A QSAR Study of Artemisinin Analogs (II)

Rajarshi Guha

Penn State University

Initial Study

- Molecules
  - Ignored all enantiomeric pairs
  - 179 molecules
- Descriptors
  - Used all the descriptors
  - Reduced pool had 64 descriptors
  - 8 HPSA descriptors present
  - 3 atom pair descriptors present

Type I Models

- The 19 best models were considered
- The highest R^2 for the TSET was .52 (9 descriptors)
- 17 out of 19 models contained 1 or more HPSA descriptors.
- An atom pair descriptor only occurred once.

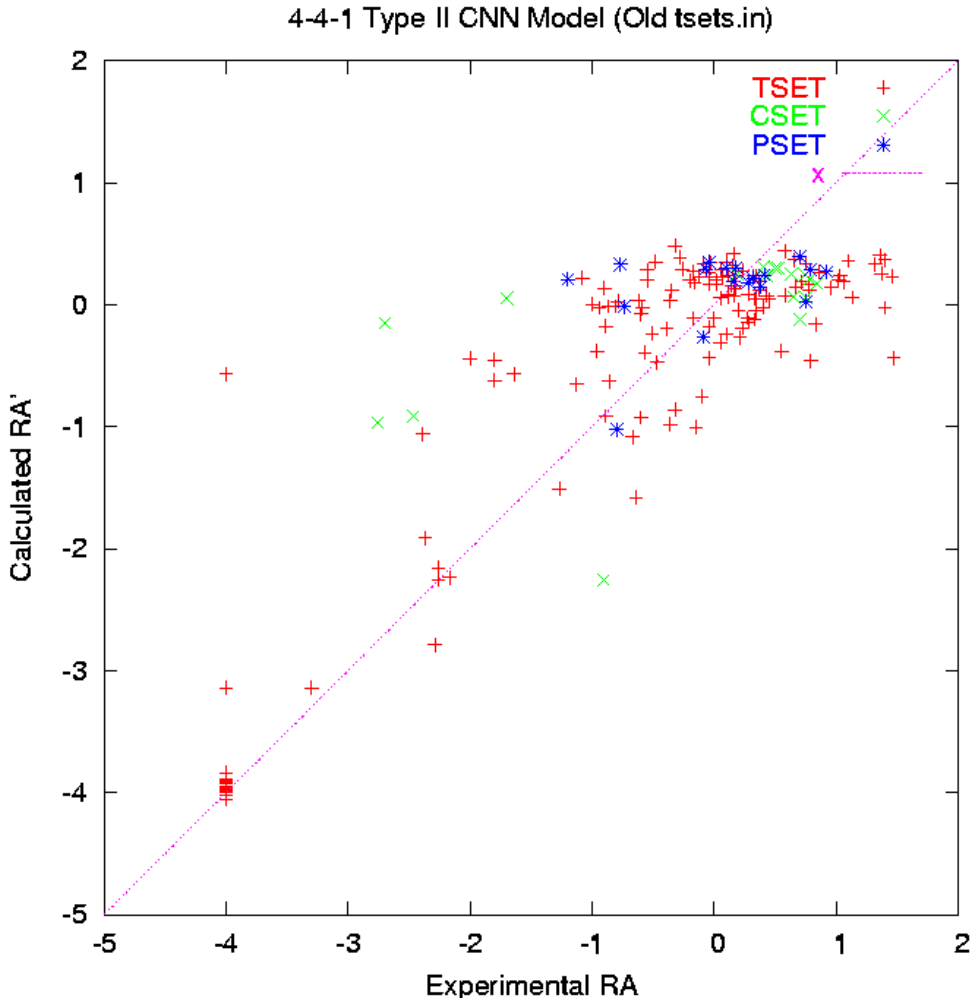
Type II Models

- 4 to 10 descriptor models were generated.

	R^2			RMSE		
Arch	TSET	CSET	PSET	TSET	CSET	PSET
4-4-1	.83	.31	.14	.69	1.09	.58
5-4-1	.85	.26	.02	.70	1.14	.71
6-4-1	.88	.38	.002	.57	1.04	.93
7-7-1	.93	.48	.0001	.45	.95	.92
8-5-1	.91	.47	.007	.51	.99	.70
9-2-1	.88	.38	.006	.60	11.05	.67
10-5-1	.93	.56	.03	.45	.92	.88

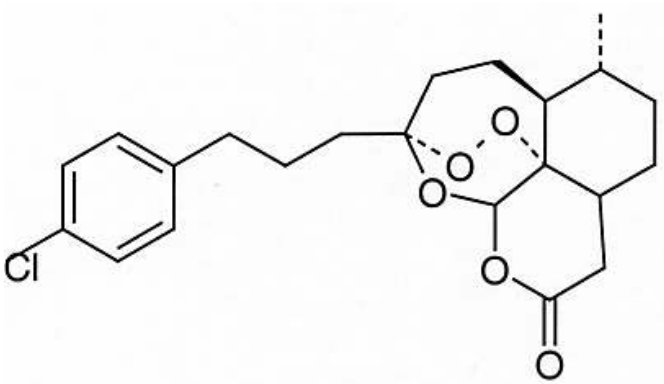
Massaging *tsets.in*

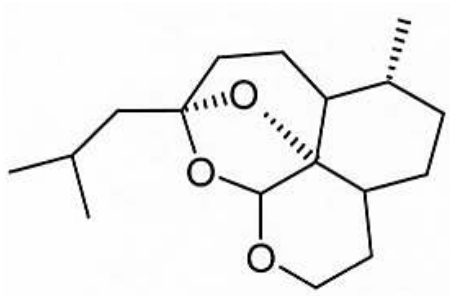
- All Type II models have plots similar to the one below

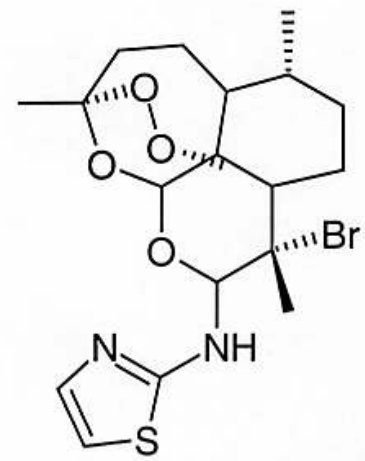


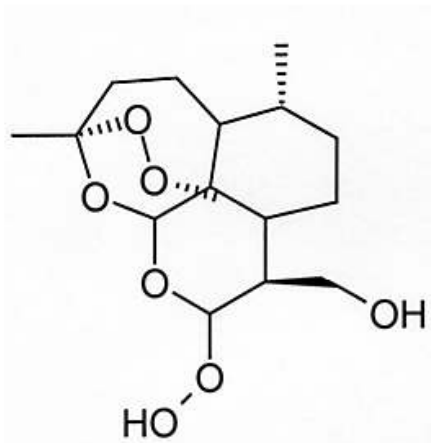
Massaging *tsets.in*

- Swapped 4 compounds of the PSET with 4 compounds of the TSET

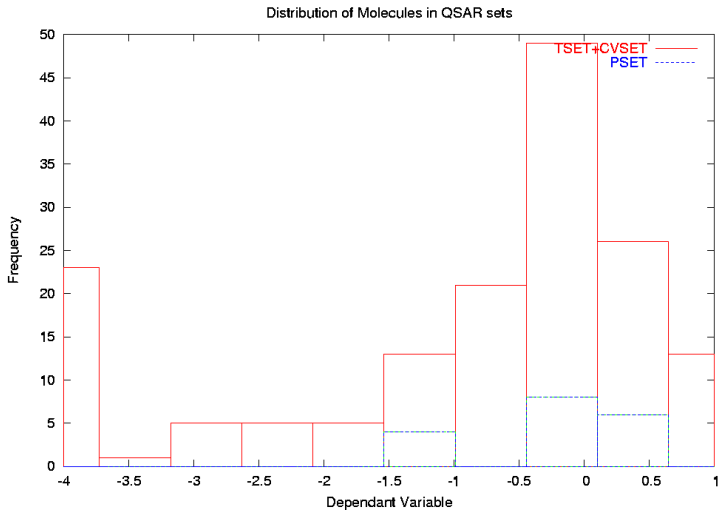




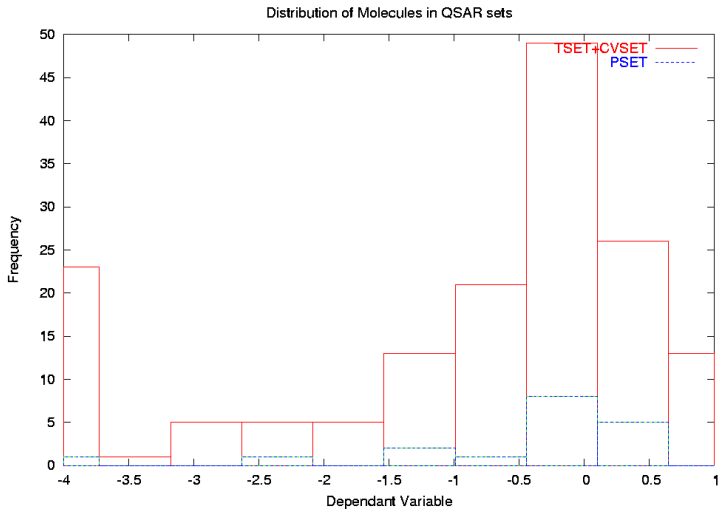




Comparison of Set Distributions



Original tsets.in



tsets.in after swapping

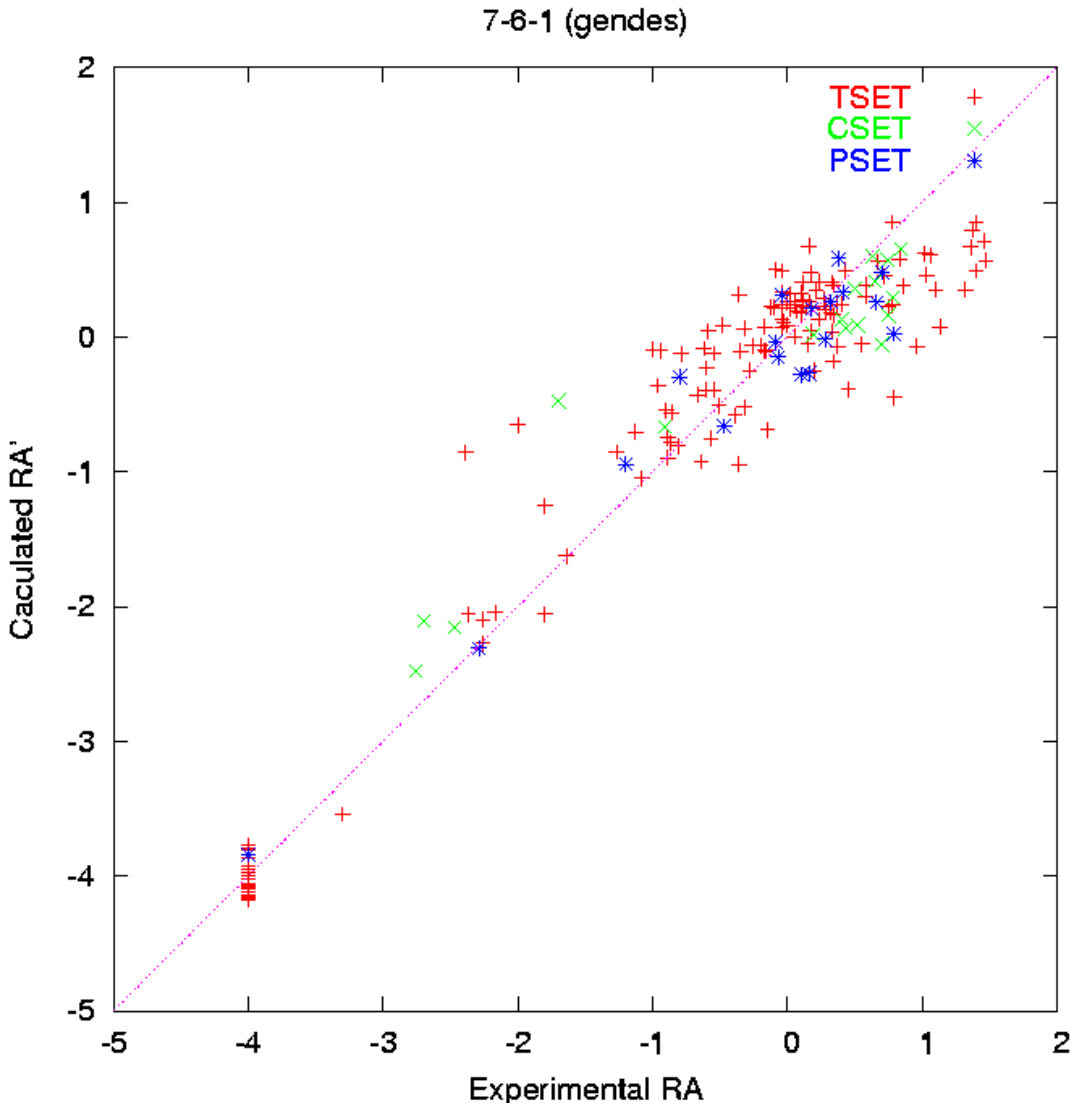
Type III Models

- Best Type III models using swapped sets

	R^2			RMSE			
Arch	TSET	CSET	PSET	TSET	CSET	PSET	Cost
7-6-1	.91	.93	.93	.47	.59	.44	.51
8-4-1	.92	.93	.61	.47	.47	.89	.48
9-6-1	.94	.86	.82	.42	.60	.79	.49
9-5-1	.93	.86	.82	.46	.58	.71	.51
6-4-1	.90	.92	.80	.52	.53	.63	.52

- Descriptors for the 7-6-1 model
  - ALLP-5, N7CH, WTPT-5, MDE-13, MDE-23, GEOM-6, THWS-1

Plot for the 7-6-1 Model



PNN Models

- Several PNN models were created with the modified tsets.in
- None of the results were very impressive

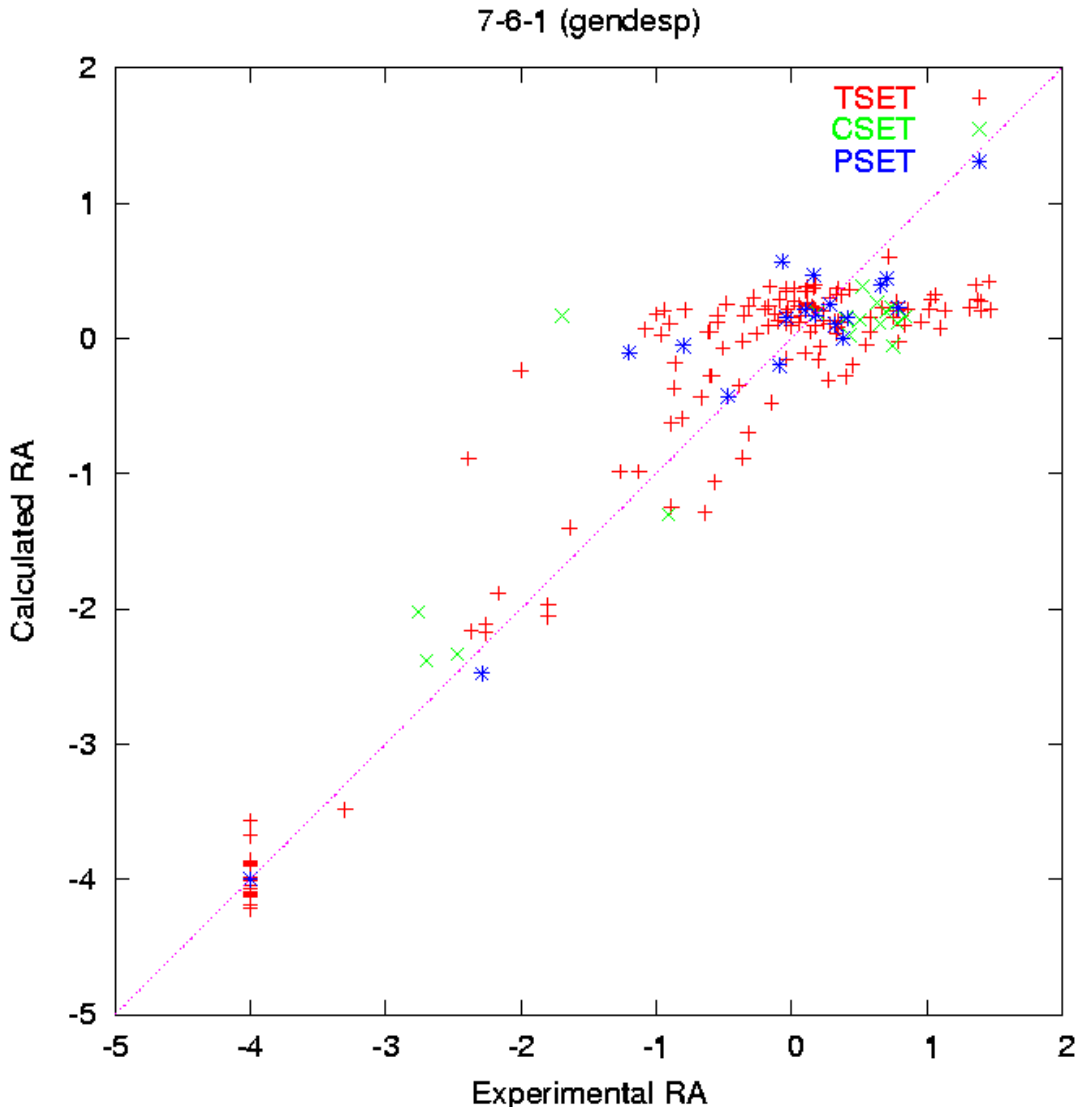
		RMSE	
Size	Cost	TSET	PSET
2	0.65	0.80	0.62
3	0.50	0.70	0.50
4	0.47	0.69	0.65
5	0.45	0.67	0.74
7	0.41	0.64	0.84

Results from *gendesp*

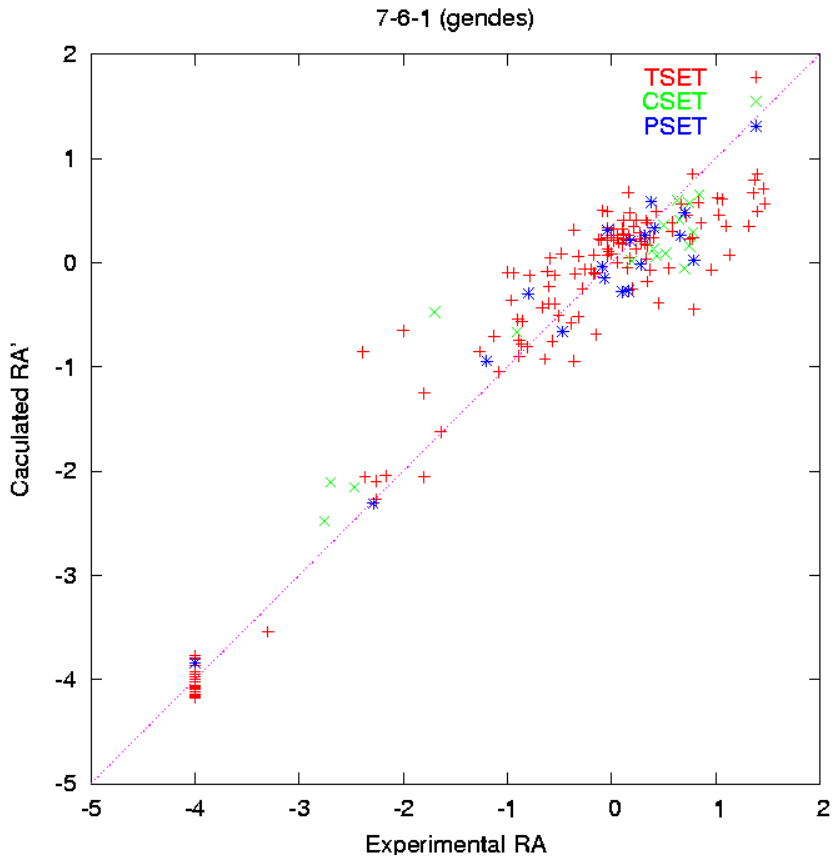
- The good Type III results could be the result of a lucky choice
- A *gendesp* run should reduce the chances of good results from a lucky tsets.in

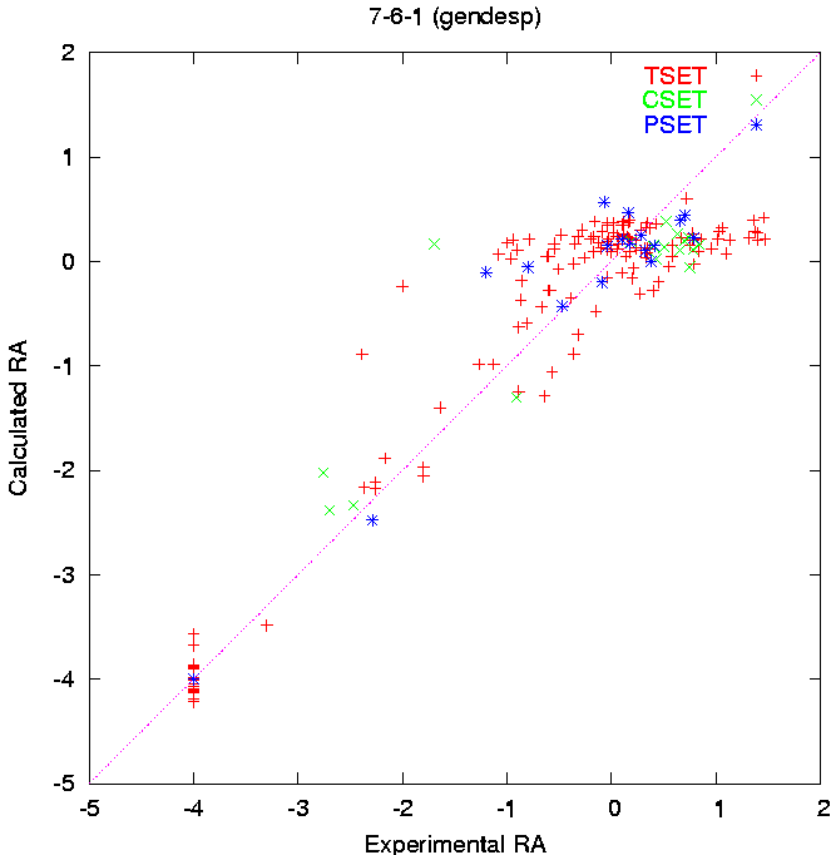
	RMSE			R^2		
Method	TSET	CVSET	PSET	TSET	CVSET	PSET
gendesp	0.54	0.69	0.46	0.89	0.78	0.87
gendes	0.47	0.59	0.44	0.93	0.91	0.93

Plot for the 7-6-1 (gendesp) Model



Comparison: gendes & gendesp





Randomization Tests

- Random Training, CV and Prediction Sets
- Architecture: 7-6-1

RMSE			R^2		
TSET	CVSET	PSET	TSET	CVSET	PSET
0.47	0.53	1.31	0.92	0.85	0.48

Randomization Tests

- Scrambled Dependant Variable
- Architecture: 7-6-1

RMSE			R^2		
TSET	CVSET	PSET	TSET	CVSET	PSET
0.89	1.06	1.52	0.77	0.51	0.34

Conclusions

- It appears that setbin.py was lucky in generating a set which performed well
- To prevent this form of bias affecting the final models, *gendesp* should be run in preference to *gendes*