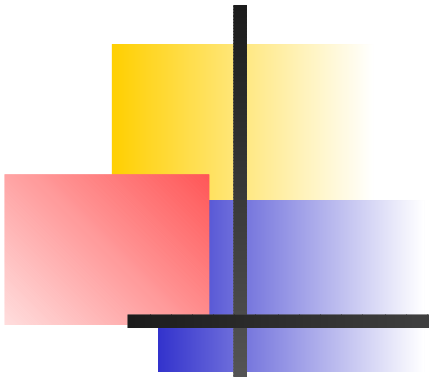

A Comparison of Set Generation Methods

Rajarshi Guha

Penn State University

The Set Generation Study

- Investigates how sets generated by different methods perform
- The methods are: KSOM, Tropsha Method and Activity Binning
- The pcDHFR dataset was used to test the generated QSAR sets.
- Only Type III models are summarized

Preliminaries

- Both the KSOM and the Tropsha method require external descriptor sets to generate the sets
- The following Dragon descriptor sets were used
  - BCUT & 2D Autocorrelation
  - BCUT & Galvez topological indices
  - Getaway
  - MoRSE & 2D Autocorrelation
  - MoRSE & Getaway
  - MoRSE & WHIIM

Preliminaries

- The Tropsha method requires the user to specify the *Disimilarity Level* (DL)
- DL for each Dragon set were chosen such that the PSET was 10% of the total dataset and TSET+CVSET was the remainder
- The CVSET was randomly selected from the TSET

Results: Activity Binning

- These are the original published results
- Best Model: 10-6-1

RMSE			R^2		
TSET	CVSET	PSET	TSET	CVSET	PSET
0.45	0.49	0.66	0.83	0.78	0.66

Results: KSOM

- The MoRSE - WHIM & MoRSE - 2D Auto Correlation sets appeared to give good results
- In both cases the architectures were simpler
- RMS errors were comparable

Arch	RMSE			R^2		
	TSET	CVSET	PSET	TSET	CVSET	PSET
5-3-1[a]	0.63	0.63	0.68	0.68	0.60	0.64
6-5-1[b]	0.60	0.61	0.65	0.75	0.78	0.67

[a] MoRSE - 2D Autocorrelation

[b] MoRSE - WHIM

Results: Tropsha Method

- The results are not remarkable
- *Best* results were obtained from the Getaway set
- None of the architectures were significantly better than the original

Arch	RMSE			R^2		
	TSET	CVSET	PSET	TSET	CVSET	PSET
8-5-1[a]	0.51	0.56	0.63	0.75	0.80	0.67

[a] Getaway

Summary of Results

Method	Arch	RMSE			R^2		
		TSET	CVSET	PSET	TSET	CVSET	PSET
AB[a]	10-6-1	0.45	0.49	0.66	0.83	0.78	0.66
KSOM	5-3-1	0.63	0.63	0.68	0.68	0.60	0.64
KSOM	6-5-1	0.60	0.61	0.65	0.75	0.78	0.67
TM[b]	8-5-1	0.51	0.56	0.63	0.75	0.80	0.67

[a] Activity Binning

[b] Tropsha Method