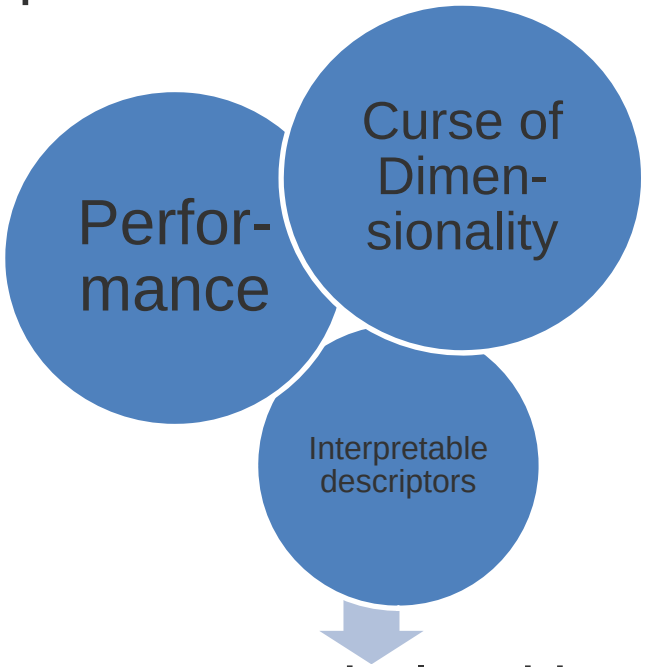
Novel Methods for Graph Mining in Databases of Small Molecules

Andreas Maunz, andreas@maunz.de

Retreat Spitzingsee, 05.-06.04.2011

Tradeoff Overview

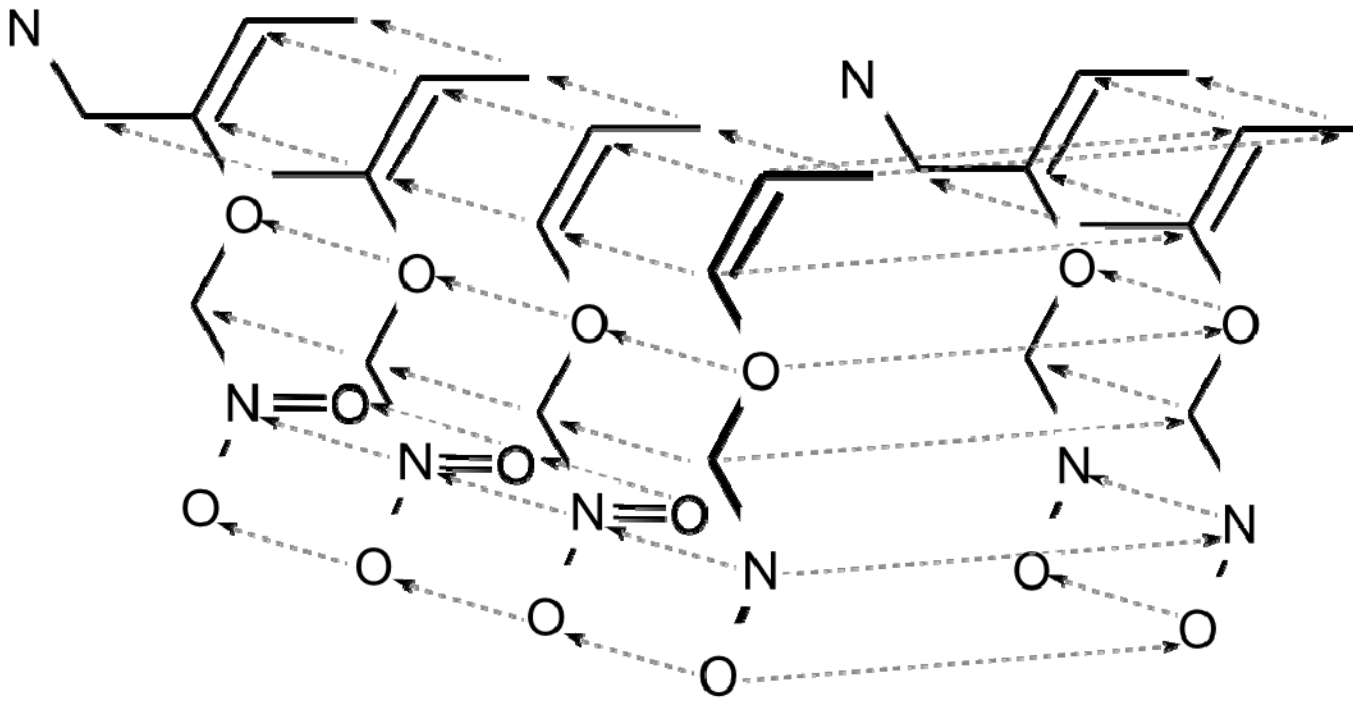
Data Mining Exercise: Find patterns / motifs in large graph databases as descriptors for classification or regression.



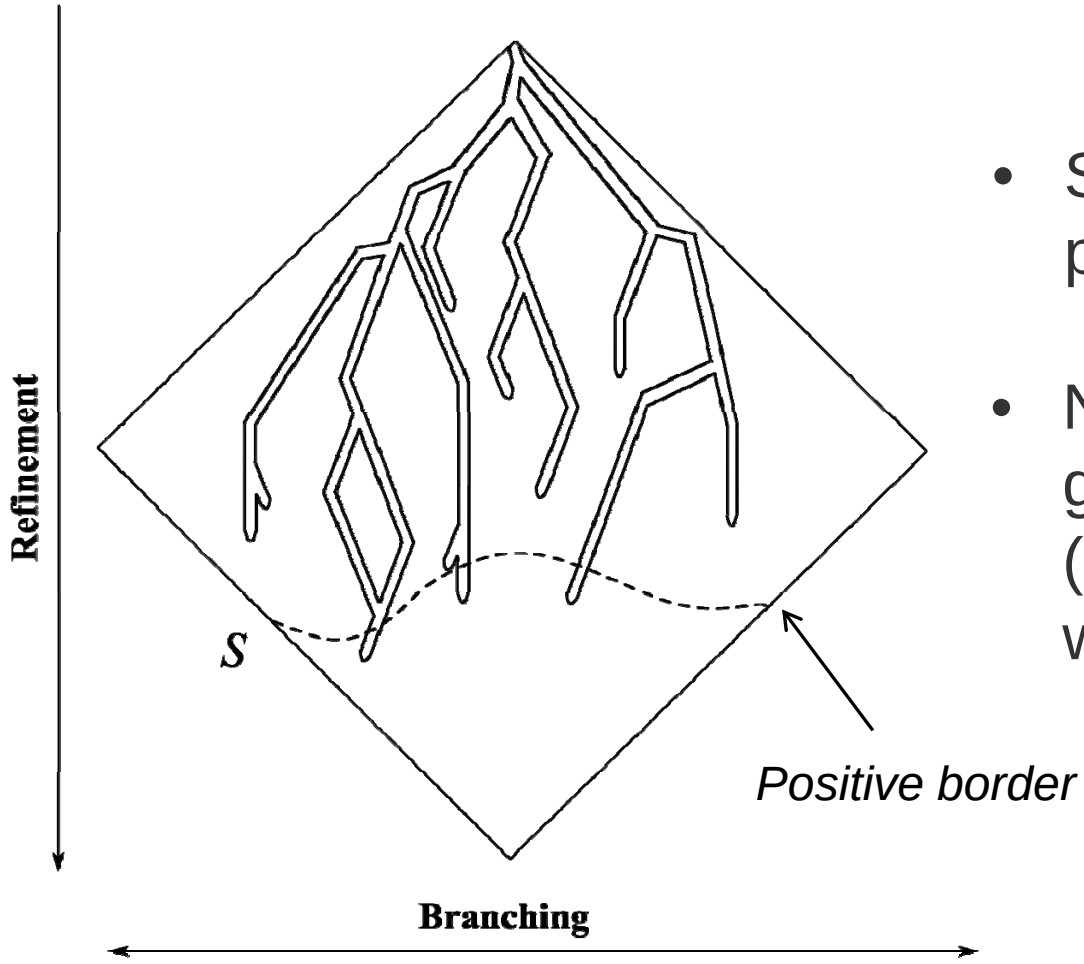
Integrated Algorithms
BBRC [1], LAST-PM [2]

Hypothesis Spaces

Frequent Subgraph Mining is anti-monotonic w.r.t. refinement.
Subgraphs form a partial order.

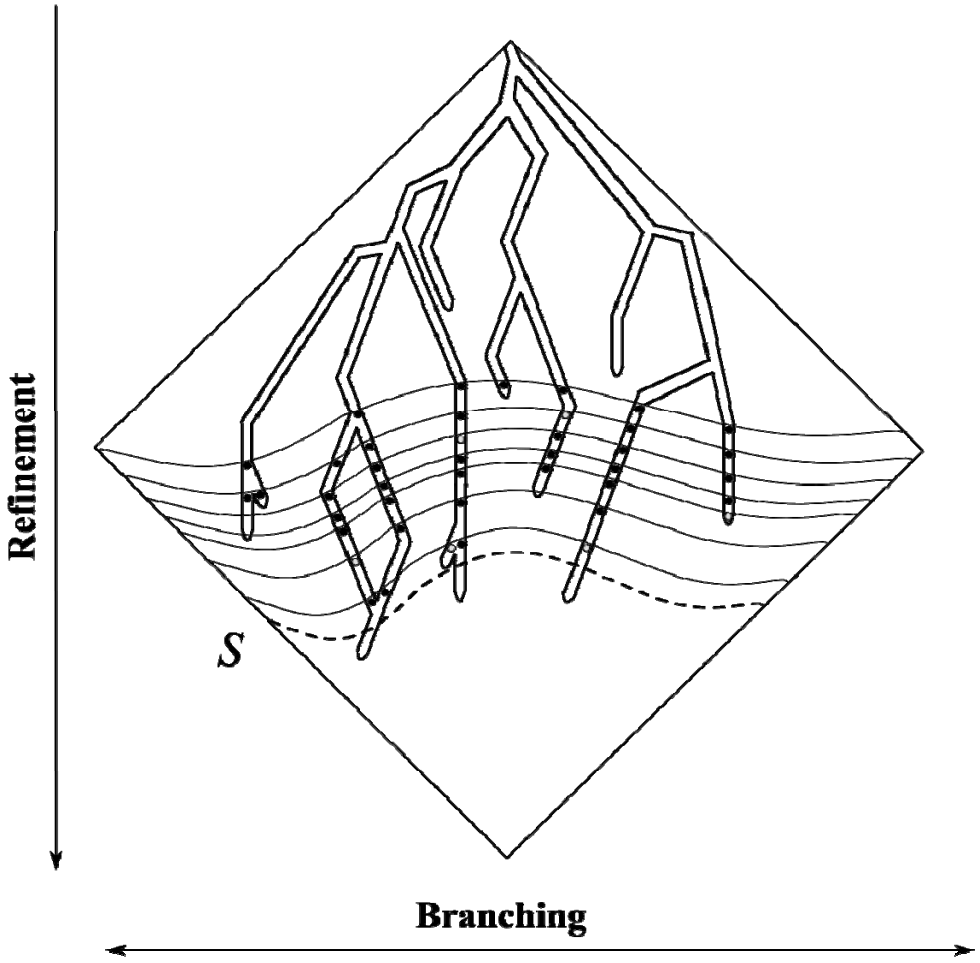


Hypothesis Spaces



- Schematic depiction of the partial order.
- Not restricted to sub-graphs.
  (applies to general sets with the subset relation)

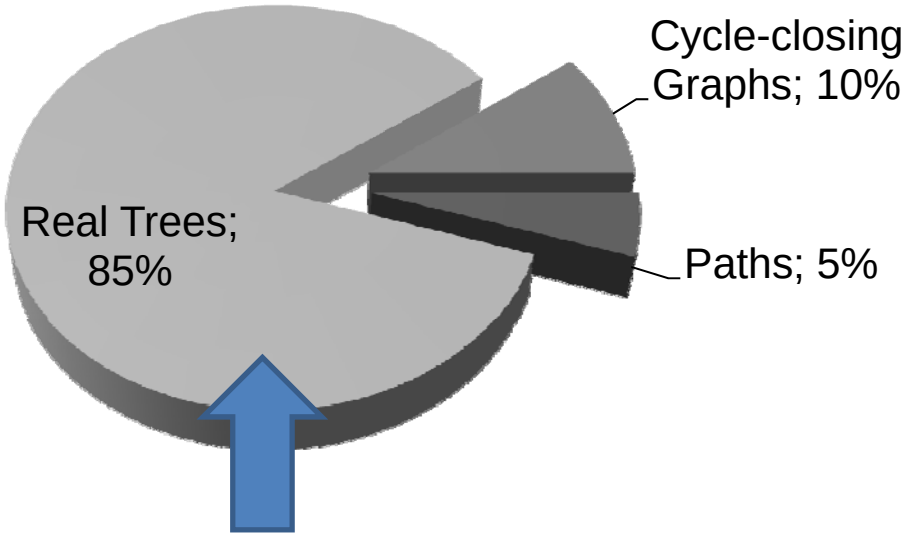
OPEN Patterns



- State of the art compression method.
- Yields one subgraph for each level of frequency.
- Subsumes the set of all frequent subgraphs in terms of occurrences (losless compression).

Tree Mining

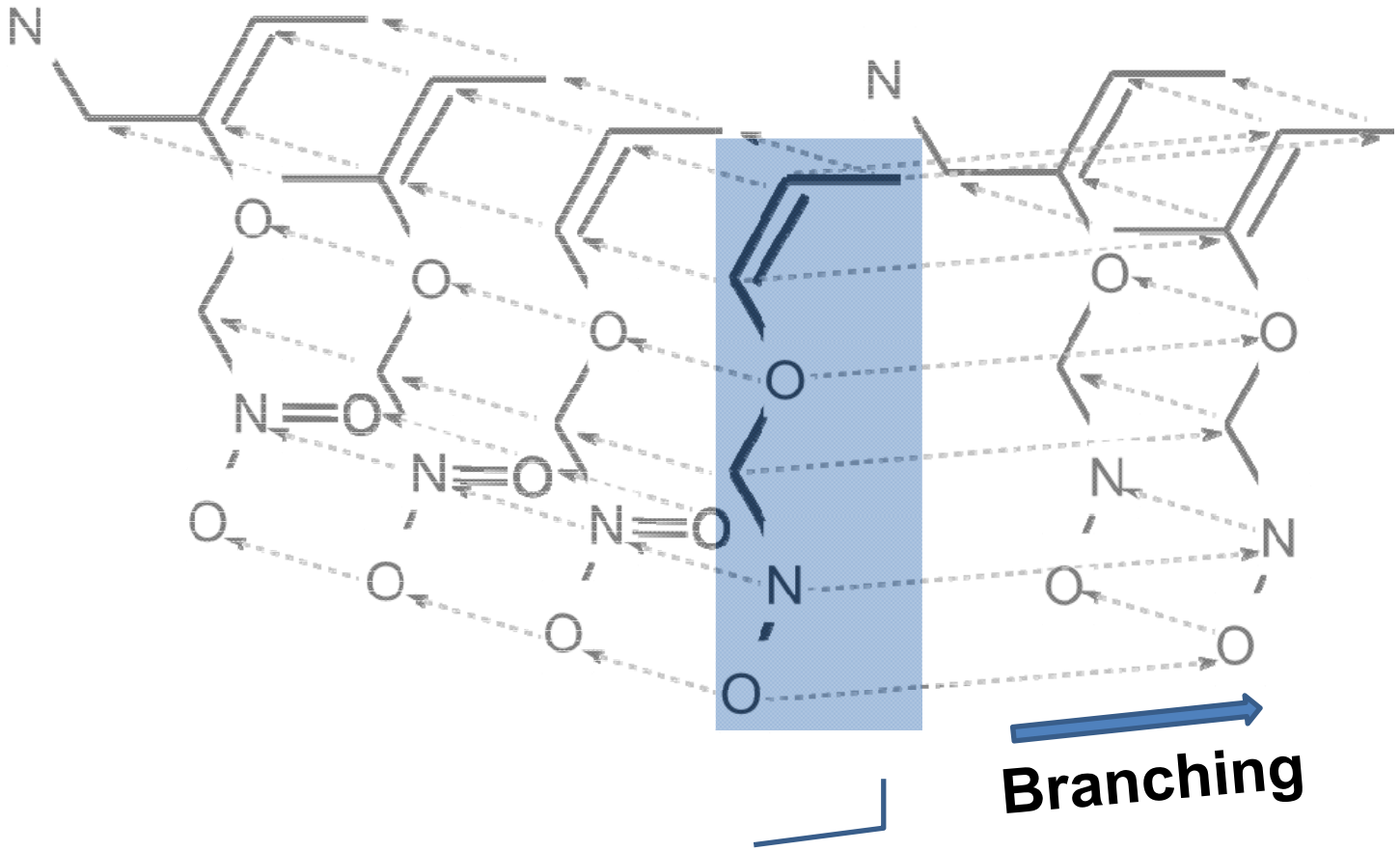
Typical Substructure Frequencies for Databases of Small Molecules



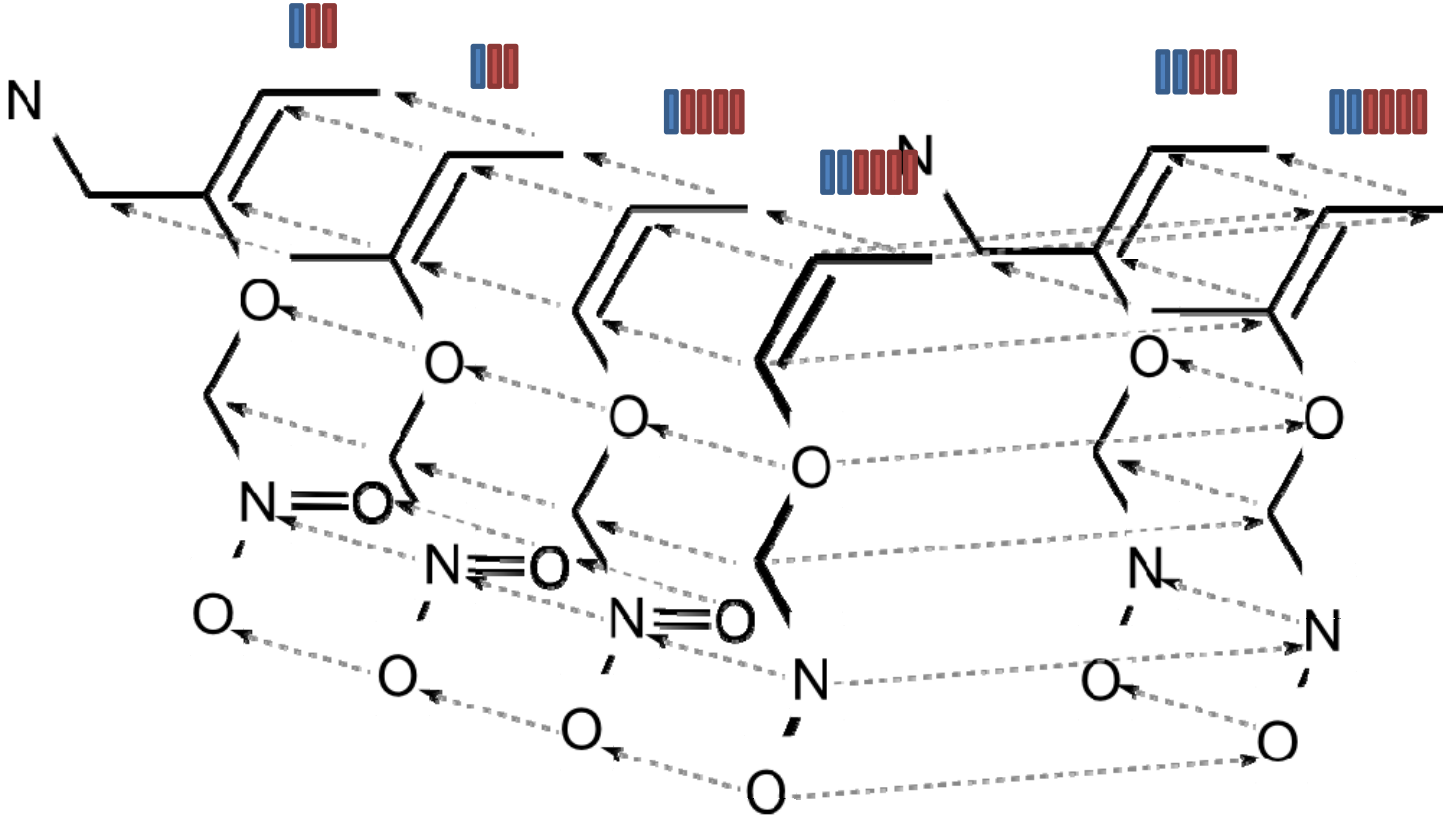
How to break up this large piece?

Backbone Refinement Class Mining (BBRC)

The search space of Trees

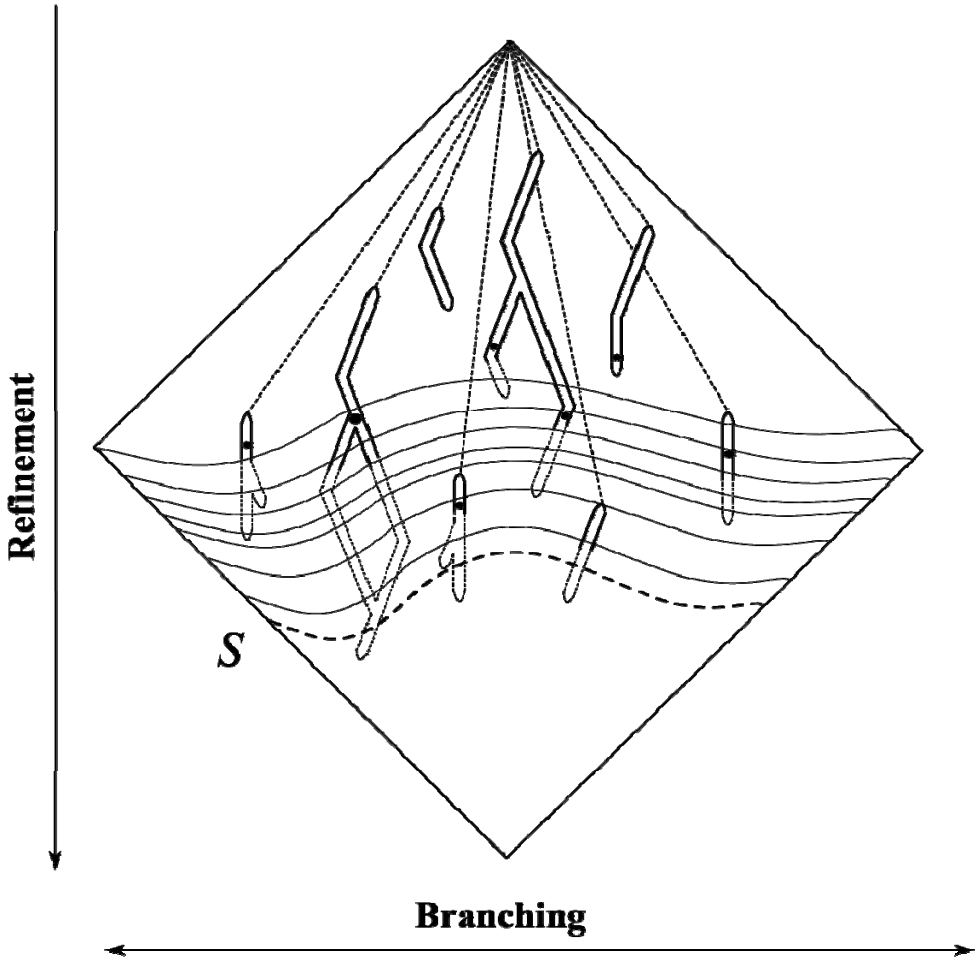


The search space of χ^2 values



Can derive **upper bound** for χ^2 values of specializations of any subgraph [14].

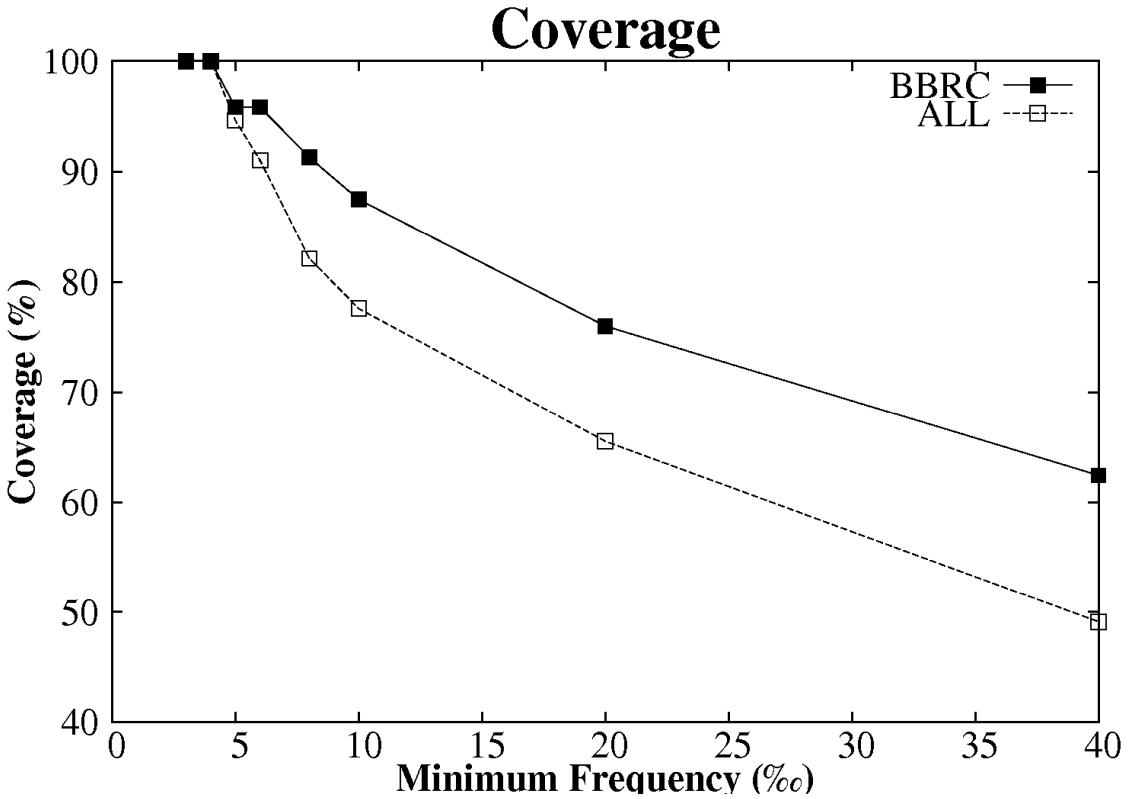
BBRC Hypothesis Space



- Forms **classes** by prohibiting backbone changes during refinement.
- Search is **structurally partitioned**.
- Uses the **convex χ^2 measure** to find most significant member of each class (allows for efficient pruning).

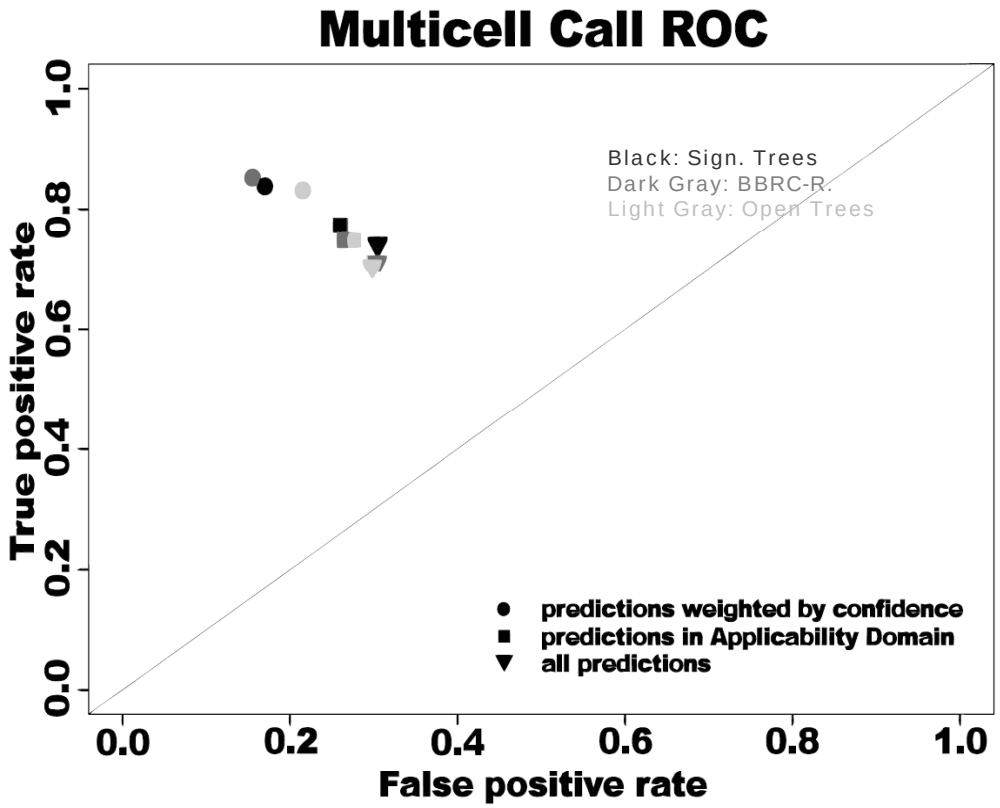
BBRC Experiments

Coverage compared to all frequent and significant Trees.



BBRC Experiments

Accuracy, Sensitivity, Specificity

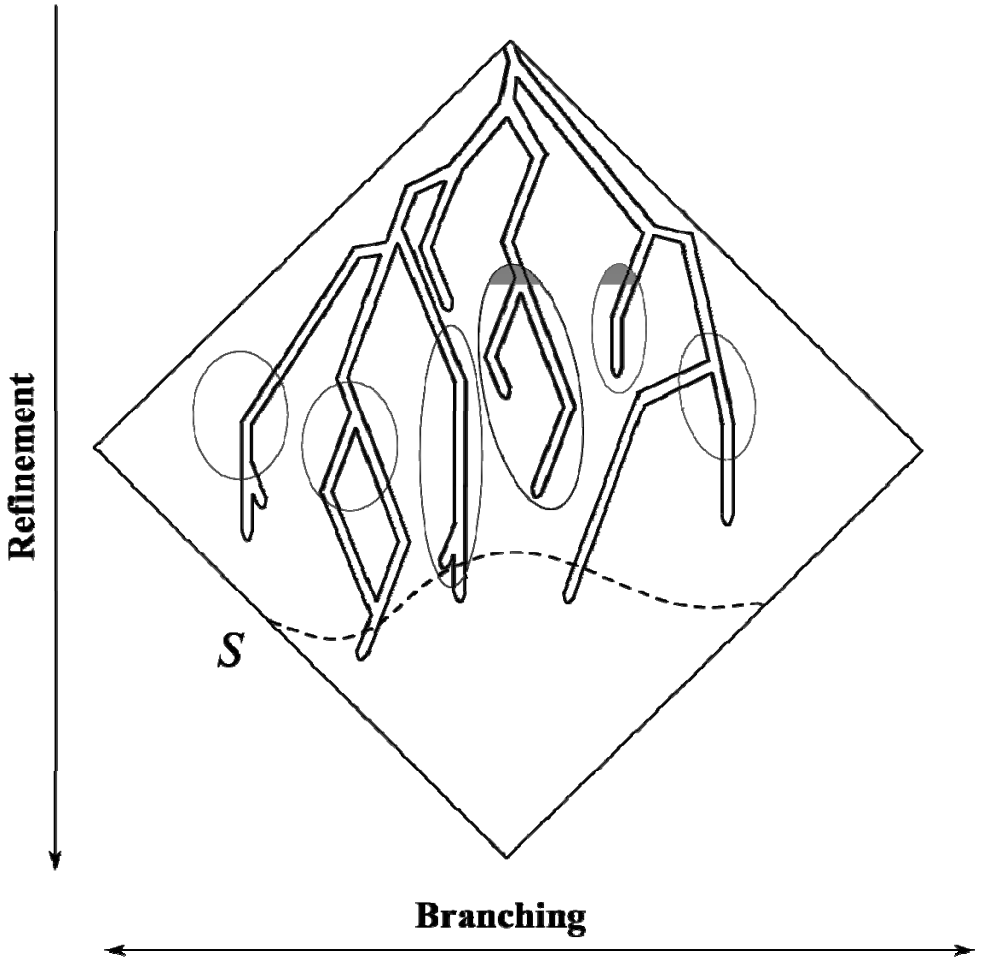


Nearest neighbor predictions
all: all predictions
AD: top 80% confidence predictions
wt.: predictions weighted by confidence

		Sign. Tr.	Open Tr.	BBRC-R.
SM	all	74.6	**75.5**	74.6
	AD	**80.7**	80.6	79.4
	wt.	**86.8**	84.5	85.4
RC	all	64.4	64.5	**67.2**
	AD	70.0	68.7	**70.4**
	wt.	81.8	80.0	**82.2**
MoC	all	**73.3**	71.5	71.7
	AD	75.7	74.4	**76.5**
	wt.	**83.7**	80.8	82.0
MuC	all	**71.9**	70.2	70.3
	AD	**75.6**	73.5	74.1
	wt.	83.5	81.3	**84.9**

Latent Structure Pattern Mining (LAST-PM)

LAST-PM Hypothesis Space



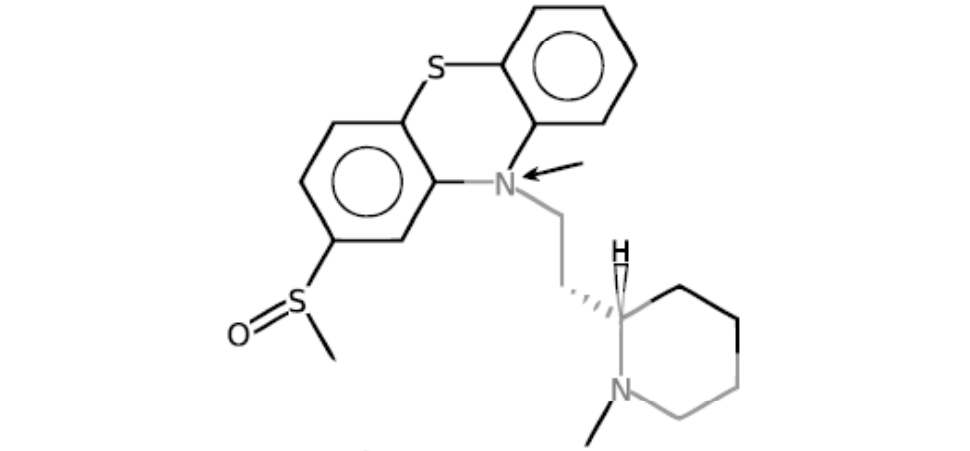
- Observation: Subgraphs significant for the same target class form *patches* in the search space.
- Subgraphs within a patch are often very similar, patches may have branches.
- **Idea:** Merge patch members to a single pattern.

LAST-PM Rationale

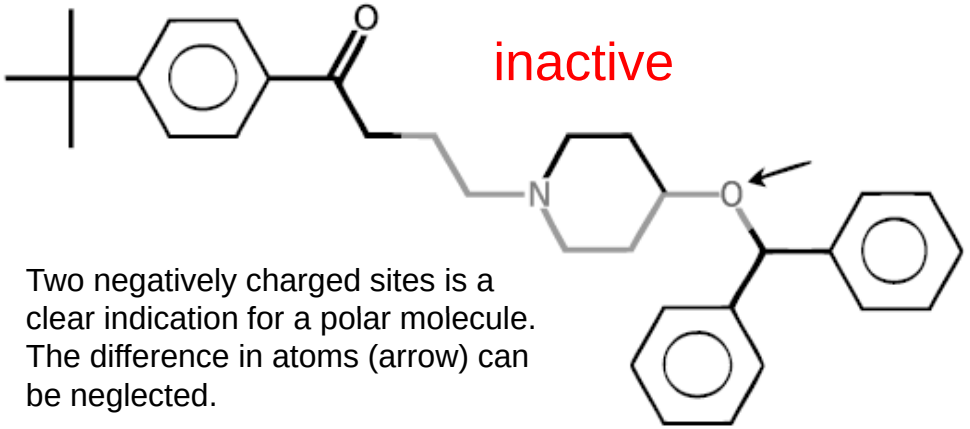
Chemical Example:

Conclusion: The electrostatic properties decrease the ability of both drugs to cross membranes in the body, such as the blood-brain barrier.

Indeed, the two molecules are inactive for the blood-brain barrier endpoint, see Hu et al., *J Chem Inf, 2005* **[3]**.



inactive

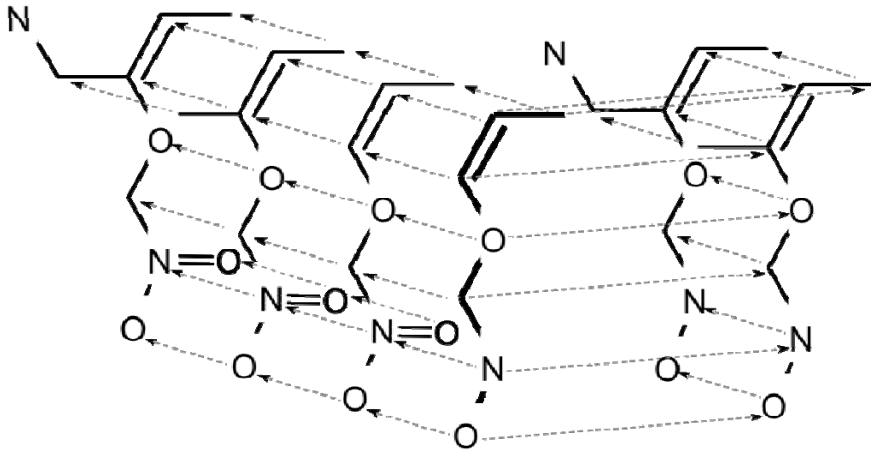


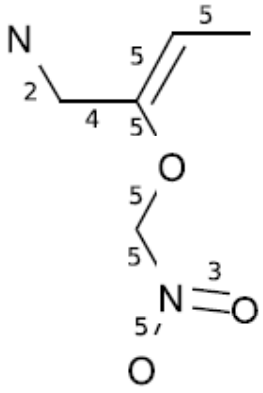
Two negatively charged sites is a clear indication for a polar molecule. The difference in atoms (arrow) can be neglected.

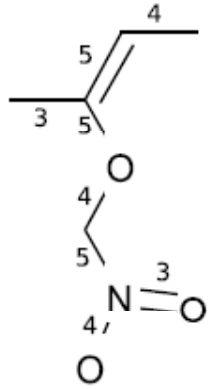
inactive

LAST-PM Properties

LAST-PM executes a three-step pipeline on ground Patterns:



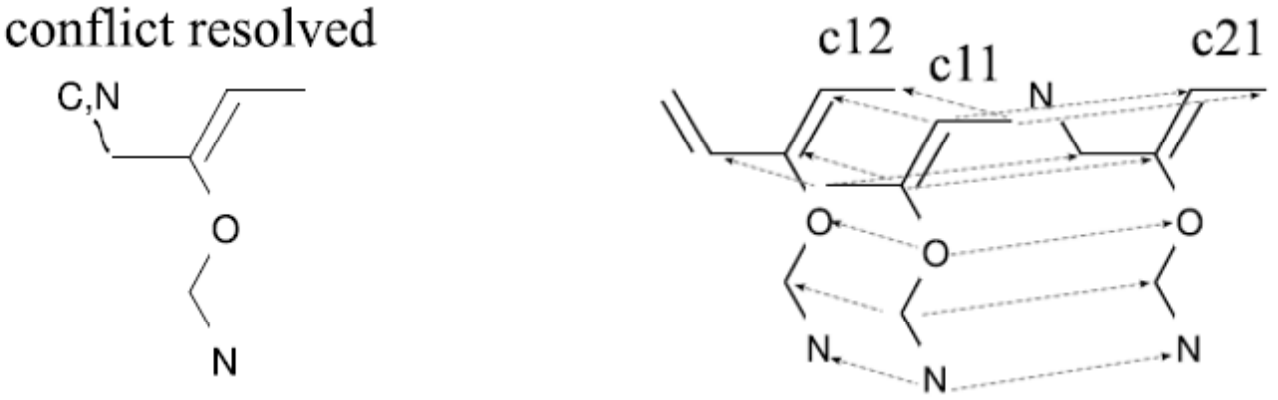




The latent structure pattern contains characteristics of the *ensemble* of ground Patterns, such as weight, *core* size , and „heavy“ regions.

LAST-PM Properties

Conflicting Patterns:
Distinct nodes/edges with the same position.



Resolved by logical OR ⇨ Ambiguous positions.

LAST-PM Experiments

ID	Biological Endpoint	Study Reference
yoshida	Bioavailability	Yoshida and Topliss, *J Med Chem, 2000* **[9]**
nctrer	Estrogen Receptor	Fang *et al.*, *Chem Res Tox, 2001* **[5]**
bloodbarr	Blood-Brain Barrier	Hu et al., *J Chem Inf, 2005* **[3]**, Hou and Xu, *J Chem Inf, 2003* **[10]**

LAST-PM Experiments

Crossvalidation: repeated 10-fold.

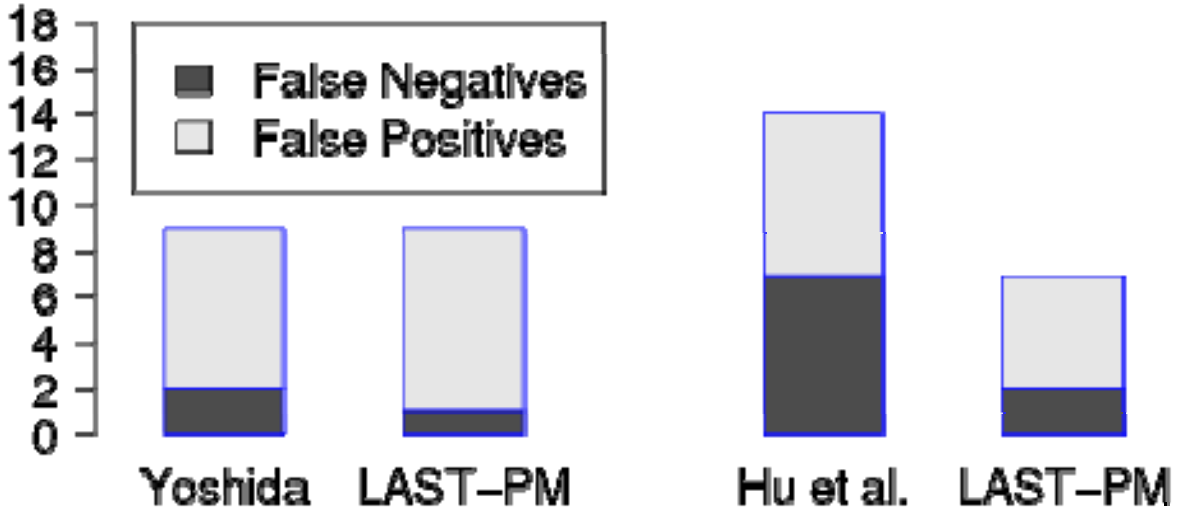
Dataset	LAST-PM	ALL	BBRC	MOSS	SLS
	%Test	%Test	%Test	%Test	%Test
bloodbarr	72.20	70.49	68.50	67.49	70.4
nctrer	80.22	79.13	80.22	77.17	78.4
yoshida	69.81	65.19	65.96	66.46	63.8

Dataset	LAST-PM	ALL	Ratio
	# Patterns (t)	# Patterns (t)	#Patterns/t
bloodbarr	249 (1.23s)	1613 (0.36s)	0.51
nctrer	193 (12.49s)	22942 (0.13s)	0.81
yoshida	124 (0.28s)	462 (0.09s)	0.84

<1: positive tradeoff

LAST-PM Experiments

Performance on yoshida and bloodbarr test sets (Hu *et al.*).



- bloodbarr (Hou and Xu):
  LAST-PM improved phys-chem descriptor model
  in 10-fold CV.

Summary

Backbone Refinement Class Representatives

- Structurally heterogeneous descriptors, compression by structural invariant (backbone constraint)
- Good dataset coverage, robust against increasing minimum frequencies
- Applicable to large-scale graph databases through a novel statistical pruning technique
- Discriminative potential similar to complete set of trees, but significantly better than open trees.

Summary

Latent Structure Pattern Mining

- Describing latent patterns in the data.
- More expressive than most other subgraph descriptors, including the complete set of ground Patterns from which they were derived.
- Very good performance on external test sets that have been difficult for structural Patterns.
- Favorable tradeoff between Pattern reduction and runtime.

Acknowledgements

Supported by the EU seventh framework programme under contract no Health-F5-2008-200787 (**OpenTox**).

http://www.opentox.org

Visit the project websites (source code, API, how to reproduce results,…):

http://bbrc.maunz.de | **http://last-pm.maunz.de**

Thank you for your attention. I will be happy to answer your questions.

References

[4] J. Kazius, S. Nijssen, J. Kok, T. Baeck, and A. P. Ijzerman. Substructure Mining Using Elaborate Chemical Representation. J Chem Inf Model, 46:597-605, 2006.

[15] Hu Li, Chun Wei Yap, Choong Yong Ung, Ying Xue, Zhi Wei Cao, and Yu Zong Chen. Effect of Selection of Molecular Descriptors on the Prediction of Blood-Brain Barrier Penetrating and Nonpenetrating Agents by Statistical Learning Methods. Journal of Chemical Information and Modeling, 45(5):1376-1384, Aug 2005.

[3] Siegfried Nijssen and Joost N. Kok. A Quickstart in Frequent Structure Mining can make a Difference. In KDD '04: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 647-652, New York, NY, USA, 2004. ACM.

[1] Andreas Maunz, Christoph Helma, and Stefan Kramer. Large-Scale Graph Mining Using Backbone Refinement Classes. In KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 617-626, New York, NY, USA, 2009. ACM.

[5] Heiko Hofer, Christian Borgelt, and Michael R. Berthold. Large Scale Mining of Molecular Fragments with Wildcards. Intell. Data Anal, 8(5):495-504, 2004.

[6] Ulrich Rückert and Stefan Kramer. Optimizing Pattern Sets for Structured Data. In ECML '07: Proceedings of the 18th European Conference on Machine Learning, pages 716-723, Warsaw, Poland, 2007. Springer-Verlag, Berlin-Heidelberg.

[7] Fumitaka Yoshida and John G. Topliss. QSAR Model for Drug Human Oral Bioavailability. Journal of Medicinal Chemistry, 43(13):2575-2585, Jun 2000.

[8] T. J. Hou and X. J. Xu. ADME Evaluation in Drug Discovery. 3. Modeling Blood-Brain Barrier Partitioning Using Simple Molecular Descriptors. Journal of Chemical Information and Computer Sciences, 43(6):2137-2152, Oct 2003.

[9] X. Yan and J. Han. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02), page 721, Washington, DC, USA, 2002. IEEE Computer Society.

References

[10] X. Yan and J. Han. CloseGraph: Mining Closed Frequent Graph Patterns. In KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 286-295, New York, NY, USA, 2003. ACM.

[11] L. Schietgat, F. Costa, J. Ramon, and Luc De Raedt. Effective Pattern construction by maximum common subgraph sampling. In Machine Learning Journal 2010.

[12] M. Al Hasan, V. Chaoji, S. Salem, J. Besson, and M. Zaki. Origami: Mining Representative Orthogonal Graph Patterns. ICDM 2007. Seventh IEEE International Conference on Data Mining, pages 153-162, Oct. 2007.

[13] A. Maunz, C. Helma, and S. Kramer. Efficient mining for structurally diverse subgraph patterns in large molecular databases. In Machine Learning Journal 2010.

[14] S. Morishita and J. Sese. Traversing Itemset Lattices with Statistical Metric Pruning. In Symposium on Principles of Database Systems, pages 226-236, 2000.

[2] A. Maunz, C. Helma, T. Cramer, and S. Kramer. Latent Structure Pattern Mining. In Proceedings of the European Conference of Machine Learning 2010,.