COMP-238
Model Applicability Domains: When Can I Use my Model?
235th National Meeting of The American Chemical Society
New Orleans, LA
April 9, 2008

Combining global and local approaches to model domain applicability

Rajarshi Guha, School of Informatics, Indiana University

David T. Stanton, Modeling & Simulations Department, Corporate Research, Procter & Gamble

The Big Picture

- A molecule that is not similar to the TSET is probably going to be poorly predicted
  - We shouldn't rely on such a prediction
- If a molecule is similar to the TSET, it may or may not be well predicted
  - Can we rely on such a prediction?
  - Why should a molecule similar to the TSET be poorly predicted?

Traditional Approaches

- Most approaches to model applicability consider the similarity to TSET
- "Extra" features in a molecule may be invisible to the model descriptors
  - The extra features cause the observed property to differ from that of similar molecules in the TSET
  - We see the effect of this as a large residual

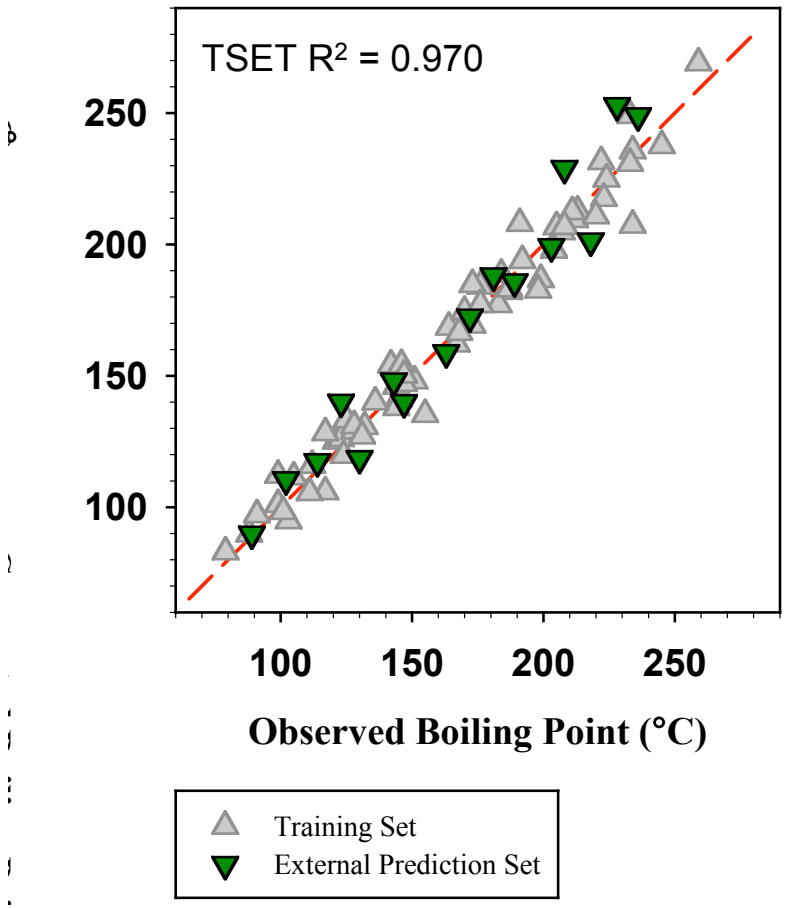
*How do we decide that a molecule, similar to the training set, is **actually** similar to the training set*

Simple Example – Boiling Point Model

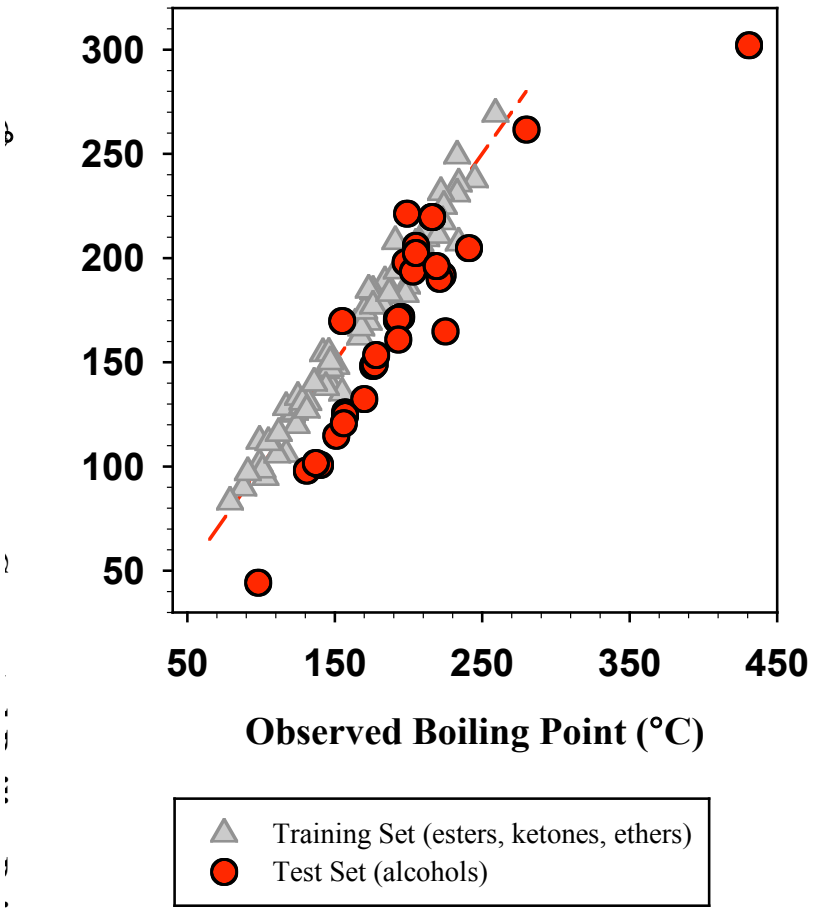
- C_4 – C_{12} Non-aromatic, containing at least 1 oxygen
  - Model development: 80 esters, ethers, ketones
    - 64 in Training set
    - 16 in Validation set
  - New External Prediction Set: 32 alcohols
- 6-variable model
  - R^2 = 0.970
  - Good fit for Validation set (r = 0.973)
- Applied model to predict BP for alcohols

Simple Boiling Point Model - Results

Model Development Results



External Prediction Results

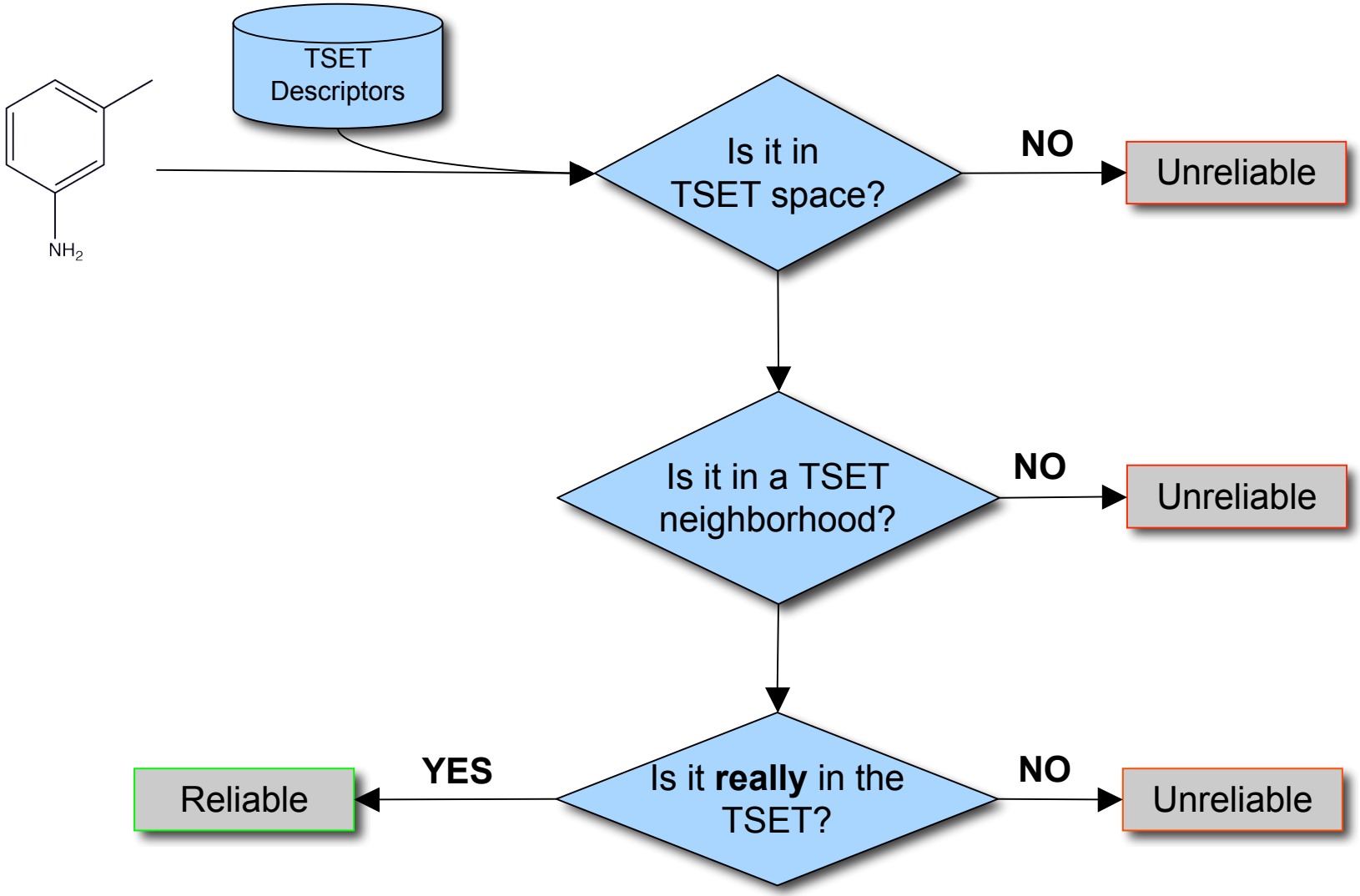


The model under predicts most of the new observations

Indiana University School of informatics

P&G Research and Development

A Hierarchical Strategy

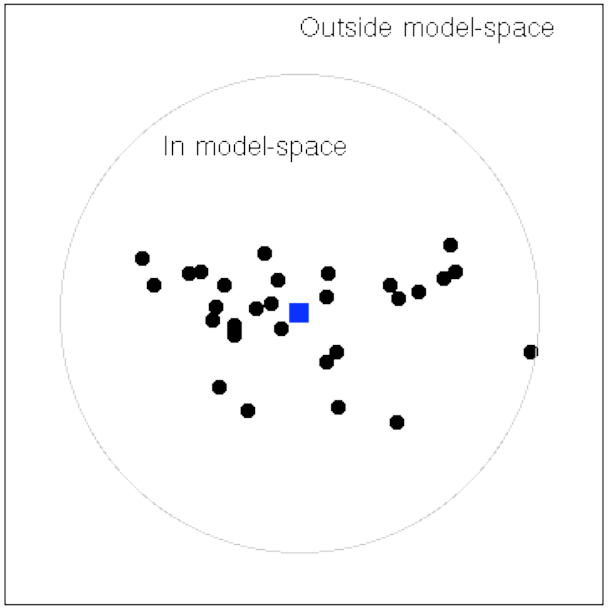


Indiana University School of informatics

P&G Research and Development

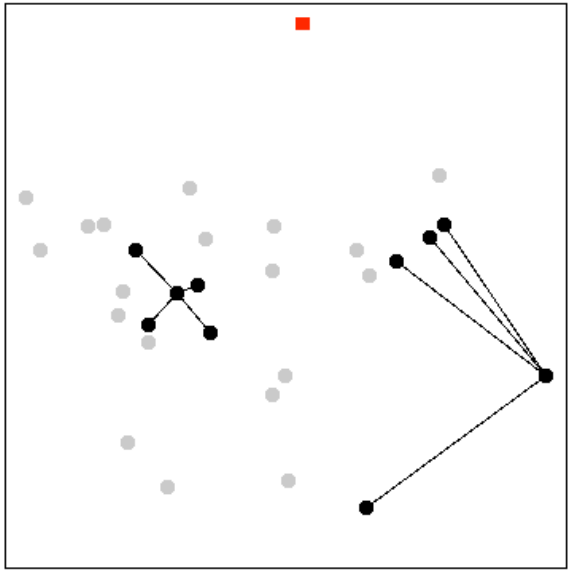
Model-Space Characterization

- Evaluate mean and std dev of the distances of the TSET points to the centroid of the TSET
  - A PSET point is *in model-space* if its distance to the centroid is within 2SD of the mean TSET distances
- Crude classification
- But we can safely ignore points that lie outside model space



Neighborhood Characterization

- Evaluate the average pairwise distance between 5-NN of each TSET point
- Summarize the whole TSET in terms of mean and standard deviation
  - For a PSET point, determine the average 5-NN distance in the TSET
  - Apply a similar classification rule as before
- Isolated TSET points can skew the mean and std dev

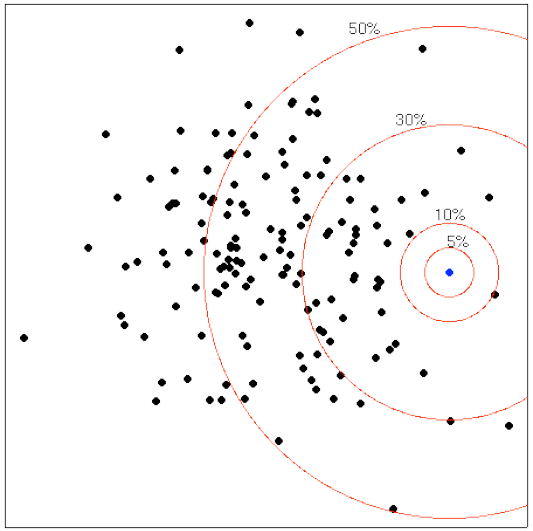


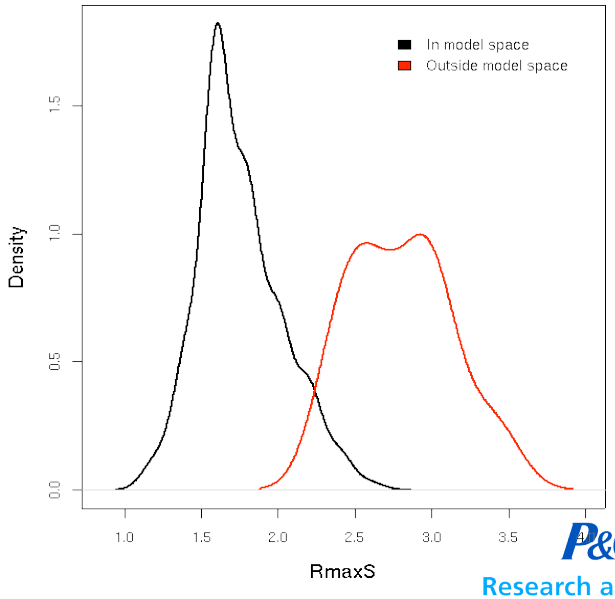
Indiana University School of informatics

P&G Research and Development

Neighborhood Characterization

- The *R*-NN curve method numerically measures the density of space around a given point
  - The measure is called $R_{max(S)}$ and higher values indicate a more sparse location
  - Makes no assumptions about the spatial distribution of points

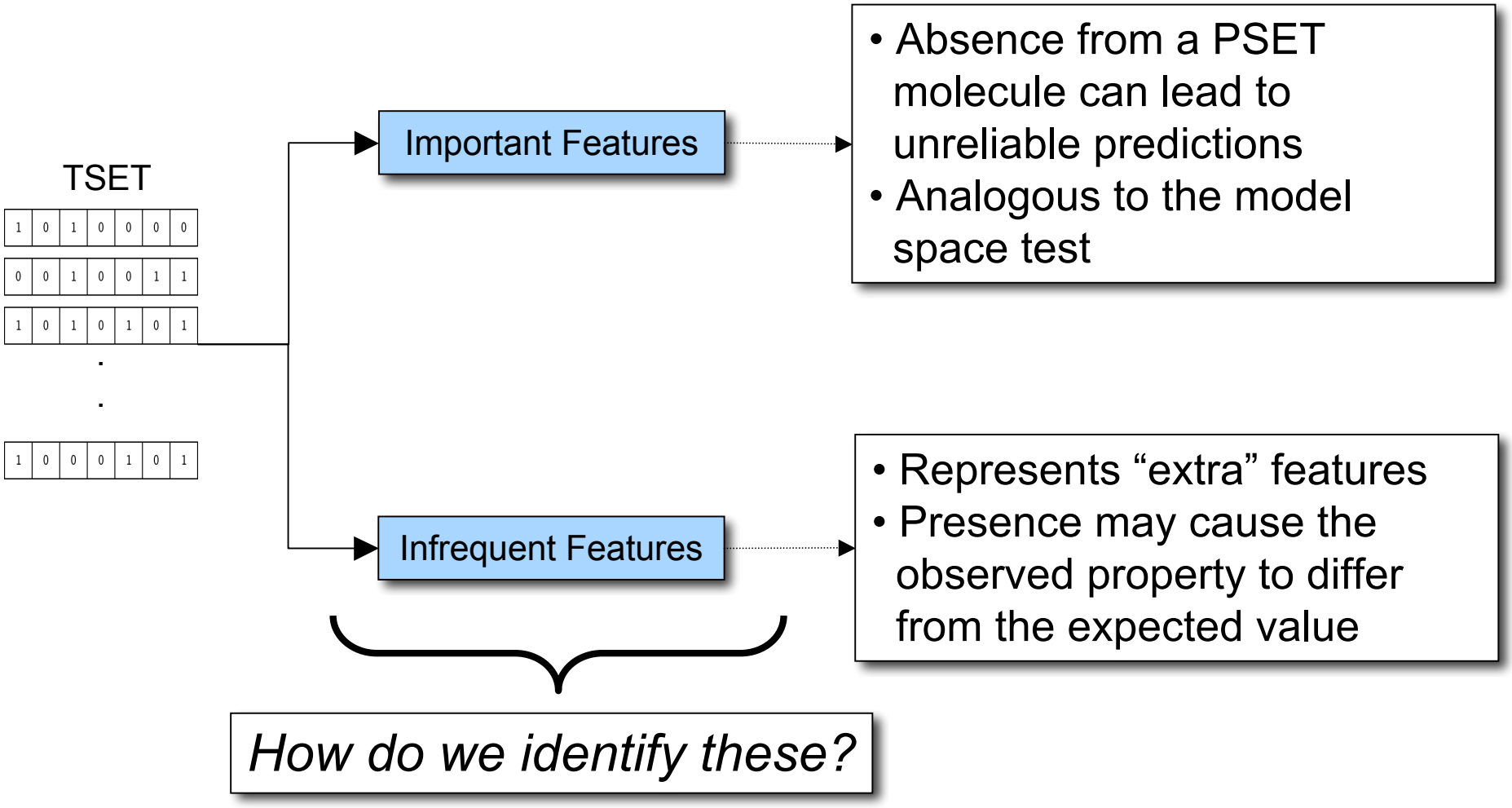




Indiana University School of informatics

P&G Research and Development

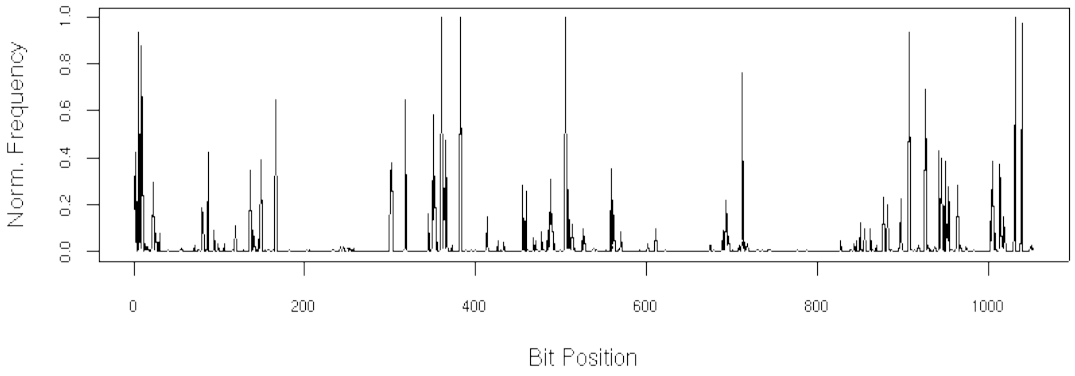
Global Features - Structural Fingerprints

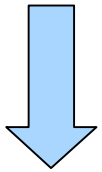


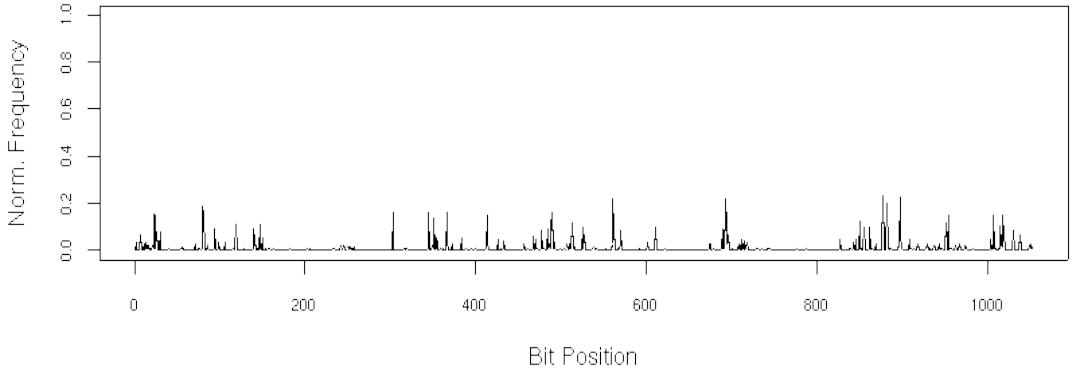
Global Features - Structural Fingerprints

TSET

1	0	1	0	0	0	0
0	0	1	0	0	1	1
1	0	1	0	1	0	1
.						
.						
.						
1	0	0	0	1	0	1







- Summarize the TSET in terms of bit frequencies
- Identify bits that have low frequencies
- Identify bits that have intermediate frequencies

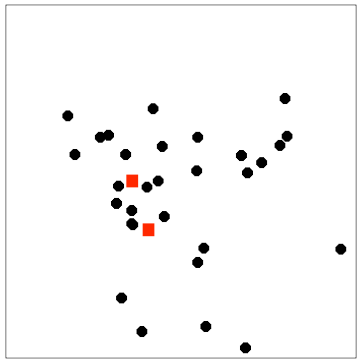
Infrequent Features

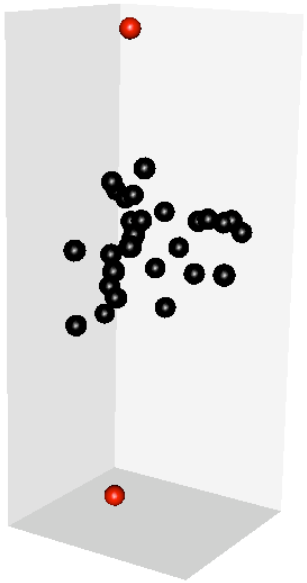
Indiana University School of informatics

P&G Research and Development

Augmenting Descriptor Spaces

- Is a PSET molecule truly within the TSET?
- Augment the model space with another dimension
  - If the distance and neighborhood classifications don't hold, maybe it really isn't in model space





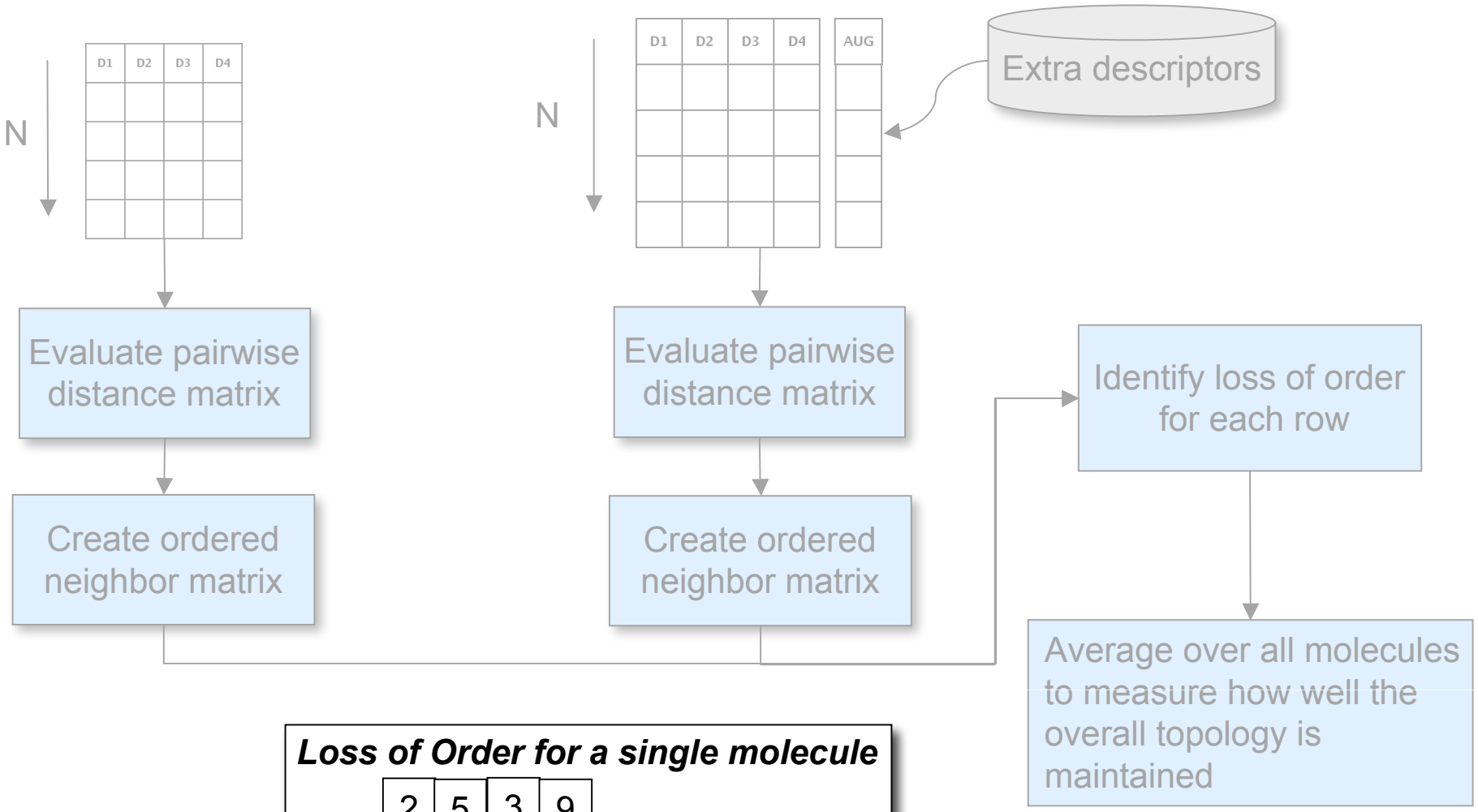
Indiana University School of informatics

P&G Research and Development

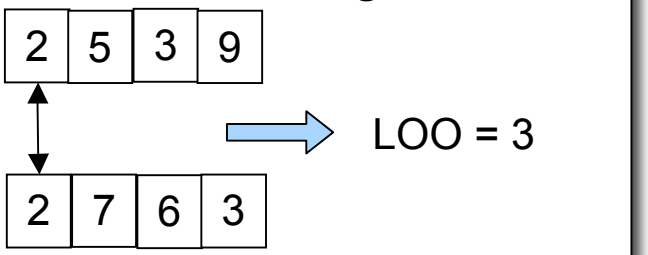
Augmenting Descriptor Spaces

- We don't know what this extra dimension should be
- We might need more than one extra dimension
- But we can provide a few constraints
  - The topology of the TSET should be the same in both spaces
  - The extra dimensions should not be highly correlated to the model descriptors
  - The distributions in the extra dimensions should not be the same for the TSET and PSET

Searching for Suitable Spaces



Loss of Order for a single molecule

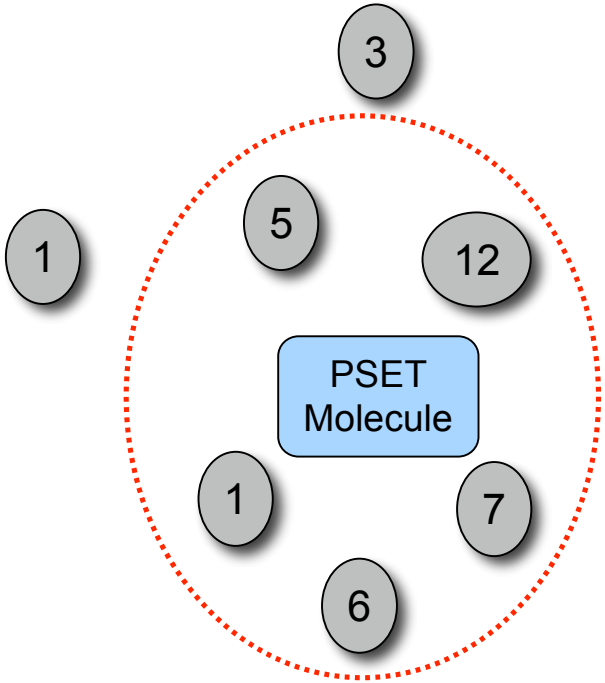


Indiana University School of informatics

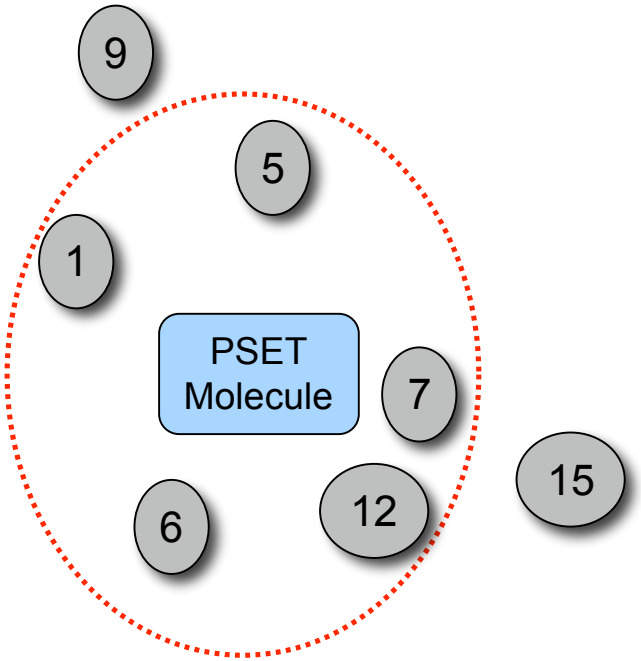
P&G Research and Development

Using Augmented Descriptor Spaces

Original Descriptor Space



Augmented Descriptor Space



- If the nearest neighbors stay the same in the augmented space, the PSET molecule is most likely similar to the TSET

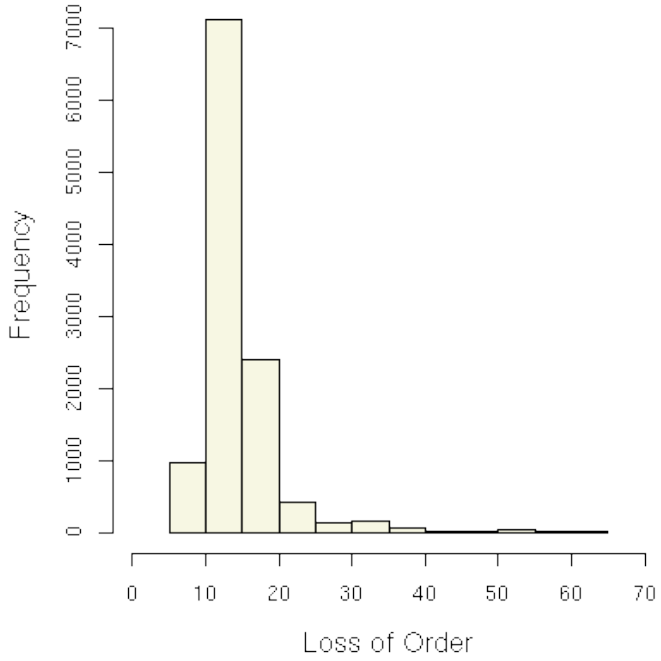
Simple BP Model - Local Assessments

	Local Model IN	*Local Model OUT*
Error IN	6	1
Error OUT	23	2

- Most of the PSET is considered to be within the TSET
- The model descriptors are not capturing molecular features that differentiate the PSET molecules from the TSET

Simple BP Model - Descriptor Augmentation

- Augmented with 2 descriptors
  - Used a brute force method to identify suitable sets
  - Most sets cause relatively low loss of order



- Chose two H-bonding descriptors
  - RPCS, RNCS
  - Very low correlation to the model descriptors
  - Non-zero values in the TSET and PSET

Indiana University School of informatics

P&G Research and Development

Simple BP Model - Global Assessments

	Local Model IN	Local Model OUT
Error IN	6	1
Error OUT	23	2

25% Correct Assessment

True Positives

	Global Model IN	Global Model OUT
Error IN	0	6
Error OUT	0	0

False Positives

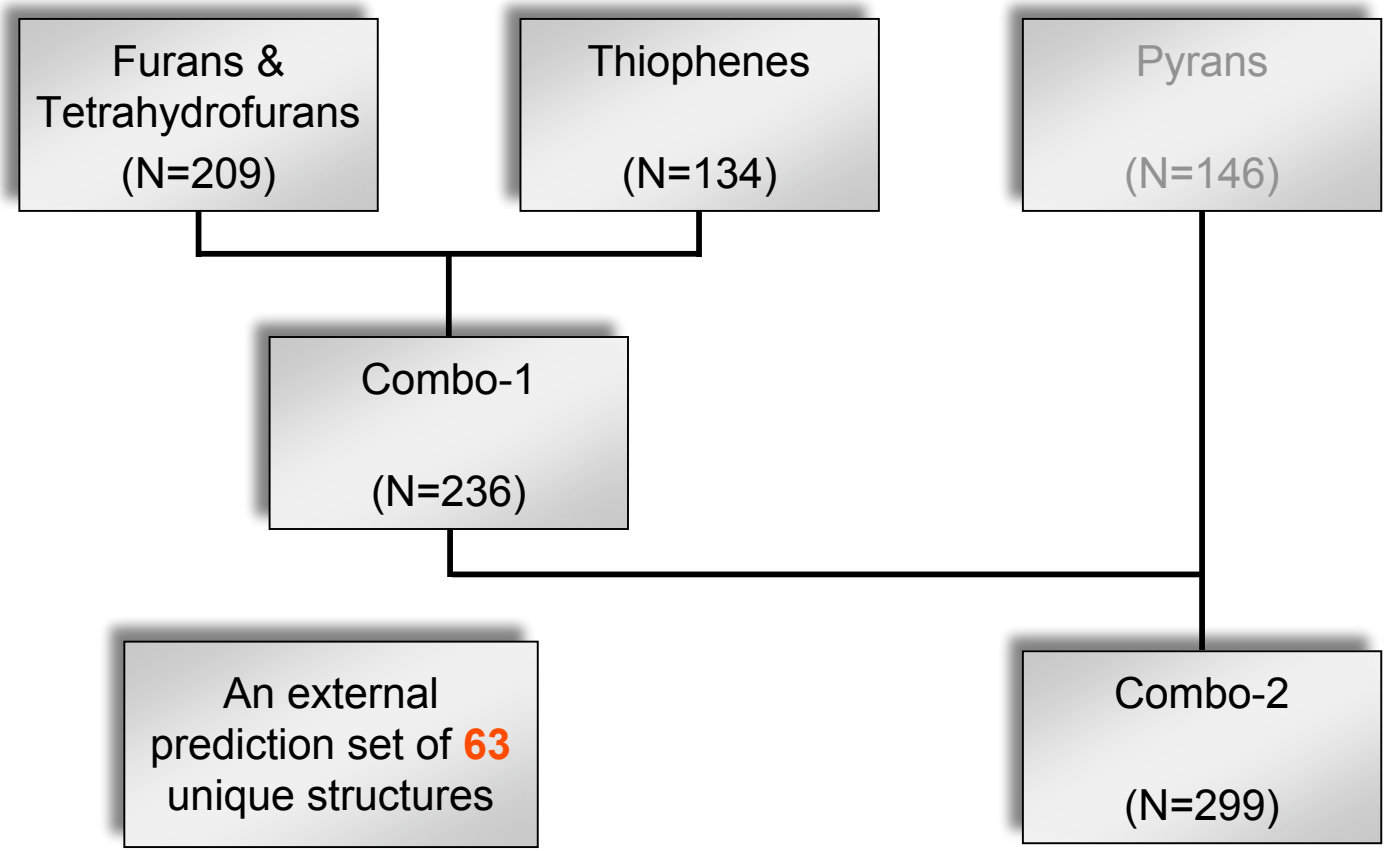
	Global Model IN	Global Model OUT
Error IN	0	0
Error OUT	0	23

Indiana University School of informatics

P&G Research and Development

Beilstein Data Set BP Models

- Published boiling point models*
  - Four very diverse training sets



Indiana University School of informatics

* *J. Chem. Inf. Comput. Sci.*, **1991**, *31*, 301-310
J. Chem. Inf. Comput. Sci., **1992**, *32*, 306-316

P&G Research and Development

Beilstein BP Models - Local Assessments

Furan / Tetrahydrofuran Model

	Local Model IN	*Local Model OUT*
Error IN	24	1
Error OUT	35	3

42.8% Correct Assessment

Combo-2 Model

	Local Model IN	*Local Model OUT*
Error IN	63	0
Error OUT	0	0

100% Correct Assessment

Indiana University School of informatics

P&G Research and Development

Beilstein BP Models - Global Assessments

Furan / Tetrahydrofuran Model

	Local Model IN	*Local Model OUT*
Error IN	24	1
Error OUT	35	3

42.8% Correct Assessment

True Positives

	Global Model IN	*Global Model OUT*
Error IN	20	4
Error OUT	0	0

False Positives

	Global Model IN	*Global Model OUT*
Error IN	0	0
Error OUT	30	5

Indiana University School of informatics

P&G Research and Development

Beilstein BP Models - Global Assessments

Combo-2 Model

	Local Model IN	*Local Model OUT*
Error IN	63	0
Error OUT	0	0

100% Correct Assessment

True Positives

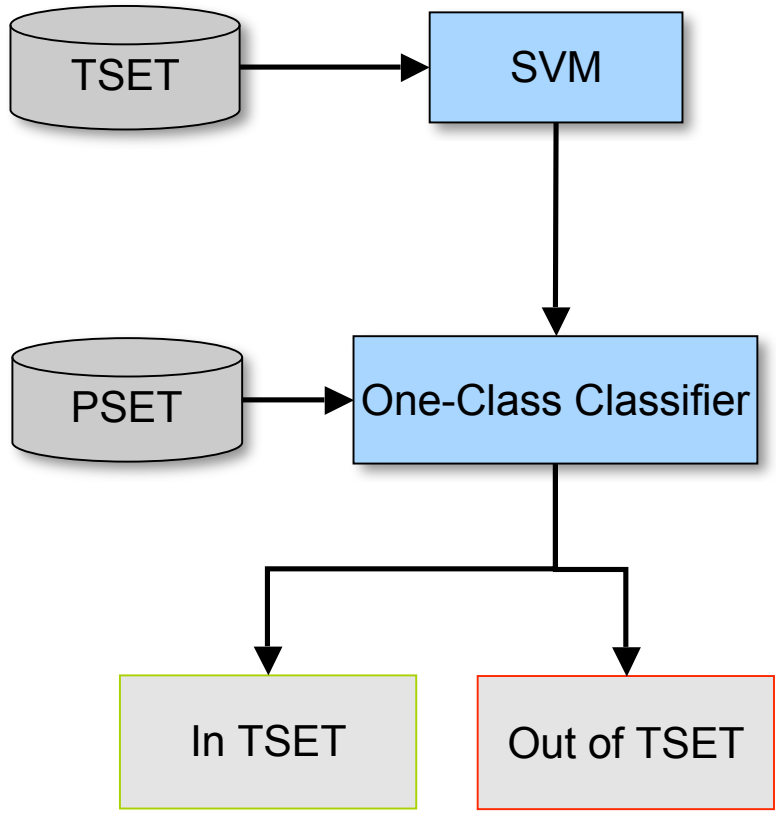
	Global Model IN	*Global Model OUT*
Error IN	42	21
Error OUT	0	0

Indiana University School of informatics

P&G Research and Development

More Sophistication?

- Model applicability is equivalent to novelty detection
- Augmented spaces are conceptually similar to the theory of SVM's
  - Go to a high-D space, to get better separation



The original problem remains - what is a set of suitable descriptors?

Indiana University School of informatics

P&G Research and Development

Scholkopf, B. et al, *Adv. Neural Inf. Proc.*, **2000**, *12*

Summary

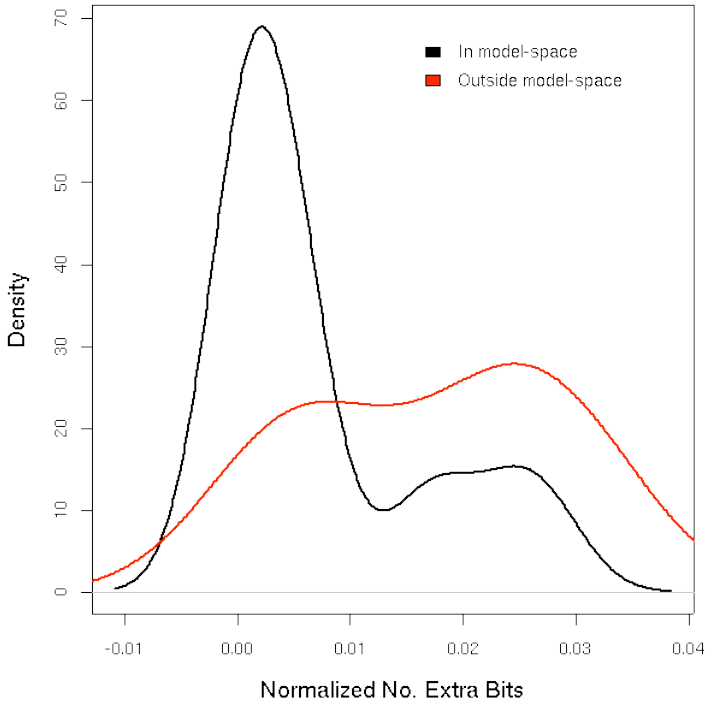
- It's not enough to just consider the model space
- So what is a good global space?
  - Augmented space approach is intuitive
  - Molecular diversity
  - Probably more important than methodology
- Fix the false positives, leave the true positives
- Becomes much harder when applied to diverse training sets

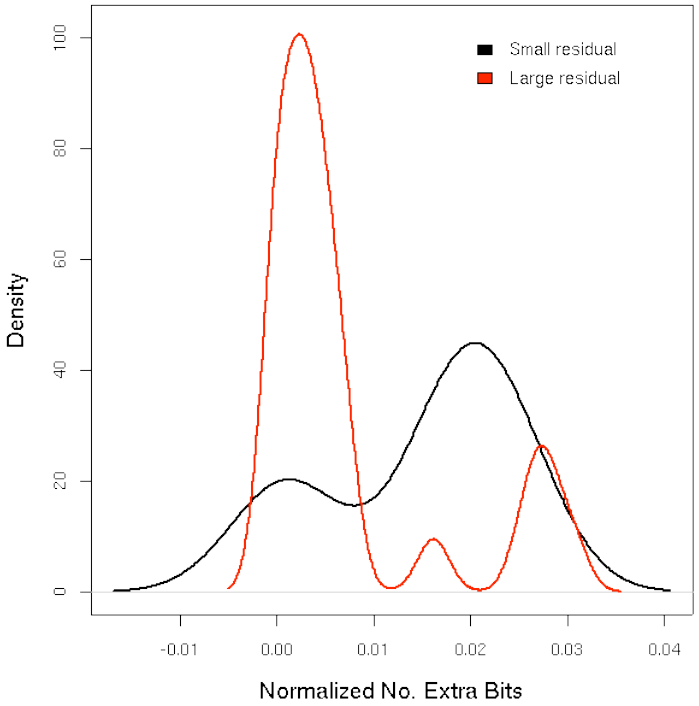
Indiana University School of informatics

P&G Research and Development

Simple BP Model - Fingerprint Classification

- Molecules that are similar to the TSET by distance have very few “extra” features



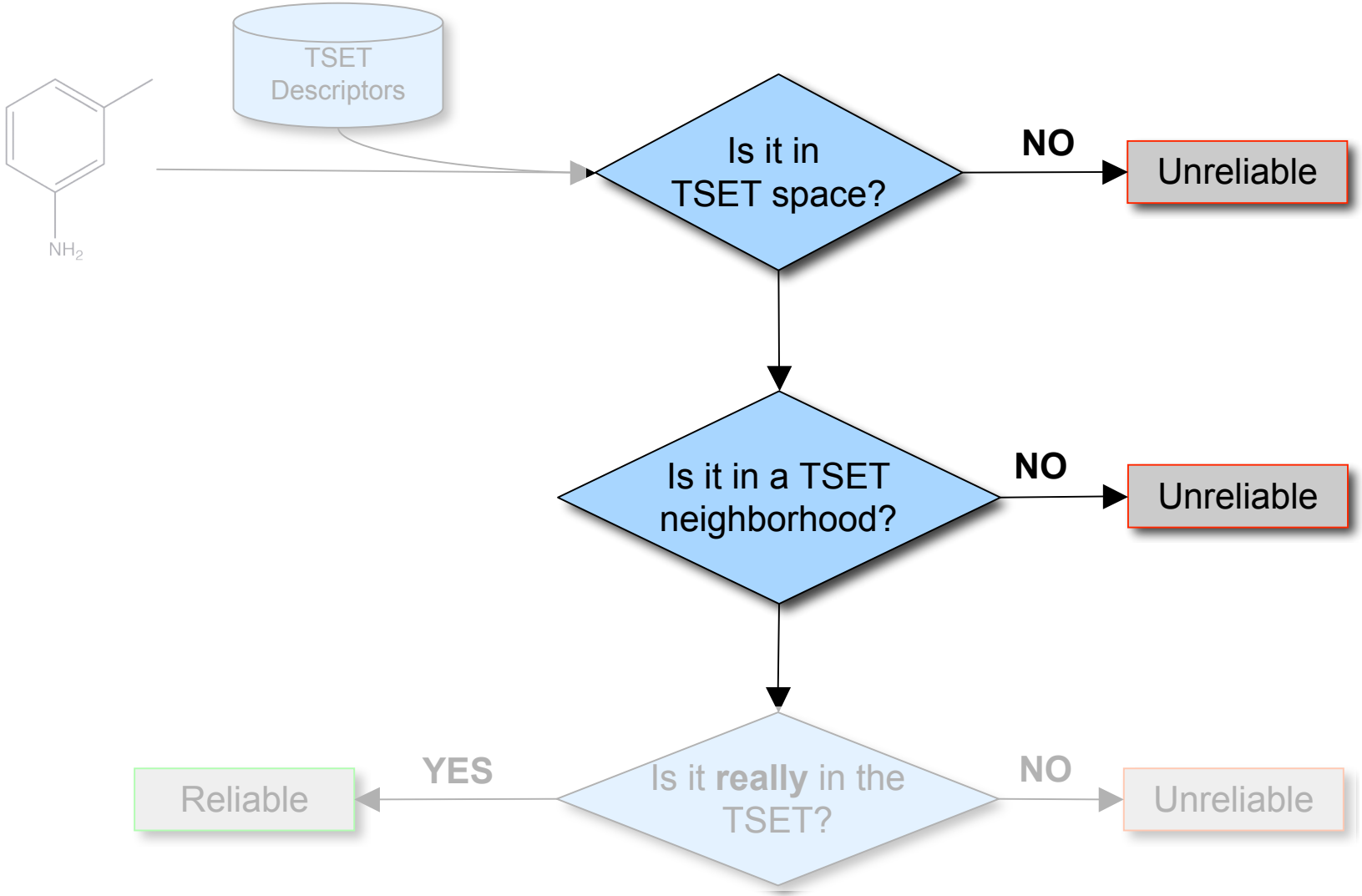


- This does not correspond well with residual classes
- Molecules with no “extra” features may still have a high residual

Indiana University School of informatics

P&G Research and Development

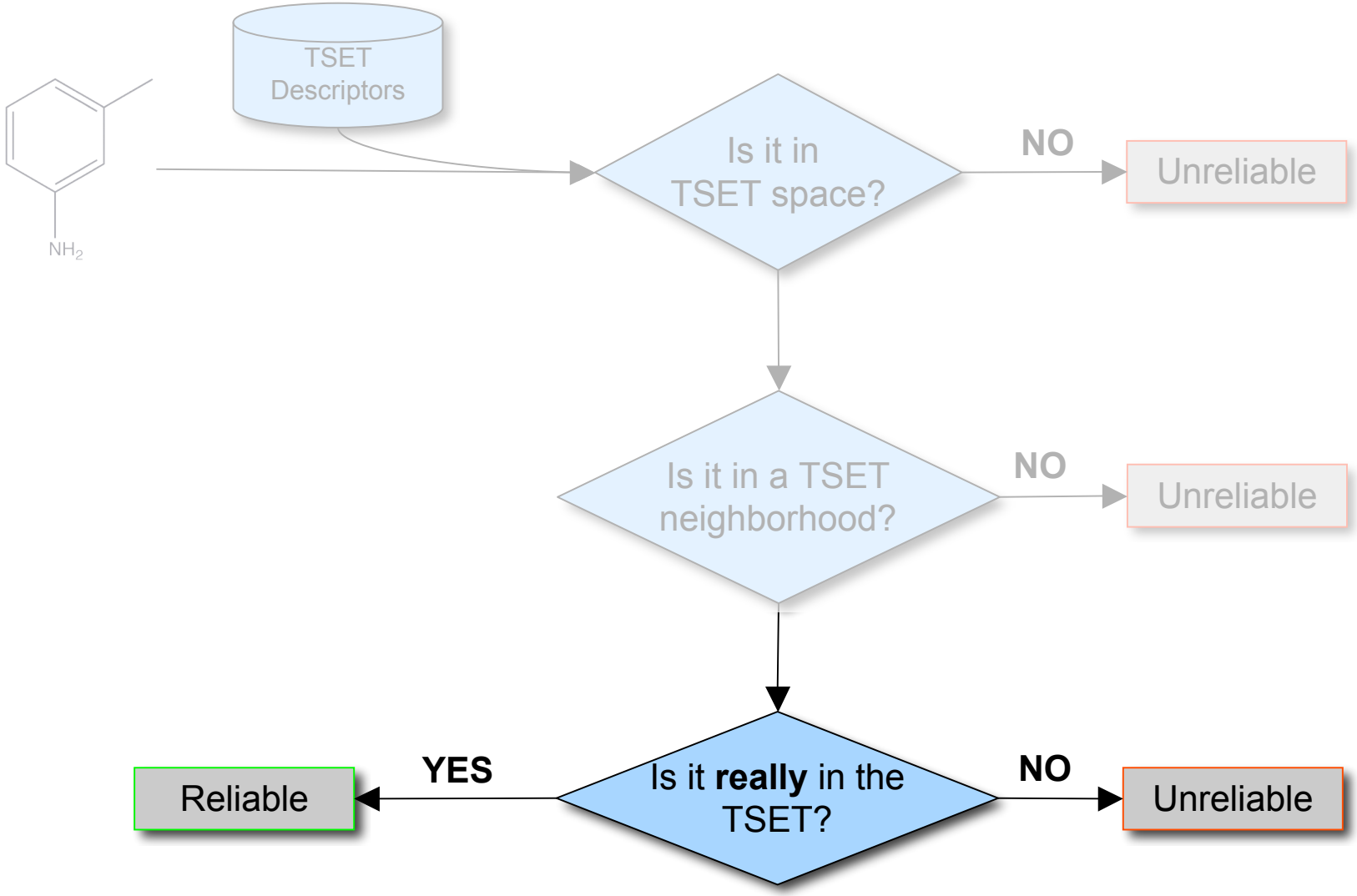
A Hierarchical Strategy - Local Information



Indiana University School of informatics

P&G Research and Development

A Hierarchical Strategy - Global Information



Simple BP Model - Descriptor Augmentation

- From a pool of 160 descriptors, we identified 3 descriptors
  - Maintained the topology of the TSET perfectly
- They were all H-bonding descriptors
  - Values for the TSET were 0, non-zero for the PSET molecules
- We consider SSAH as the augmenting descriptor

Simple BP Model - Descriptor Augmentation

- In terms of the distance to centroid class
  - All PSET molecules are outside of model space in the augmented space
- In terms of the number of common nearest neighbors
  - All PSET molecules have the same neighborhood in both spaces