Implications of Additivity and Nonadditivity for Machine Learning and Deep Learning Models in Drug Design

Karolina Kwapien (karolina.kwapien@astrazeneca.com)
AstraZeneca Sweden: AstraZeneca AB https://orcid.org/0000-0002-2003-0915
Eva Nittinger
AstraZeneca Sweden: AstraZeneca AB https://orcid.org/0000-0001-7231-7996
Jiazhen He
AstraZeneca Sweden: AstraZeneca AB https://orcid.org/0000-0001-5848-8318
Christian Margreitter
AstraZeneca Sweden: AstraZeneca AB https://orcid.org/0000-0002-5473-6318
Alexey Voronov
AstraZeneca Sweden: AstraZeneca AB https://orcid.org/0000-0003-0709-4954
Christian Tyrchan
AstraZeneca Sweden: AstraZeneca AB https://orcid.org/0000-0002-6470-984X

Research article

Keywords: Matched molecular pair analysis, nonadditivity analysis, structure-activity relationship, experimental uncertainty, machine learning, support vector machine, random forest, deep learning, graph neural network

Posted Date: February 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1180599/v1

Abstract

Matched molecular pairs (MMPs) is nowadays a commonly applied concept in drug design. It is used in many computational tools for structure activity relationship analysis, biological activity prediction or optimization of physicochemical properties. However, up to date it has not been shown in a rigorous way that MMPs, i.e. changing only one substituent between two molecules, can be predicted with high accuracy and precision in contrast to any other chemical compound pair. It is expected that any model should be able to predict such a defined change with high accuracy and reasonable precision. In this study, we examine the predictability of four classical properties relevant for drug design ranging from simple physicochemical parameters (logD and solubility) to more complex cell based ones (permeability and clearance), using different data sets and machine learning algorithms.

Our study confirms that additive data is the easiest to predict which highlights the importance of recognition of nonadditivity events and the challenging complexity of predicting properties in case of scaffold hopping. Despite of deep learning being well suited to model non-linear events, these methods do not seem to be an exception of this observation. Though, they are in general performing better than classical machine learning methods, this leaves the field with a still standing challenge.

Introduction

A Matched molecular pair (MMP) describes a pair of molecules that differs in one substituent only. Such a structural transformation is associated with a potential property change. MMP analysis is often used by medicinal chemists to compare properties in order to understand structure-activity relationship (SAR) for a series of compounds. An extension from a pair to a series of molecules that differ in a single transformation forms a Matched Molecular Series (MMS). MMS have been used to investigate automatic ways to derive a SAR similarity score[1, 2] and to predict ADME properties.[3]

The reason of the popularity of MMP analysis is its intuitivity: a particular change in a molecular structure introduces a certain change in a biological activity or physical property. However, this simple concept works only under the assumptions of linearity and additivity. Linearity means that the change in property due to a particular change in structure is constant. Additivity means that the effect of a structural change on a property is independent of other variables. It is important to take these assumptions into consideration before performing a MMP analysis or building a quantitative structure-activity relationship (QSAR) model.[4] Unfortunately most publications in the field do not report any such analysis on the data sets. We advocate that this relevant step becomes good practice in the QSAR/ML field. The use of a linear model would fail to capture the trend of nonadditive data mathematically, resulting in erroneous predictions.

Another important aspect of checking the validity of the additivity assumption is the identification of outliers. Outliers indicate so-called activity cliffs, a pair of molecules or even a single observation where a small structural change causes a significant change in a property or biological activity.[5–7] Analysis of

outliers and its understanding can lead to more efficient and effective design of molecules. The interpretation of activity cliffs is hampered by the complexity of the underlying effects and the fact that they can stem from any combination of these.[8–10] A common example of such an activity cliff is the so-called magic methyl where a single methyl group has a large effect on bioactivity or selectivity of a molecule.[11, 12]

Nonadditive data highlight critical changes in SAR and are therefore the most interesting for a medicinal chemist. Most common causes of nonadditive SAR are interactions between substituents, different binding modes and changes in protein conformation.[8, 9] Identification and analysis of nonadditive effects is important and can lead to understanding of changes in binding modes or ligand conformation. Additionally, it prevents chemists from missing good compounds and can change the direction of ligand optimization.

In this publication we examine several machine learning and deep learning algorithms to predict four properties (logD, solubility, permeability and clearance) using different data sets obtained from AstraZeneca's (AZ) internal database. First, we determine experimental uncertainty for each property as this is an upper limit for predictability of in silico models.[13, 14] Then, we perform a nonadditivity analysis (NAA) using the algorithm published by Kramer[15] to identify nonadditive datapoints. Based on this analysis we generate four data sets: 1) all data, additive and nonadditive; 2) all MMPs; 3) additive MMPs (MMPs A); and 4) nonadditive MMPs (MMPs N). By comparing the different data sets we analyze the influence of nonadditivity on the modeling and we also check if using only the MMPs is beneficial for the performance of a model. A variety of methods are considered starting from simple Partial Least Squares (PLS, serving as a benchmark), through Random Forest (RF), Support Vector Regressor (SVR), gradient-boosted trees (XGBoost) to deep learning algorithm (single and multitask deep neural networks). The quality of the models was evaluated using statistical parameters (R^2 and RMSE). Other common parameters in QSAR studies as receiver operating curves (ROC) or precision recall curves (PRC) are not taken into account as our intend is not to judge the performance in a virtual screening setting. Our aim is to evaluate the capability of machine learning methods to qualify and predict MMPs, the smallest possible compound change in a medicinal chemistry project.

Methods

Data Sets

Inhouse AstraZeneca data was used for all four properties, logD, solubility in DMSO, cell permeability and liver microsome clearance. By using inhouse data, a continuous assay set up is guaranteed for each property, to reduce the influence of systematic errors in the analysis.

All inhouse data was collected on 2020/09/14. Data was curated based on our previously developed pipeline.[4] Herein, molecules were standardized using PipelinePilot (standardization of stereoisomers, neutralization of charges, and removal of unknown stereochemistry), the canonical tautomer was

generated and kept for further analysis. All properties were converted to log values (SI Table S 1). Further data curation involved removal of unknown or uncertain ("<", ">") values and molecules with more than 70 heavy atoms (*data_all*). Subsequently, for compounds measured multiple times the median was calculated (*data_stereo*). Last, compounds with large differences between their multiple measurements (> 2.5 log units) were discarded and compounds only varying in their stereochemistry were combined, while keeping the more active compound (*Set 1*, Table 1).

Table 1. Number of (Nof) compounds (cpds) after the different curation steps (* = compounds measured ≥ 2 times).

Property	Data all | w/o outlier	Nof multi measures*	Nof stereoduplicates*	Nof cpds in Set 1
logD	215418 | 214320	18429	6510	207306
Solubility	226955 | 226189	21444	5527	219987
Permeability	18076 | 18051	2282	646	17257
Clearance	179637 | 179495	24493	5408	172947

Using the open source package mmpdb[16] all MMPs were obtained (*Set 2*, Table 2). Based on the NAA two additional sets were generated, one containing only additive compounds (*Set 3*) and one containing only nonadditive ones (*Set 4*). In order to determine (non-)additivity, a double transformation cycles (DTC) must be generated. Since not all MMPs are also in a DTC, the number of MMPs (*Set 2*) is larger than the combination of *Set 3* and *Set 4*.

For logD and solubility the size of the corresponding sets is similar, with clearance generally having slightly less compounds. The data sets for permeability are about ten times smaller. The exception is permeability *Set 4* with only 909 compounds in total.

For machine learning approaches we would expect *Set 4* to be most difficult to predict, followed by *Set 1*. *Set 3* should be easiest, since all compounds are additive.

The training and test sets for the machine learning approaches were obtained by doing a classical stratified training-test split with 0.8 and 0.2 ratio.

Table 2. Number of (Nof) compounds (cpds) in each data set (A = additive data, N = nonadditive data).

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

Experimental Uncertainty and R^2_{max}

For all selected properties data was collected for a) multiple measurements for the same compound (*data_all*) and b) measurements for compounds only differentiating in their stereochemistry (*data_stereo*). These data were used to calculate the experimental uncertainty of each respective assay.

Herein, the weighted mean was used to derive the experimental uncertainty for each property:

$$\varepsilon_{w\_mean} = \sum_{i=bin=2}^{bin=x} m_i \cdot median_i \quad (1)$$

with x being the bin where 2.5% (0.5%) of datapoints for multi measures (stereoduplicates) are included. Smaller amount of datapoints per bin only lead to an artificial increase of experimental uncertainty.

Based on the experimental uncertainty, the maximum R^2 achievable for a machine learning approach can be determined[17]:

$$R^2_{max} = 1 - \left(\frac{uncertainty\ in\ activity}{stdev\ of\ activity}\right)^2 \quad (2)$$

Nonadditivity Analysis

Nonadditivity analysis was performed to determine (non-)additivity in a compound data set. Therefore, the open-source NA analysis code published by Christian Kramer was used (available on GitHub: https://github.com/KramerChristian/NonadditivityAnalysis).[15] The code is written in Python and makes use of the cheminformatics libraries RDKit,[18] Pandas and NumPy. NA calculations are based on matched molecular square, so called double-transformation cycles, which consist of four matched

molecular pairs (four compounds) linked by two distinct transformations. The MMPs in the NA code are generated by open-source code developed by Dalke *et al.*,[16] an implementation of the MMPA algorithm developed by Hussain and Rea.[19] The NA value of each DTC is calculated as the difference in logged biological activities ($pAct_{1-4}$) of the four compounds assembling the cycle:

$$\Delta\Delta pAct = (pAct_2 - pAct_1) - (pAct_3 - pAct_4) \quad (3)$$

Machine/Deep Learning

Machine Learning Models Using Optuna

Partial Least Squares (PLS), Random Forest (RF), Support Vector Regressor (SVR) and gradient-boosted trees (XGBoost) models were build using Optuna (https://optuna.org).[20] Optuna is a hyperparameter optimization framework and forms the basis of our in-house QPTUNA framework (unpublished), that extends Optuna by adding chemoinformatics functionality. Optuna allows to specify the hyperparameter search space for a plethora of machine learning algorithms and automatically tries to optimize them with respect to a defined output metric for a specified number of trials. By using a surrogate model, such search should be more efficient than a mere random or grid search.

For each of the data sets provided, we trained a number of regressors for a minimum of 300 iterations each. This was done with cross-validation (to avoid overfitting during training) and models were then built from the entire training sets. Finally, the models were evaluated on the respective test sets.

For some of the SVR runs we had to use a “down-sampled” data set (10% of the corresponding original size) to be able to obtain optimized hyperparameters within a reasonable time frame. This was done for logD, solubility and clearance (*Set 1-3*). The rest of the sets (all permeability data sets and *Set 4* for each property) used all datapoints for hyperparameter optimization. The following steps, model training and prediction of the respective test sets, were performed on the full size sets for all properties.

Graph Neural Network Deep Learning Model

The Message Passing Neural Network (MPNN)[21] framework operates on molecular graphs with atoms as nodes and bonds as edges. There are two main phases: 1) message passing phase, in which the node information is propagated and updated across the graph in order to build a neural representation of the whole graph; 2) readout phase, when a final feature vector/representation describing the whole graph is created. Then a feed-forward neural network can be applied to this feature vector for prediction tasks.

The Directed Message Passing Neural Network (D-MPNN)[22] (available on GitHub: https://github.com/chemprop/chemprop) builds upon the MPNN framework with the difference that during the message passing phase, the directed edge information are used instead of node information.

In this study, the D-MPNN model was trained in a single task setting and a multi-task setting. In the single-task setting, the model was trained individually for each property task, while in the multi-task setting, a multi-task model was trained on the union of the training sets from all the property tasks where each

molecule has four target values. Therefore, after training, the multi-task model can predict the four properties simultaneously for the molecules of the test set.

Hyperparameter optimization was performed for each data set using Bayesian optimization (i.e. Hyperopt[23]) provided by chemprop, which finds the optimal parameters (hidden size, depth, dropout, and the number of feed-forward layers; details about the searching space can be found in chemprop) through multiple trials. In particular, 20 and 50 hyperparameter trial settings were tried in a single-task setting, which results in two models for each data set, hereafter named DNN-S_20 and DNN-S_50, respectively. For the multi-task setting, only 20 hyperparameter trial settings were tried (DNN-M_20).

During the hyperparameter optimization, the original training set in Table 2 is split into training, validation and test with the ratio 0.8, 0.1, 0.1, to find out the best parameter configuration based on RMSE metric. Then the model was trained using this parameter configuration, and the original training set in Table 2 is split into train and validation with ratio 0.8 and 0.2. Finally, the trained model was applied to the test set to obtain the predictions.

Results And Discussion

Experimental Uncertainty and R^2_{max}

The experimental uncertainty of an assay can be calculated by leveraging the information from compounds measured multiple times (Table 1). Herein, two aspects were analyzed: first, the experimental uncertainty based on compounds measured multiple times and second, the experimental uncertainty for compounds with different stereochemistry. The idea of the latter analysis was that the stereochemistry should play a minor role for the different physicochemical properties. Thus, the experimental uncertainty for those compounds should be rather small.

Table 3. Experimental uncertainty (in log units) and expected R^2_{max} estimated for each property.

Property	$\varepsilon_{w\ mean}$ for multi measures	$\varepsilon_{w\ mean}$ for stereoduplicates	R^2_{max}
logD	0.10	0.07	0.993
Solubility	0.26	0.15	0.935
Permeability	0.22	0.10	0.936
Clearance	0.12	0.15	0.947

Table 3 summarized the experimental uncertainties as well as the resulting maximum R^2 values (SI Figures S1-4). Solubility has the highest experimental uncertainty for multi measurements, followed by permeability, resulting for both assays in a 2-fold variability of the measured value. As expected, stereoduplicates show very low experimental uncertainties. As expected only for clearance this trend is not true, stereoduplicates display a similar experimental uncertainty. logD has the lowest experimental uncertainty with 0.07 log units. Thus, using machine learning approached almost ideal performance is theoretically possible (R^2_{max} = 0.993).

In the following, the experimental uncertainties are used for cut-offs for the NAA. Herein, compounds with a nonadditivity value greater than two times the experimental uncertainty are classified as nonadditive.

Nonadditivity Analysis

NAA allows the classification of compounds into additive and nonadditive ones. Herein, a prerequisite is the composition of matched molecular squares. These are used to determine whether a cycle is additive or nonadditive.

Table 4. NAA results for each property (Nof = number of, cpds = compounds, * = significance threshold determined by two times the experimental uncertainty).

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 %)

Surprisingly, logD, solubility and clearance all have more than 12% nonadditive compounds (Table 4). In our previous study of nonadditivity in bioactivity data 9% (5%) of compounds were nonadditive for inhouse (and public ChEMBL) data. Compared to this, the amount of nonadditivity found here is significantly larger. The reasons might be manifold and different for each property,[4] e.g. in the case of logD solubility might play an important role, as in the case of solubility crystal packing seems to be important. Permeability displays an exception with only 5% of compounds being classified as nonadditive (Table 4).

The results of the NAA were used to generate the data sets for machine learning.

Machine/Deep Learning

Table 5 and Table 6 present R^2 and RMSE obtained for all algorithms, data sets and properties discussed in this work. The results are also presented visually for *Set 3* (only additive data) in Figure 1 and Figure 2.

Table 5. R^2 (for test set) for all algorithms, data sets and properties discussed in this work.

Property	Data	Model						
		PLS	RF	SVR	XGBoost	DNN-S_20	DNN-S_50	DNN-M_20
logD	Set 1 (all data)	0.52	0.63	0.65	0.76	0.91	0.91	0.90
	Set 2 (all MMPs)	0.52	0.64	0.66	0.76	0.91	0.91	0.90
	Set 3 (MMPs A)	0.55	0.67	0.58	0.77	0.95	0.95	0.95
	Set 4 (MMPs N)	0.53	0.60	0.74	0.69	0.84	0.84	0.82
Solubility	Set 1 (all data)	0.36	0.46	0.46	0.56	0.67	0.67	0.68
	Set 2 (all MMPs)	0.36	0.48	0.47	0.57	0.68	0.68	0.68
	Set 3 (MMPs A)	0.43	0.61	0.46	0.68	0.78	0.79	0.80
	Set 4 (MMPs N)	0.23	0.28	0.32	0.32	0.41	0.42	0.43
Permeability	Set 1 (all data)	0.46	0.56	0.63	0.57	0.65	0.68	0.71
	Set 2 (all MMPs)	0.48	0.59	0.66	0.62	0.69	0.70	0.75
	Set 3 (MMPs A)	0.64	0.71	0.83	0.68	0.82	0.84	0.85
	Set 4 (MMPs N)	0.11	0.21	0.18	0.20	0.24	0.18	0.41
Clearance	Set 1 (all data)	0.27	0.40	0.38	0.48	0.57	0.57	0.61
	Set 2 (all MMPs)	0.28	0.42	0.39	0.50	0.58	0.59	0.62
	Set 3 (MMPs A)	0.37	0.52	0.37	0.54	0.71	0.72	0.75
	Set 4 (MMPs N)	0.21	0.32	0.37	0.33	0.34	0.35	0.37

Table 6. RMSE (for test set) for all algorithms, data sets and properties discussed in this work.

Property	Data	Model						
		PLS	RF	SVR	XGBoost	DNN-S_20	DNN-S_50	DNN-M_20
logD	Set 1 (all data)	0.86	0.75	0.72	0.61	0.37	0.37	0.39
	Set 2 (all MMPs)	0.84	0.73	0.71	0.59	0.36	0.36	0.38
	Set 3 (MMPs A)	0.72	0.62	0.70	0.52	0.24	0.23	0.24
	Set 4 (MMPs N)	0.86	0.79	0.64	0.70	0.51	0.51	0.54
Solubility	Set 1 (all data)	0.83	0.76	0.77	0.69	0.60	0.60	0.58
	Set 2 (all MMPs)	0.82	0.74	0.75	0.67	0.58	0.58	0.58
	Set 3 (MMPs A)	0.71	0.59	0.70	0.54	0.45	0.44	0.42
	Set 4 (MMPs N)	0.90	0.87	0.85	0.85	0.79	0.78	0.77
Permeability	Set 1 (all data)	0.63	0.57	0.53	0.57	0.51	0.49	0.46
	Set 2 (all MMPs)	0.60	0.54	0.49	0.52	0.47	0.46	0.42
	Set 3 (MMPs A)	0.44	0.40	0.30	0.42	0.31	0.30	0.28
	Set 4 (MMPs N)	0.79	0.74	0.76	0.75	0.73	0.76	0.64
Clearance	Set 1 (all data)	0.45	0.41	0.42	0.38	0.35	0.35	0.33
	Set 2 (all MMPs)	0.44	0.40	0.41	0.37	0.34	0.34	0.32
	Set 3 (MMPs A)	0.39	0.34	0.39	0.33	0.27	0.26	0.25
	Set 4 (MMPs N)	0.47	0.44	0.42	0.43	0.43	0.43	0.42

The analysis of R^2 and RMSE shows the same accuracy ranking for all the models when comparing different data sets: *Set 3* > *Set 2* > *Set 1* > *Set 4*. This is valid for all the properties considered in this work with the exception of logD modelled using PLS and SVR. Predictive models are most accurate for additive data sets (*Set 3*) while nonadditive data (*Set 4*) is least well predictable. Mixed data sets with both additive and nonadditive data (*Set 1* and *Set 2*) being ranked in the middle. There is just a small difference in R^2 and RMSE values between *Set 1* (all data) and *Set 3* (MMPs), indicating that using only MMPs instead of all datapoints does not improve the models significantly.

Figure 1 shows R^2 and RMSE values obtained for *Set 3* (only additive data) using different models for all the properties examined here. Deep learning algorithms usually give the best results (lowest RMSE and highest R^2), followed by XGBoost and then RF. The worst performance is observed for the benchmark PLS.

SVR does not perform significantly better than PLS with the exception of permeability prediction. Permeability is the only property for which the full size of *Set 3* was used for hyperparameter optimization, for the rest of the properties we had to reduce the number of datapoints (see details in the machine/deep learning section). The reduction in set size might be the reason for poor performance of SVR because this algorithm is very volatile to the hyperparameters. In either case even in the fully trained permeability set it does not outperform the deep learning models. Thus, it is reasonable to assume that the potential real performance lies somewhere in between. Another aspect of SVR is its tendency to overtrain. Comparison among all the models presented in Figure 1 shows the most striking difference for SVR (particularly for logD, solubility and clearance) with a significant drop in R^2 from training to test set. In our experience SVR usually performs very well on the training set, but then gives much poorer results on the test set. It cannot be excluded that SVR in general does not perform well on data sets with non-linear error distributions. Nevertheless we did not investigate the matter further as this study is not a comparison between different methods, but rather concerned with the general question of their ability to model small changes in molecules.

In terms of the deep learning models, multi-task modelling improves over single-task for all studied properties except for logD. More hyperparameter trial setting (20 compared to 50) usually makes the models only slightly better and it does not outperform the multi-task modelling (DNN-S_20 and DNN-S_50 in Table 5 and Table 6).

Figure 2 displays correlation between R^2 and RMSE and shows different slopes for each property trendline. The lowest RMSE and R^2 are for clearance, while the highest for logD, with permeability and solubility being placed in between. For most properties the correlation between R^2 and RMSE is linear, with some variation observed for logD *Set 3* (only additive data). The exception here is observed again for permeability *Set 4* which is the smallest set in our analysis (Table 2).

Comparison between additive and nonadditive data (Table 5 and Figure 2) reveals that even deep learning methods have problems with nonadditivity. It can be clearly seen in Figure 2 that only for logD the R^2 range is similar for additive and nonadditive data, while for the other properties it is much shifted towards lower values (below 0.45). This is understandable, there should be no impact of nonadditivity in logD as such as it is a bulk property. The observed nonadditivity might be the result of random experimental errors. For the rest of the studied properties many factors can introduce nonadditivity, like crystal packing in case of solubility, and efflux, sticking to the membranes and metabolism for cell based properties (permeability and clearance) among others.

Figure 3 displays the correlation between measured and predicted solubility values using random forest (RF) and one of the deep learning models (DNN-S_20) for additive and nonadditive data. As expected, the deep learning model performs generally better than RF. This improved performance is observed throughout the whole range of values, for all data sets and properties (SI Figure S5-7).

Conclusions

In this publication, we have evaluated the implications of matched molecular pairs and (non-)additivity on machine learning and deep learning models. We hypothesized that due to the small molecular changes captures in matched molecular pairs, these should be easier to predict than non MMPs. As expected, data sets with only additive datapoints are easiest to predict, as opposed to data sets with only nonadditive datapoints. Mixed data sets with both additive and nonadditive data being ranked in the middle. The sole reduction from all datapoints to MMPs only, does not lead to significant increase in predictability. Using only additive data thus leads to an improvement.

Comparison among properties shows the best performance for logD, followed by solubility, permeability, and clearance. This is in accordance with the complexity of the physicochemical property. In terms of models, deep learning methods give the best results with lowest RMSE and highest R^2. However, our study indicates that even deep learning algorithms have problems with nonadditivity. This highlights the importance of recognition of nonadditive events before building a QSAR/ML model. Moreover, nonadditive data reveal critical changes in SAR and are therefore the most interesting for a medicinal chemist.

Abbreviations

NA: Nonadditivity; AZ: AstraZeneca; SAR: Structure-Activity Relationship; QSAR: Quantitative Structure-Activity Relationship; ML: Machine learning; MMP: Matched Molecular Pair; PLS: Partial Least Squares; RF: Random Forest; SVR: Support Vector Regressor.

Declarations

Acknowledgements

We would like to thank Dea Gogishvili for her work on nonadditivity analysis and the preparation of the data curation pipeline. Jiazhen He thanks the PostDoc program at AstraZeneca.

Authors' contributions

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

Availability of data and materials

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

- S1: Additional figures and tables.
- The Jupyter notebook for data preparation and NA analysis is available on GitHub (https://github.com/MolecularAI/NonadditivityAnalysis).

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

Matched molecular pair generation code was made available by Andrew Dalke on GitHub (https://github.com/rdkit/mmpdb).

Competing interests

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

Funding

Not applicable

References

1. Ehmki ESRR, Kramer C (2017) Matched Molecular Series: Measuring SAR Similarity. J Chem Inf Model 57:1187–1196. https://doi.org/10.1021/acs.jcim.6b00709
2. Tyrchan C, Evertsson E (2017) Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations. Comput. Struct. Biotechnol. J. 15:86–90
3. Awale M, Riniker S, Kramer C (2020) Matched Molecular Series Analysis for ADME Property Prediction. J Chem Inf Model 60:2903–2914. https://doi.org/10.1021/acs.jcim.0c00269
4. Gogishvili D, Nittinger E, Margreitter C, Tyrchan C (2021) Nonadditivity in public and inhouse data: implications for drug design. J Cheminform 13:. https://doi.org/10.1186/s13321-021-00525-z
5. Dimova D, Bajorath J (2016) Advances in Activity Cliff Research. Mol Inform 35:181–191. https://doi.org/https://doi.org/10.1002/minf.201600023
6. 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. https://doi.org/10.1021/jm400147j
7. Hu H, Bajorath J (2020) Introducing a new category of activity cliffs combining different compound similarity criteria. RSC Med Chem 11:132–141. https://doi.org/10.1039/C9MD00463G
8. 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

9. 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. https://doi.org/10.1021/acs.jmedchem.8b00713
10. 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. https://doi.org/10.1016/J.JMB.2010.02.007
11. 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. https://doi.org/https://doi.org/10.1002/anie.201303207
12. Leung CS, Leung SSF, Tirado-Rives J, Jorgensen WL (2012) Methyl Effects on Protein–Ligand Binding. J Med Chem 55:4489–4500. https://doi.org/10.1021/jm3003697
13. 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
14. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P (2013) Comparability of Mixed IC50 Data - A Statistical Analysis. PLoS One 8:e61007. https://doi.org/10.1371/journal.pone.0061007
15. Kramer C (2019) Nonadditivity Analysis. J Chem Inf Model 59:4034–4042. https://doi.org/10.1021/acs.jcim.9b00631
16. 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
17. Sheridan RP, Karnachi P, Tudor M, et al (2020) Experimental Error, Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure-Activity Relationship Models? J Chem Inf Model 60:1969–1982. https://doi.org/10.1021/acs.jcim.9b01067
18. Landrum G (2006) RDKit: Open-source cheminformatics
19. 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
20. Akiba T, Sano S, Yanase T, et al (2019) Optuna: A Next-generation Hyperparameter Optimization Framework. Proc ACM SIGKDD Int Conf Knowl Discov Data Min 2623–2631. https://doi.org/10.1145/3292500.3330701
21. Gilmer J, Schoenholz SS, Riley PF, et al (2017) Neural Message Passing for Quantum Chemistry
22. Yang K, Swanson K, Jin W, et al (2019) Analyzing Learned Molecular Representations for Property Prediction. J Chem Inf Model 59:3370–3388. https://doi.org/10.1021/acs.jcim.9b00237
23. Bergstra J, Yamins D, Cox DD (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures

Figures

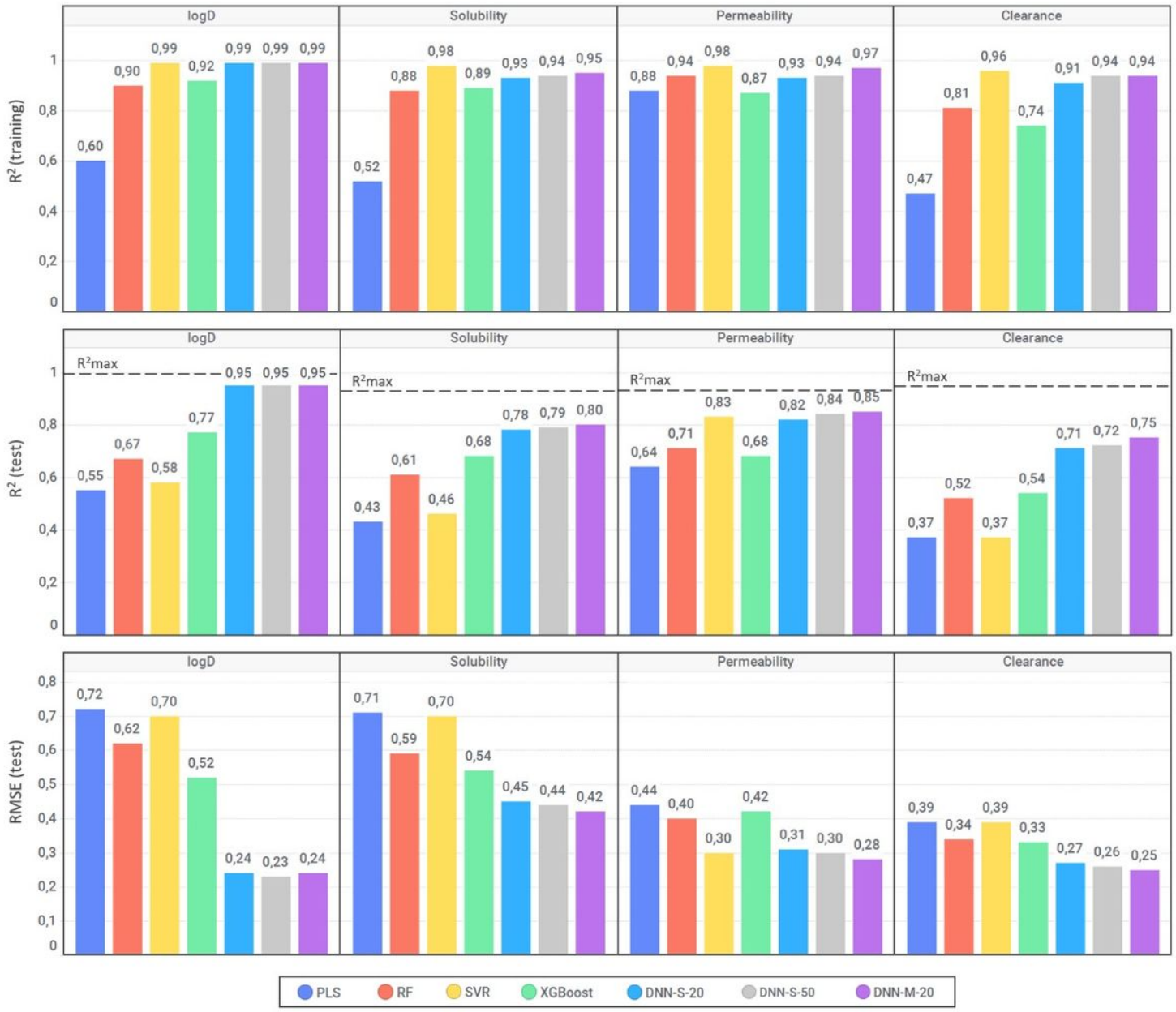

Figure 1

R^2 and RMSE for *Set 3* (only additive data). Comparison of different models and endpoints. R^2_{max} (dashed line) is the upper limit for R^2 derived from experimental uncertainty (Table 3).

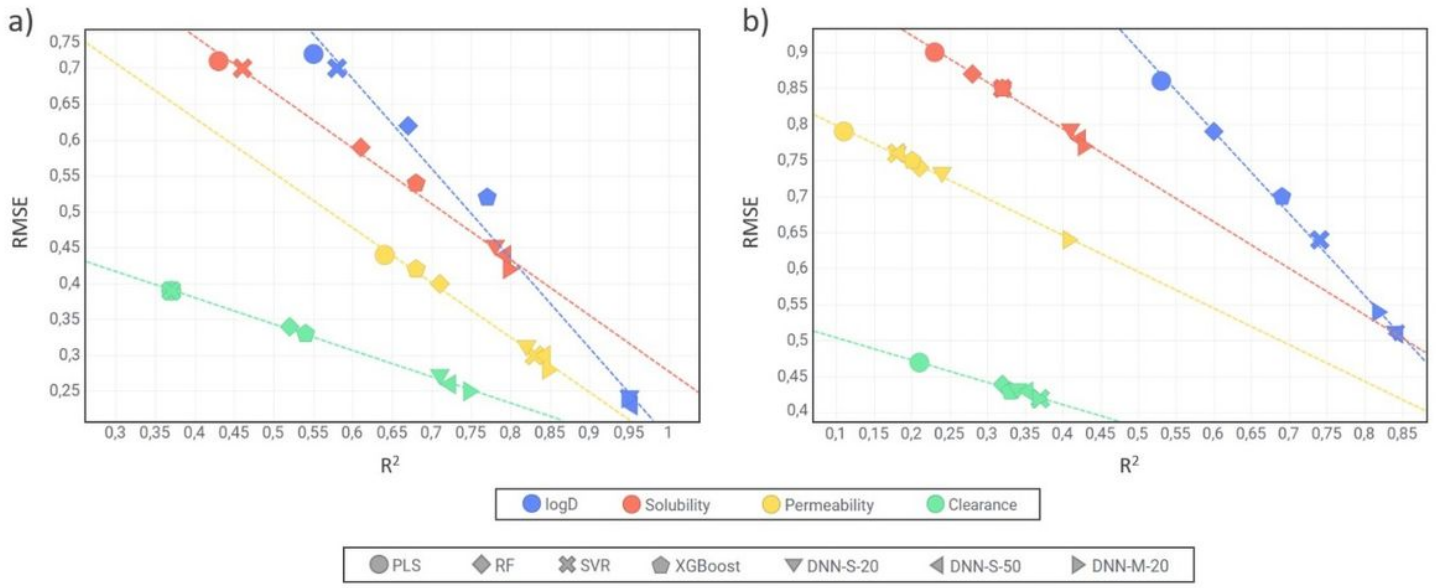

Figure 2

R^2 against RMSE for test a) *Set 3* (only additive data) and b) *Set 4* (only nonadditive data). Comparison of different models and endpoints.

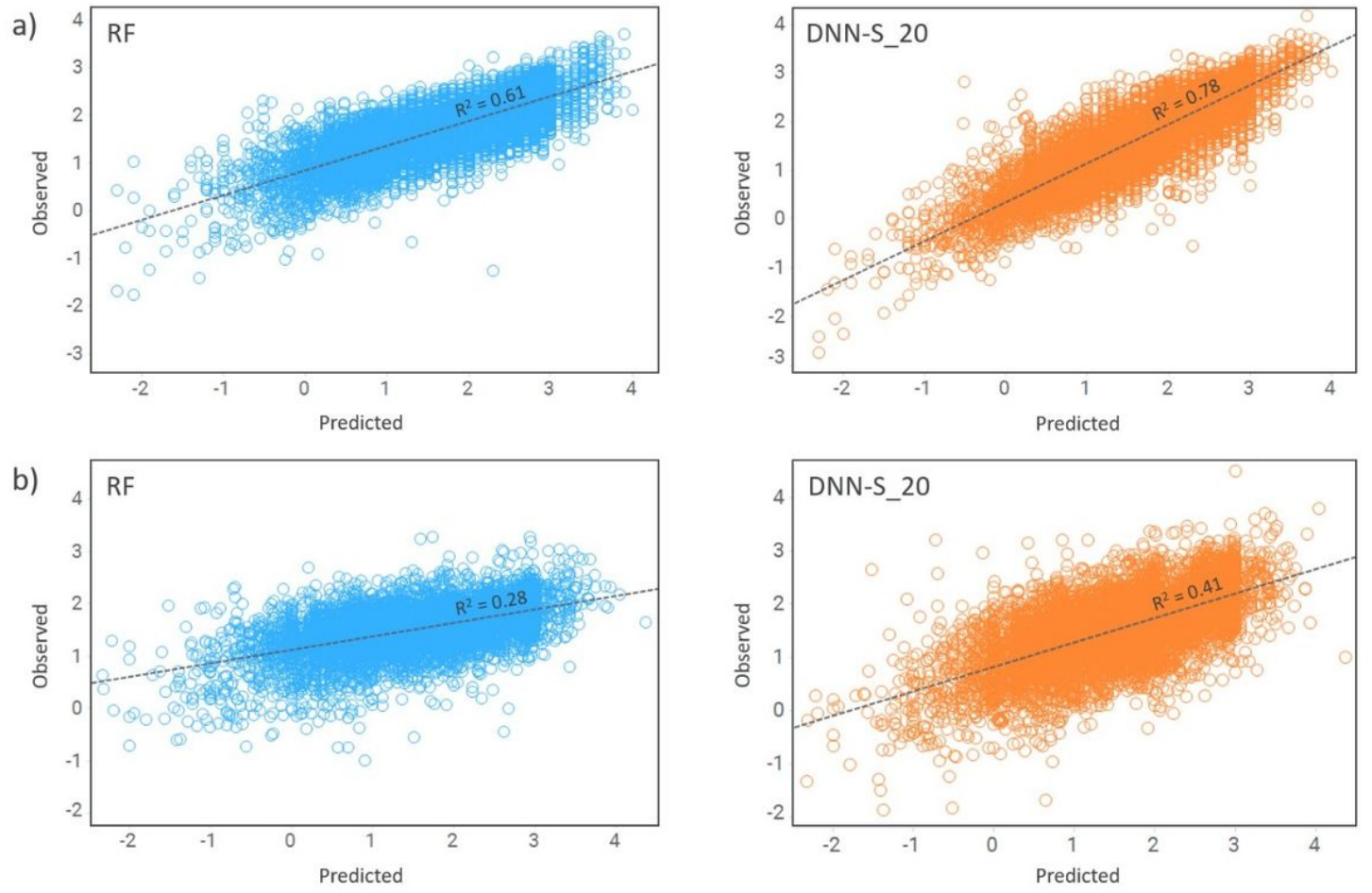

Figure 3

Predicted versus measured values for solubility. Comparison between RF (blue) and DNN-S_20 (orange) for a) *Set 3* (only additive data) and b) *Set 4* (only nonadditive data). The values are in log units.

Supplementary Files

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

- abstract2.png
- SupportingInformation7dec.docx