Testing Biases in QSAR Set Composition

or

Predicting Values for the Whole Dataset

Rajarshi Guha

Penn State University

Terminolgy

Some terms that need to be defined

- **QSAR Sets**: A collective term for the training, cross validation (CV) and prediction sets.
- **Proto QSAR Sets**: Groups of molecules that are used to generate multiple QSAR sets by shuffling
- **All Sets Procedeure**: The method by which a specific CNN model is tested to remove bias in QSAR set composition

CNN Models

- CNN models are built using 3 sets.
- The training set is used to build the model
- The CV set is used to check on quality of the model during training
- The prediction set is used to test the ability of the model to predict values for unseen cases
- All sets are mutually exclusive

Problems With This Approach

- It is possible that the QSAR sets with which a good model is built are a lucky combination
- QSAR sets composed differently might lead to a model with poorer statistics
- This method only provides predictions for a subset of the whole dataset

Goals of a Better Method

- Remove the bias inherent in a single QSAR set combination
- Obtain multiple predictions for each memeber of the dataset

The All Sets Procedeure

- The method is based on the ensemble technique for CNN models
- The dataset is grouped into N groups
- The groups are combined to give N QSAR sets
- Each group acts as the prediction set once
- Thus each molecule in the dataset is predicted once
- The procedure is repeated multiple times randomly generating the initial grouping of the dataset
- As a result each molecule in the dataset is predicted multiple times and the average can be taken as the final predicted value

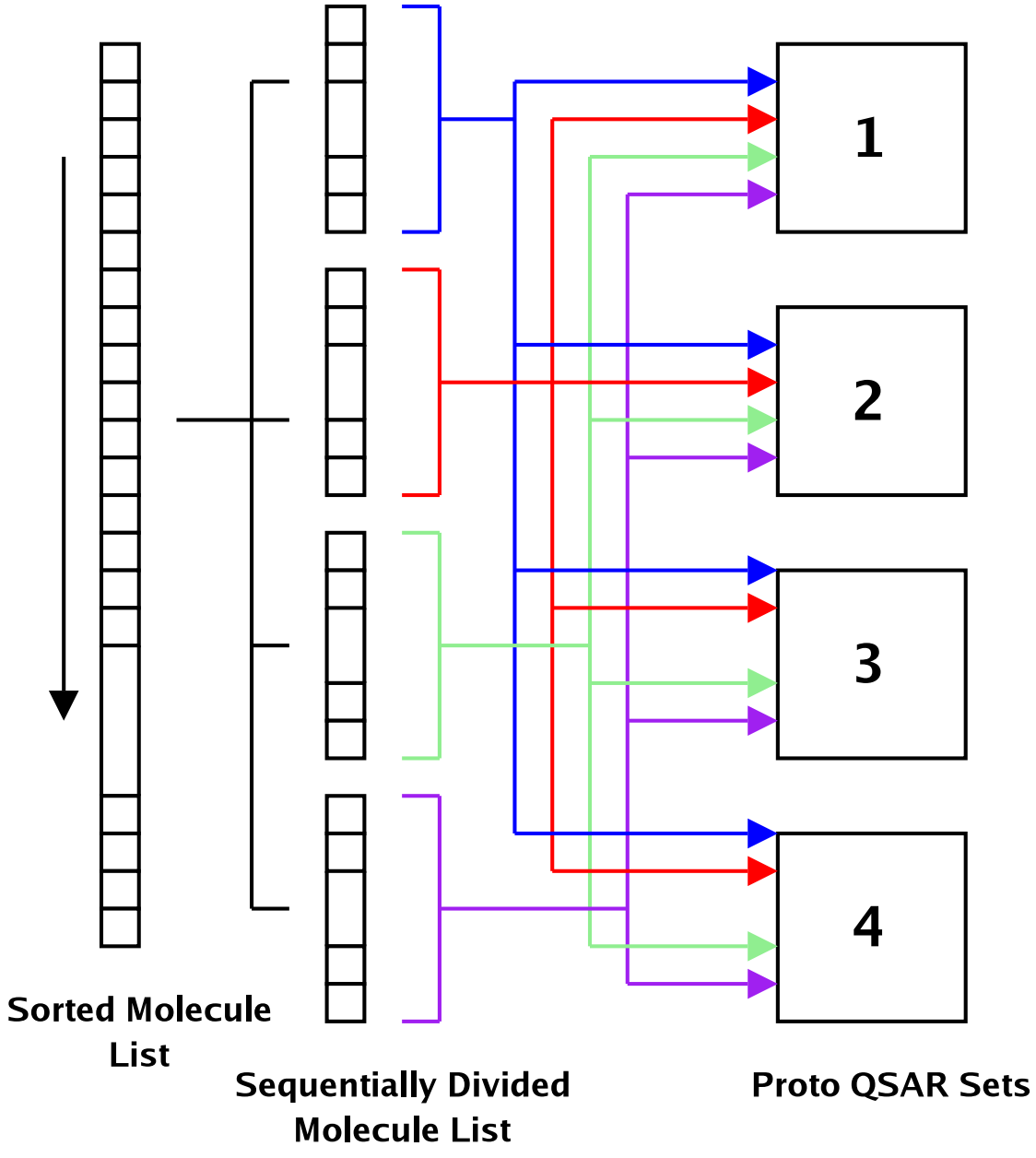
The Procedure in Detail

- Assume that we want a leave 25% out procedure.
- Thus we should have 4 prediction sets (and corresponding CV & training sets)
- The dataset is sorted according to the value of the activities.
- To make the length of the dataset a multiple of 4 molecules from the end of the list are removed into the group called **remainder**.

The Procedure in Detail

- The dataset is then divided into 4 groups, by placing the first 25% into the first group, say A, the next 25% into the next group, say B and so on
- Next, we create 4 empty groups - the proto QSAR sets, labelled **pqs1**, **pqs2** etc
- The molecules from group A are distributed randomly into **pq*** so that any given molecule is in only one proto QSAR set
- Repeat for groups B, C and D
- Distribute the molecules considered as **remainder** randomly into **pq***

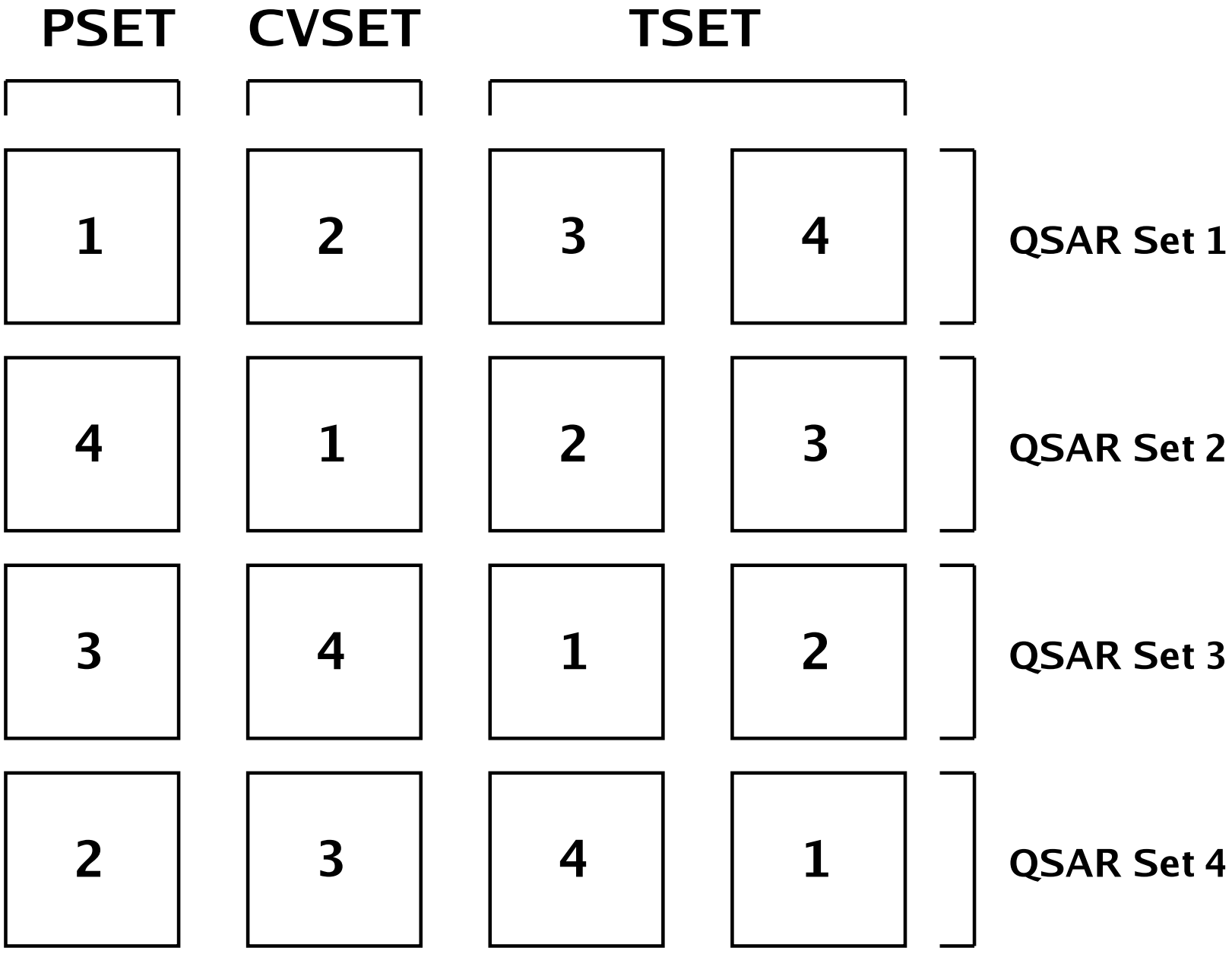
The Procedure in Detail



Proto QSAR Sets to QSAR Sets

- In general the first proto QSAR set (**pqs1**) is the prediction set, the second the CV set and the remainder are combined to give the training set.
- A CNN model is built using the resultant QSAR sets.
- Next, the sets are right rotated by one position so that the first set is now **pqs4**.
- A CNN model is built with the QSAR sets resulting from this ordering.
- This process is repeated two more times, each time right rotating the order of the proto sets once.
- **Result**: Each element of the data set is placed in the prediction set once.

Proto QSAR Sets to QSAR Sets



The Procedure in Detail

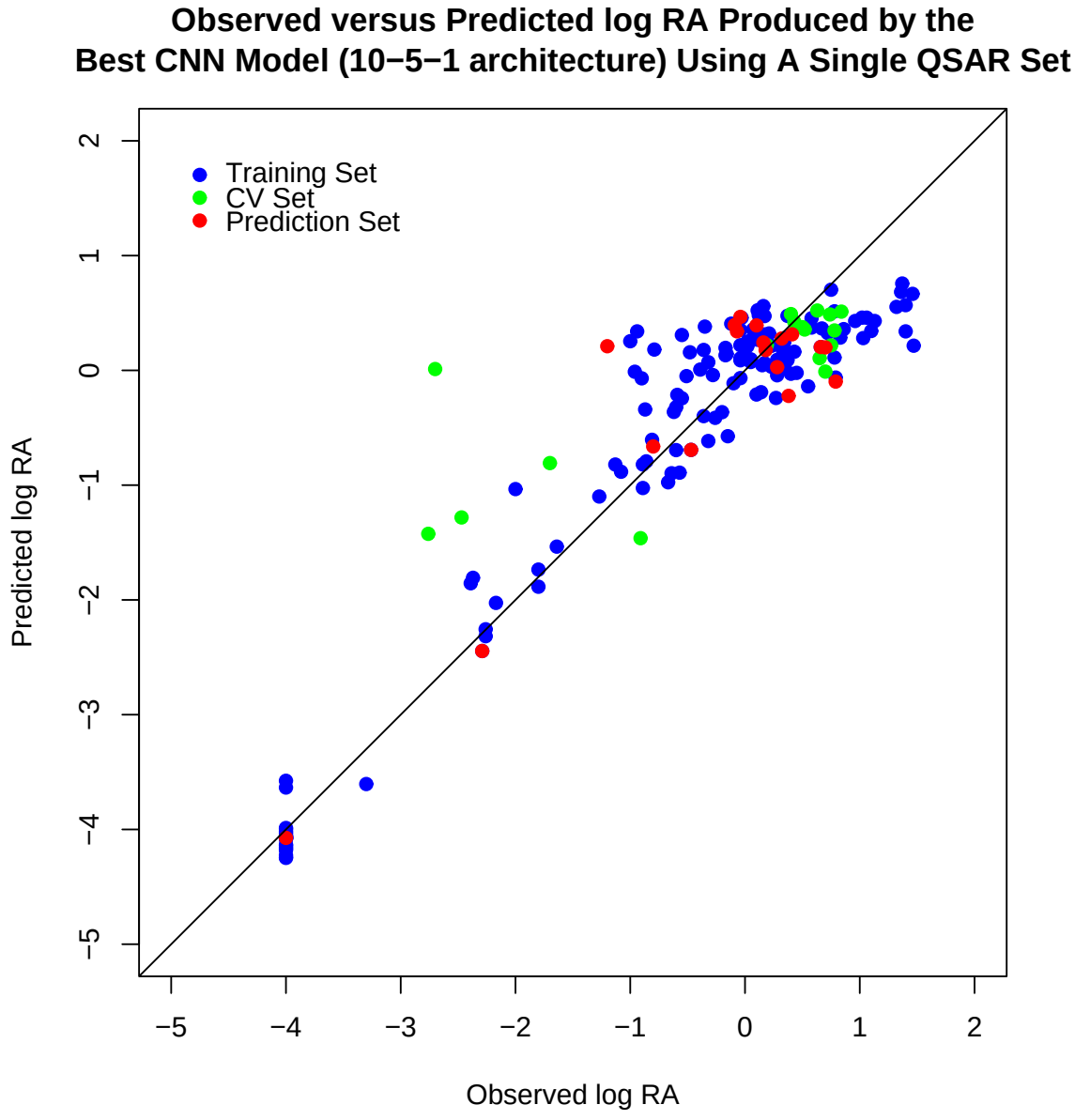
- The result of this procedure is to generate a prediction for each element of the dataset
- The whole procedure is then repeated multiple times
- Each time, the distribution of the divided dataset into the proto QSAR sets is random

An Example - Single QSAR Set

- A dataset of 179 molecules was chosen
- The range of the dependent variable was -4.00 to 1.47
- A CNN model was built using 10 descriptors using a 10-5-1 architecture.

RMSE			R^2		
TSET	CSET	PSET	TSET	CSET	PSET
0.46	0.93	0.62	0.93	0.62	0.82

An Example - Single QSAR Set

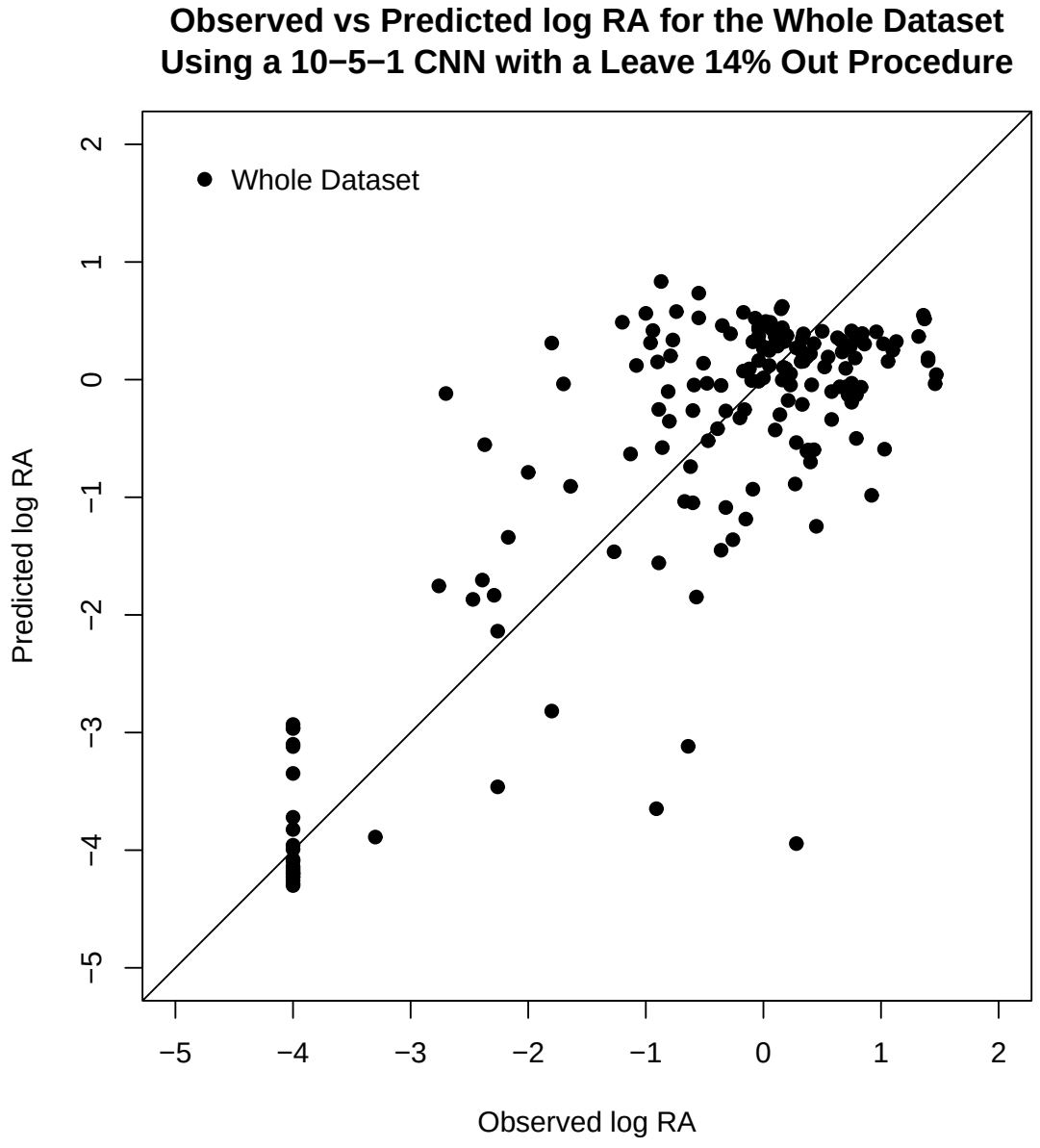


An Example - Leave 14& Out Method

- A leave 14% procedure resulted in 7 different prediction sets and hence 7 models.
- The procedure was carried out 3 times
- Each element of the dataset was predicted three times:
- The RMSE and R^2 of the averaged predictions for the whole dataset was 0.84 and 0.84 respectively.
- The model was thus regenerated with a total of 21 different QSAR sets

	RMSE			R^2		
	TSET	CSET	PSET	TSET	CSET	PSET
Mean	0.53	0.68	0.89	0.88	0.81	0.69
Std. Dev	0.05	0.07	0.18	0.02	0.04	0.11

An Example - Leave 14& Out Method



Observations

- Clearly, when multiple QSAR sets are considered, the results degrade
- The RMSE & R^2 for the prediction set in the original (single QSAR set) method are significantly better
- The fact that carrying out a leave n% out procedure degrades the RMSE & R^2 values for the predictions indicates the previous results were possibly due to a lucky QSAR set composition

Conclusion

- The leave n% out procedure allows us to make multiple predictions for the whole dataset, which can be averaged
- Since the QSAR sets are randomly generated each time, we are assured of removing biases due to set composition