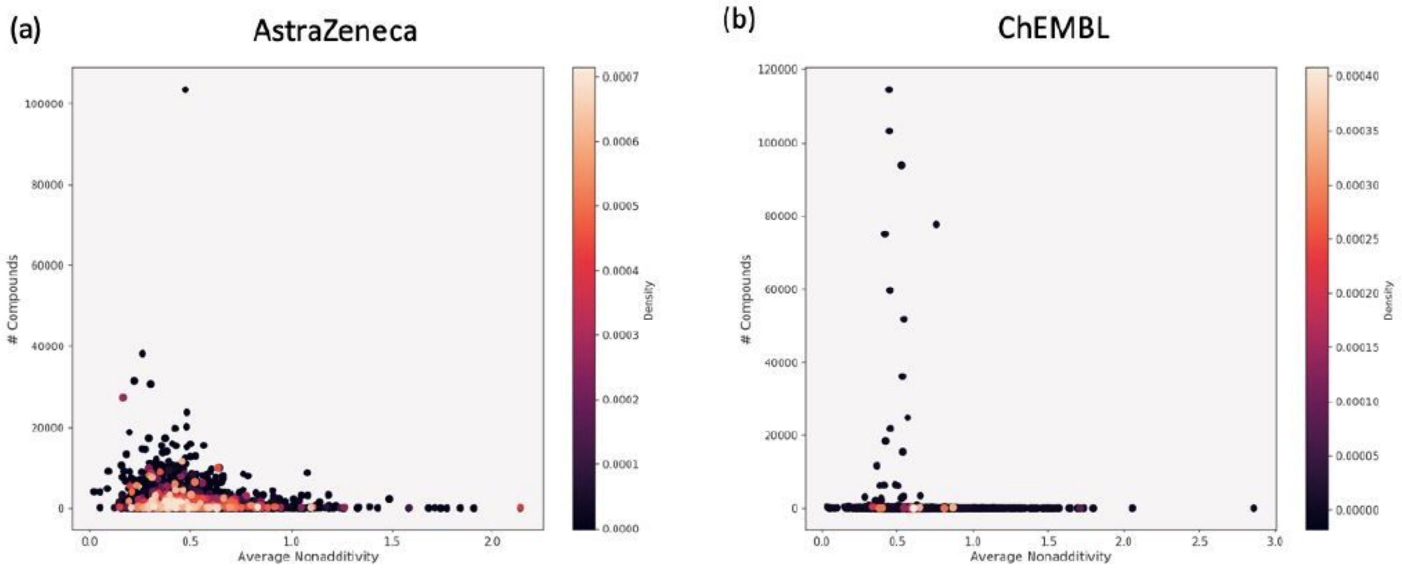


Figure 7

Density distribution of the tests showing significant NA from AZ (a) and ChEMBL (b) based on the average NA and the number of compounds in each test.

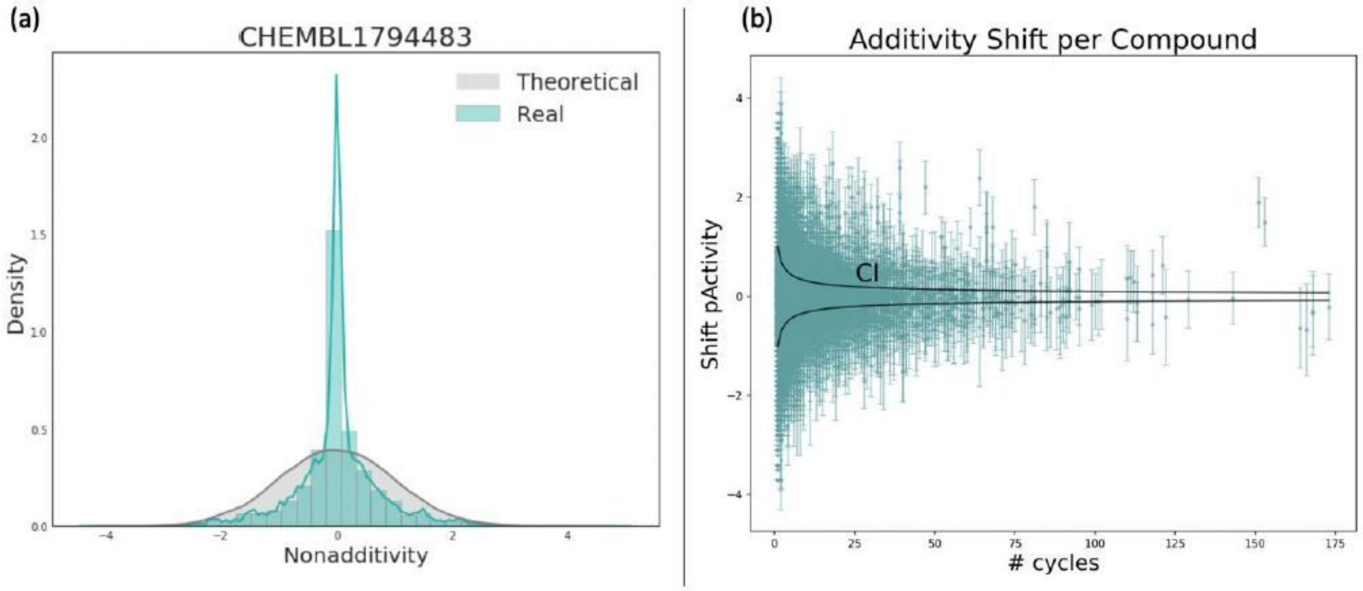


Figure 8

(a) Theoretical NA distribution expected from an experimental uncertainty of 0.5 log units (grey line), and an actual NA distribution for CHEMBL1794483 test (Blue). (b) The average additivity shifts per compound and the standard deviation of the shift for the CHEMBL1794483 data set. Black lines show the confidence interval (CI = 95%) indicating the area where the compounds should appear in case of additivity given the selected threshold of experimental uncertainty (0.5 log units in this case).

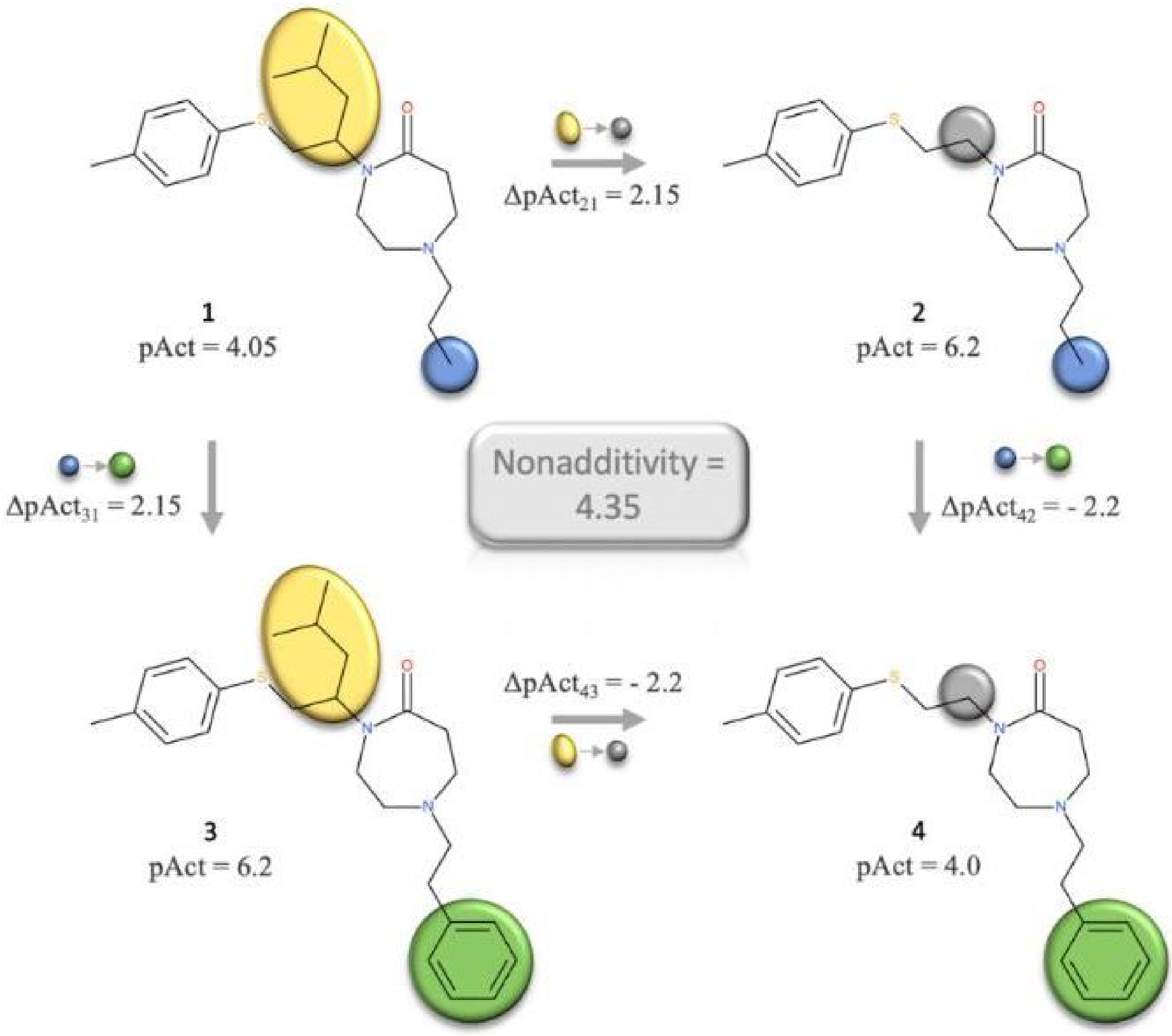

$\Delta pAct_{21} = 2.15$
1
pAct = 4.05
2
pAct = 6.2
$\Delta pAct_{31} = 2.15$
Nonadditivity = 4.35
$\Delta pAct_{42} = - 2.2$
$\Delta pAct_{43} = - 2.2$
3
pAct = 6.2
4
pAct = 4.0

Figure 9

The DTC from CHEMBL1794483 data set with one of the highest NA score (4.35).

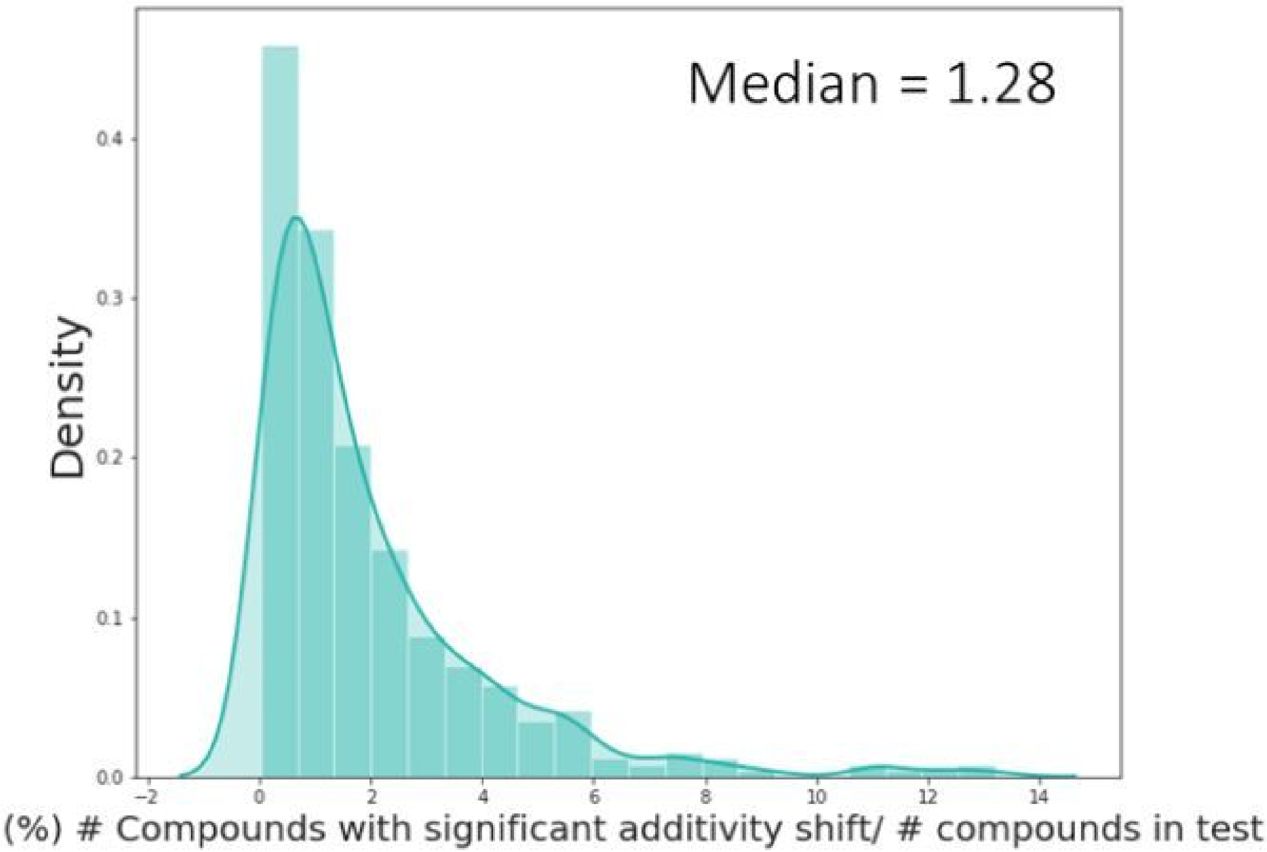


Figure 10

Distribution of NA compounds (%) and the number of DTCs (%) in ChEMBL tests that show NA.

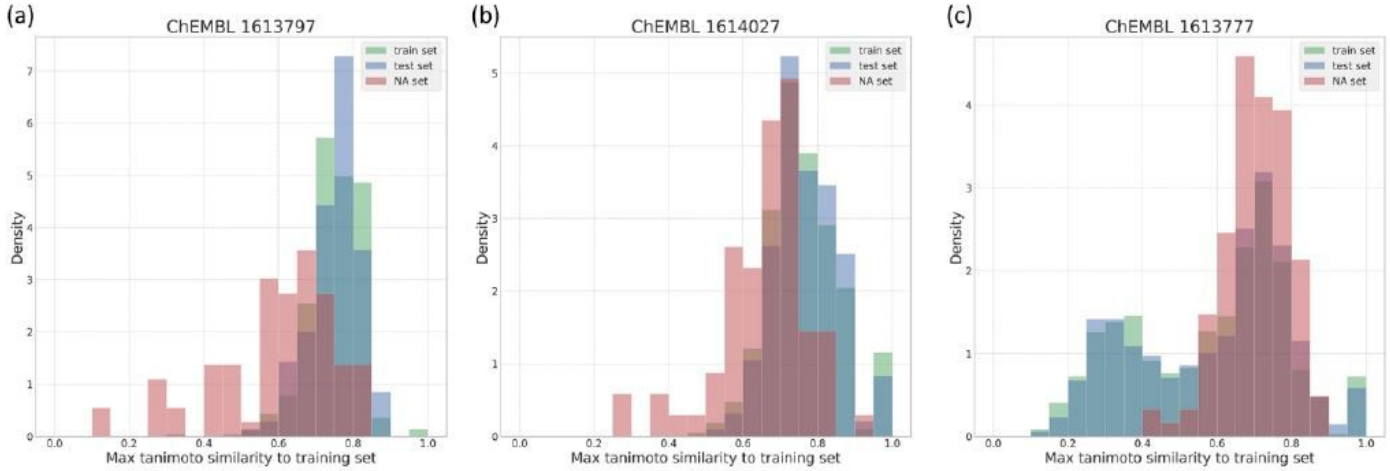


Figure 11

Overlay of tanimoto similarity distributions for training (green), and both test data sets, i.e. additive (blue) and NA (red). Tanimoto similarity was calculated using ECFP6. For training set similarity the identity for the molecule was excluded for its similarity calculation.

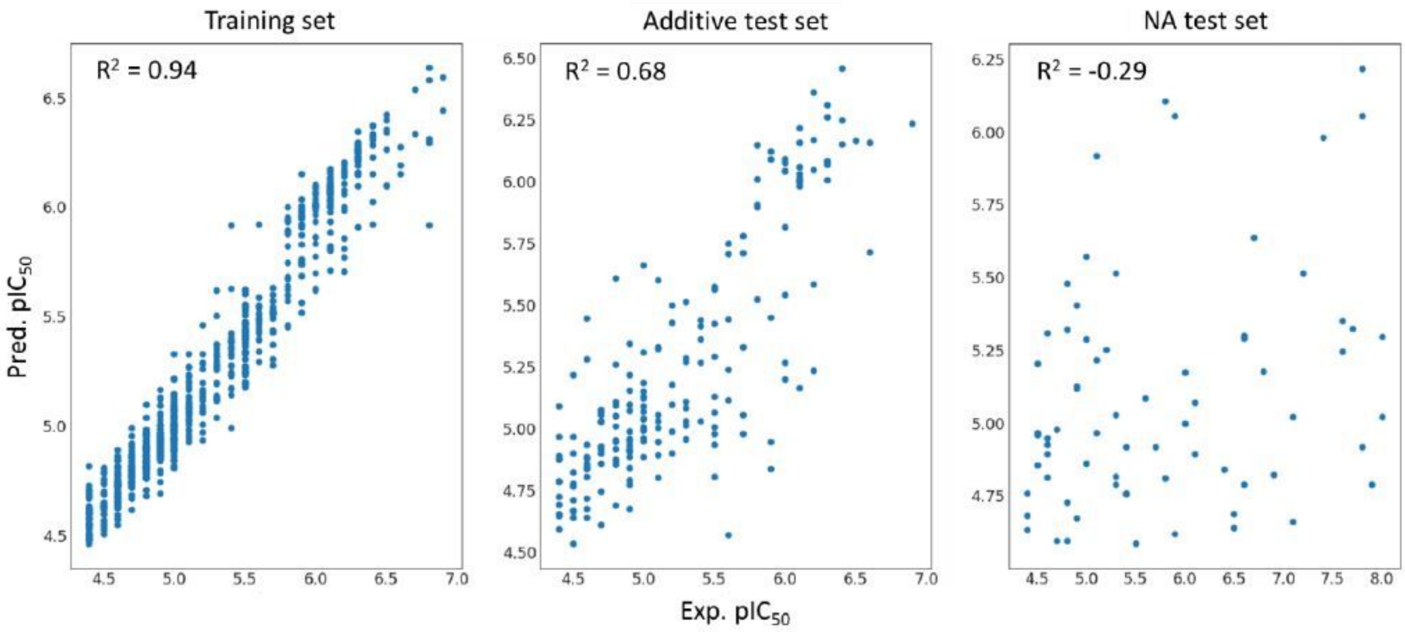


Figure 12

Correlation plots with RF predictions for ChEMBL1614027.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NAASI.pdf