Interreg
ITALIA-SLOVENIJA
UNIONE EUROPEA
EVROPSKA UNIJA
TRAIN
Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj

Semi-supervised multi-target prediction for analysis of screening data

Dragi Kocev

Jurica Levatić, Michelangelo Ceci, Sašo Džeroski

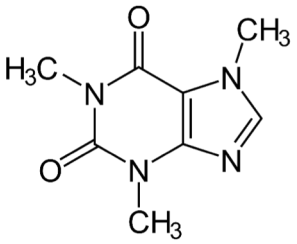
May 13-17, 2019, Bled, Slovenia

IJS
DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

What is semi-supervised learning?

Supervised learning: Labeled data → Predictive model

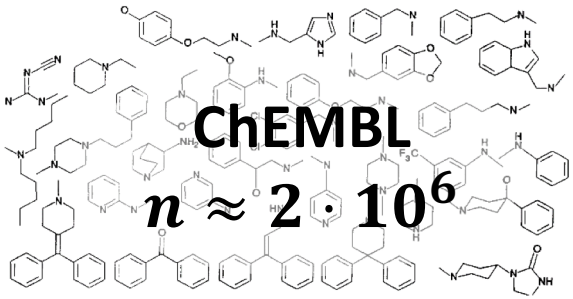
Semi-supervised learning: Labeled + Unlabeled data → (Better) Predictive model



Labeled examples

Hard to get

Scarce



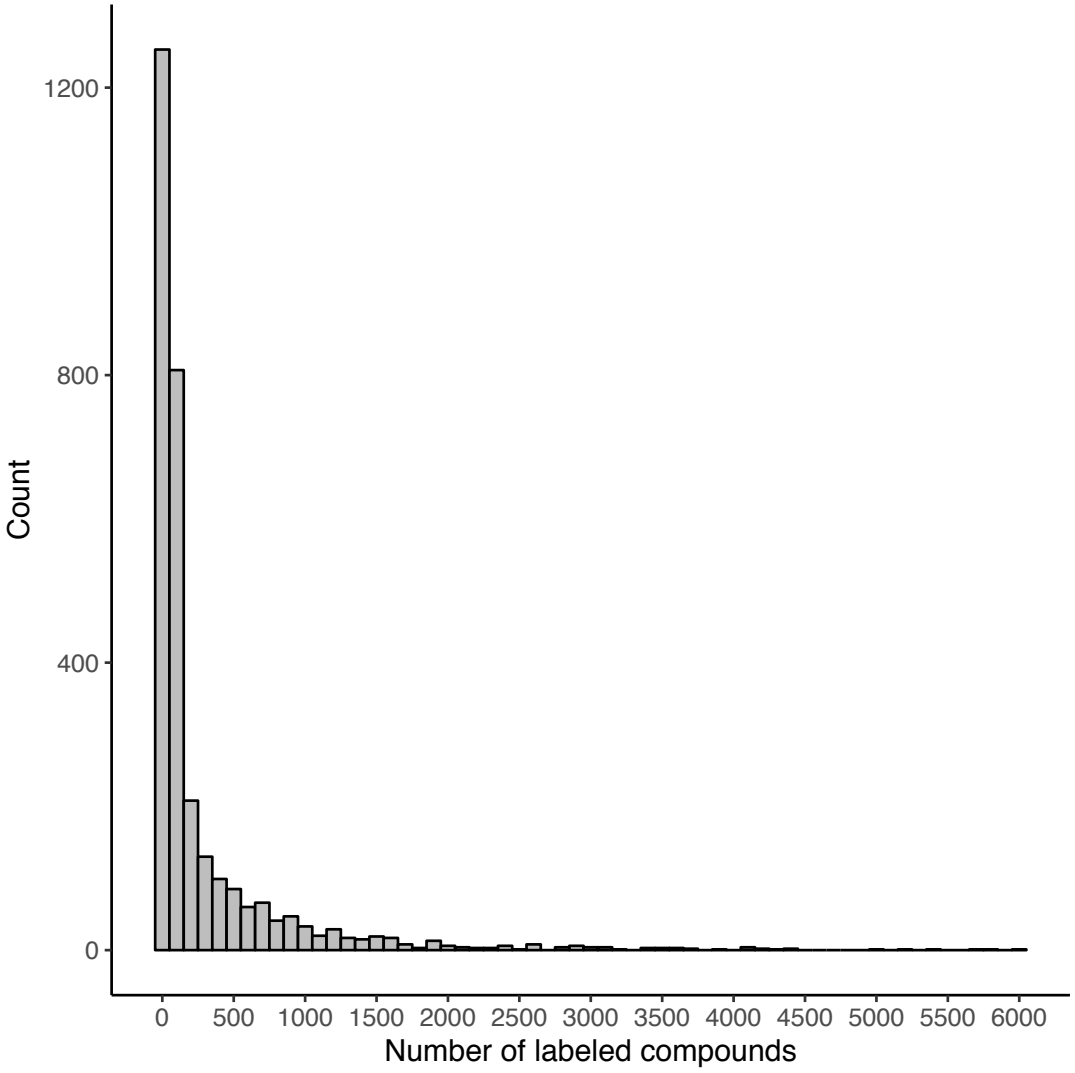
?

Unlabeled examples

Easy to get

Abundant

Why semi-supervised learning?



Histogram of dataset sizes (in terms of number of labelled compounds) for 3047 biological targets extracted from the ChEMBL database shows that, for a vast majority of targets, less than 100 compounds is labeled.

QSAR datasets available at OpenML

outline

- Introduction
- (Semi-supervised) predictive clustering trees (PCTs)
- Evaluation and illustrative examples
- Conclusions

The task of semi-supervised learning

Given:

- An input (descriptive) space X
- A output (or target) space Y
- A set of **labeled** examples $E_l = \{(x_i, y_i) : x_i \in X, y_i \in Y, 1 \leq i \leq N_l\}$
- A set of **unlabeled** examples $E_u = \{x_i : x_i \in X, 1 \leq i \leq N_u\}$
- A quality criterion q

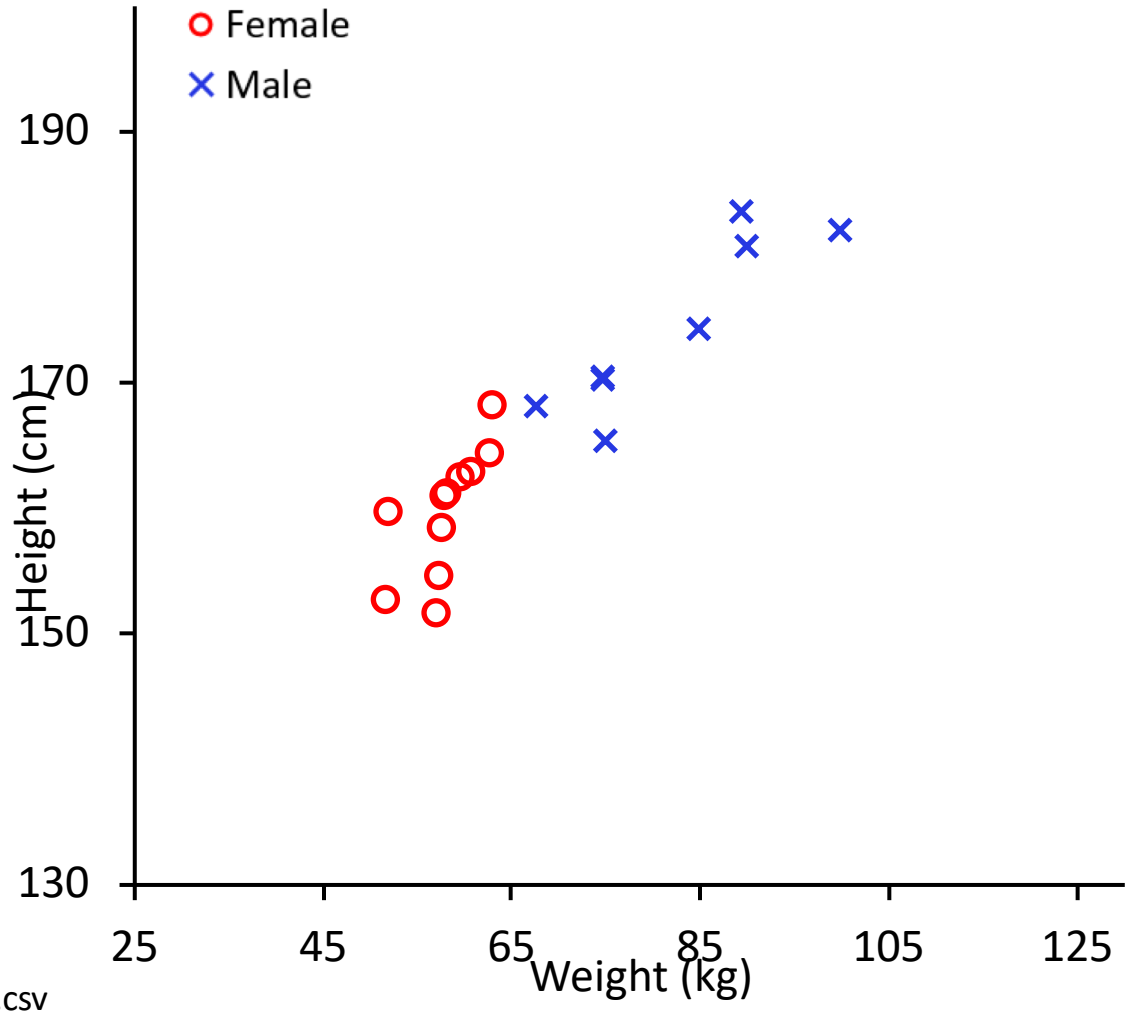
Find: A function $f : X \rightarrow Y$ such that f maximizies q

Goal: Achieve better performance than only with labeled data E_l

How unlabeled data can help?

Task: Predict gender from person's weight and height

Data[1]: 10000 entries of height, weight and gender

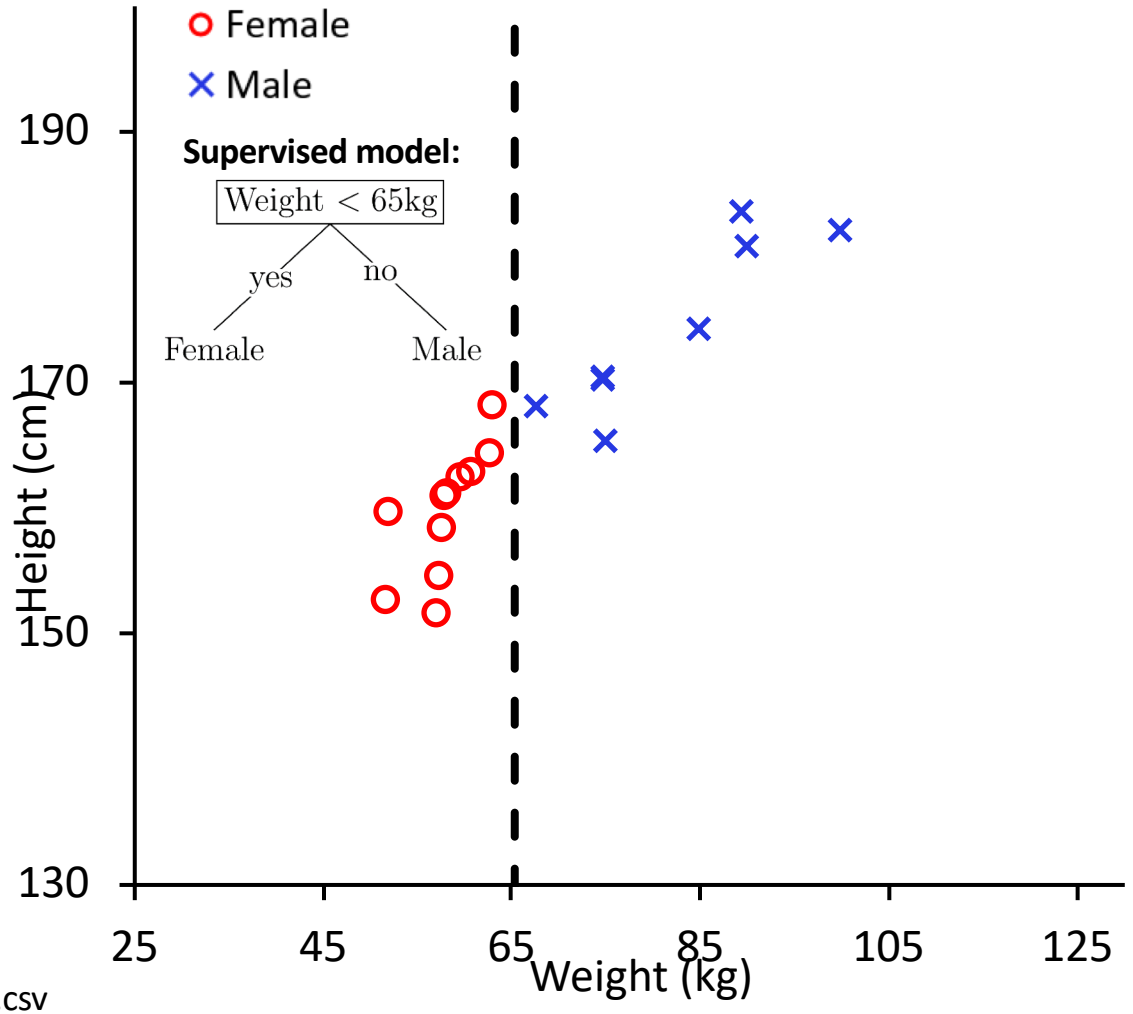


[1]https://helloacm.com/data/gender-height-weight.csv

How unlabeled data can help?

Task: Predict gender from person's weight and height

Data[1]: 10000 entries of height, weight and gender



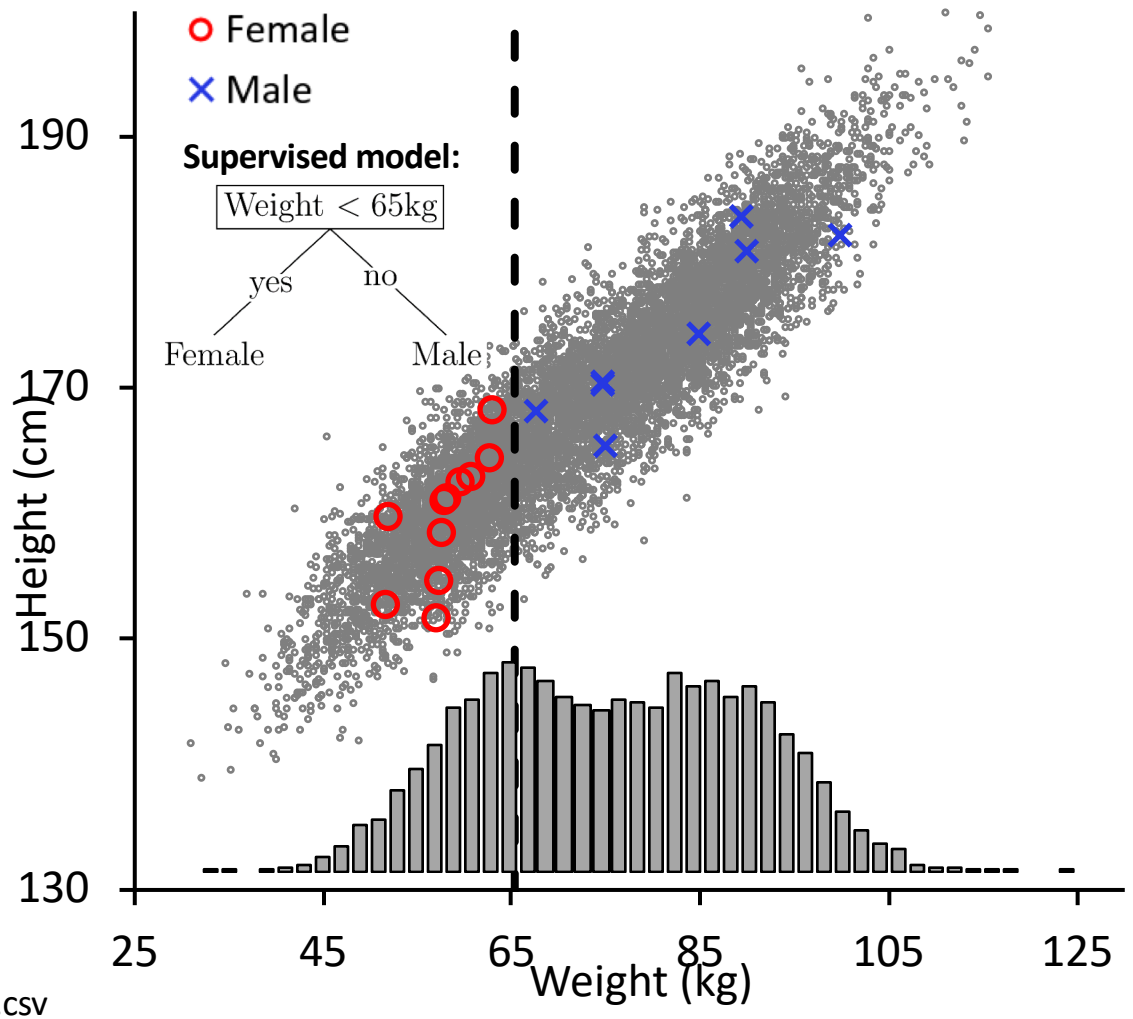
[1]https://helloacm.com/data/gender-height-weight.csv

How unlabeled data can help?

Task: Predict gender from person's weight and height

Data[1]: 10000 entries of height, weight and gender

Assumptions about the labels with respect to the structure of unlabeled data



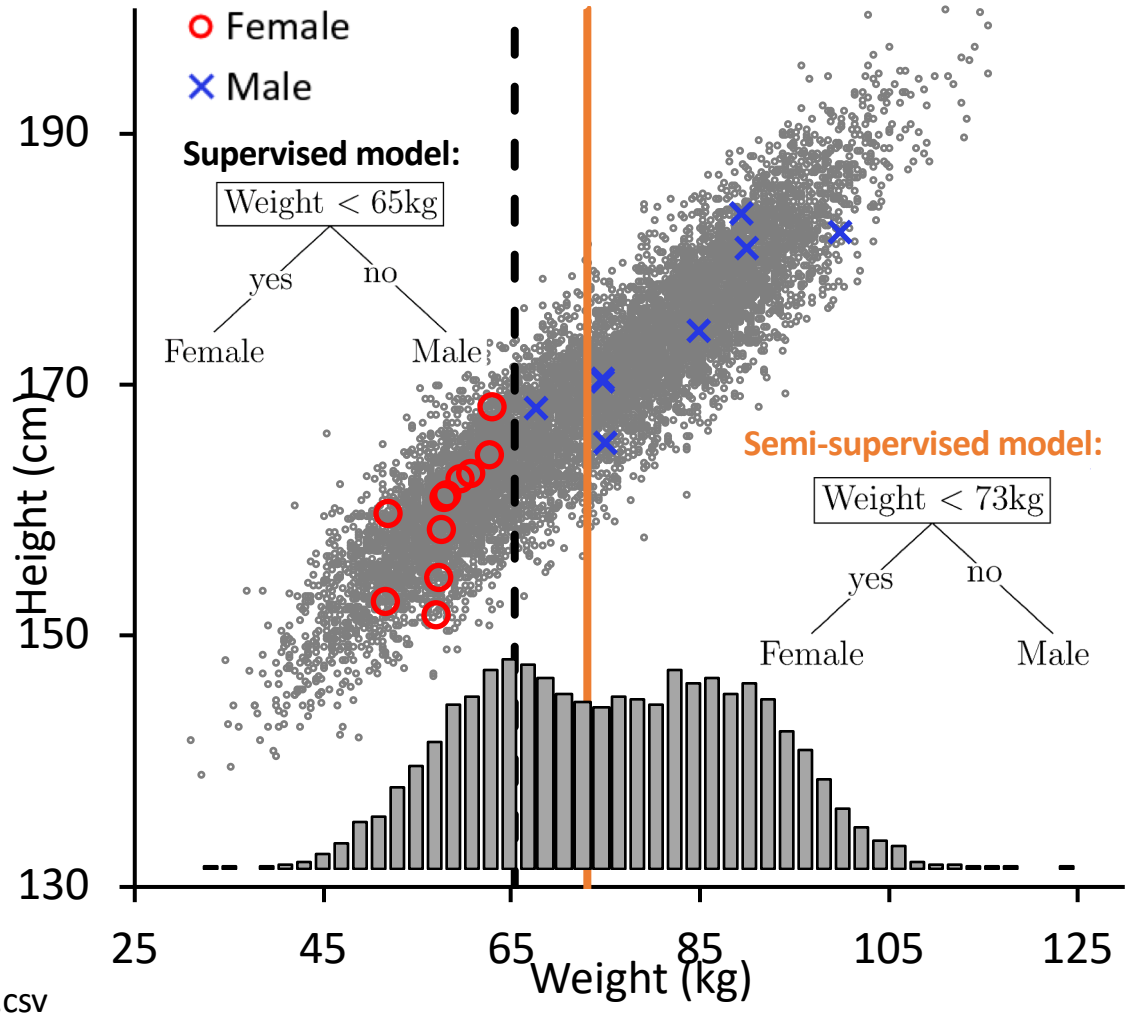
[1]https://helloacm.com/data/gender-height-weight.csv

How unlabeled data can help?

Task: Predict gender from person's weight and height

Data[1]**:** 10000 entries of height, weight and gender

Assumptions about the labels with respect to the structure of unlabeled data



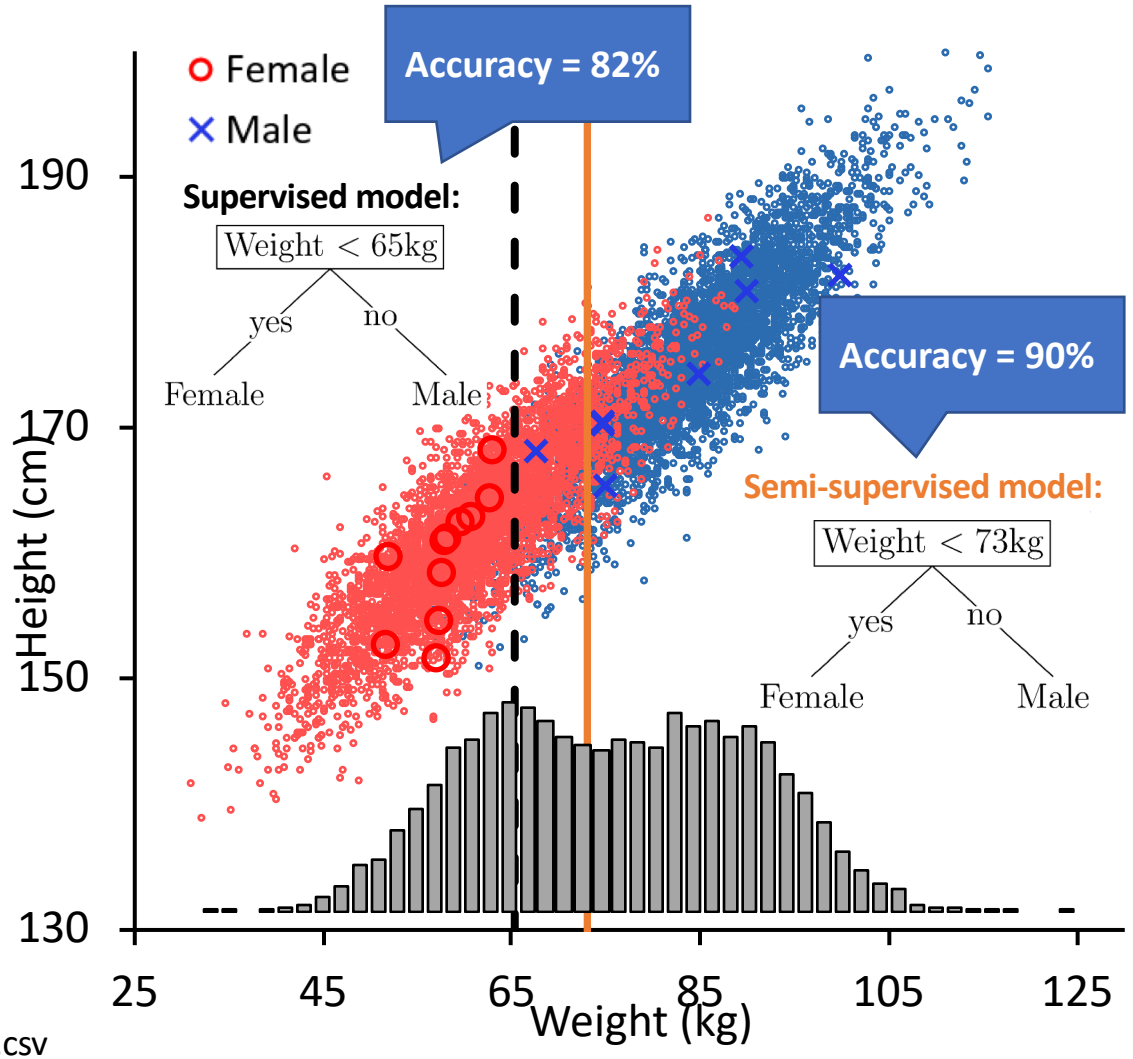
[1]https://helloacm.com/data/gender-height-weight.csv

How unlabeled data can help?

Task: Predict gender from person's weight and height

Data[1]**:** 10000 entries of height, weight and gender

Assumptions about the labels with respect to the structure of unlabeled data



[1]https://helloacm.com/data/gender-height-weight.csv

Why multi-target prediction?

Primitive outputs: $Y \subseteq \mathbb{R}$ (regression), $Y \subseteq \mathbb{N}$ (classification)

Multi-target prediction: tuple of values, potentially present hierarchical dependencies of the values

Applications:

- Gene function prediction
- Gene – disease, drug/compound – gene, drug side effects, ...

Global or local models:

(Global) methods that take the structure into account are better!

SSL for classification tasks

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	Yes
Example 2	2	FALSE	0.08	0.07	**?**
Example 3	1	FALSE	0.08	0.07	**?**
Example 4	2	TRUE	0.49	0.69	Yes
Example 5	3	TRUE	0.49	0.69	No
Example 6	4	FALSE	0.08	0.07	**?**
…	…				…

SSL for regression tasks

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	0.84
Example 2	2	FALSE	0.08	0.07	**?**
Example 3	1	FALSE	0.08	0.07	0.11
Example 4	2	TRUE	0.49	0.69	**?**
Example 5	3	TRUE	0.49	0.69	**?**
Example 6	4	FALSE	0.08	0.07	0.78
…	…				…

SSL for multi-label classification

	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	?	?	?
Example 2	2	FALSE	0.08	0.07	0	1	1
Example 3	1	FALSE	0.08	0.07	?	?	?
Example 4	2	TRUE	0.49	0.69	1	0	1
Example 5	3	TRUE	0.49	0.69	?	?	?
Example 6	4	FALSE	0.08	0.07	1	0	0
…	…				…	…	…

SSL for multi-target regression

	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	?	?	?
Example 2	2	FALSE	0.08	0.07	0.56	0.99	7.59
Example 3	1	FALSE	0.08	0.07	?	?	?
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	?	?	?
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
…	…				…	…	…

Existing SSL methods for SOP

I. Methods for a specific SOP task:

- **MLC:** Graph based (Chen 2008; Zha 2009; Wang 2011; Kong 2013; Wang 2014, 2016), kNNs (d Lucena 2015), Co-training (Xu 2014), Binary relevance (Švec 2014), Boosting (Zhao 2015), SVMs (Wu 2013)
- **HMLC:** Spectral graph transducer (Ceci 2008), Self-training (Santos 2014)
- **MTR:** Gaussian processes for computer vision (Navaratnam 2007)

II. Methods for several SOP tasks:

- **MLC + HMLC:** SVMs (Altun 2006; Brefeld 2007; Li 2014), Co-training (Brefeld 2006), Conditional Random Fields (Wang 2009; Subramanya 2010; Dhillon 2011), kNNs (Jiang 2016; Du 2017), Hybrid discriminative-generative (Suzuki 2007), Graph based (Hu 2010)
- **MLC + MTR:** Kernelized Bayesian matrix factorization (Gönen 2014)
- **HMLC + MTR:** Input Output Kernel Regression (Brouard 2016)
- **MLC + HMLC + MTR: This talk**

Limitations of the existing methods

1) **Can handle only specific type(s) of structured output(s)**
   - Mostly nominal types
2) **High Computational complexity**
   - Conditional Random Fields, SVMs, Graph and kernel based methods
3) **Difficult to use for non-experts**
   - The user needs to define task-specific kernels
4) **None of the existing methods produce interpretable models**
   - Important in knowledge discovery aspect of predictive modeling
5) **Limited application and evaluation**
   - Evaluated only on specific domains and/or very few datasets
   - Advantages as compared to supervised methods are not clear

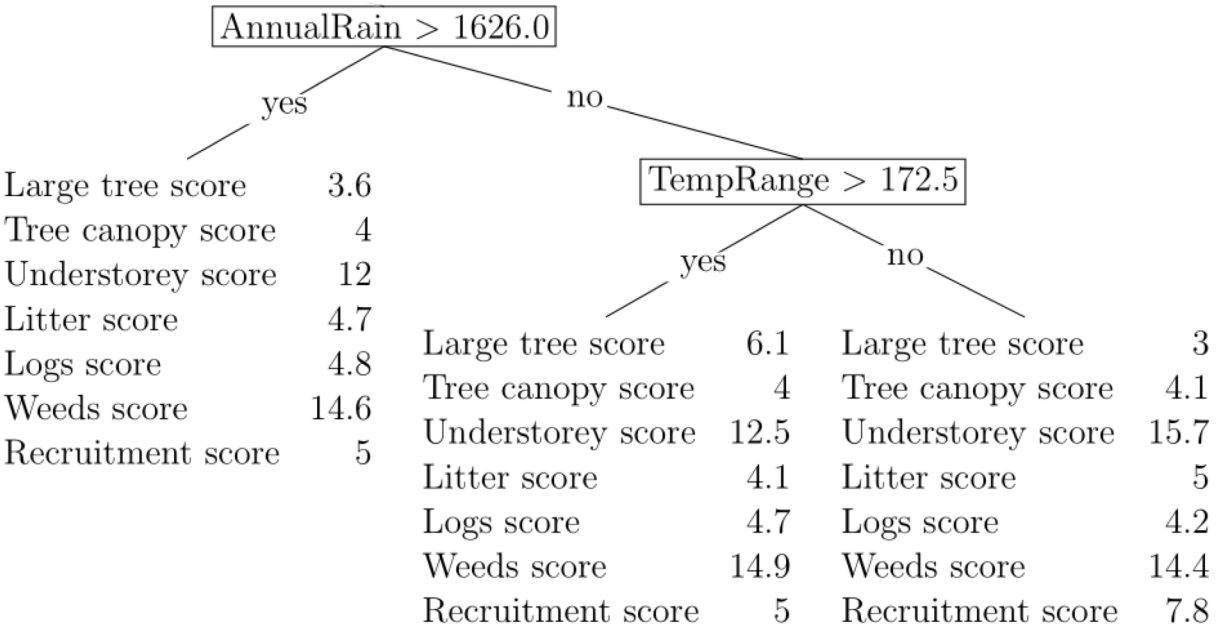
outline

- Introduction
- (Semi-supervised) predictive clustering trees (PCTs)
- Evaluation and illustrative examples
- Conclusions

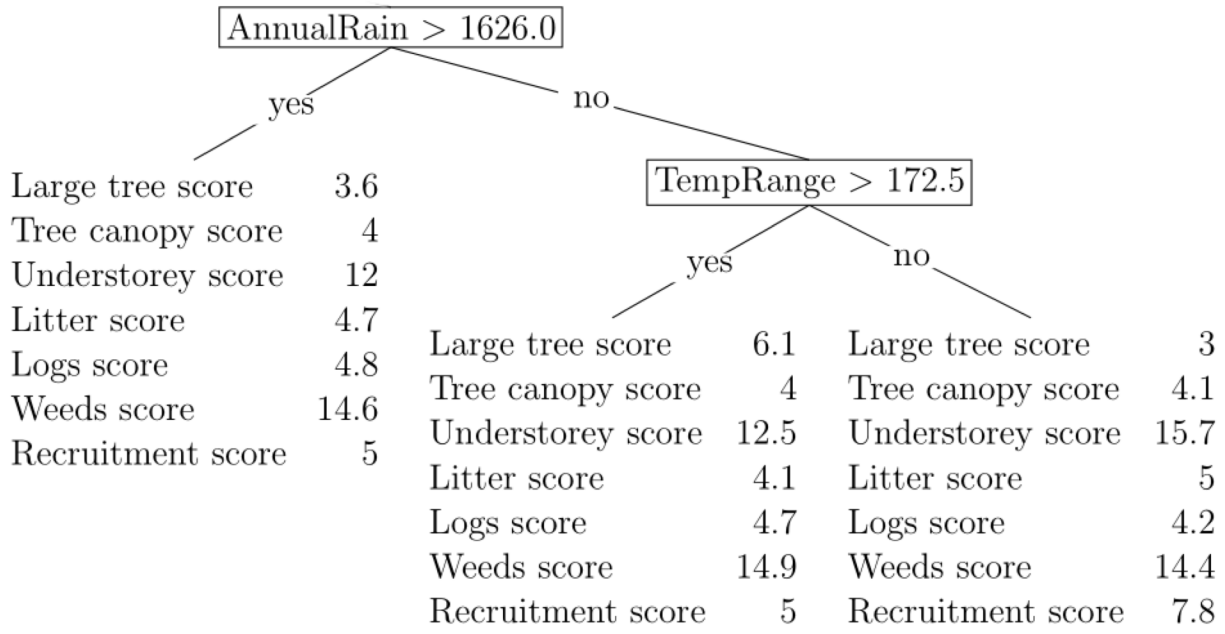
Predictive clustering trees

- Generalization of decision trees towards various tasks, including MTR, MLC and HMLC
- Computationally efficient
- Easy to use Interpretable models

<u>Easily extendable towards SSL for SOP!</u>



Supervised PCTs



Considers only output space in supervised learning!
(Clusters are coherent only in output space)

Variance function

- Evaluates splits

Prototype function

- Calculates predictions

PCTs instantiations

- Multi-target regression
  - Prototype: Average
  - Variance: $$Var(E) = \sum_{i=1}^{T} Var(Y_i)$$
- Multi-target classification/Multi-label classification
  - Prototype: Probability distribution and Majority vote
  - Variance: $$Var(E) = \sum_{i=1}^{T} Gini(E, Y_i) \text{ or } Var(E) = \sum_{i=1}^{T} Entropy(E, Y_i)$$
- Hierarchical multi-label classification
  - Prototype: Average with a threshold for class membership
  - Hierarchy type: tree or DAG
  - Variance: $$Var(E) = \frac{1}{|E|} \cdot \sum_{E_i \in E} d(L_i, \overline{L})^2,$$
    $$d(L_1, L_2) = \sqrt{\sum_{i=1}^{|L|} \omega(c_i) \cdot (L_{1,i} - L_{2,i})^2}, \omega(c_i) = \omega_0 \cdot \omega(par(c_i))$$

Semi-supervised PCTs

Variance function: Variance of **output** space **+** Variance of **input** space

$$Var_f(E,Y,X) = w \cdot Var_f(E,Y) + (1-w) \cdot Var_f(E,X)$$

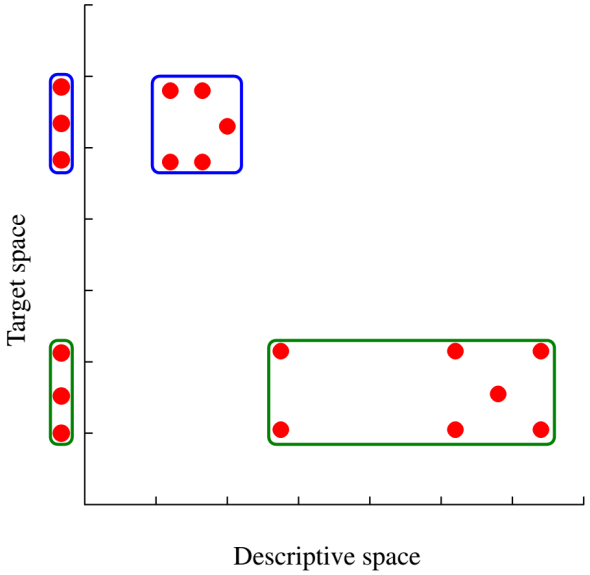
$\boldsymbol{w \in [0,1]}$ = controls the amount of supervision:

$w = 0$	$0 < w < 1$	$w = 1$
Unsupervised	Semi−supervised	Supervised

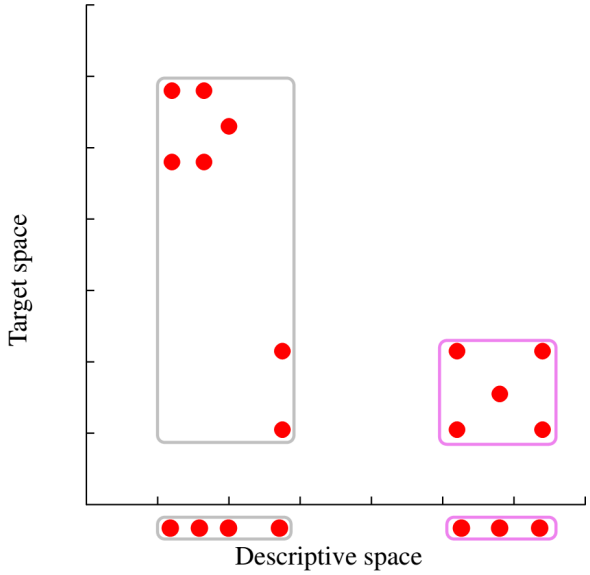
- $Var_f(E,Y)$ and $Var_f(E,X)$ extended to handle unlabeled data
- Resolved mixing different variances: numeric/nominal/hierarchical

Predictive clustering

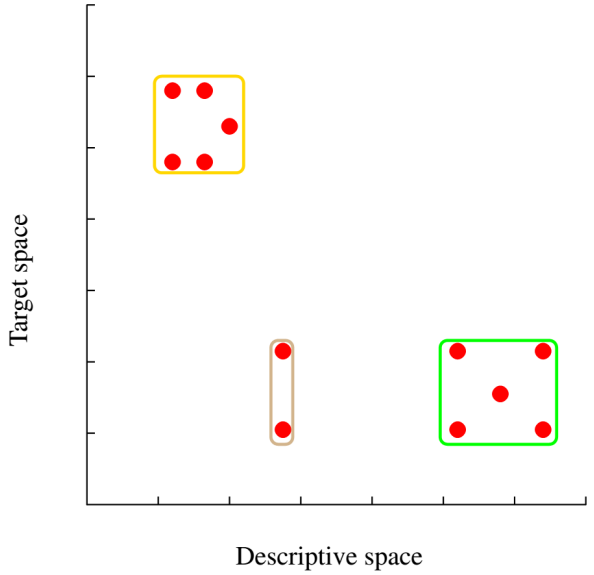
- Clusters are coherent in both input and output spaces



Predictive modelling
Supervised



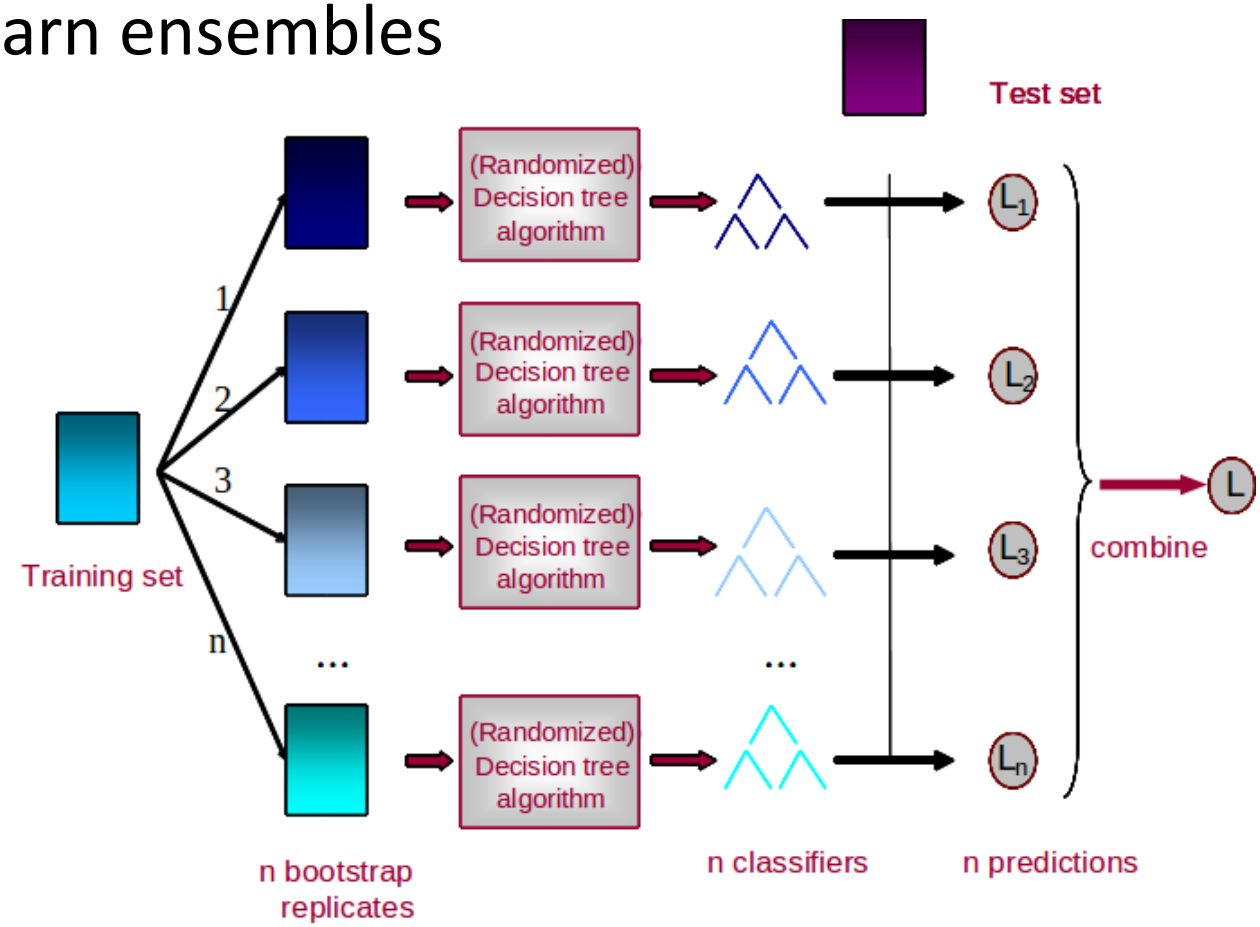
Clustering
Unsupervised



Predictive Clustering

Ensembles of semi-supervised PCTs

- Once we have developed SSL PCTs, it is fairly easy to learn ensembles



outline

- Introduction
- (Semi-supervised) predictive clustering trees (PCTs)
- Evaluation and illustrative example
- Conclusions

Experimental evaluation

1) **Predictive performance**
   - Can we improve over supervised PCTs?
   - Influence of the amount of labeled data?
2) **Influence of the w parameter**
   - How it affects the performance?
3) **Influence of the unlabeled data**
   - is it necessary to improve?
4) **Interpretability and model sizes**
5) **Predictive performance for tasks with primitive outputs**

Experimental setup

Comparison: Supervised PCTs (PCT) and Random Forests (RF)

Datasets: ecology, economy, biology, astronomy, text, audio, images...

- **Multi-target prediction:** MTR, MLC and HMLC
- **Primitive output:** binary, multi-class classification and regression

} 12 datasets each task

Labeled data: 50, 100, 200, 350, 500 labeled examples

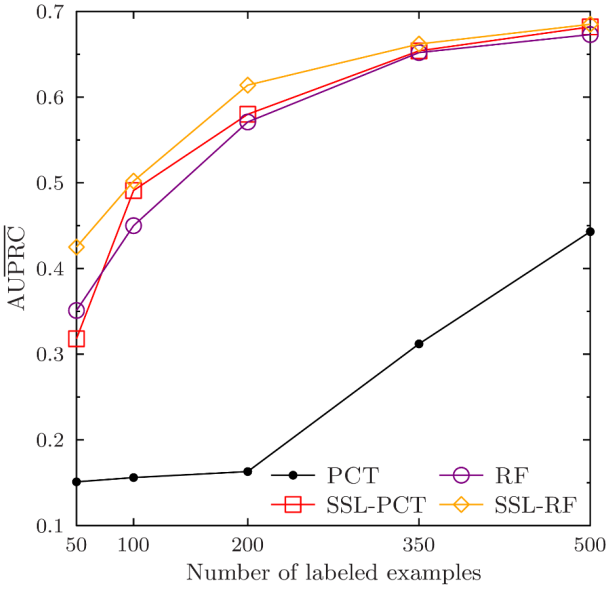
- Selected at random, the rest is unlabeled
- 10 random repetitions

Evaluation: Unlabeled data = Test set

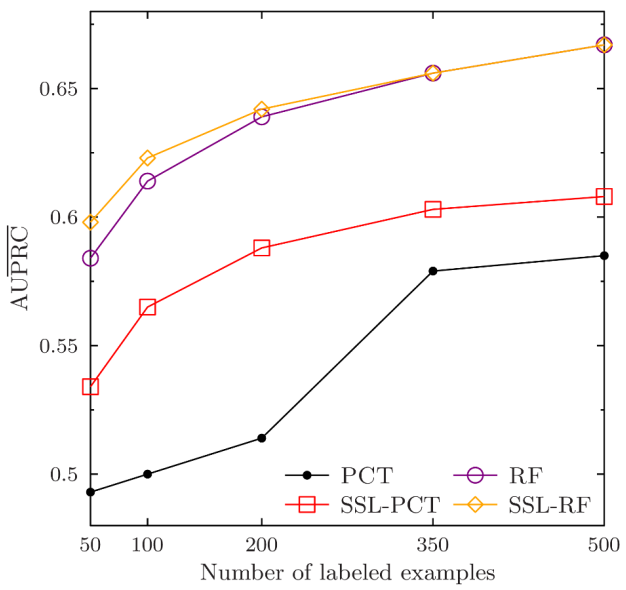
w **parameter:** optimized via internal 3-fold cross validation

Predictive performance (examples)

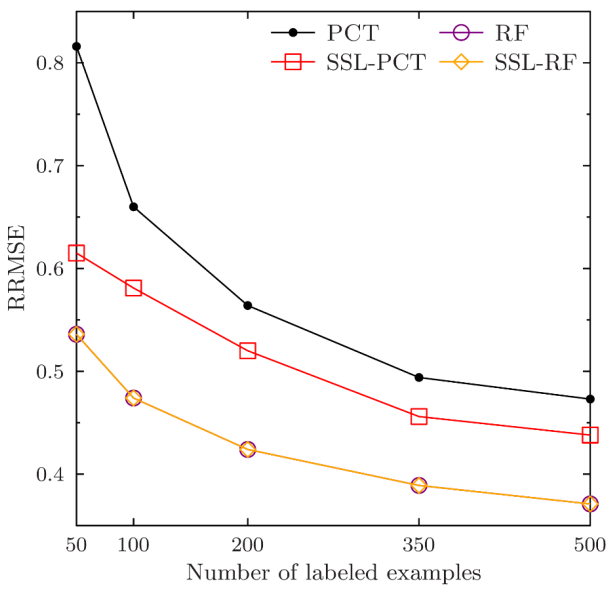
MLC
(Medical dataset)



HMLC
(Enron dataset)



MTR
(RF1 dataset)



Statistical analysis

p-values of Wilcoxon paired signed rank test ($\alpha = 0.05$)*

Methods			Number of labeled examples				
			50	100	200	350	500
Multi-target regression							
PCT	vs.	SSL-PCT	0.093	**0.022**	**0.028**	**0.022**	**0.009**
RF	vs.	SSL-RF	0.959	0.445	0.445	0.333	0.445
Multi-label classification							
PCT	vs.	SSL-PCT	**0.013**	**0.008**	**0.008**	0.093	0.053
RF	vs.	SSL-RF	0.241	0.415	0.262	0.308	0.575
Hierarchical multi-label classification							
PCT	vs.	SSL-PCT	0.834	0.093	**0.028**	**0.028**	**0.028**
RF	vs.	SSL-RF	0.345	0.345	0.249	0.345	0.345

*In all tests, semi-supervised algorithms have better sum of ranks

Influence of the $\boldsymbol{w}$ parameter

Unlabeled data can hurt the performance

- No semi-supervised method is universally good

180 experiments	**Wins**	**Ties**	**Loses**
SSL-PCT vs. PCT	67%	25%	8%

- **Average relative improvement** over PCTs is **43%** (degradation **8%**)

$\boldsymbol{w}$ provides safety mechanism!

- $\boldsymbol{w}$ needs to be optimized for every dataset/amount of labeled data

Influence of the unlabeled data

PCT^{D+T} : supervised variant of SSL-PCTs

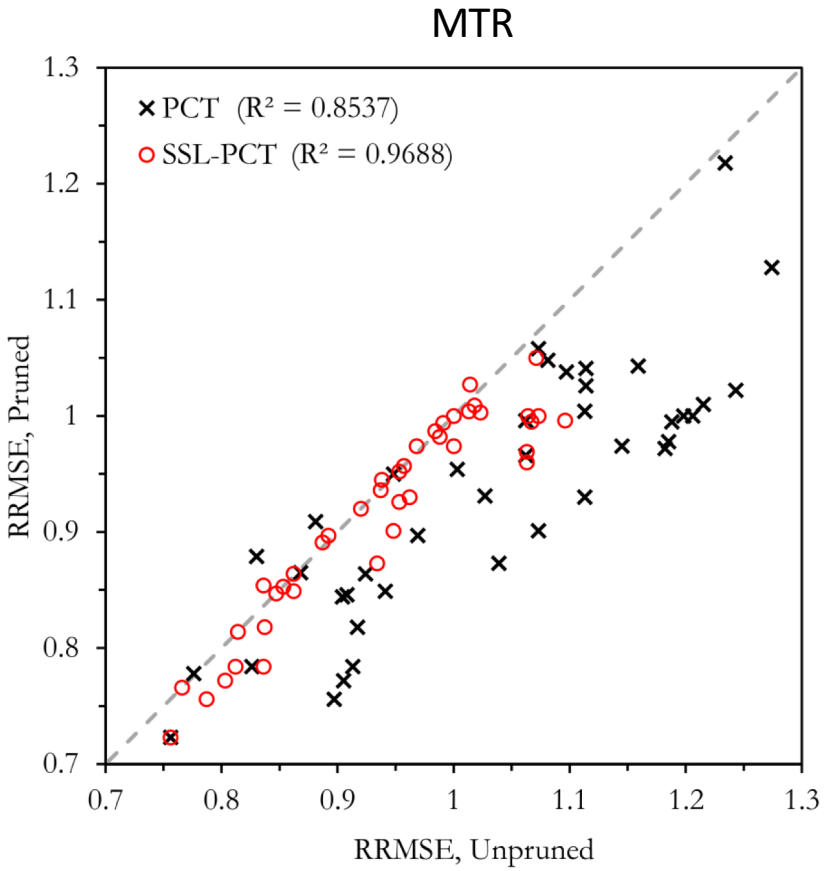
- Considers both input and output space
- It is not supplied with unlabeled data

180 experiments	**Wins**	**Ties**	**Loses**
SSL-PCT vs. PCT^{D+T}	41%	46%	13%
SSL-PCT vs. PCT	67%	25%	8%

Average relative improvement of PCT^{D+T} over PCTs is **5%** **Unlabeled data are the principal component of SSL-PCTs!**

Interpretability and model sizes

- SSL-PCTs produce readily interpretable models
- The only such SSL method for MTP
- SSL-PCTs can even enhance interpretability of PCTs
  - smaller model size
- SSL-PCTs less affected by pruning
  - overfit less than PCTs



SSL-PCTs for primitive outputs

p-values of Wilcoxon paired signed rank test ($\alpha = 0.05$)*

Methods			Number of labeled examples					
			25	50	100	200	350	500
Binary classification								
PCT	vs.	SSL-PCT	**0.009**	0.388	0.066	**0.005**	**0.019**	**0.019**
RF	vs.	SSL-RF	0.529	0.192	**0.002**	0.099	0.093	**0.012**
Multi-class classification								
PCT	vs.	SSL-PCT	0.248	0.084	**0.014**	**0.007**	0.192	0.081
RF	vs.	SSL-RF	0.563	**0.011**	**0.011**	**0.003**	**0.004**	**0.02**
Regression								
PCT	vs.	SSL-PCT	**0.011**	**0.01**	**0.004**	0.367	0.48	0.583
RF	vs.	SSL-RF	**0.008**	**0.065**	**0.008**	**0.023**	**0.034**	0.126

*In all tests, semi-supervised algorithms have better sum of ranks

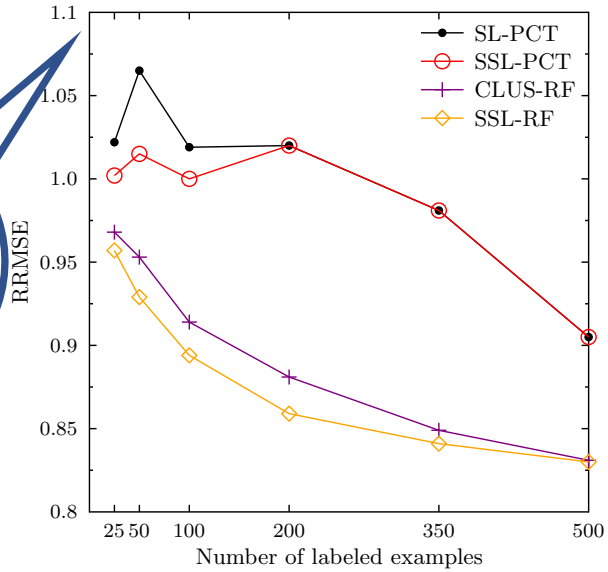
Illustrative study on QSAR datasets

Dataset	Domain	N	D/C
Neurokinin 1 receptor (NK1)	QSAR	2446	1024/0
Glycogen synthase kinase-3 alpha (GSK3A)	QSAR	1211	1024/0
Rho-associated protein kinase 2 (ROCK2)	QSAR	1521	1024/0
Human immunodeficiency virus type 1 protease (HIV-1)	QSAR	4442	1024/0

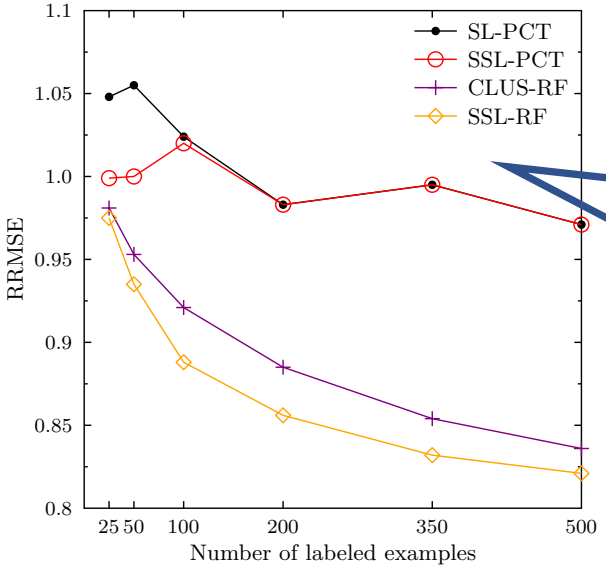
FCFP molecular fingerprints

Performance results

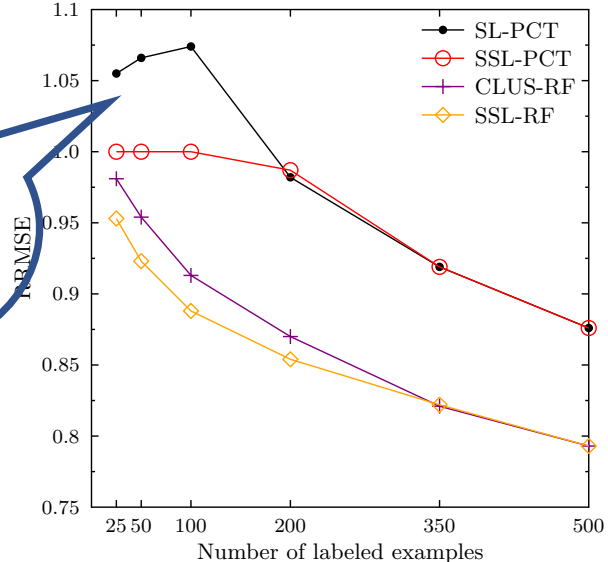
NK1



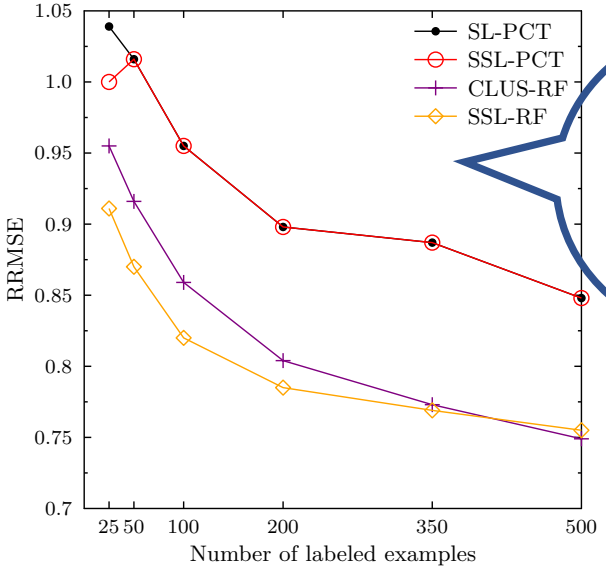
HIV-1



GSK3A



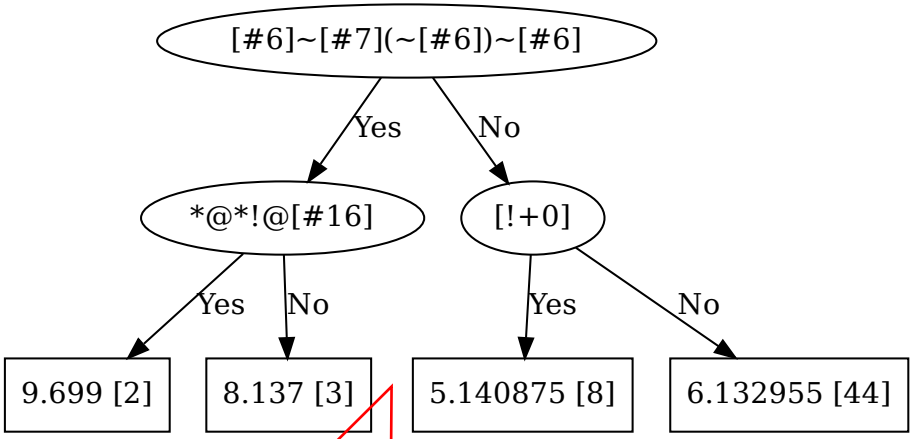
ROCK2

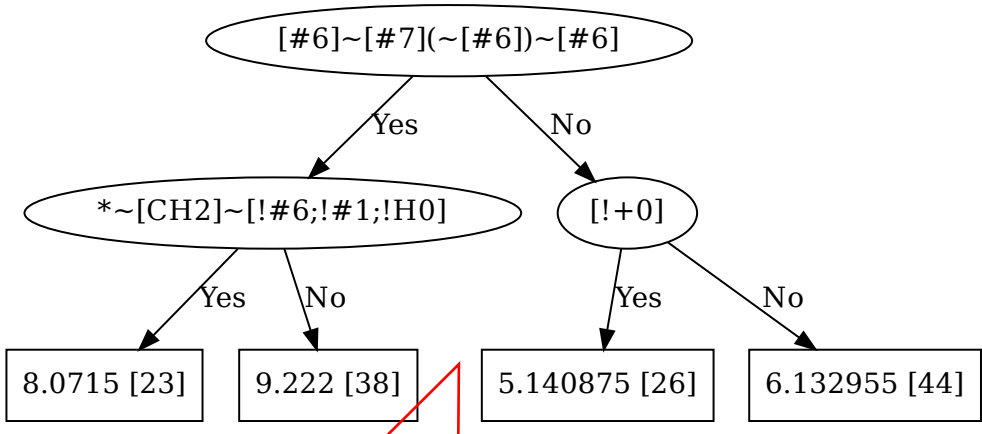


Interpretability potential

- Focus on farnesyltransferase (FTase)
- 57 compounds that inhibit FTase in *Saccharomyces cerevisiae S288c*
- Extracted 74 other compounds with unknown inhibitory property (and Tanimoto similarity > 0.8)
- MACCS structural keys fingerprints calculated with the RDKit library
- The fingerprints are binary vectors of length 166, where each bit corresponds to a specific SMARTS pattern

Obtained PCTs

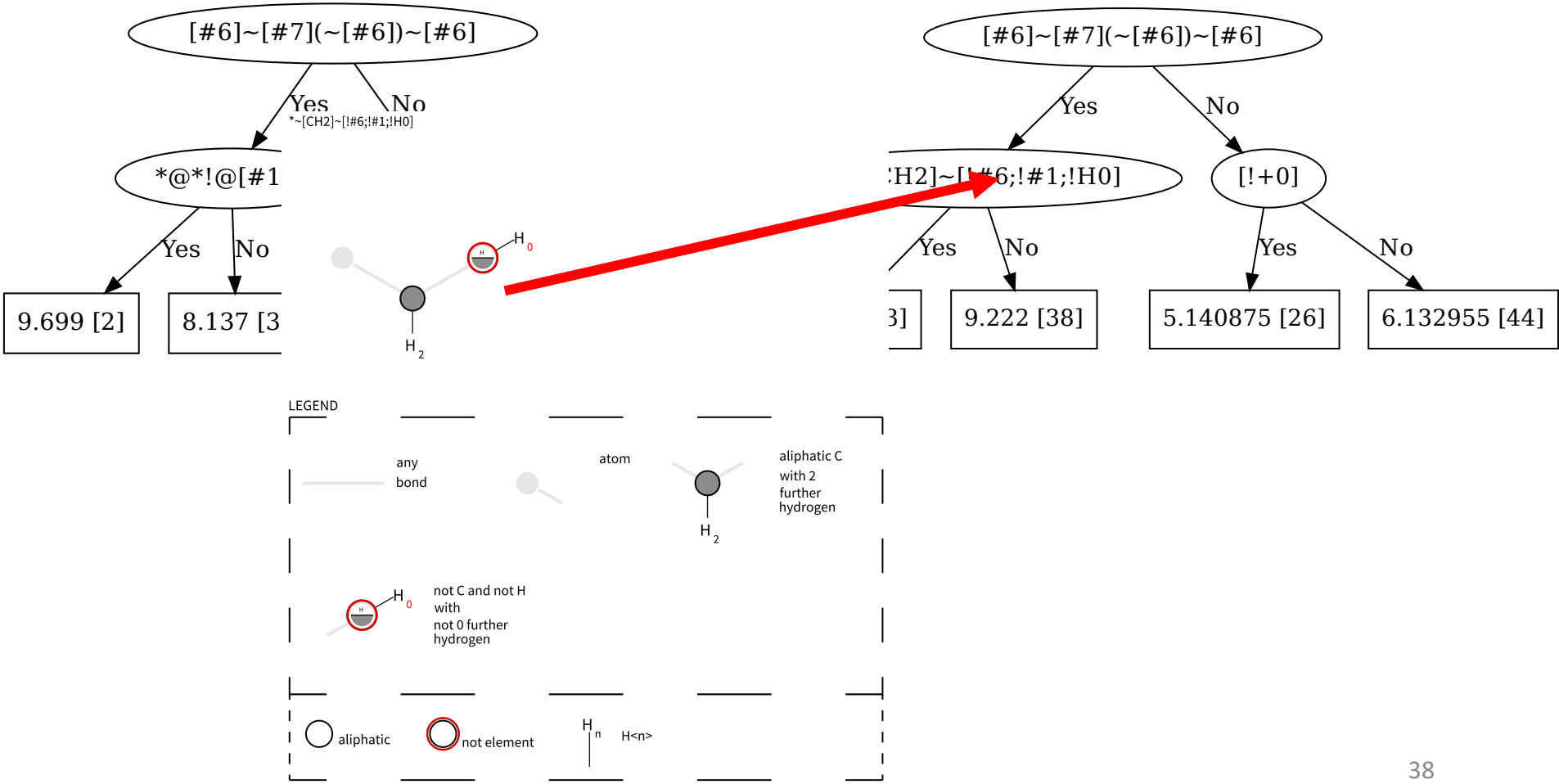




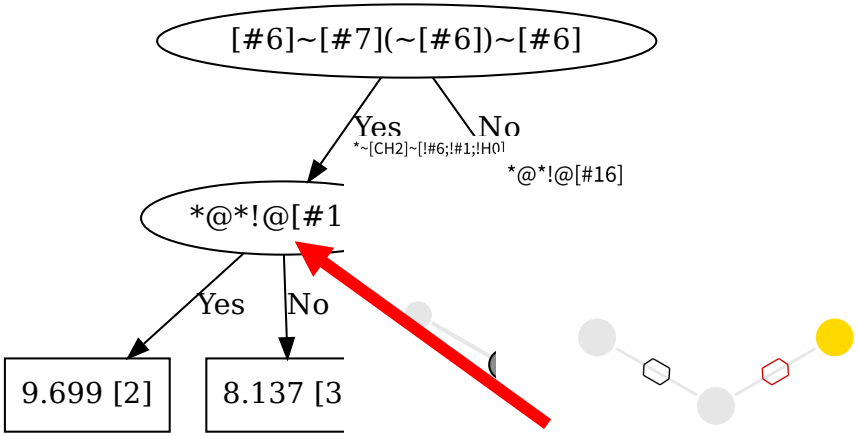
Supervised PCT
RMSE: 0.764

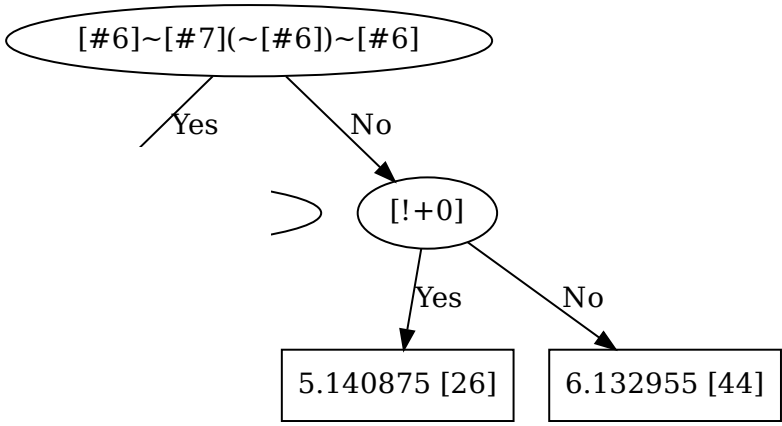
Semi-supervised PCT
RMSE: 0.683

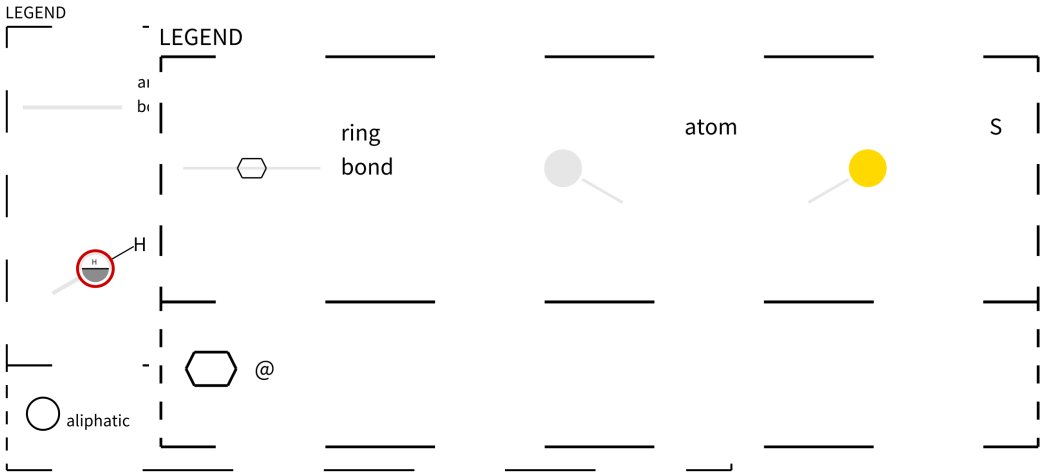
Obtained PCTs



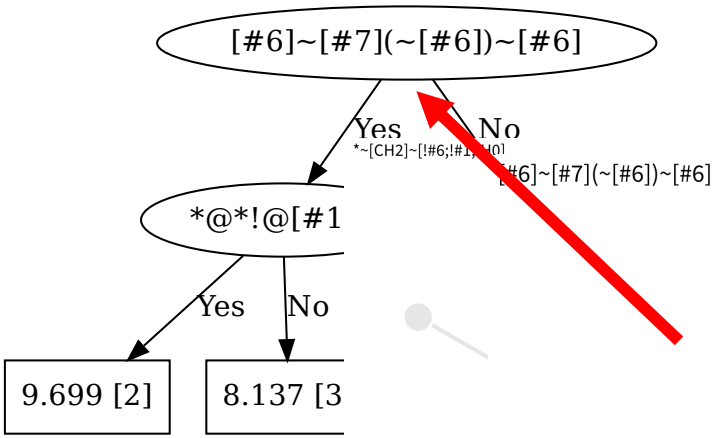
Obtained PCTs

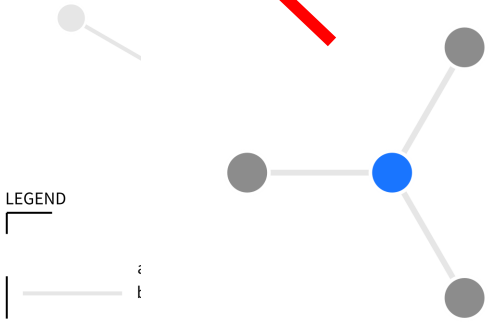


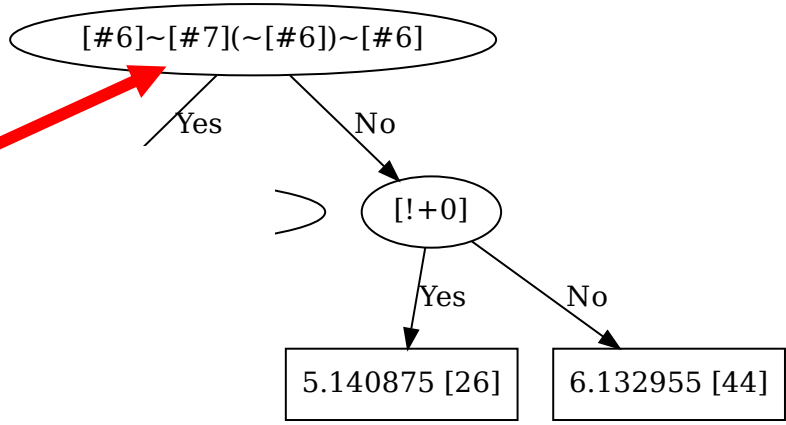


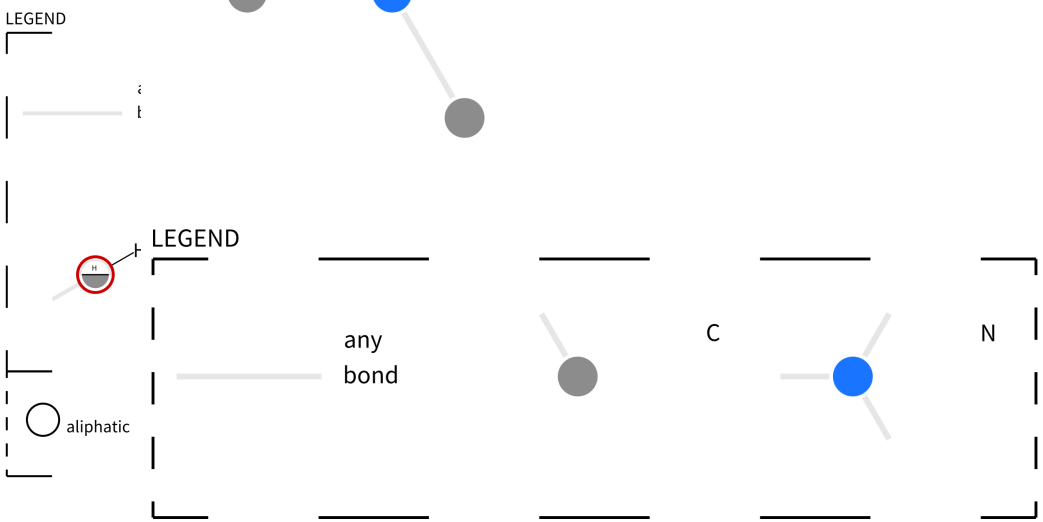


Obtained PCTs

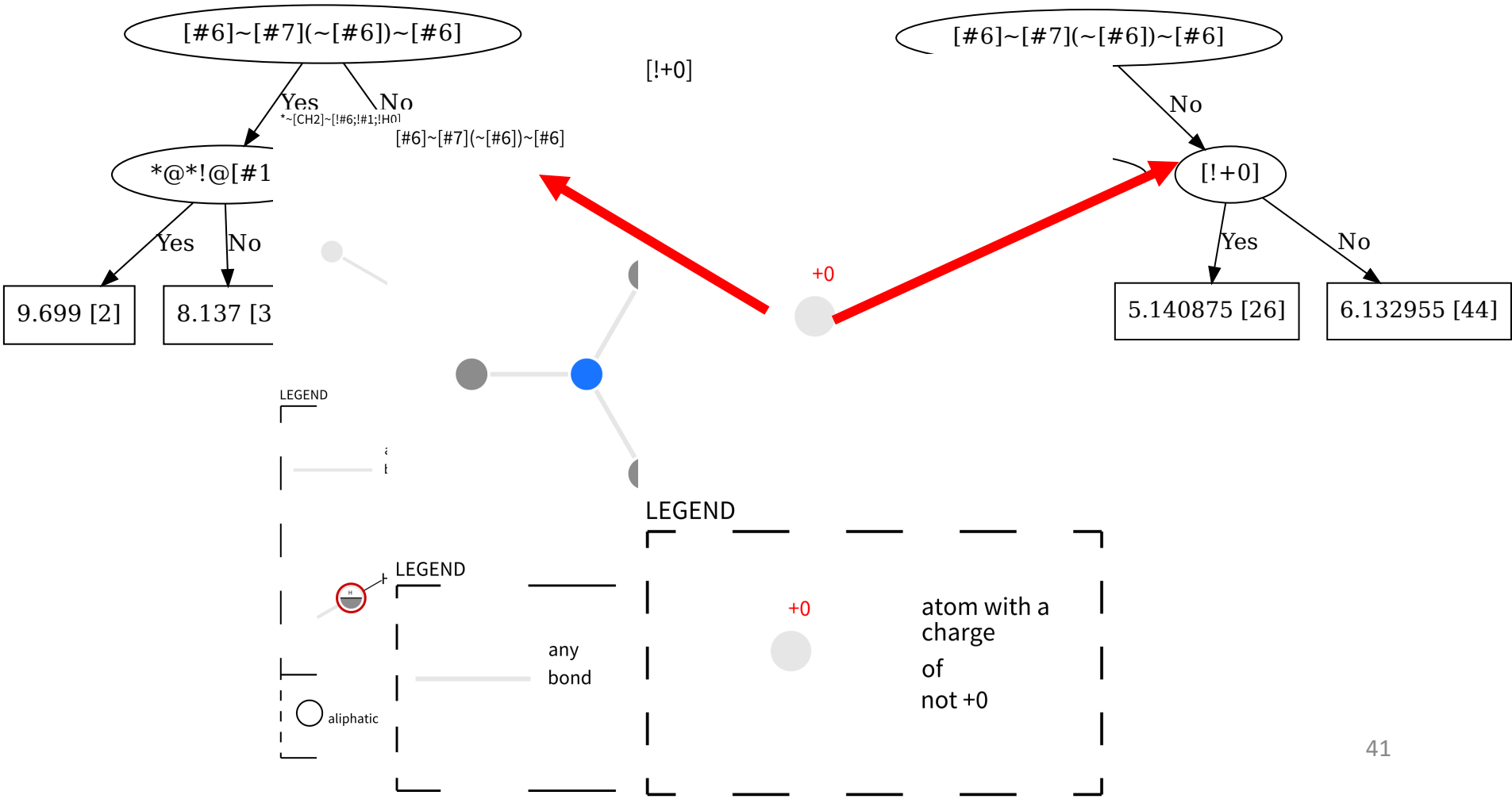








Obtained PCTs



outline

- Introduction
- (Semi-supervised) predictive clustering trees (PCTs)
- Evaluation and illustrative examples
- Conclusions

Conclusions

- Versatile in terms of MTP tasks (and also primitive outputs)
- Improve predictive performance of supervised PCTs and overfit less
- Highly useful in practice („safety mechanism“, easy to use)
- Performance improvement does not entirely translates to the ensemble setting
- Interpretable models (even can enhance interpretability)

Questions?