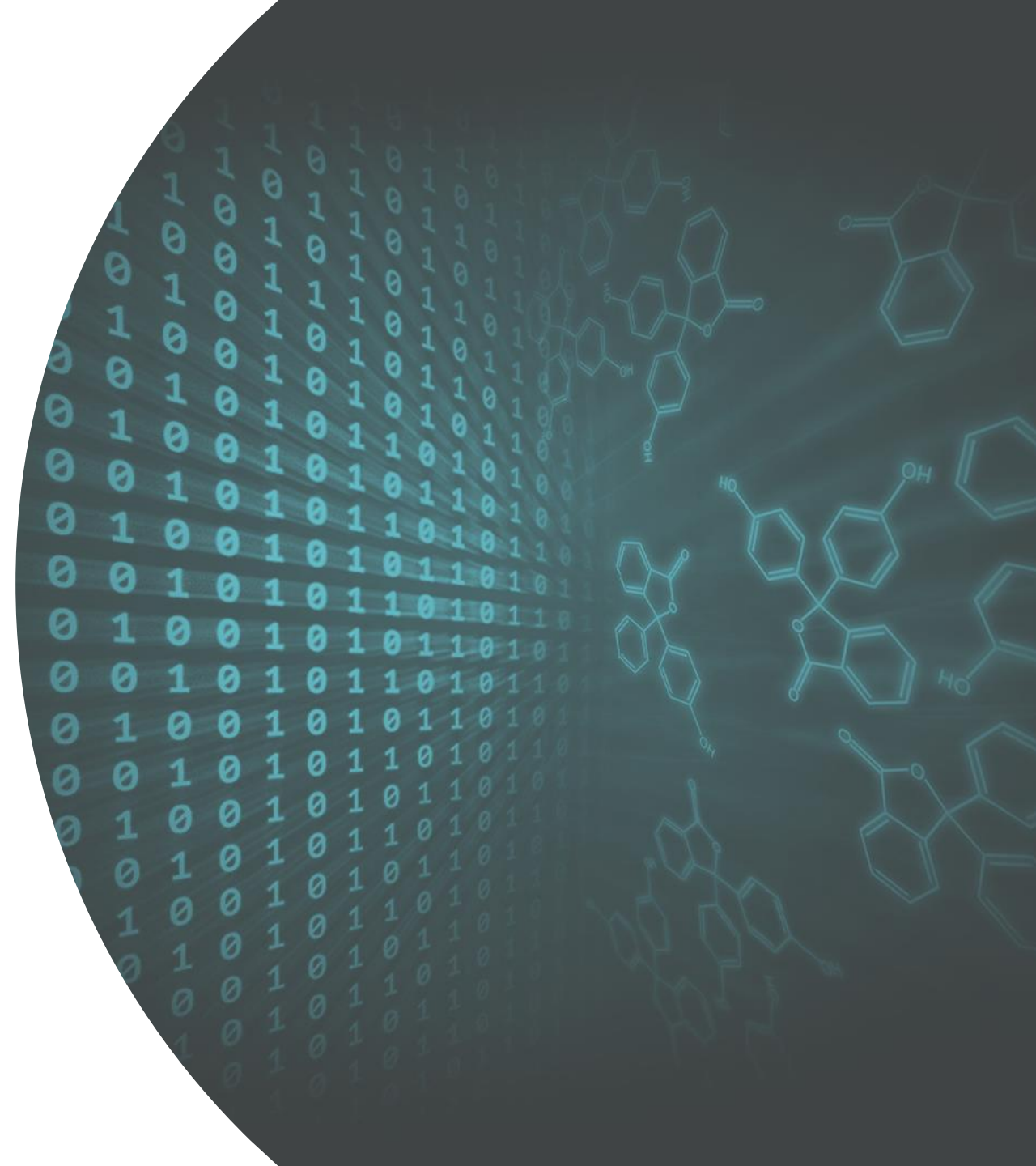


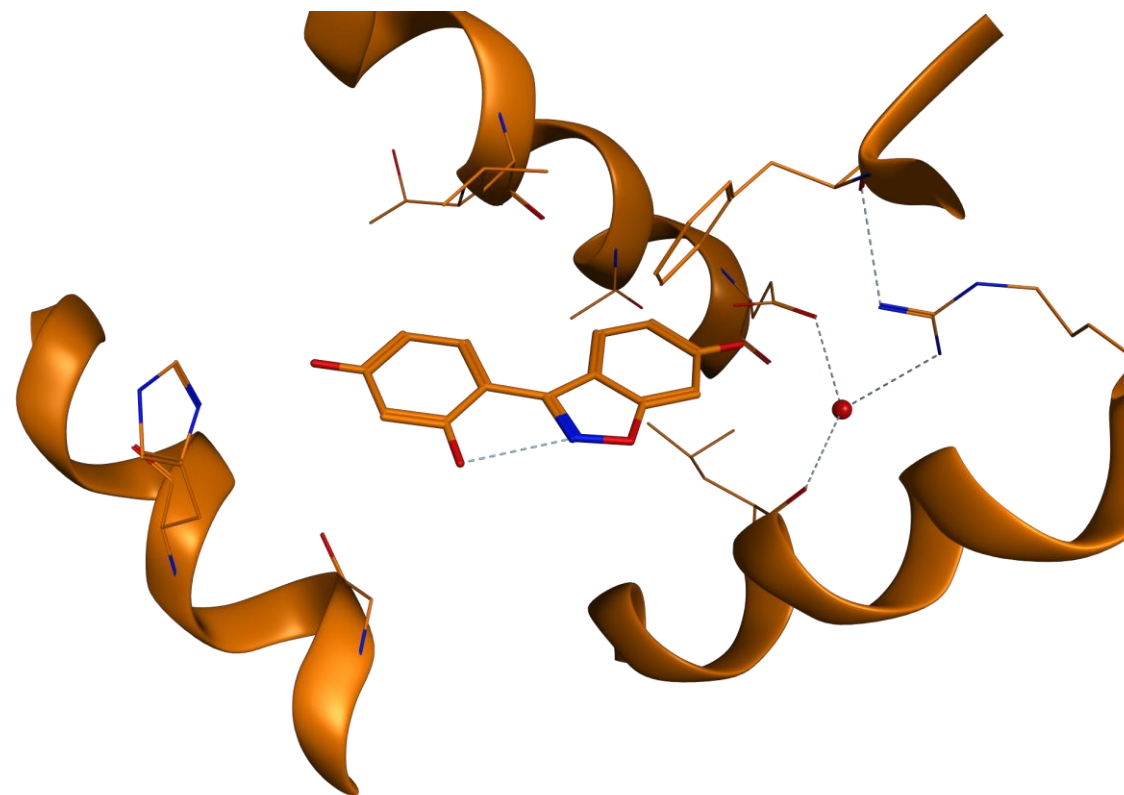
The Influence of Nonadditivity on Machine Learning and Deep Learning Models

Eva Nittinger



Where does Nonadditivity Occure?

- Assumptions:
 - Similarity principle: “*Compounds with similar structure have similar activities*”
 - Linearity and additivity in the chemical space
 - Precondition for extrapolation and prediction of unknown data from known data

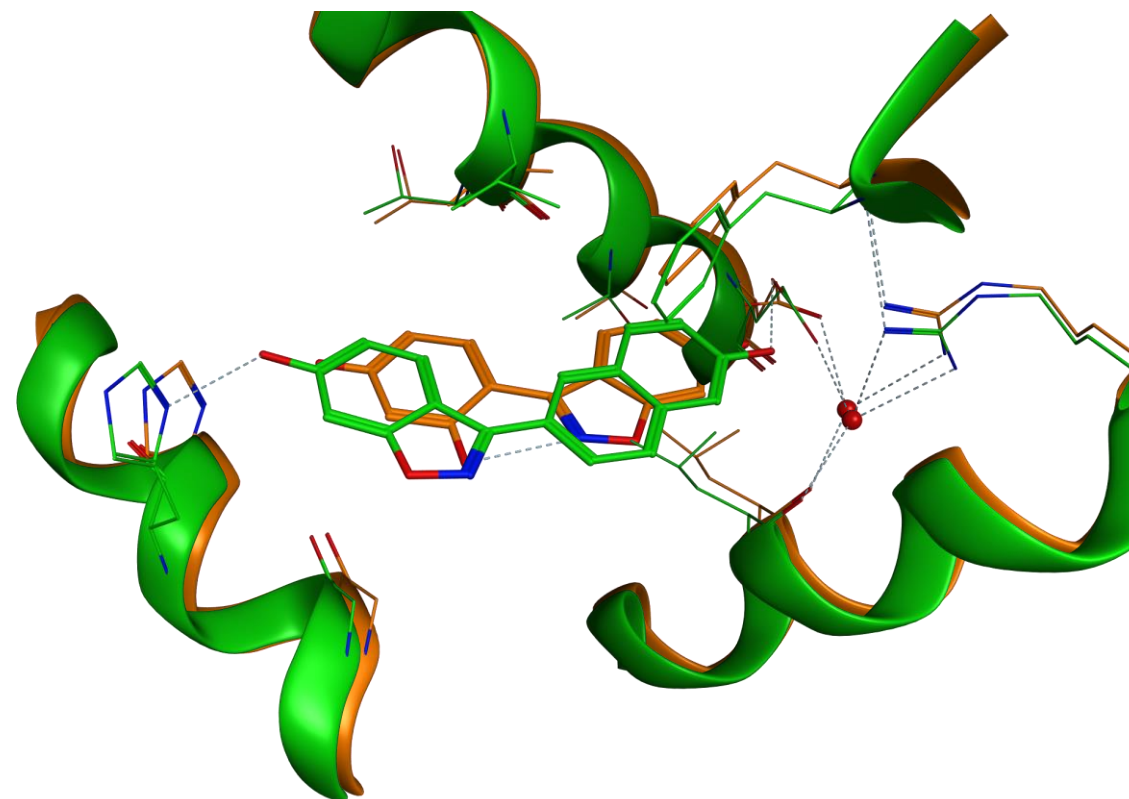


Estrogen receptor β ligands (1u3q, 1u3s)



Where does Nonadditivity Occure?

- Assumptions:
 - Similarity principle: “*Compounds with similar structure have similar activities*”
 - Linearity and additivity in the chemical space
 - Precondition for extrapolation and prediction of unknown data from known data

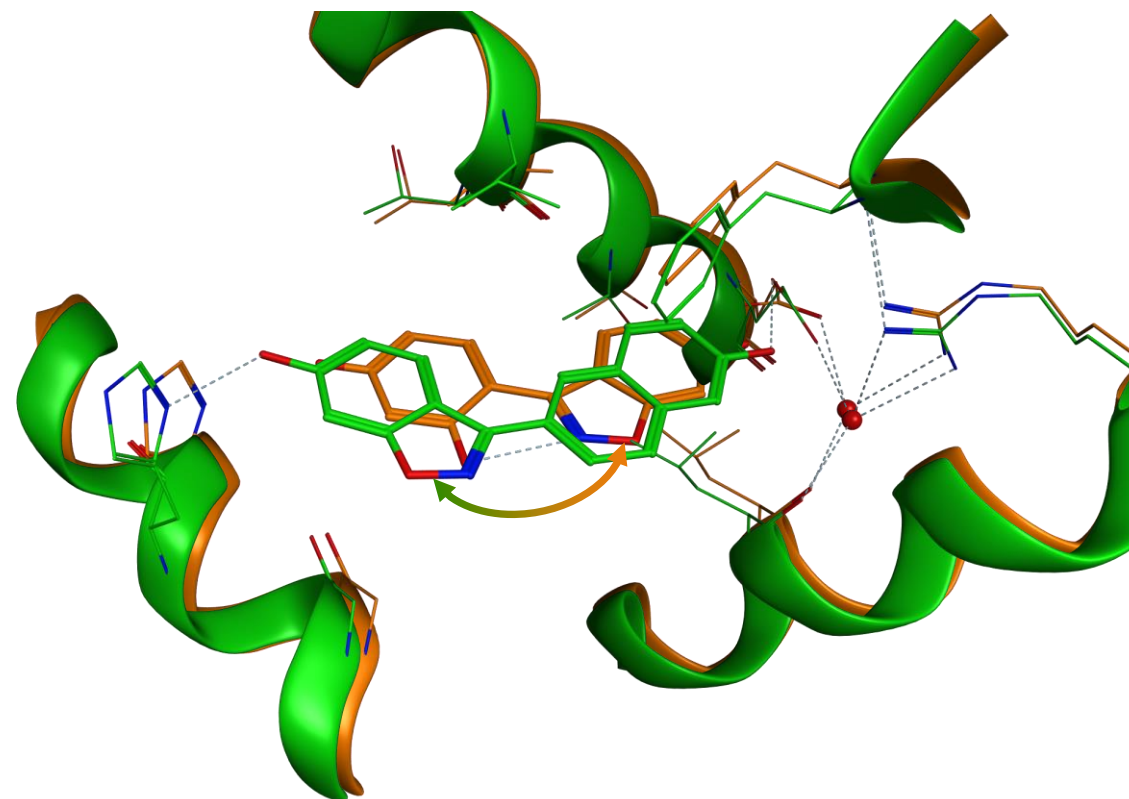


Estrogen receptor β ligands (1u3q, 1u3s)



Where does Nonadditivity Occure?

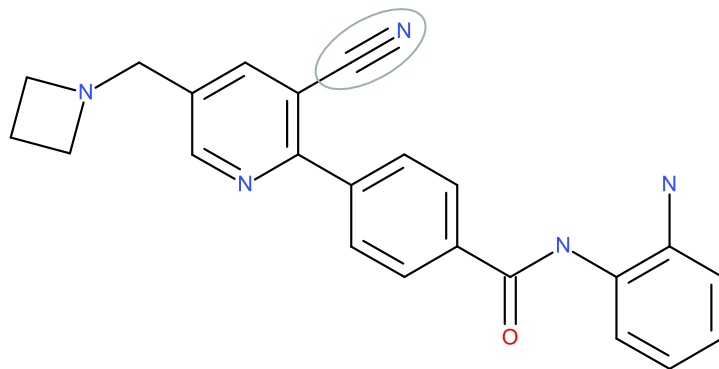
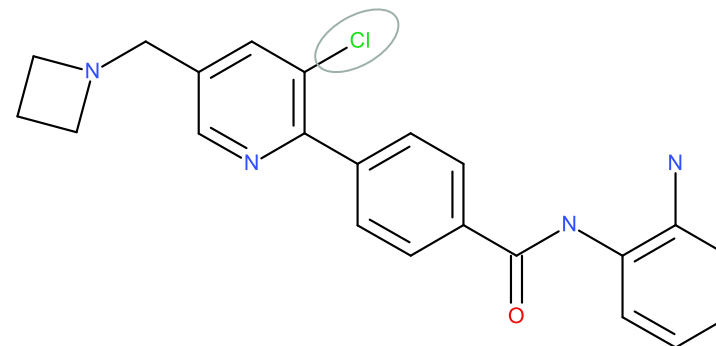
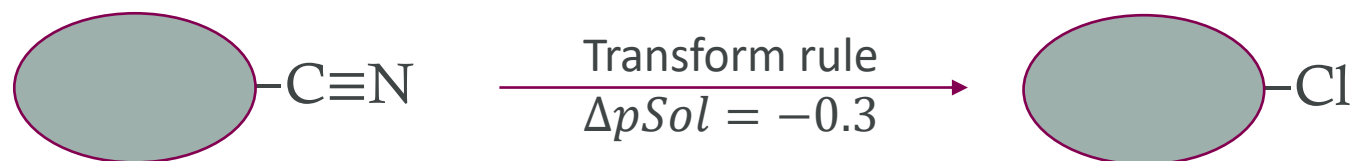
- Assumptions:
 - Similarity principle: “*Compounds with similar structure have similar activities*”
 - Linearity and additivity in the chemical space
 - Precondition for extrapolation and prediction of unknown data from known data



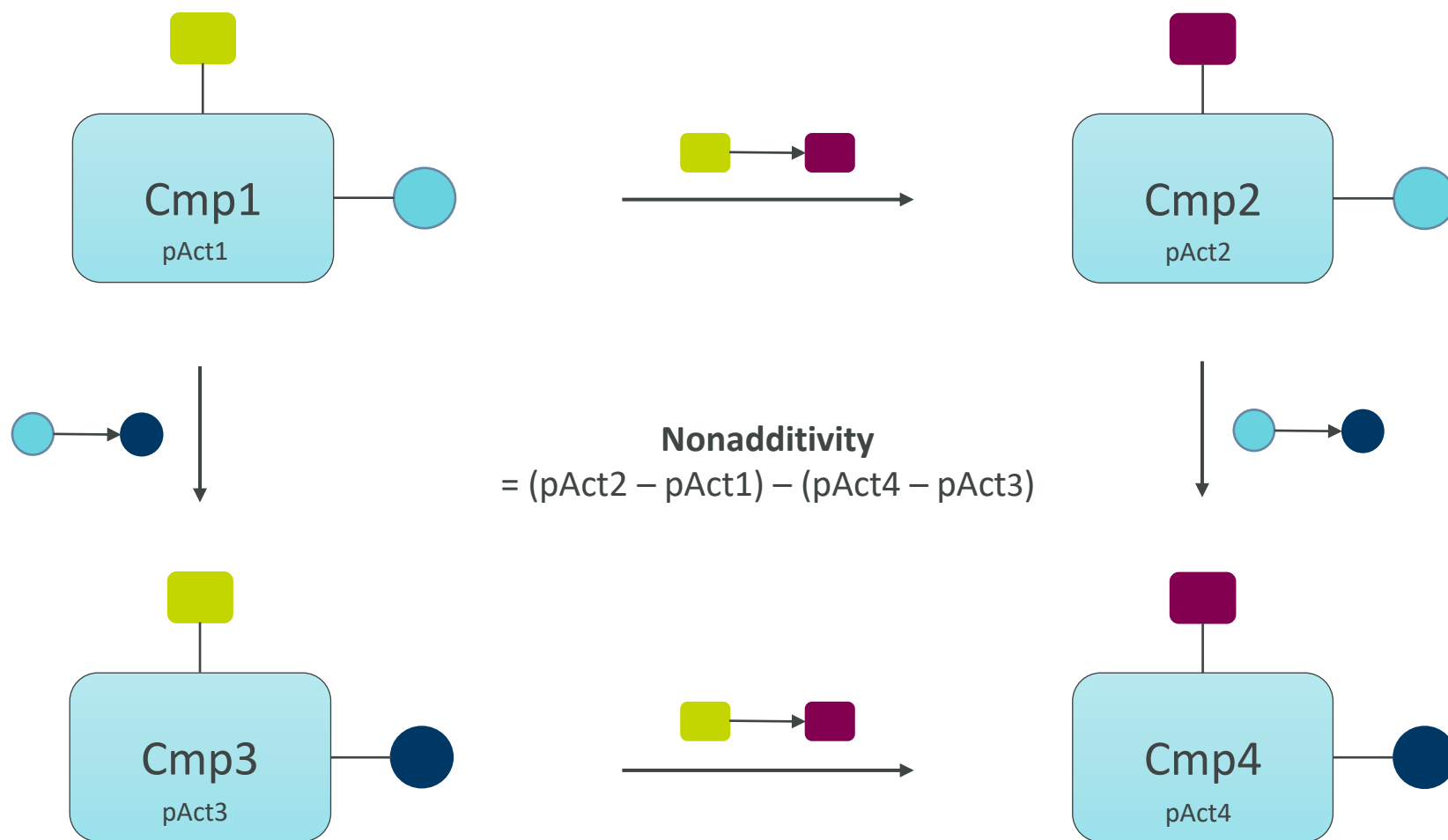
Estrogen receptor β ligands (1u3q, 1u3s)



NonAdditivity Analysis – What is it and How to Calculate it?


$$pSol = 3.2$$

$$pSol = 2.9$$


NonAdditivity Analysis – What is it and How to Calculate it?




1. Kramer, C. et al. Strong Nonadditivity as a Key Structure-Activity Relationship Feature: Distinguishing Structural Changes from Assay Artifacts. J. Chem. Inf. Model. 2015.



Method and Data Selection

- NAA analysis is based on binding assays
 - How often does NA occur?
- MMP analysis was performed on phys-chem properties
 - Can we predict (non)additivity?
- Nonadditivity analysis performance
 - NAA GitHub code provided by C. Kramer¹
 - Based on MMP analysis open-source code by A. Dalke²
 - Implementation of MMPA algorithm from Hussain and Rea³

NAA analysis data

	AstraZeneca	ChEMBL
Initial nof assays	22,317	1,125,387
Initial nof measurements	76,663,091	15,504,603
↓ Filtering and Cleaning 		
Final nof assays	6,224	13,620
Final nof measurements		3,625,044
Final nof compounds	1,221,623	799,860

MMP analysis data

	Nof cpds w w/o outlier	# multi measures	# stereo- duplicates
LogD	215418 214320	18429	6510
Solubility	226955 226189	21444	5527
Permeability	18076 18051	2282	646
Clearance	179637 179495	24493	5408

1. Kramer, C. Nonadditivity Analysis. J. Chem. Inf. Model. **2019**.

2. Dalke, A. *et al.* J. Chem. Inf. Model. **2018**.

3. Hussain, J.; Rea, C. J. Chem. Inf. Model. **2010**.0



Relevance of Experimental Uncertainty

$$\Delta\Delta pAct = \Delta\Delta pAct_{true} + \Delta\Delta pAct_{noise}$$

$$\Delta\Delta pAct_{noise}$$

$$= \sqrt{var(\varepsilon 1) + var(\varepsilon 2) + var(\varepsilon 3) + var(\varepsilon 4)}$$

$$= \sqrt{4 \cdot var(\varepsilon)} = 2\sigma_{\varepsilon}$$

Experimental uncertainty estimate

- 0.5 log units for public data¹
- 0.2 - 0.3 log units for in-house pActivity data²



Experimental uncertainty threshold as indicator for NA

1. Kramer, C. et al. The Experimental Uncertainty of Heterogeneous Public Ki Data. J. Med. Chem. 2012.
2. Kalliokoski, T. et al. Comparability of Mixed IC50 Data—a Statistical Analysis. PLoS One 2013.

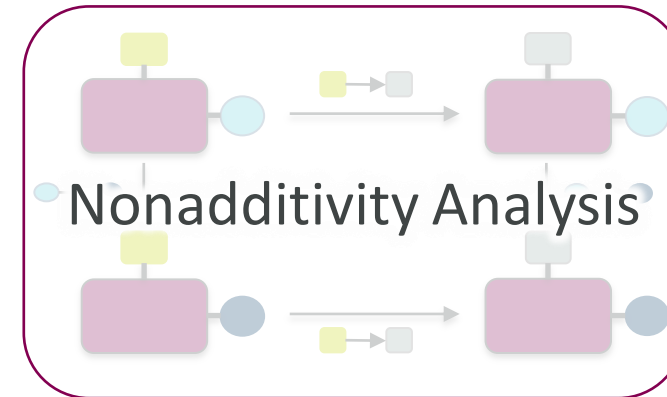


1

NAA Results

How do **inhouse** and **public** data compare?

How often does nonadditivity occur? Is it significant or neglectable?

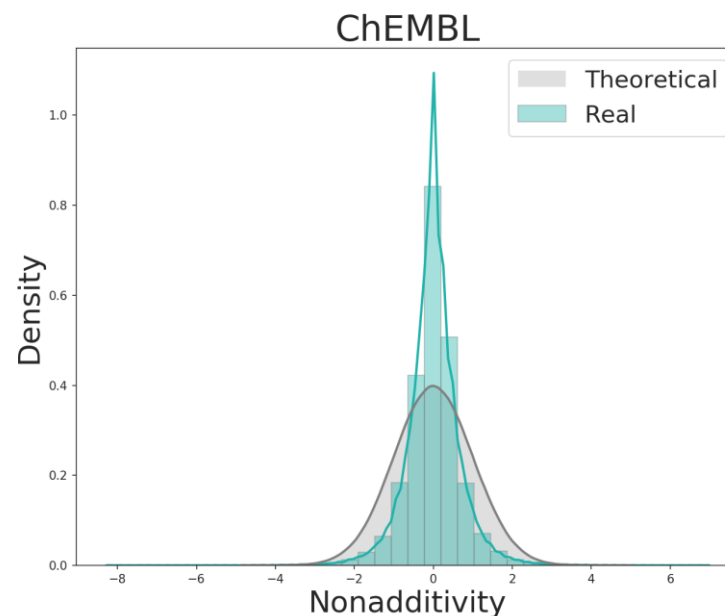
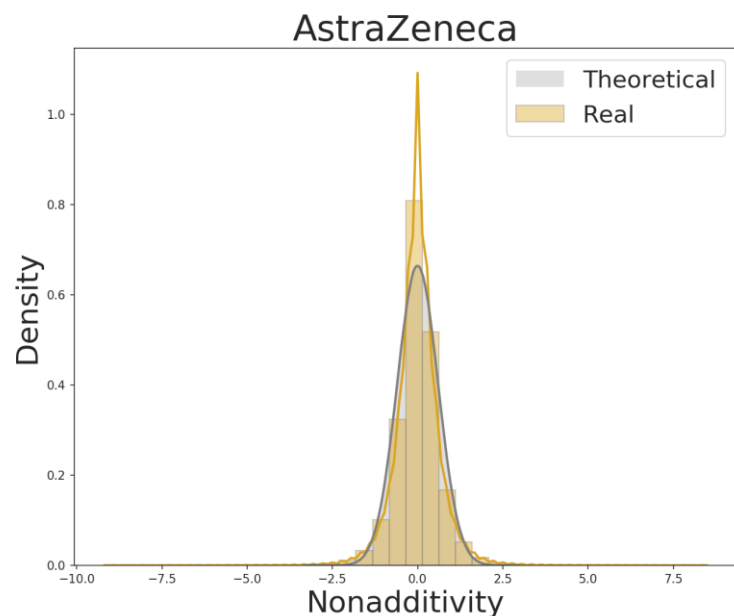


How often can nonadditivity be observed in tests/DTC/compounds?

How does nonadditivity influence machine learning?



Nonadditivity – Comparison of Inhouse and Public Data

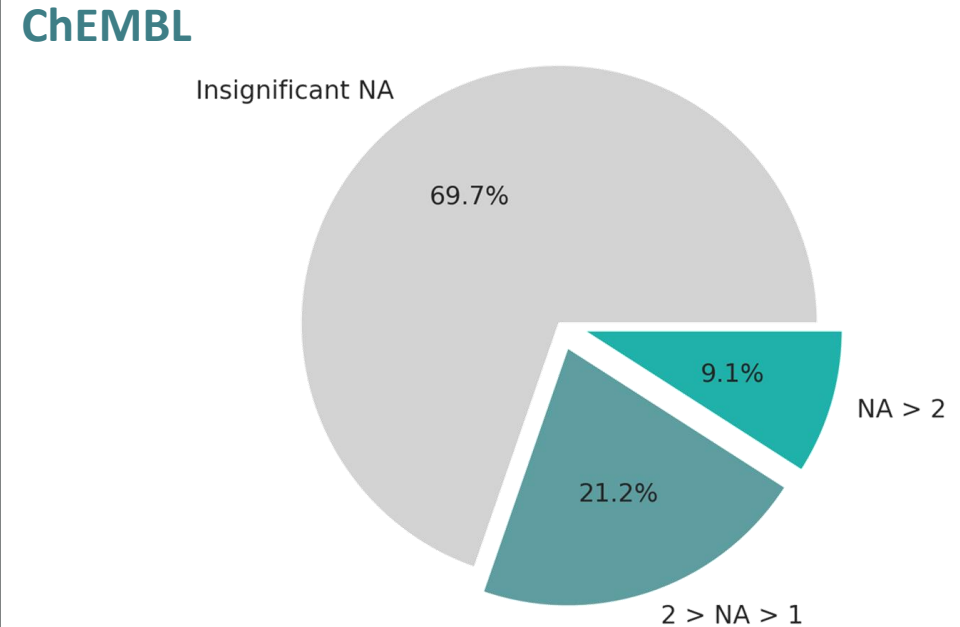
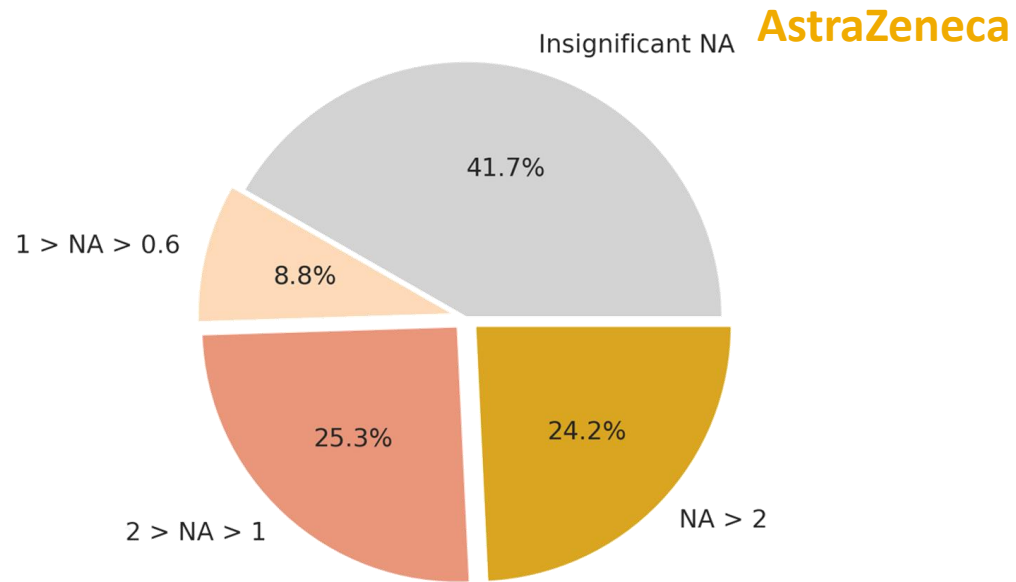


	AstraZeneca	ChEMBL
#Obs	3,053,055	1,246,975
Mean \pm std	0 ± 0.65	0 ± 0.68
Variance	0.42	0.46
Skewness	0	-0.01
Kurtosis	3.13	4.52

- Non-normal distribution for both data sets
 - Kurtosis, i.e. 'tailedness' is significantly large
- NA distributions are not different from each other



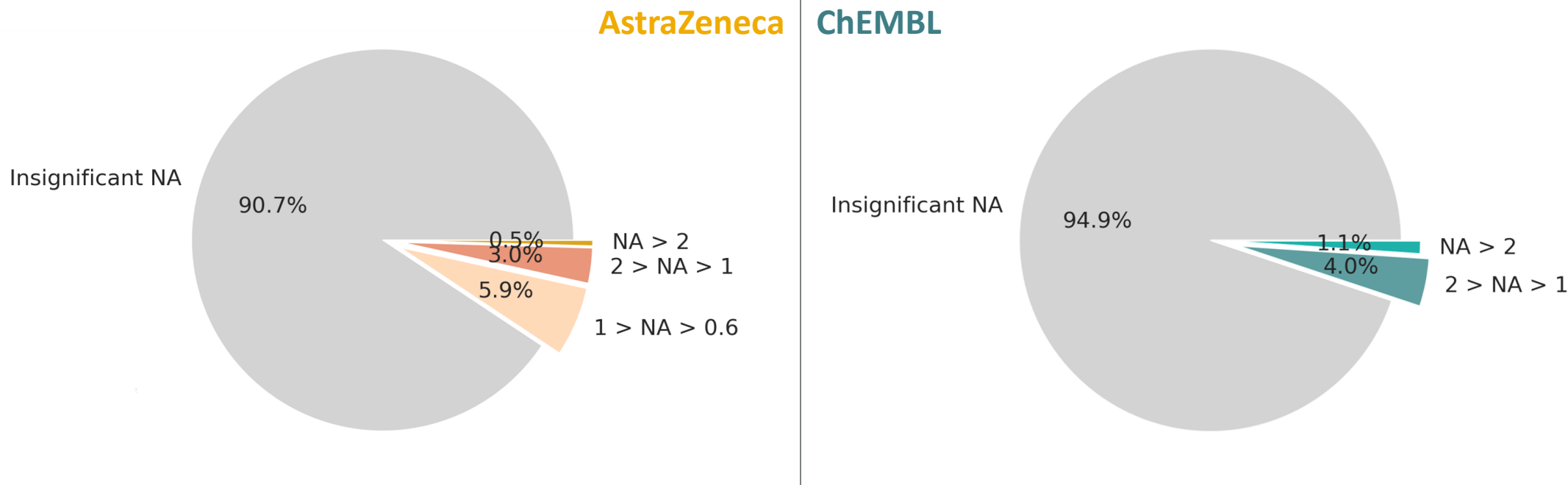
Data Comparison – Nonadditivity in Tests



- Inhouse data: 1 out of 4 tests shows strong NA
- Public data: 1 out of 10 tests shows strong NA



Data Comparison – Nonadditivity in Compounds

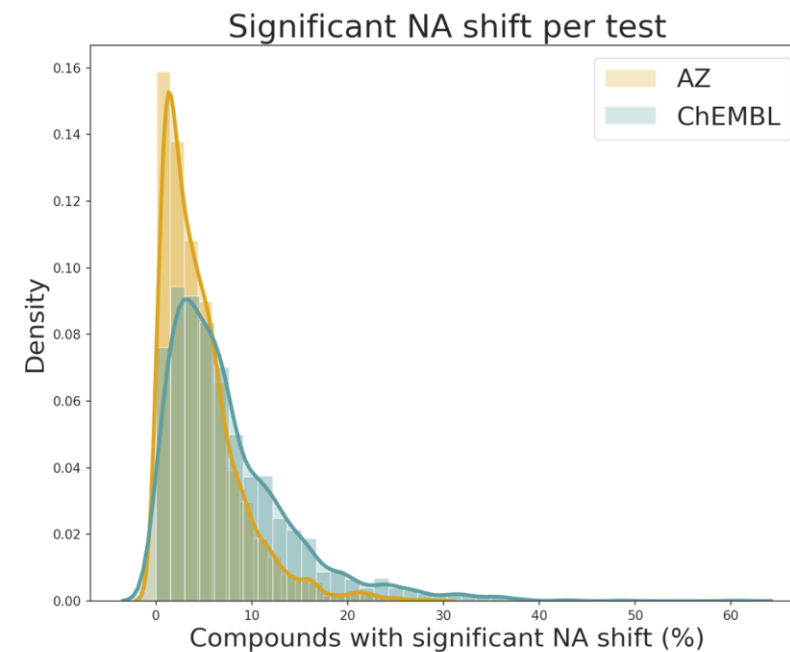


- Inhouse data: 9.4% shows significant NA
- Public data: 5.1% shows significant NA



Nonadditivity – Conclusion Part I

- AZ data indicates that nonlinearity frequently occurs in assays
 - It has to be examined carefully: derive structural explanation or reveal measurement errors
- Less nonlinearity observations in ChEMBL
 - Maybe due to the different cut-off for experimental uncertainty
 - Because the assays often have less compounds, and thus less matched squares
 - Publication bias, i.e. negative data are less often reported



Influence of Nonadditivity on Machine Learning

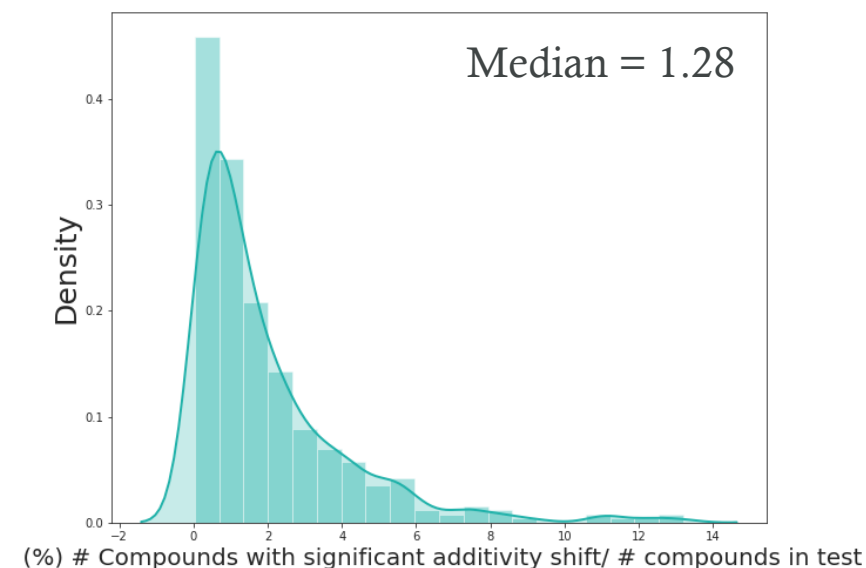
- **Automatic Generation of Machine Learning Models**

- Optuna¹ framework for automatic extensive hyper parameter optimization
- SVM and RF as robust baseline models²
- 500 trial runs with 5-fold cross-validation

ChEMBL data	# Cpd	# Cpd with significant NA (%)	# Cycles	# Cycles with significant NA (%)
1613797	772	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)

- Four model setups:

1: No NA, 2: Q1, 3: Median, 4: Q3

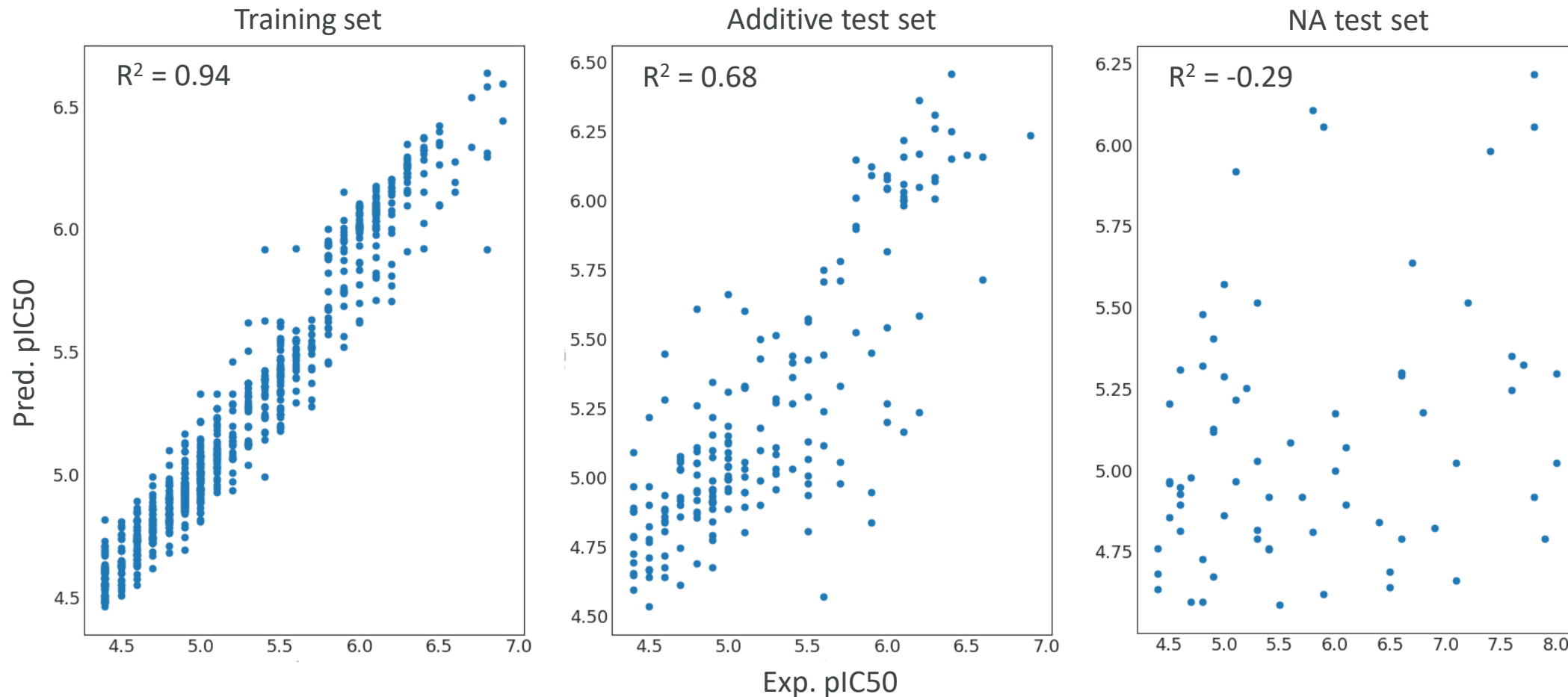


1. Akiba, T. et al. Optuna: A Next-Generation Hyperparameter Optimization Framework. ACM SIGKDD, 2019.
2. Pedregosa, F. et al. Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res. 2011.



Effect of Nonadditivity on Machine Learning Models

- RF performance for ChEMBL1614027 (#2,892)



Effect of Nonadditivity on Machine Learning Models

ChEMBL data (#measures)	SVM				RF			
	Test r^2 (RMSE)		Test MCC		Test r^2 (RMSE)		Test MCC	
	A*	NA#	A*	NA#	A*	NA#	A*	NA#
1613797 (772)	0.05 (0.33)	↘ -0.35 (1.22)	0.14	0.07	0.06 (0.33)	↘ -0.27 (1.19)	0.06	0.22
1614027 (1024)	0.68 (0.34)	↘ -0.29 (1.26)	0.54	0.08	0.68 (0.34)	↘ -0.29 (1.26)	0.53	0.20
1613777 (3511)	0.24 (0.69)	↘ -0.47 (1.33)	0.49	0.00	0.24 (0.69)	↘ -0.37 (1.29)	0.40	-0.01

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

- Consistent drop in r^2 and rise in RMSE from additive to NA test data
 - Both for SVM and RF
- Binary classification: drop for majority in MCC



Effect of Nonadditivity on Machine Learning Models

ChEMBL data (#measures)	SVM				RF			
	Test r^2 (RMSE)		Test MCC		Test r^2 (RMSE)		Test MCC	
	A*	NA#	A*	NA#	A*	NA#	A*	NA#
1613797 (772)	0.05 (0.33)	↘ -0.35 (1.22)	0.14 ↘	0.07	0.06 ↘ (0.33)	-0.27 (1.19)	0.06 ↗	0.22
1614027 (1024)	0.68 (0.34)	↘ -0.29 (1.26)	0.54 ↘	0.08	0.68 ↘ (0.34)	-0.29 (1.26)	0.53 ↘	0.20
1613777 (3511)	0.24 (0.69)	↘ -0.47 (1.33)	0.49 ↘	0.00	0.24 ↘ (0.69)	-0.37 (1.29)	0.40 ↘	-0.01

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

- Consistent drop in r^2 and rise in RMSE from additive to NA test data
 - Both for SVM and RF
- Binary classification: drop for majority in MCC



Effect of Nonadditivity on Machine Learning Models

ChEMBL data	RF (MCC for test)			
	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

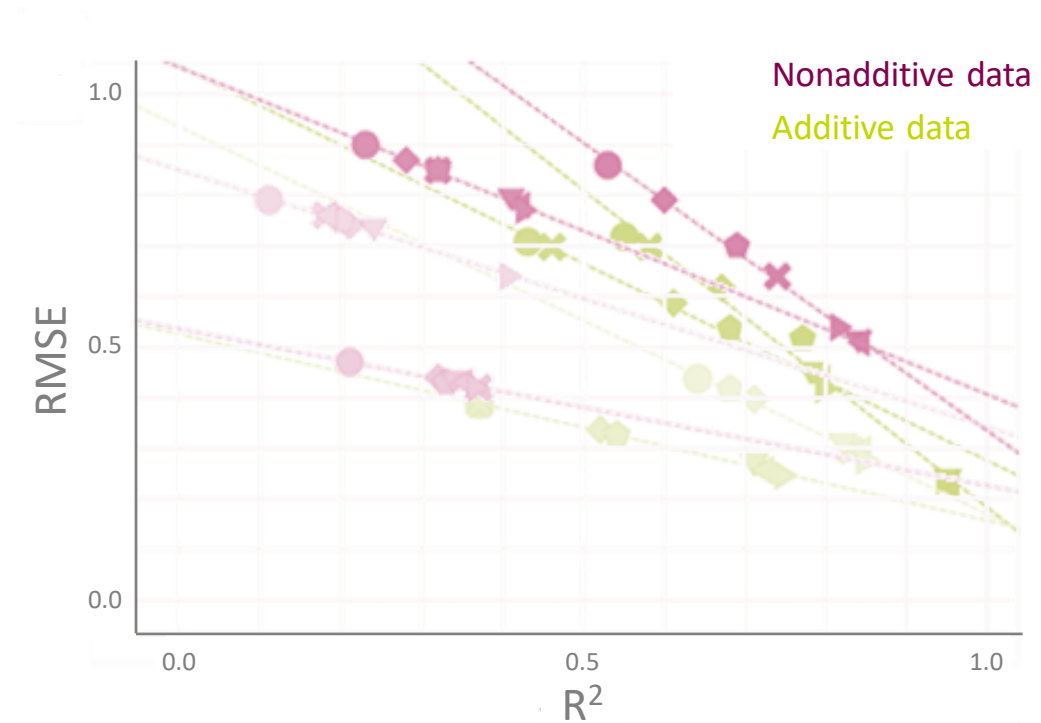
* Test set size for Q0 differs from Q1/Median/Q3.

- Adding different levels of NA data to the training
 - No significant differences for the different training sets
 - Reasons:
 1. Difficulty to learn from NA examples
 2. Too few examples included in training -> but realistic number as would be expected



2

MMP Results



Experimental Uncertainty



What are the experimental errors for inhouse phys-chem properties?



How reliable is the data?

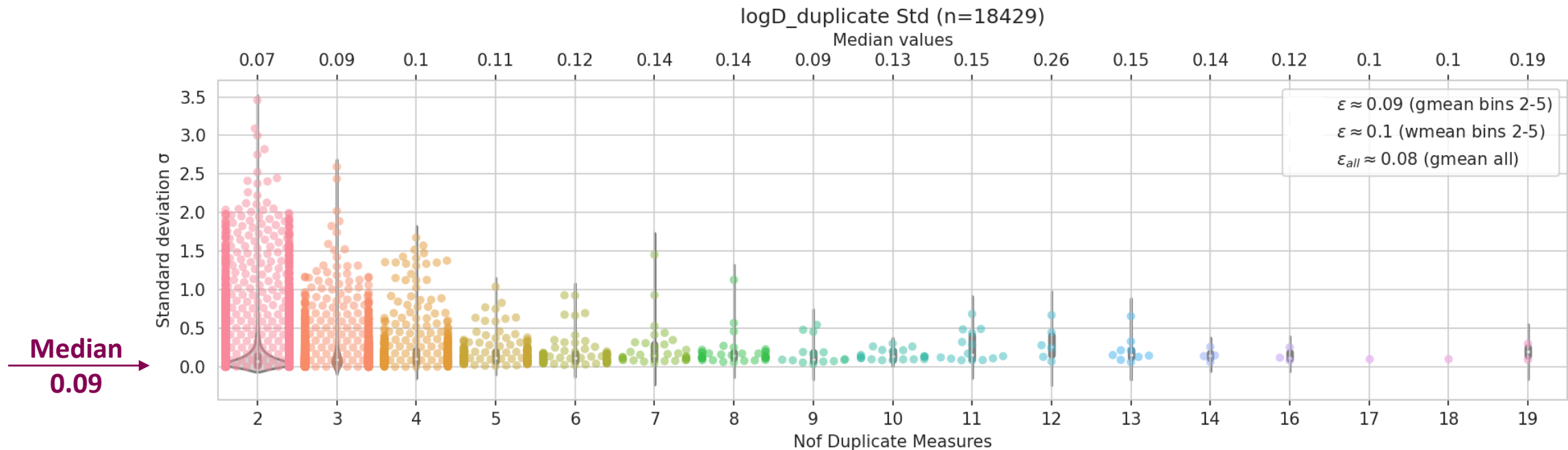


What can I expect from predictive models?



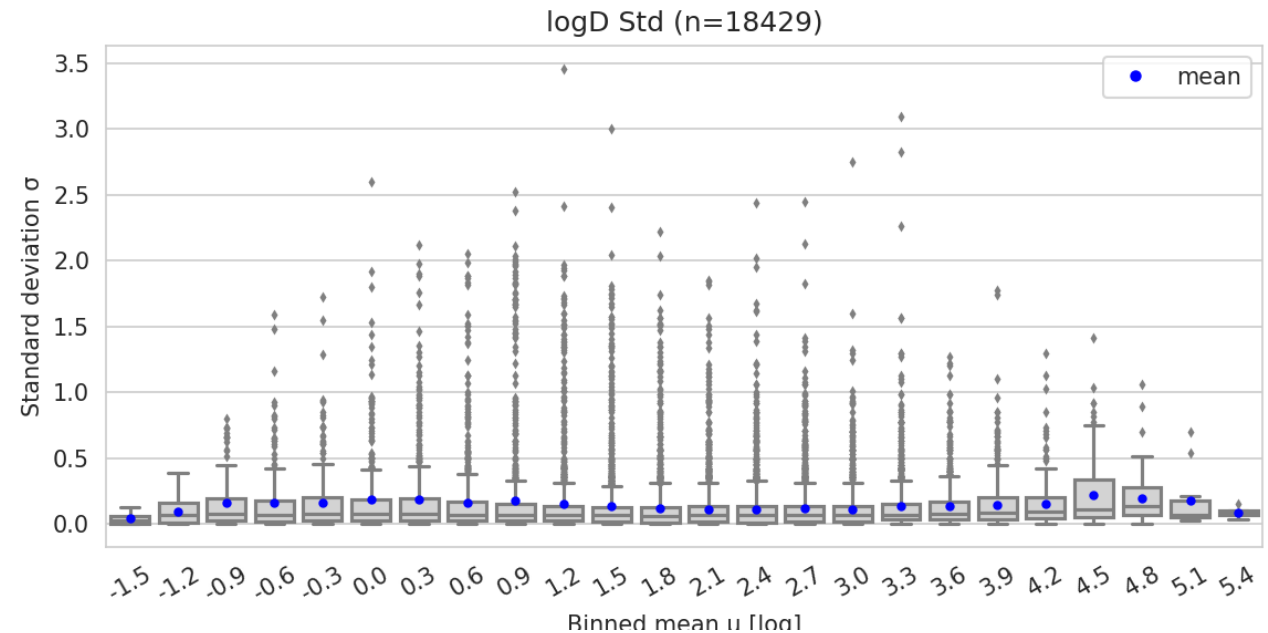
Experimental Error Estimate for logD

1. Binning by number of measurements available
2. Calculation of standard deviation
3. Generation of boxplots



Standard Deviation of Assay is **Nonlinear**

- Standard deviation varies for the experimental range
- Careful examination of error estimate necessary
- Regression model may only make sense for a defined experimental range



Conclusions – R^2_{\max}

	Exp. Uncertainty for Multi Measures	$R^2_{\max} = 1 - \left(\frac{\text{uncertainty in activity}}{\text{stdev of activity}} \right)^2$
LogD	0.1	0.993
Solubility	0.26 (~2 fold)	0.935
Permeability	0.22 (~2 fold)	0.936
Clearance – Hu Mics	0.12	0.947

- Models for all assays can achieve an R^2 of > 0.9
 - Measurements of stereo duplicates are slightly more consistent, i.e. have a smaller error, than duplicate measurements
- A model for logD assay could achieve an almost perfect $R^2 \sim 0.99$



Hypothesis

MMPs are the easiest changes and thus should be predictable

MMP Data

Property	Nof cpds	Nof cycles	Cpds with significant NA*
logD	207306	191605	25318 (12.21 %)
Solubility	219987	184116	28072 (12.76 %)
Permeability	17257	13977	916 (5.31 %)
Clearance	172947	121941	21750 (12.58 %)

* significance threshold determined by two times the experimental uncertainty

4 data sets

- Set 1 – all data
- Set 2 – MMPs
- Set 3 – additive MMPs
- Set 4 – nonadditive MMPs

ML/DL methods

- Qptuna
 - PLS, RF, SVR, XGBoost
- Directed Message Passing Neural Network (D-MPNN)
 - Single and multi-task setting

➤ 112 model trained



MMP Data

- Qptuna model training*
 - 300 iterations per model
 - 3-fold cross validation on training to avoid overfitting
 - Selection of best parameters and retraining on full data set
- DNN model training
 - Single task: training on individual property data
 - Multi task: training on union of property data
 - Hyperparameter optimization using Bayesian optimization provided by chemprop

Property	Data	Nof cpds	Training	Test
logD	Set 1 (all data)	207306	165844	41462
	Set 2 (all MMPs)	187162	149729	37433
	Set 3 (MMPs A)	47380	37904	9476
	Set 4 (MMPs N)	24775	19820	4955
Solubility	Set 1 (all data)	219987	175989	43998
	Set 2 (all MMPs)	196451	157160	39291
	Set 3 (MMPs A)	45976	36780	9196
	Set 4 (MMPs N)	27650	22120	5530
Permeability	Set 1 (all data)	17257	13805	3452
	Set 2 (all MMPs)	14612	11689	2923
	Set 3 (MMPs A)	4443	3554	889
	Set 4 (MMPs N)	909	727	182
Clearance	Set 1 (all data)	172947	138357	34590
	Set 2 (all MMPs)	155043	124034	31009
	Set 3 (MMPs A)	33755	27004	6751
	Set 4 (MMPs N)	21471	17176	4295

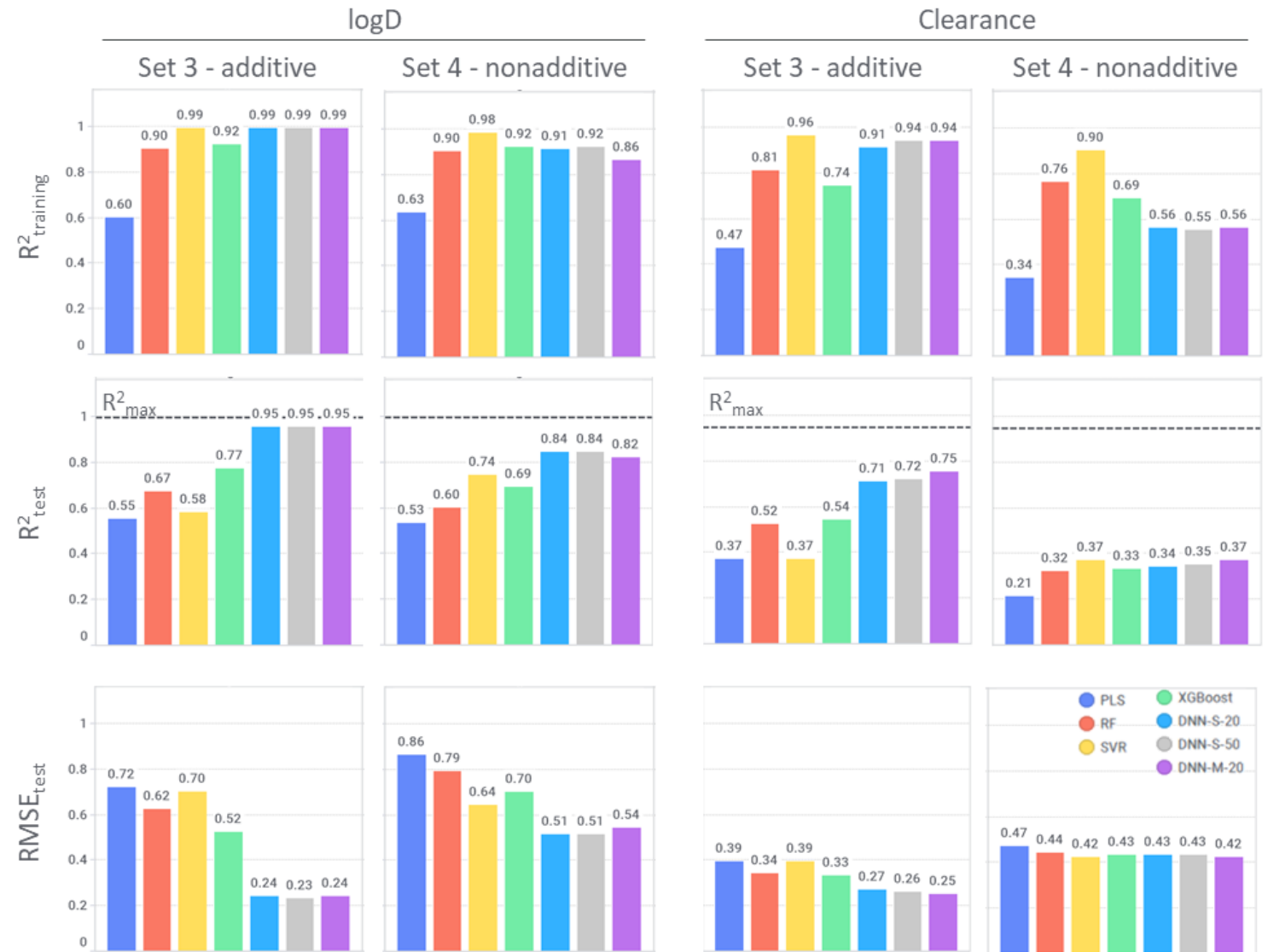
A – additive data; N – nonadditive data

* logD, solubility and clearance SVR runs had to be downsampled due to time-consumption.



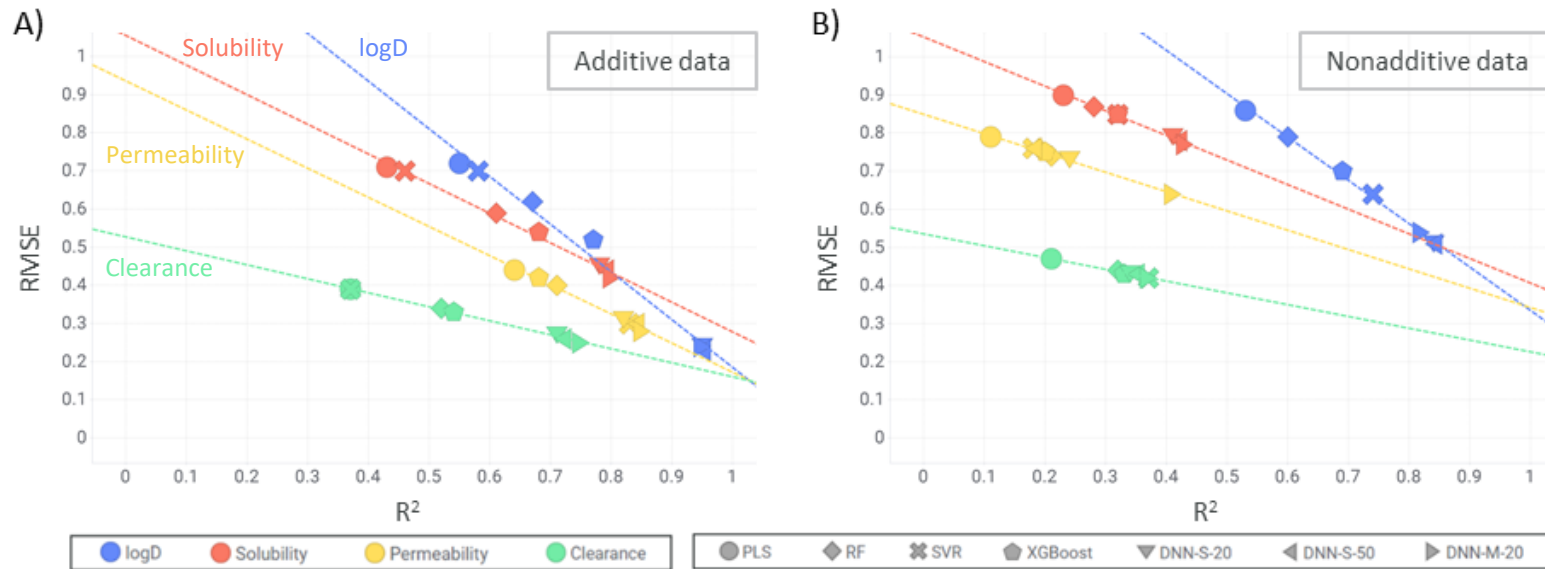
ML/DL Results

- Benchmark model PLS has worst performance
- DL models give best results (highest R^2 , lowest RMSE)
- logD: nonadditivity has lower effect on performance
- Clearance: greater drop in performance due to nonadditive data
 - Similar results for solubility and permeability ($R^2 < 0.43$)



MMP – Results

Performance Metric for Different ML/DL Models



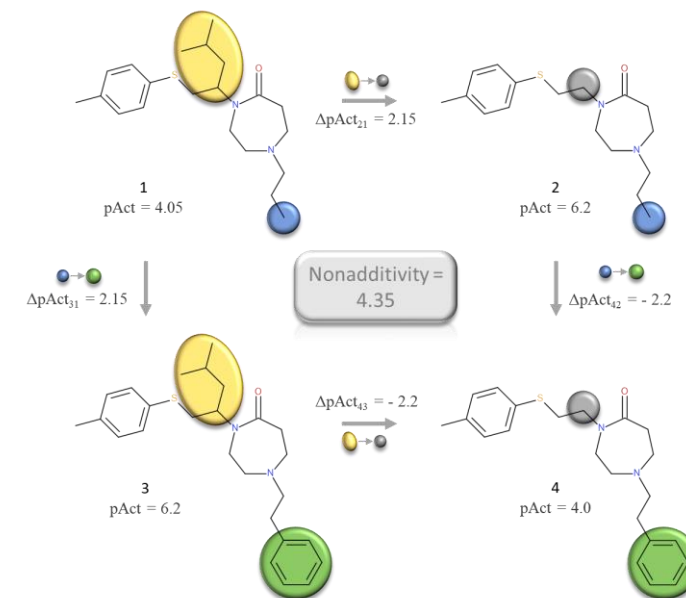
- R^2 and RMSE are significantly worse for nonadditive data

Non-linear models also fail on NA data!



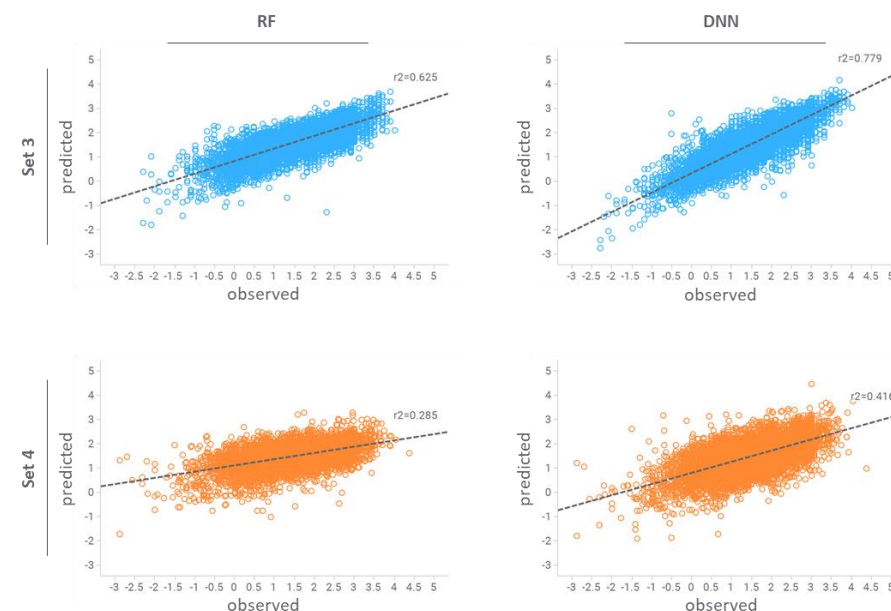
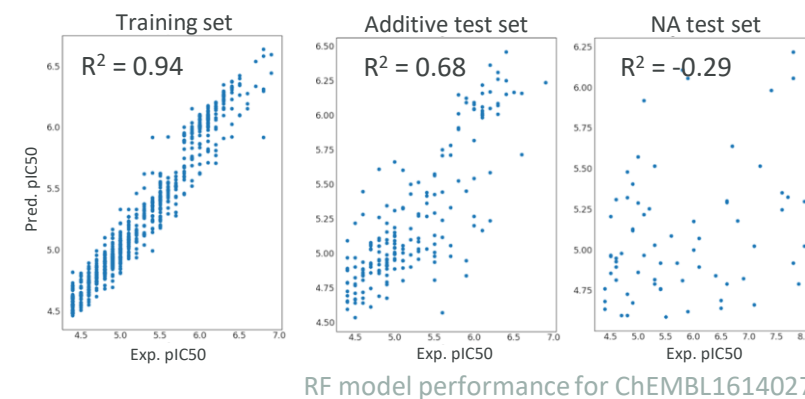
Conclusions and Future Work

- Detection of non-linearity in data
 - Important for further use, i.e. model building
- Significant number of compounds with NA in public and inhouse data
- ChEMBL data shows fewer NA
 - Reasons may be the lower number of compounds/test or the different experimental uncertainty cut-off
- NA data cannot be correctly predicted easily in ML models
 - DL, i.e. non-linear, models also fail to predict nonadditivity



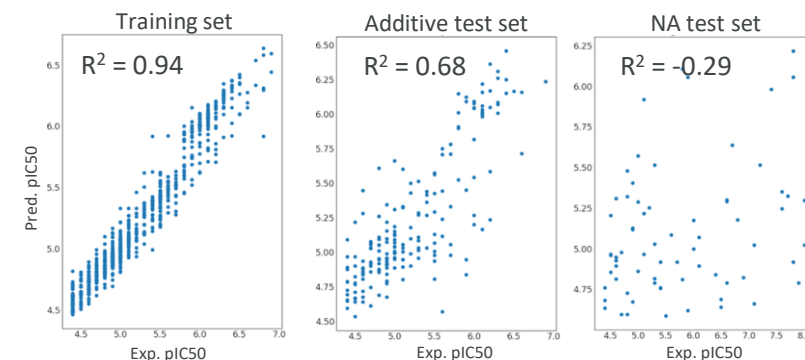
Conclusions and Future Work

- Detection of non-linearity in data
 - Important for further use, i.e. model building
- Significant number of compounds with NA in public and inhouse data
- ChEMBL data shows fewer NA
 - Reasons may be the lower number of compounds/test or the different experimental uncertainty cut-off
- NA data cannot be correctly predicted easily in ML models
 - DL, i.e. non-linear, models also fail to predict nonadditivity

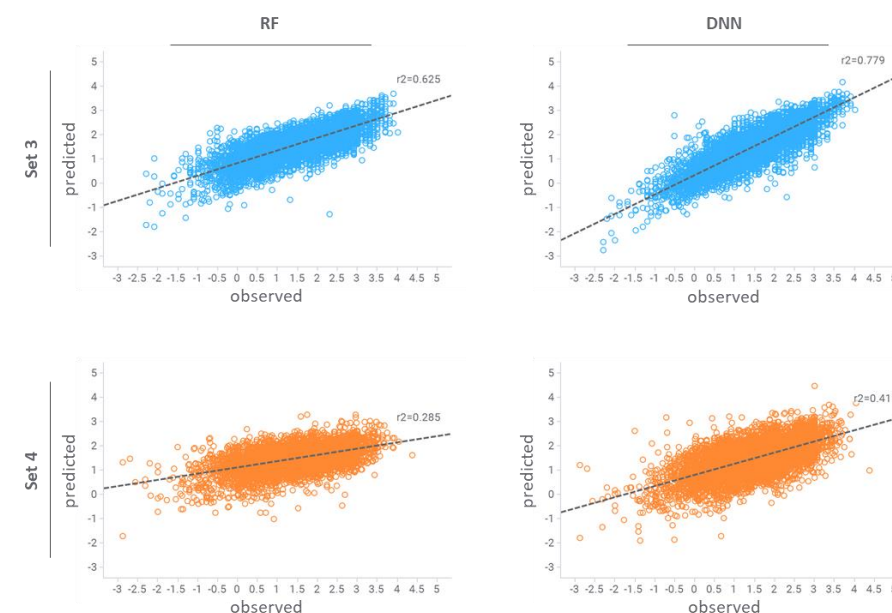


Conclusions and Future Work

- Detection of non-linearity in data
 - Important for further use, i.e. model building
- Significant number of compounds with NA in public and inhouse data
- ChEMBL data shows fewer NAs than public and inhouse data
 - **NA analysis should be considered regularly during CADD and for training of ML models.**
- NA data cannot be correctly predicted easily in ML models
 - DL, i.e. non-linear, models also fail to predict nonadditivity



RF model performance for ChEMBL1614027



Gogishvili, D.; Nittinger, E.; Margreitter, C.; Tyrchan, C. [Nonadditivity in Public and Inhouse Data: Implications for Drug Design](#). J. Cheminform. 2021, 13 (1).

Kwapien, K.; Nittinger, E.; He, J.; et al. Implications of Additivity and Nonadditivity for Machine Learning and Deep Learning Models in Drug Design, submitted.



Acknowledgements
&
Thank you!

Dea Gogishvili

Karolina
Kwapién

Christian Tyrchan

Werngard Czechitzky

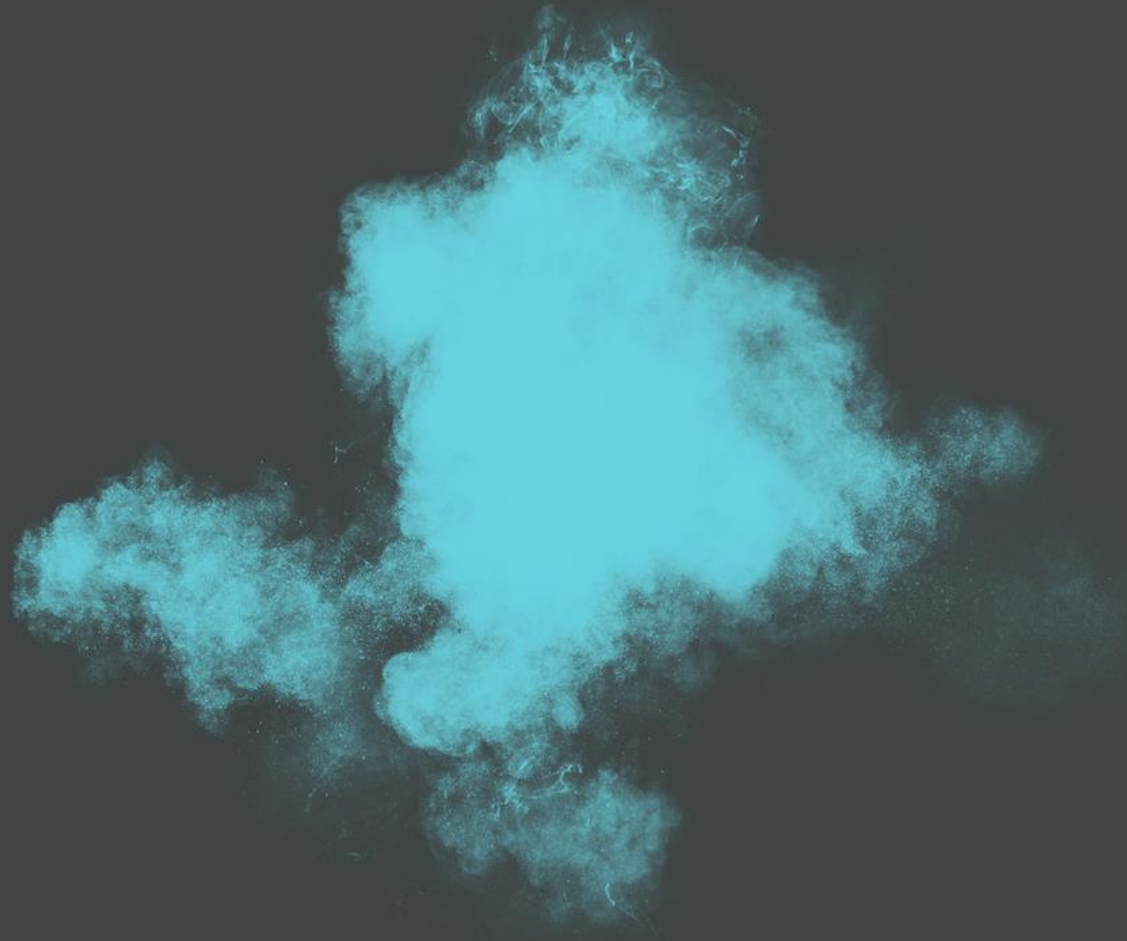
Christian
Margreitter

Jiazhen He

Alexey Voronov



Supplementary

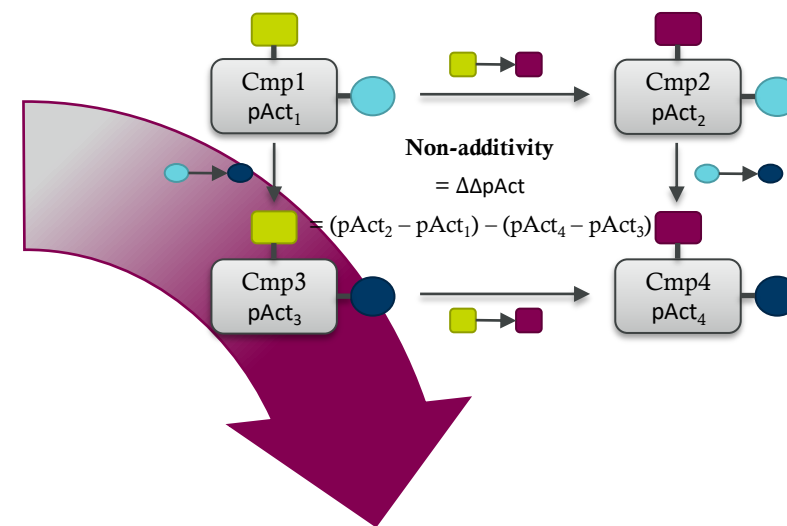


Non-additivity and its influence on ML performance

NA plays a significant role and has to be considered on a regular basis in CADD

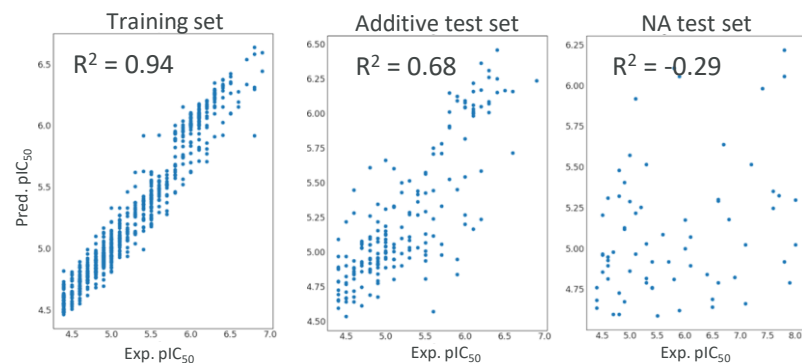
Assumption:

- Similarity principle: “Compounds with similar structure have similar activities”
- Linearity and additivity in the chemical space
 - Precondition for extrapolation and prediction of unknown data from known data



Influence of NA on ML

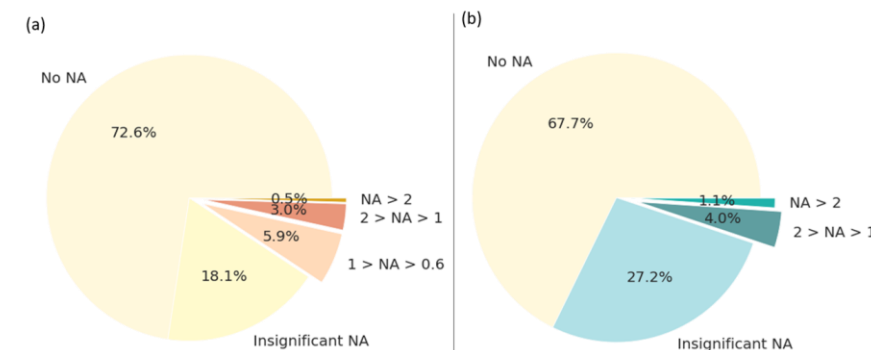
- Data with NA cannot be predicted accurately
- Model performance does not increase with NA training data



RF model performance for ChEMBL1614027 (1024 data points).

NA in public and inhouse data

- 9.3% of inhouse and 5.1% of public of compounds show significant NA



NA distribution among all unique compounds.

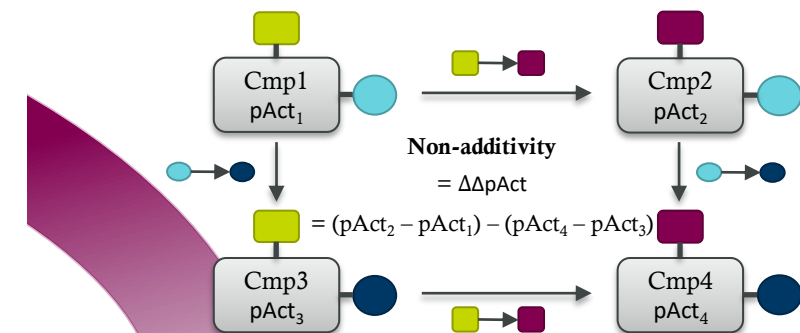


The Influence of Nonadditivity on ML and DL Models

Even nonlinear model cannot accurately model NA data.
NA has to be considered on a regular basis in CADD.

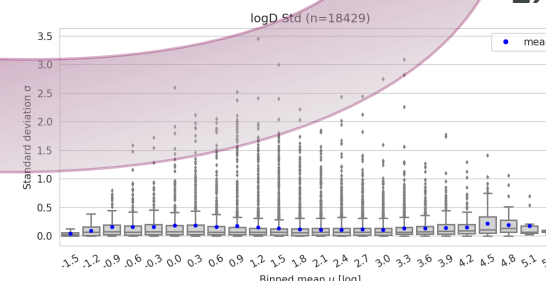
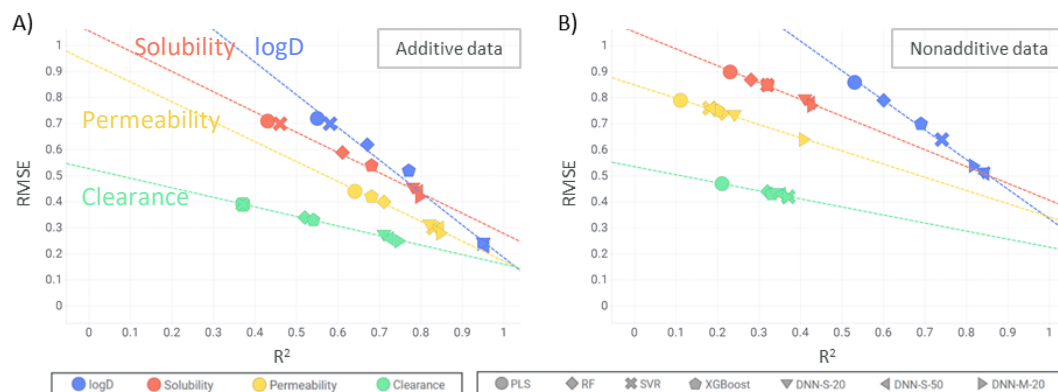
Assumption:

- Similarity principle: "Compounds with similar structure have similar activities"
- Linearity and additivity in the chemical space
 - Precondition for extrapolation and prediction of unknown data from known data



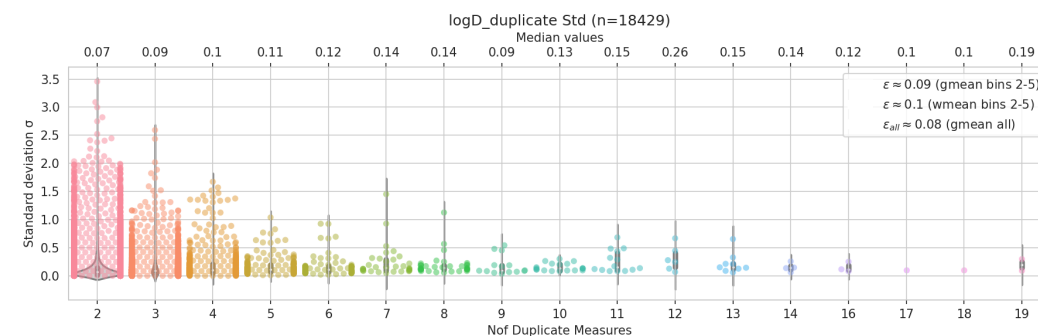
Influence of NA on ML and DL

- Significant rise in RMSE and lower R^2 for NA data
- DL models that are nonlinear cannot model NA data



Experimental Uncertainty & R^2_{max}

- Standard deviation varies for the experimental range
- Models for all assays can achieve R^2 of > 0.9



Confidentiality Notice

This file is private and may contain confidential and proprietary information. If you have received this file in error, please notify us and remove it from your system and note that you must not copy, distribute or take any action in reliance on it. Any unauthorized use or disclosure of the contents of this file is not permitted and may be unlawful. AstraZeneca PLC, 1 Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0AA, UK, T: +44(0)203 749 5000, www.astrazeneca.com

