Machine Learning for Environmental & Life Sciences

Sašo Džeroski
Jozef Stefan Institute, Ljubljana, Slovenia

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

The basic Machine Learning task: Predictive modeling

- Classification

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	Yes
Example 2	2	FALSE	0.08	0.07	Yes
Example 3	1	FALSE	0.08	0.07	No
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	Yes
…	…				…

- Regression

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	0.84
Example 2	2	FALSE	0.08	0.07	0.75
Example 3	1	FALSE	0.08	0.07	0.11
Example 4	2	TRUE	0.49	0.69	0.52
Example 5	3	TRUE	0.49	0.69	0.35
Example 6	4	FALSE	0.08	0.07	0.78
…	…				…

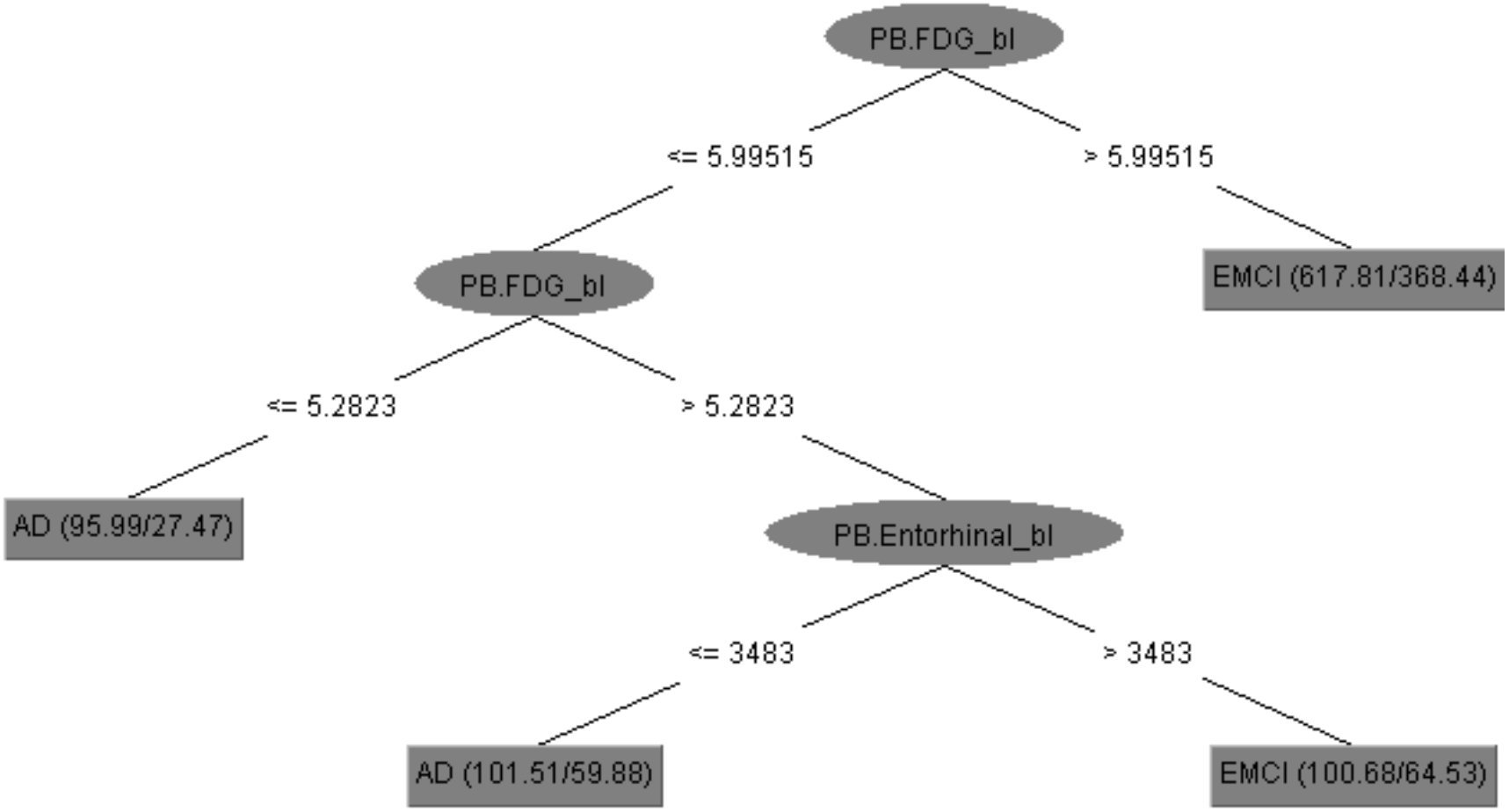
An example task of Predictive Modelling: Medical diagnosis

- Predictive models focus on a target variable and predict its value from the values of input variables
- Classical problem: Medical diagnosis
- An example: Neurodegenerative diseases
- Target variable: Diagnosis; Possible values:
  - CN - Cognitively Normal (0)
  - SMC - Significant Memory Concern
  - EMCI - Early Mild Cognitive Impairment
  - LMCI - Late Mild Cognitive Impairment
  - AD - Alzheimer‘s Disease (4)

Example task: Descriptive vars.; Biomarkers for Alzheimer's

1. APOE4 – Genetic variations of APOE4 related gene
2. FDG – Positron emission tomography (PET) imaging results with [^{18}F]fluorodeoxyglucose
3. AV45 – Positron emission tomography (PET) imaging results with [^{18}F]-labeled amyloid imaging agent AV45
4. Ventricles
5. Hippocampus
6. WholeBrain
7. Entorhinal
8. Fusiform – Fusiform gyrus
9. MidTemp – Middle Temporal Gyrus
10. ICV – Intracerebral volume [Volumetric data 4-10]

Example: Decision tree for diagnosis



Another example of single-target predictive modeling (classification)

Task: Habitat suitability modeling

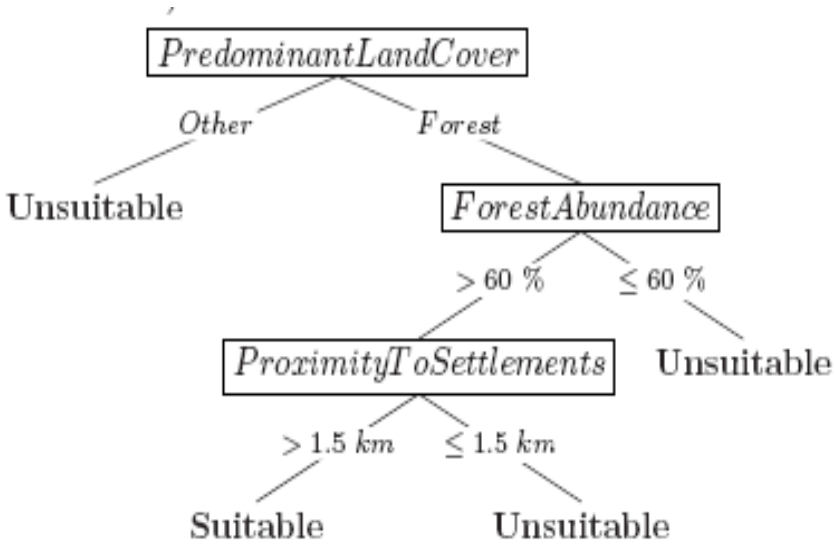
Input: Data on locations and habitat suitability

Location	PLC	FOREST-ABUNDANCE	PTS	OtherEnvVariables	BBH
l1	Forest	80	21.4	...	Yes
l2	Forest	66	13.9	...	Yes
l3	Forest	55	50.0	...	No
l4	Forest	72	1.2	...	No
l5	Grassland	6	19.1	...	No
l6	Grassland	0	11.4	...	No
l7	Wetland	3	5.8	...	No
l8	Water	0	3.9	...	No

Another example of single-target predictive modeling (classification)

Task: Habitat suitability modeling

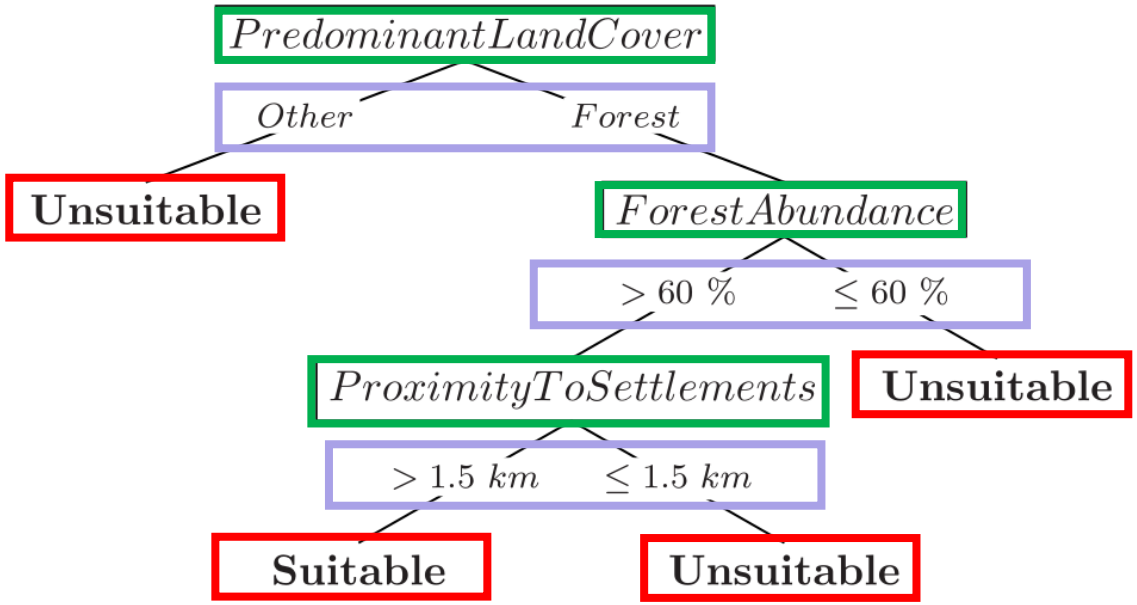
Output: Habitat suitability model




```
IF   PREDOMINANT-LAND-COVER = Forest
AND  FOREST-ABUNDANCE > 60%
AND  PROXIMITY-TO-SETTLEMENTS > 1.5 km
THEN BrownBearHabitat = Suitable
```

What is a decision tree?

- Hierarchicaly structured predictive model
- Nodes – correspond to (environmental) variables
- Arcs – possible values of the variables
- Leafs – predictions for the target variable



Making a Prediction with a Decision Tree

Take as input values of attributes/ independent vars.

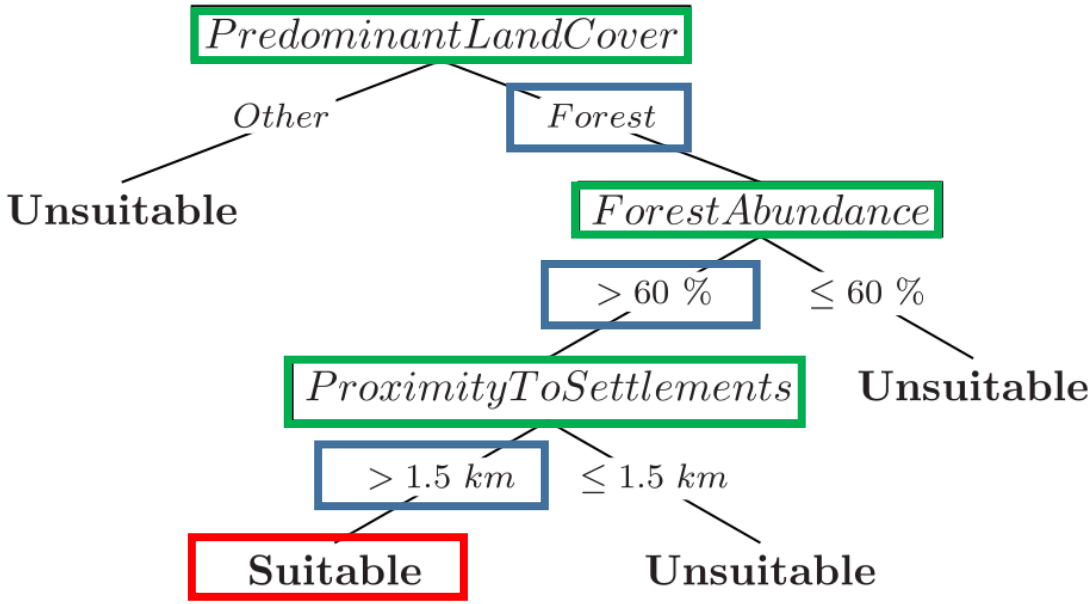
Follow branches according

to the values of these

Until you reach a leaf

PLC	FOREST-ABUNDANCE	PTS	BBH
Forest	80	21.4	?

Yes



Machine Learning of Decision Trees

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

Top-Down Induction of Decision Trees

To construct a tree T from a training set S:

- If **all the examples belong to the same class C**, construct a leaf labeled C

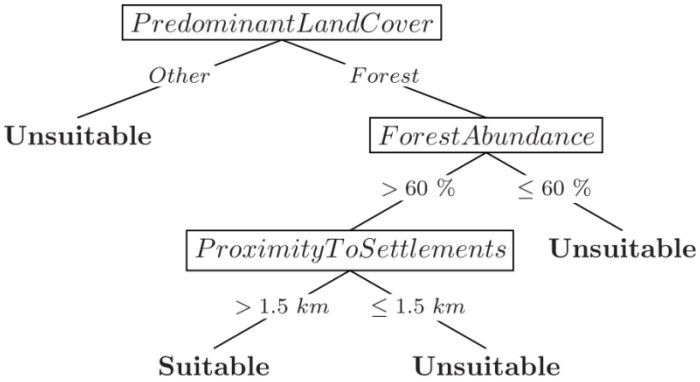
- Otherwise:
  - Select the best attribute A with values v1, …, vn, which **reduces the most the impurity of the target**
  - Partition S into S1, …, Sn according to A
  - Recursively construct subtrees T1 to Tn for S1 to Sn
  - Result: a tree with root A and subtrees T1, …, Tn

TDIDT Illustrated

Input: Set of learning examples S

1) Find the best split t (attribute value which results in the biggest reduction of variance considering the target variable)
2) Partition the data S into partitions S_v according to t
3) For each partition, if stopping criteria met (e.g., all of the examples in partition are of the same class), make a leaf, assign a (prototype) class to leaf
4) Otherwise, repeat 1) for each node

Location	PLC	FOREST-ABUNDANCE	PTS	OtherEnvVariables	BBH
l1	Forest	80	21.4	...	Yes
l2	Forest	66	13.9	...	Yes
l4	Forest	72	1.2	...	No



Mining Big and Complex Data: Dimensions of Complexity

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

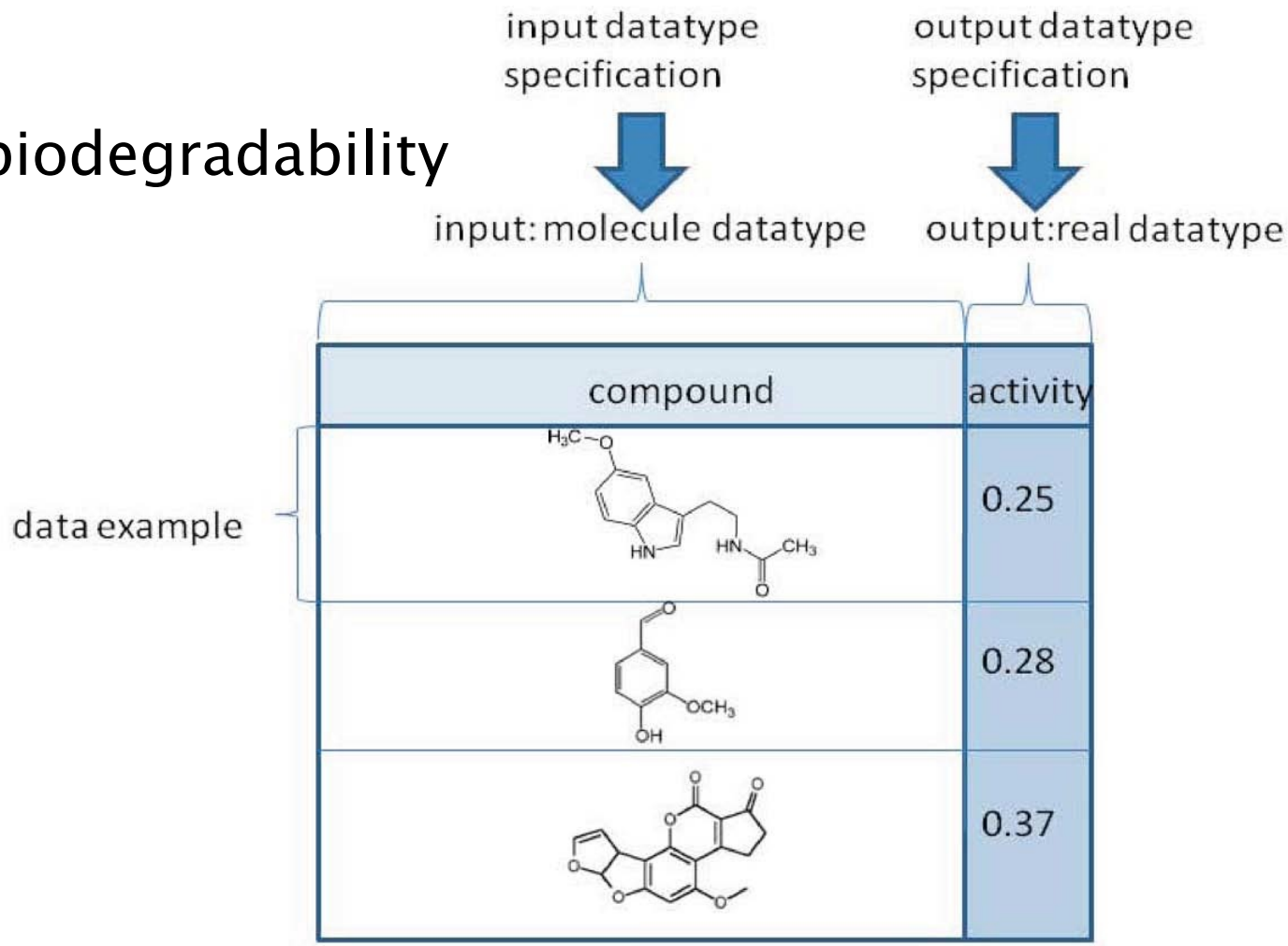
Mining Big and Complex Data

- What is big and complex data?
  - Volume & Velocity (Data Streams)
  - Variety (Structured Inputs and Structured Outputs)
- Variety:
  - Different types of data, different tasks of data mining
- MTP is a special case of structured output prediction
  - But you can have more complex outputs thah in MTP
- Combination with other dimensions of complexity
  - Semi-supervised …
  - Data streams
  - Networked data

Big Data: Variety - Structured Input

Example:

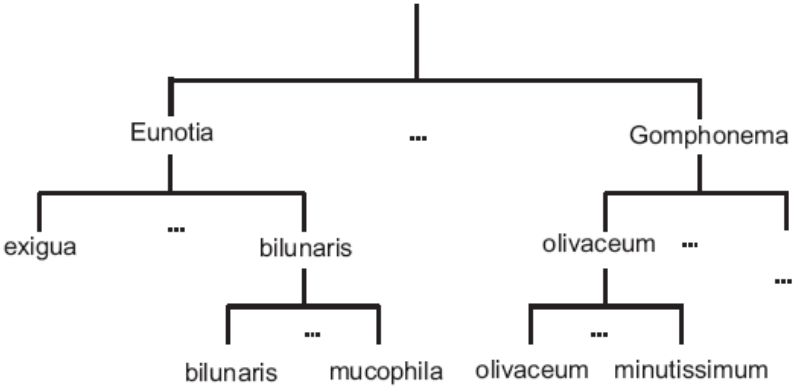
Predicting biodegradability



Big Data: Variety - Structured Output

- Hierarchical classification
- Taxonomic classification of diatoms
- From microscopic images
- Taking into account the taxonomy of diatoms

image	features/descriptors						taxonomy
	Heuristic shape descriptors						
	48	24	59	66	37	…	olivaceum
	36	25	53	45	15	…	minutissimum
	35	25	56	52	19		exigua
…	…	…	…	…	…	…	…



Predictive modeling: Structured output

- The input is the same as for the classical task of predictive modelling: A vector of feature values

- The output is not a single scalar value, but rather a data structure, e.g., a tuple of values:
  - A vector of scalar targets (Multi-target prediction)
  - If targets binary, multi-label classification
  - A hierarchy of binary targets organized in a hierarchy (Hierarchical multi-label classification)

- Other data types possible: sets, sequences
  - E.g., a (time) series of real values

Big Data: Volume & Velocity

- Large number of columns (high dimensionality)
  - Need feature ranking/selection
- Large number of rows (massive data)
  - Need efficient data mining methods
- Streaming rows (data streams)
  - Need incrementality: Not all data available simultaneously
  - Data instances arrive at **high velocities**, in a **specific order** and their number is **potentially arbitrarily large**
  - The **underlying concept** (distribution) governing the data **can change (concept drift)**
  - We need **fast processing** (due to the high velocity)
  - The large and potentially infinite number of examples demands **economical management of available memory**

Data streams: Regression

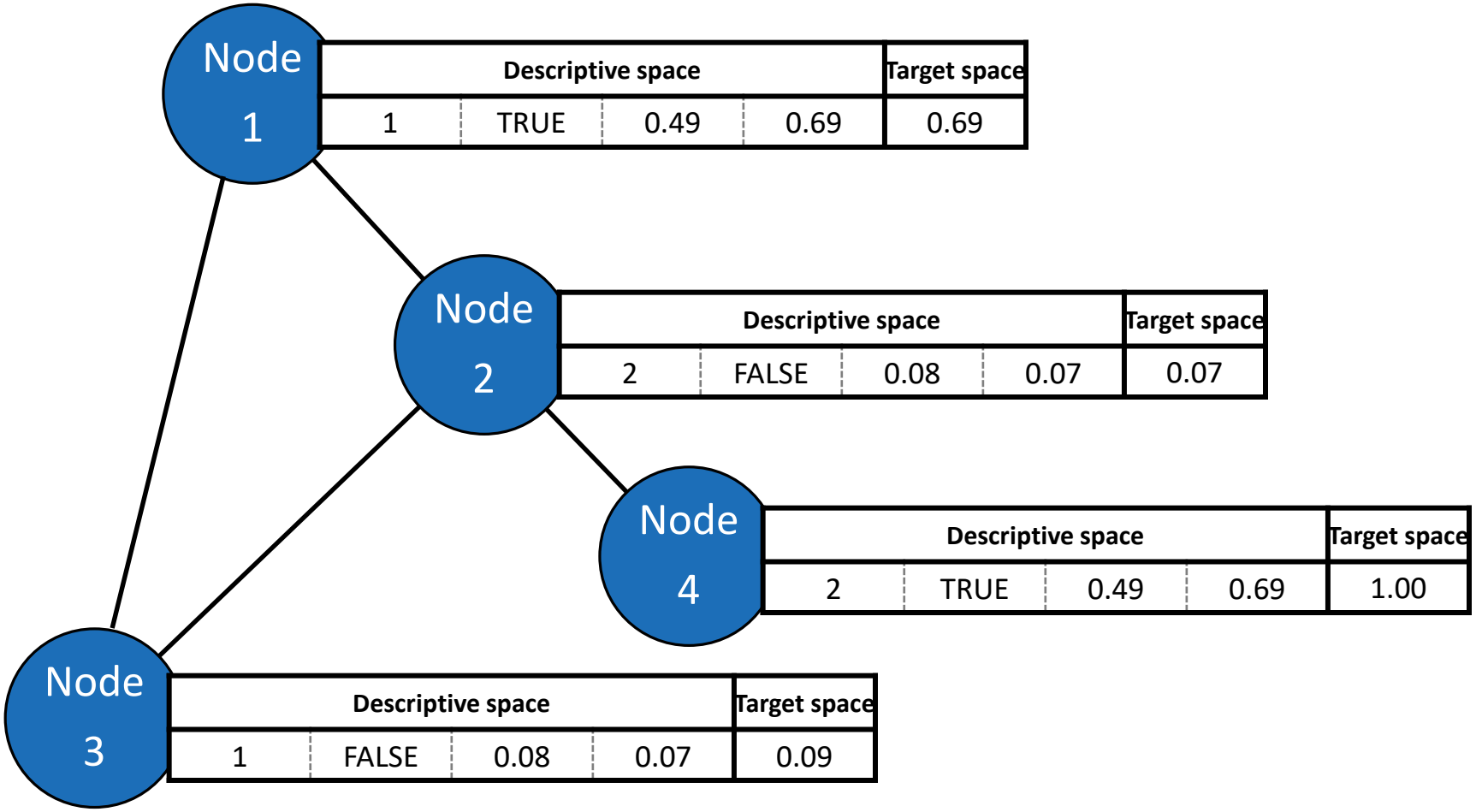
	Descriptive space				Target space
…	…				…
Example n+5	**1**	**TRUE**	**0.49**	**0.69**	**0.45**
Example n+1	4	FALSE	0.08	0.07	0.12
Example n+2	6	FALSE	0.08	0.07	1.54
Example n+3	8	TRUE	0.00	1.00	3.12
Example n+4	6	TRUE	0.00	0.00	0.05
…	…				…

Semi-supervised learning: Classification and regression

	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	**?**
…	…				…

	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
…	…				…

Data in context: Spatio-temporal, network



The Different Tasks of Multi-Target Prediction

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

Weather prediction

- STC: Predicting the outlook (sunny, overcast, rain)
- STR: Predicting the temperature (in degrees Celsius)
- MTP: Predicting the weather
  - Outlook
  - Temperature
  - Humidity
  - Quantity of precipitation …

Multi-target prediction

- Classification

	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	Yes	Blue	Rain
Example 2	2	FALSE	0.08	0.07	Yes	Green	Sun
Example 3	1	FALSE	0.08	0.07	Yes	Blue	Cloudy
Example 4	2	TRUE	0.49	0.69	Yes	Green	Sun
Example 5	3	TRUE	0.49	0.69	No	Blue	Sun
Example 6	4	FALSE	0.08	0.07	Yes	Red	Cloudy
…	…				…	…	…

- Regression

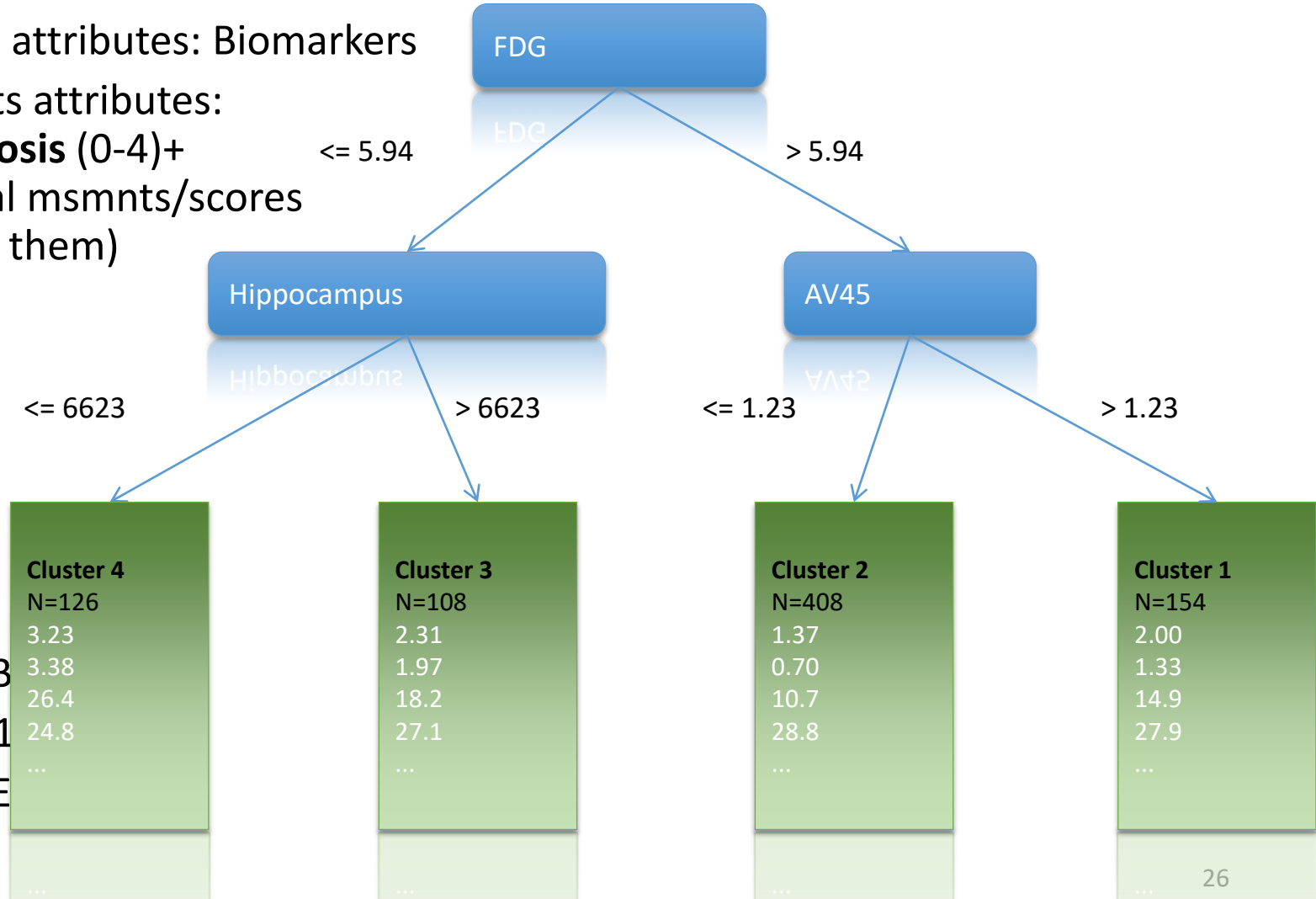
	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	0.68	0.60	3.91
Example 2	2	FALSE	0.08	0.07	0.56	0.99	7.59
Example 3	1	FALSE	0.08	0.07	0.10	1.69	7.57
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	0.11	3.51	2.50
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
…	…				…	…	…

Example MTR task: Target vars.; Clinical scores for Alzheimer's

1. CDRSB – Clinical Dementia Rating Sum of Boxes
2. ADAS13 – AD assessment scale
3. MMSE – Mini Mental State Examination
4. RAVLT (immediate, learning, forgetting, perc. forgetting) – Rey Auditory Verbal Learning Test (4 features)
5. FAQ – Functional Assessment Questionnaire
6. MOCA – Montreal Cognitive Assessment
7. Ecog**Pt** (Memory, Language,Visuospatial Abilities,Planning,Organization,Divided Attention, Total score) – Everyday cognition questionnaire – filled in by patient (7 features)
8. Ecog**SP** (Memory, Language,Visuospatial Abilities,Planning,Organization,Divided Attention, Total score) – Everyday cognition questionnaire – filled in by study parter (7 features)

Example MTR model

- Descr. attributes: Biomarkers
- Targets attributes: **diagnosis** (0-4)+ clinical msmnts/scores (23 of them)



- DX
- CDRSB
- ADAS1
- MMSE
- …

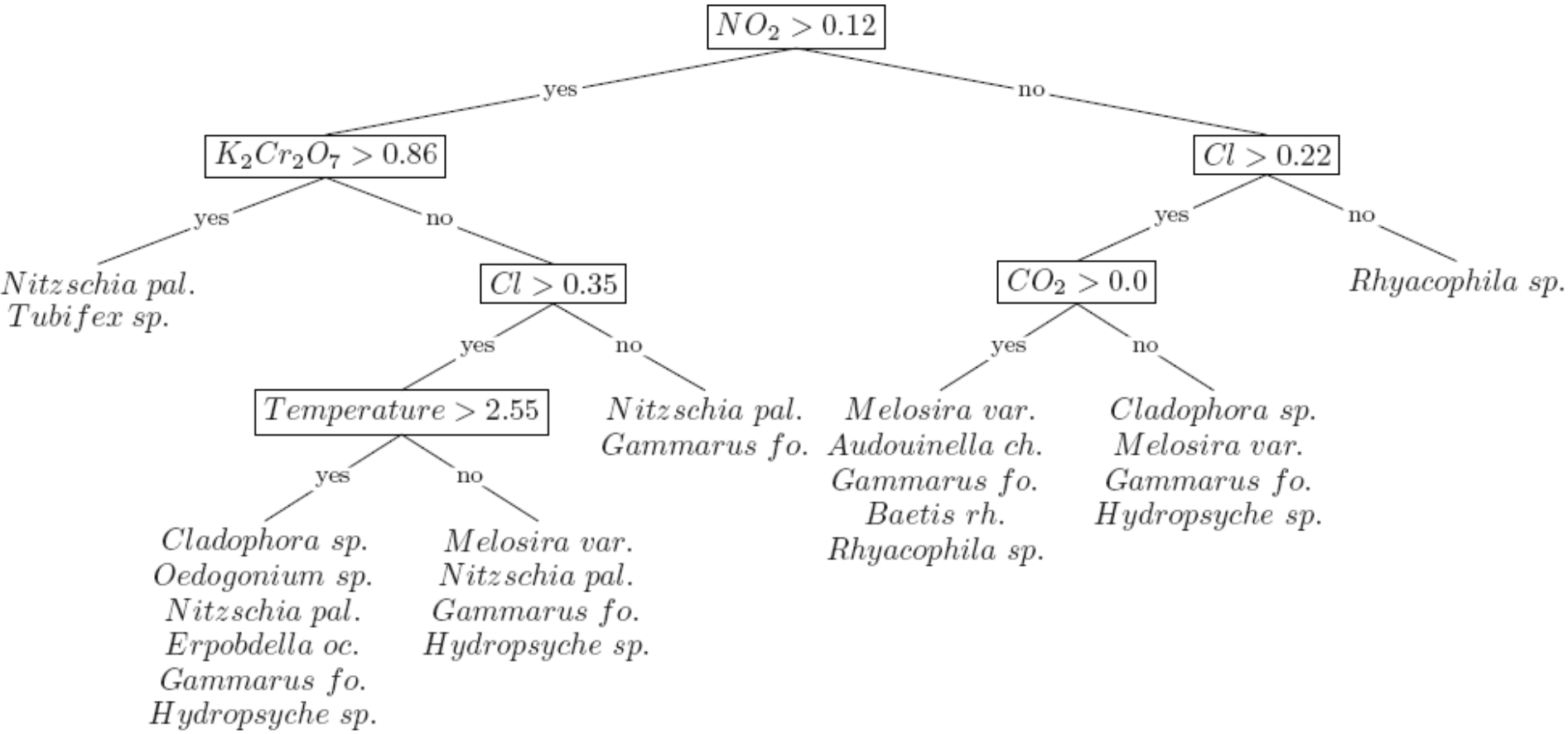
Multi-Target Classification & **Multi-Label** Classification

- Learning models that simultaneously predict several nominal/**binary** target variables
- Input: A vector of descriptive variables
- Output: A vector of several nominal/**binary** targets

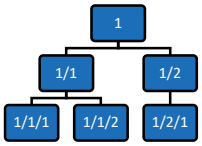
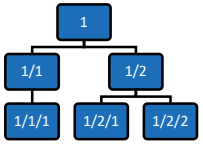
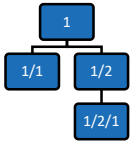
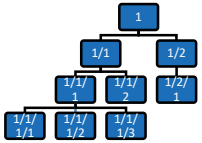
	Descriptive variables						Target variables													
Sample ID	Temperature	$K_2Cr_2O_7$	NO_2	Cl	CO_2	...	*Cladophora sp.*	*Gongrosira incrustans*	*Oedogonium sp.*	*Stigeoclonium tenue*	*Melosira varians*	*Nitzschia palea*	*Audouinella chalybea*	*Erpobdella octoculata*	*Gammarus fossarum*	*Baetis rhodani*	*Hydropsyche sp.*	*Rhyacophila sp.*	*Simulim sp.*	*Tubifex sp.*
ID1	0.66	0.00	0.40	1.46	0.84	...	1	0	0	0	0	1	1	0	1	1	1	1	1	1
ID2	2.03	0.16	0.35	1.74	0.71	...	0	1	0	1	1	1	1	0	1	1	1	1	1	0
ID3	3.25	0.70	0.46	0.78	0.71	...	1	1	0	0	1	0	1	0	1	1	1	0	1	1

Multi-Label Classification Example

- A decision tree for multi-label classification

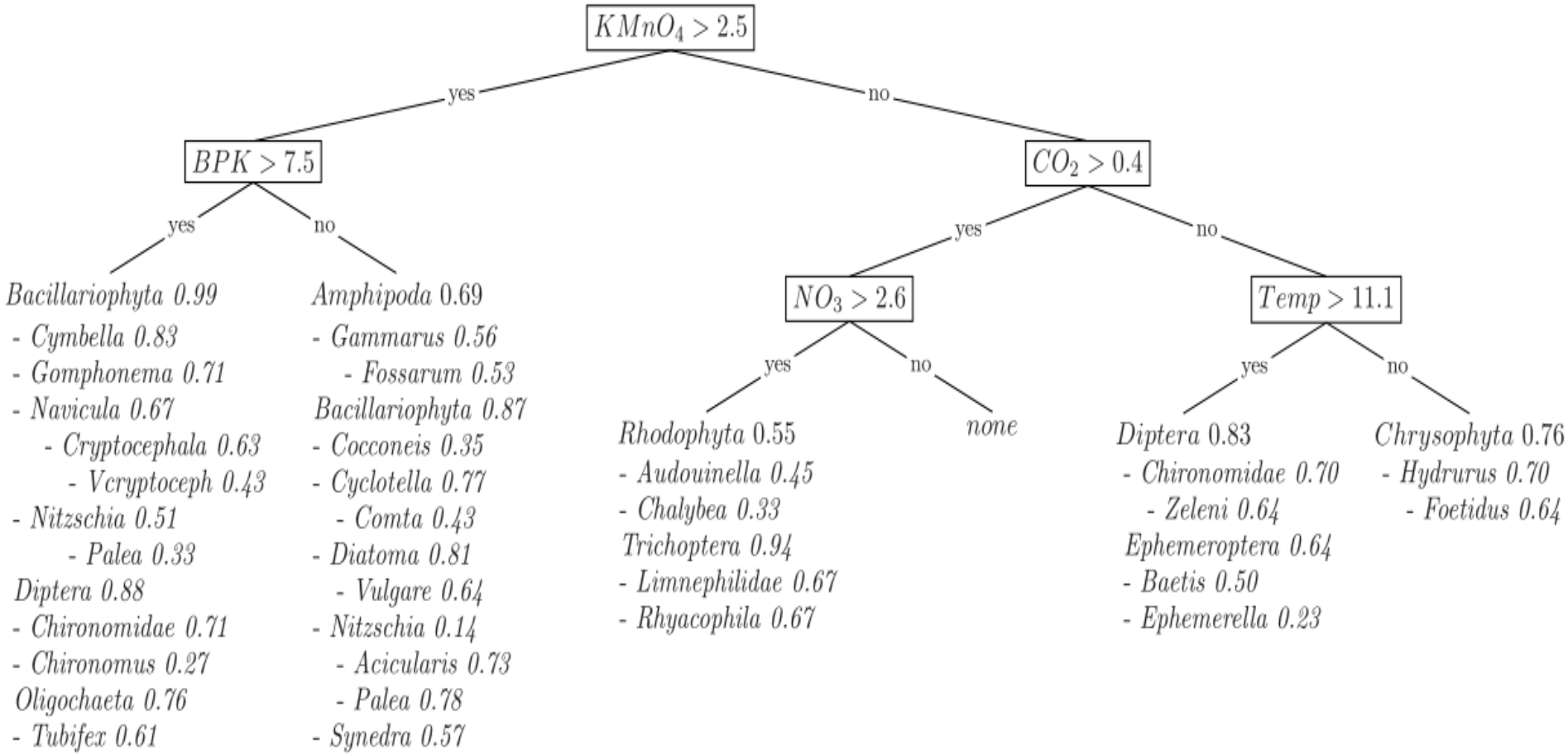


Hierarchical multi-label classification

	Descriptive space				Target space
Example 1	1	TRUE	0.49	0.69	
Example 2	2	FALSE	0.08	0.07	
Example 3	1	FALSE	0.08	0.07	
Example 4	2	TRUE	0.49	0.69	
…	…				…

Hierarchical multi-label classif.

- Predicting community structure (consider taxonomy)



Mining Big and Complex Data: Combining Complexities

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

SSL+SOP: Incomplete Annotations

Semi-supervised multi-target regression

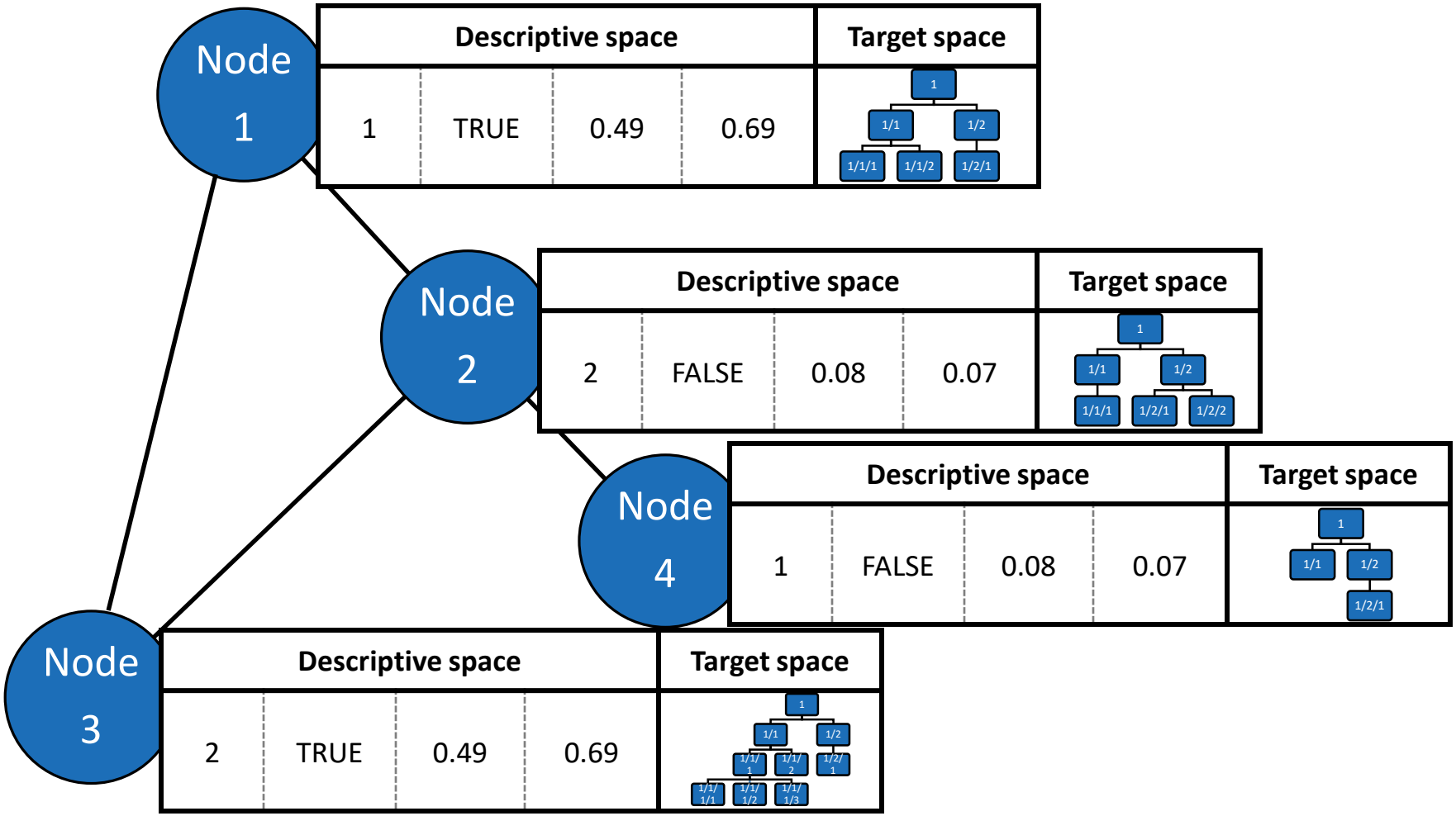
	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	**?**	0.60	3.91
Example 2	2	FALSE	0.08	0.07	0.56	0.99	7.59
Example 3	1	FALSE	0.08	0.07	**?**	**?**	**?**
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	0.11	**?**	**?**
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
…	…				…	…	…

Data streams: (MT) Regression

	Descriptive space				Target space
…	…				…
Example n+5	**1**	**TRUE**	**0.49**	**0.69**	**0.45**
Example n+1	4	FALSE	0.08	0.07	0.12
Example n+2	6	FALSE	0.08	0.07	1.54
Example n+3	8	TRUE	0.00	1.00	3.12
Example n+4	6	TRUE	0.00	0.00	0.05
…	…				…

	Descriptive space				Target space		
…	…				…		
Example n+5	**6**	**TRUE**	**0.49**	**0.69**	**0.58**	**0.00**	**7.91**
Example n+1	4	FALSE	0.08	0.07	0.10	1.69	7.57
Example n+2	6	FALSE	0.08	0.07	0.08	0.77	8.86
Example n+3	8	TRUE	0.00	1.00	0.11	3.51	2.50
Example n+4	6	TRUE	0.00	0.00	0.43	2.10	8.09
…	…				…	…	…

Network +SOP: HMC



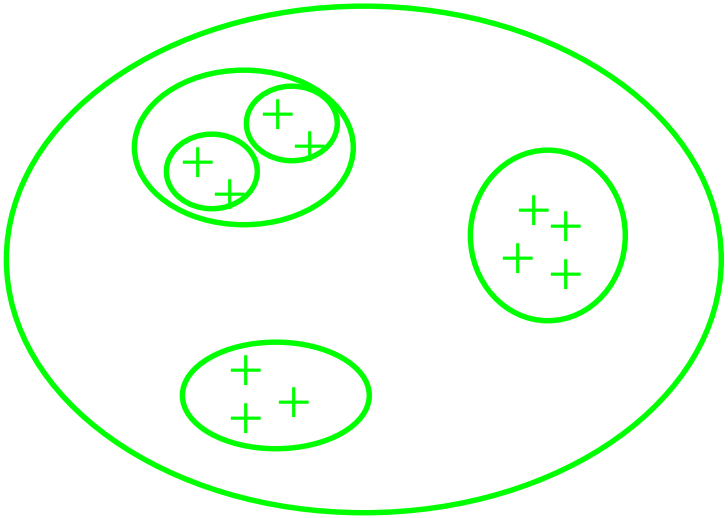
Predictive Clustering for Multi-Target Prediction

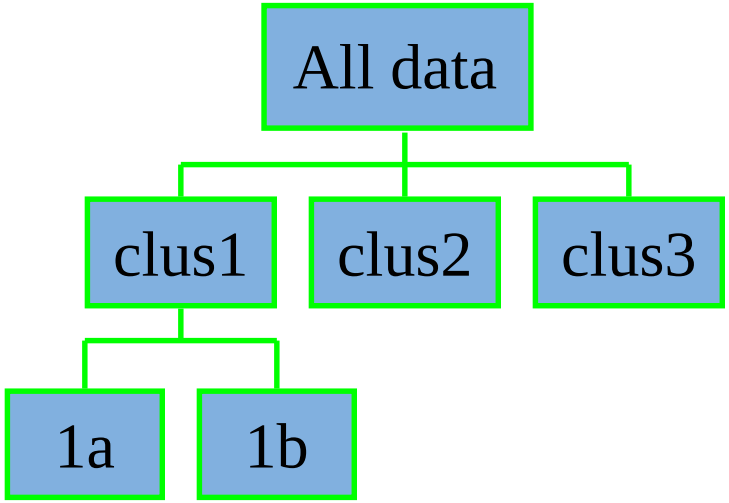
Interreg
ITALIA-SLOVENIJA
UNIONE EUROPEA
EVROPSKA UNIJA
TRAIN
Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj

Clustering

Partition a set of objects into clusters of similar objects

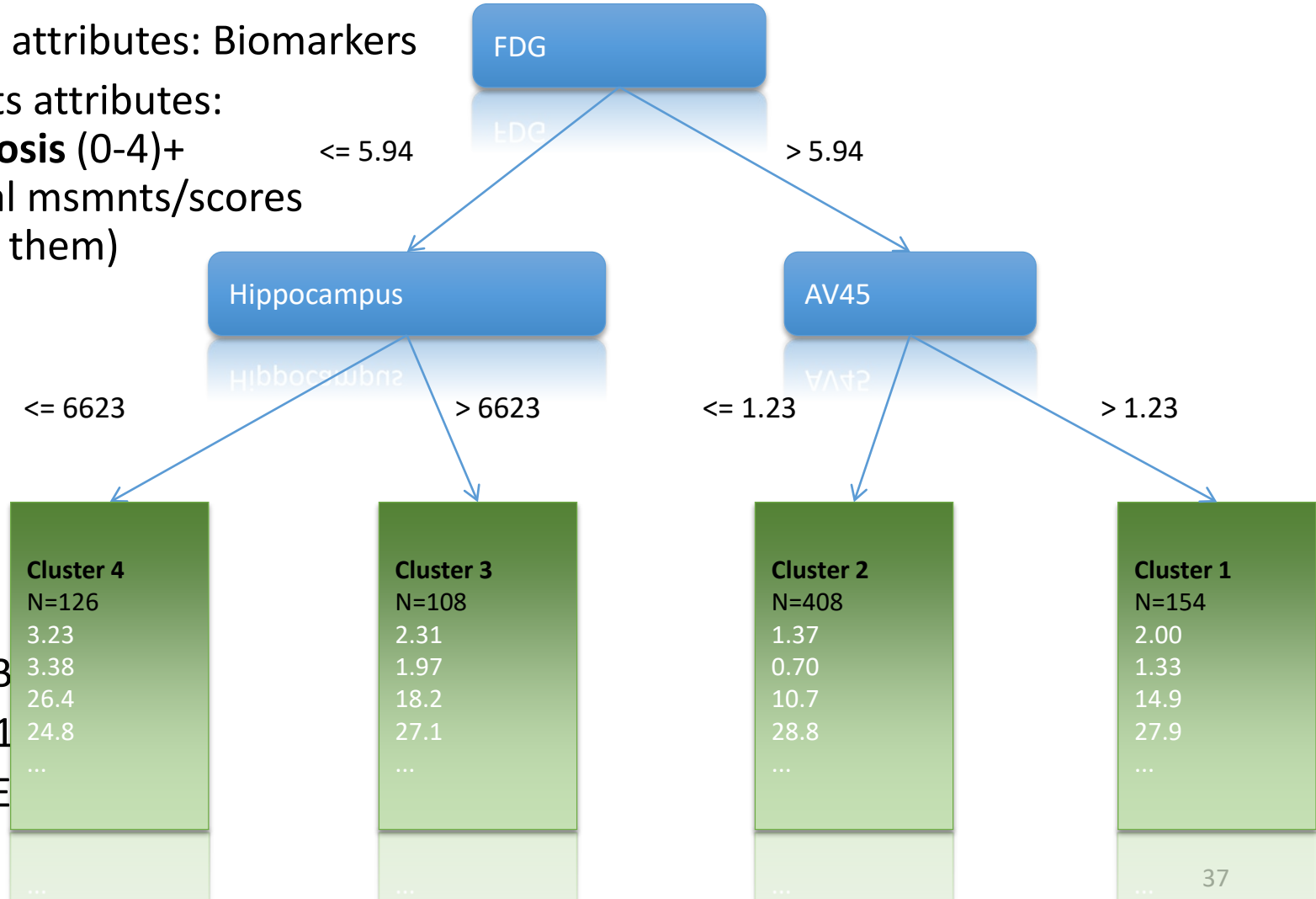
- High similarity of objects within individual clusters, low similarity between objects from different clusters
- Minimize intra-cluster variance (ICV)
- Distance/similarity measure in the example space





Example predictive clustering tree

- Descr. attributes: Biomarkers
- Targets attributes: **diagnosis** (0-4)+ clinical msmnts/scores (23 of them)



- DX
- CDRSB
- ADAS1
- MMSE
- …

Top-down induction of PCTs

To construct a tree T from a training set S:

- If **the examples in S have low variance**, construct a leaf labeled *target(prototype(S))*
- Otherwise:
  - Select the best attribute A with values v1, …, vn, which **reduces the most the variance** (*measured according to a given distance function d*)
  - Partition S into S1, …, Sn according to A
  - Recursively construct subtrees T1 to Tn for S1 to Sn
  - Result: a tree with root A and subtrees T1, …, Tn

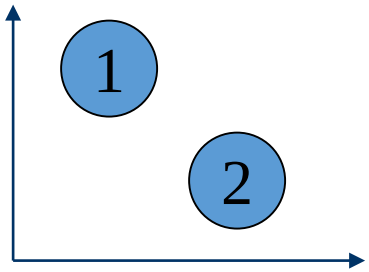
Learning PCTs

- Recursively partition data set into subsets (clusters) with low intra-cluster variance
  - Variance = avg. squared distance to prototype

$$ICV(S) = \sum_{y_j \in S} d(y_j, p(S))^2$$

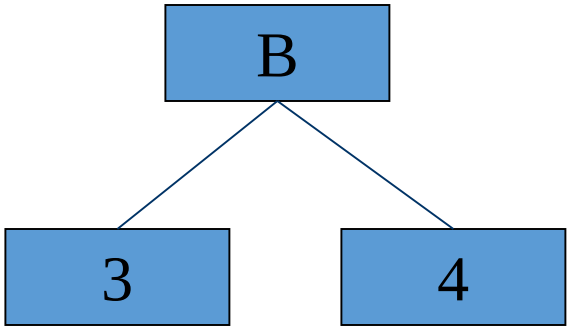
- For the variance, the distance is measured
  - In standard clustering, along all dimensions
  - In prediction, along a single target dimension
  - In predictive clustering, along a structured target, e.g., several target dimensions

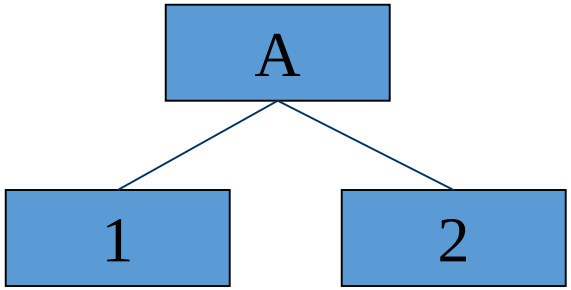
Clustering:

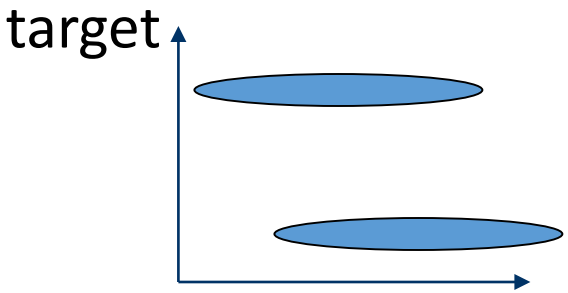


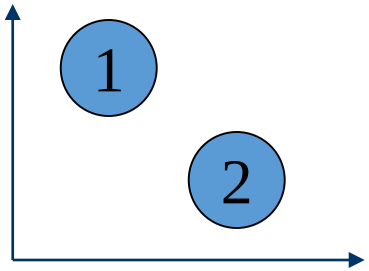
Data divided into clusters 1 and 2 coherent along two dimensions

Prediction:









B divides data into clusters coherent along single *target*

Predictive clustering: A divides data into clusters 1 and 2 coherent along two dimensions

Selecting the best test in a PCT

- Select the test that maximizes variance reduction
- Calculated in line 4

procedure BestTest(E)

1: $(t^*, h^*, \mathcal{P}^*) = (none, 0, \emptyset)$
2: **for each** possible test t **do**
3: $\quad \mathcal{P} =$ partition induced by t on E
4: $\quad h = Var(E) - \sum_{E_i \in \mathcal{P}} \frac{|E_i|}{|E|} Var(E_i)$
5: $\quad$ **if** $(h > h^*) \wedge$ Acceptable$(t, \mathcal{P})$ **then**
6: $\quad\quad (t^*, h^*, \mathcal{P}^*) = (t, h, \mathcal{P})$
7: **return** $(t^*, h^*, \mathcal{P}^*)$

Multi-target regression

- The variance function for MTR
- Is the sum of the variances
- Across all targets

$$Var(E) = \sum_{i=1}^{T} Var(Y_i).$$

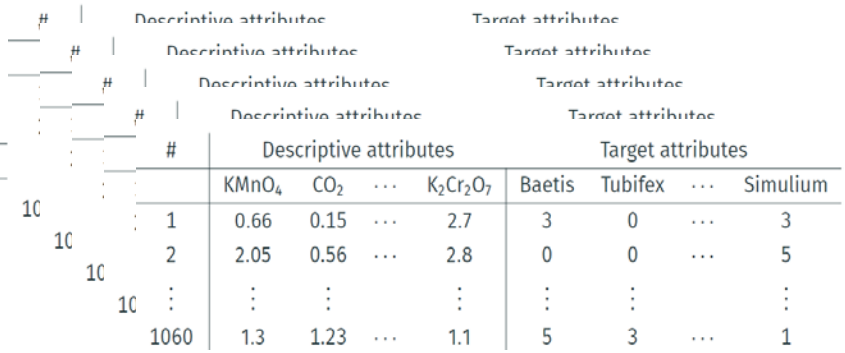
- Normalization is in order
- So that variances are comparable across targets

Ensembles of PCTs

- An ensemble is a set of predictive models, whose predictions are combined [to achieve performance better than that of individual/base predictors]

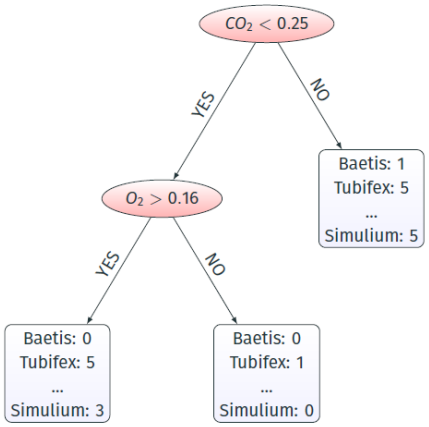
#	Descriptive attributes				Target attributes			
	$KMnO_4$	CO_2	⋯	$K_2Cr_2O_7$	Baetis	Tubifex	⋯	Simulium
1	0.66	0.15	⋯	2.7	3	0	⋯	3
2	2.05	0.56	⋯	2.8	0	0	⋯	5
⋮	⋮	⋮		⋮	⋮	⋮		⋮
1060	1.3	1.23	⋯	1.1	5	3	⋯	1

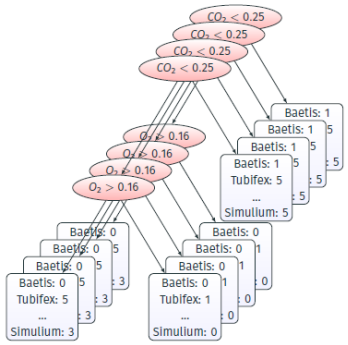
A single decision tree



#	Descriptive attributes				Target attributes			
	$KMnO_4$	CO_2	⋯	$K_2Cr_2O_7$	Baetis	Tubifex	⋯	Simulium
1	0.66	0.15	⋯	2.7	3	0	⋯	3
2	2.05	0.56	⋯	2.8	0	0	⋯	5
⋮	⋮	⋮		⋮	⋮	⋮		⋮
1060	1.3	1.23	⋯	1.1	5	3	⋯	1

An ensemble of decision trees





Relating the Environment and the Biota: From Habitat models to Community composition

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

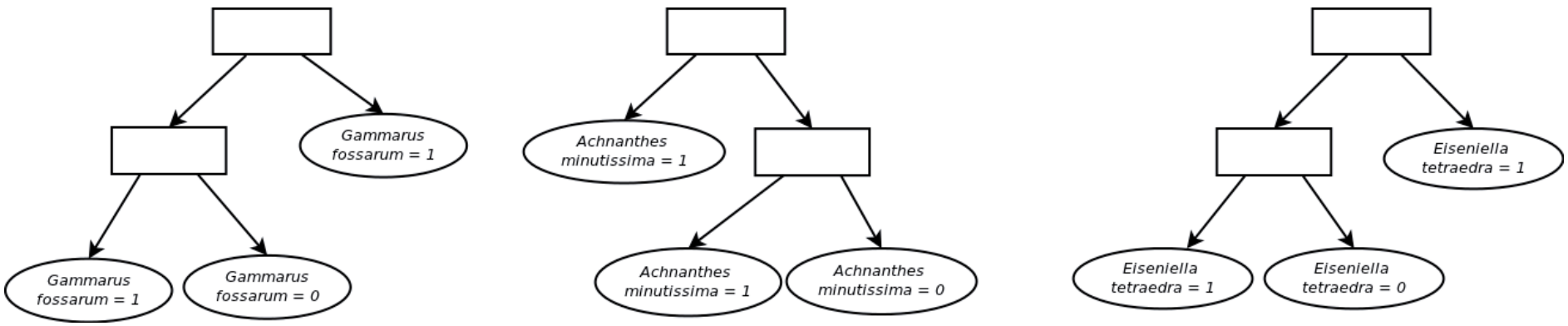
Environment <-> Biota

- Predict the biota (or specific components of it)
- At a given site
- From characteristics of the environment at the site
- E.g. predict river water biota from water properties

Sample ID	Descriptive variables						Target variables													
	Temperature	$K_2Cr_2O_7$	NO_2	Cl	CO_2	...	*Cladophora sp.*	*Gongrosira incrustans*	*Oedogonium sp.*	*Stigeoclonium tenue*	*Melosira varians*	*Nitzschia palea*	*Audouinella chalybea*	*Erpobdella octoculata*	*Gammarus fossarum*	*Baetis rhodani*	*Hydropsyche sp.*	*Rhyacophila sp.*	*Simulim sp.*	*Tubifex sp.*
ID1	0.66	0.00	0.40	1.46	0.84	...	1	0	0	0	0	1	1	0	1	1	1	1	1	1
ID2	2.03	0.16	0.35	1.74	0.71	...	0	1	0	1	1	1	1	0	1	1	1	1	1	0
ID3	3.25	0.70	0.46	0.78	0.71	...	1	1	0	0	1	0	1	0	1	1	1	0	1	1

Habitat modeling

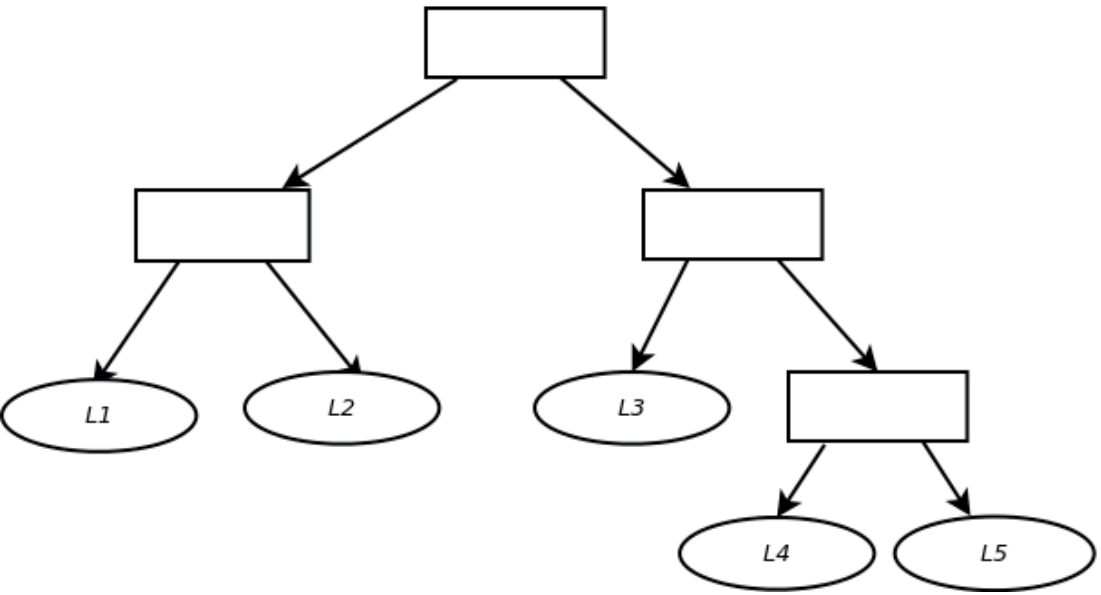
- Model the presence & absence (abundance) of each species separately



- Binary Classification (Regression)

Predicting species composition

- One model for **all the species at once**



L1:
Gammarus fossarum: 0
Achnanthes minutissima: 1
Eiseniella tetraedra: 1

L2:
Gammarus fossarum: 0
Achnanthes minutissima: 1
Eiseniella tetraedra: 0

L3:
Gammarus fossarum: 1
Achnanthes minutissima: 1
Eiseniella tetraedra: 1

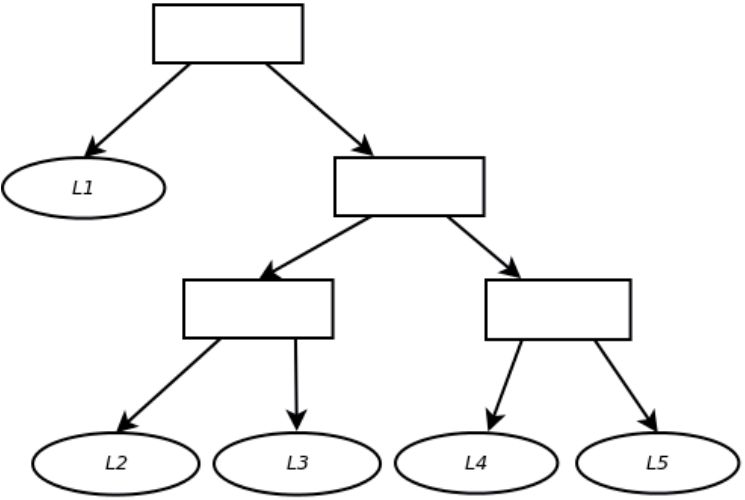
L4:
Gammarus fossarum: 1
Achnanthes minutissima: 1
Eiseniella tetraedra: 0

L5:
Gammarus fossarum: 0
Achnanthes minutissima: 1
Eiseniella tetraedra: 1

- **Multi-target classification/regression**

Predicting community structure

- One model for all of the species at once, additionally using the taxonomical hierarchy



L1:

Amphipoda : 1
Gammarus : 1
Gammarus fossarum : 1
Gammarus lacustris : 0

Bacillariophyta : 1
Achnanthes : 1
Achnanthes minutissima: 1
Eiseniella : 0
Eiseniella tetraedra: 0

L3:

Amphipoda : 1
Gammarus : 1
Gammarus fossarum : 0
Gammarus lacustris : 1

Bacillariophyta : 1
Achnanthes : 1
Achnanthes minutissima: 1
Eiseniella : 0
Eiseniella tetraedra: 0

L5:

Amphipoda : 1
Gammarus : 1
Gammarus fossarum : 1
Gammarus lacustris : 1

Bacillariophyta : 1
Achnanthes : 0
Achnanthes minutissima: 0
Eiseniella : 1
Eiseniella tetraedra: 1

L2:

Amphipoda : 1
Gammarus : 1
Gammarus fossarum : 1
Gammarus lacustris : 1

Bacillariophyta : 0
Achnanthes : 0
Achnanthes minutissima: 0
Eiseniella : 0
Eiseniella tetraedra: 0

L4:

Amphipoda : 1
Gammarus : 1
Gammarus fossarum : 1
Gammarus lacustris : 0

Bacillariophyta : 1
Achnanthes : 1
Achnanthes minutissima: 1
Eiseniella : 1
Eiseniella tetraedra: 1

- Hierarchical multi-label classification

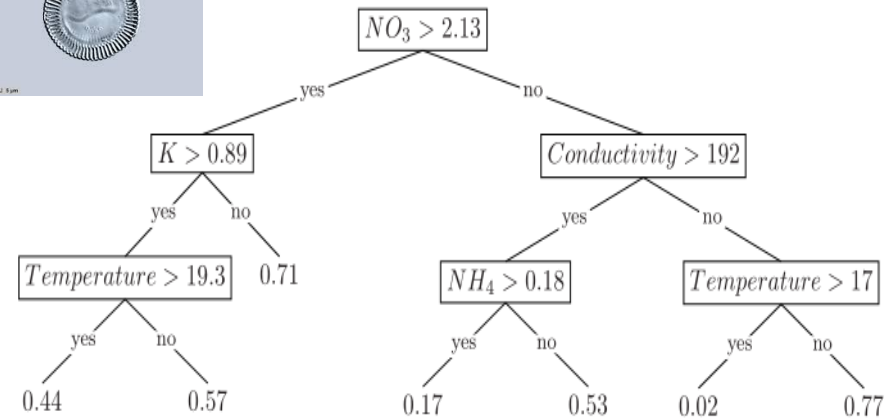
Slovenian rivers

- 1.060 samples
- 16 physical and chemical props. of water, 491 species
- data collected in 1990-1995

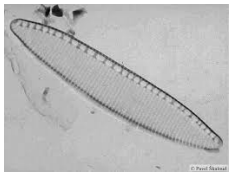
ephemeroptera
- *ephemeroptera_acantrella*
  - *ephemeroptera_acantrella_sinaica*
- *ephemeroptera_baetidae*
- *ephemeroptera_baetis*
  - *ephemeroptera_baetis_alpinus*
  - *ephemeroptera_baetis_buceratus*
  - *ephemeroptera_baetis_fuscatus*
  - *ephemeroptera_baetis_muticus*
  - ***ephemeroptera_baetis_rhodani***
  - *ephemeroptera_baetis_scambus*
  - *ephemeroptera_baetis_vernus*
- *ephemeroptera_ecdyonurus*
  - *ephemeroptera_ecdyonurus_forcipula*
  - *ephemeroptera_ecdyonurus_helveticus*
  - *ephemeroptera_ecdyonurus_insignis*
  - *ephemeroptera_ecdyonurus_torrentis*
  - *ephemeroptera_ecdyonurus_venosus*
- *ephemeroptera_electrogena*
  - *ephemeroptera_electrogena_lateralis*
  - *ephemeroptera_electrogena_quadrilineata*

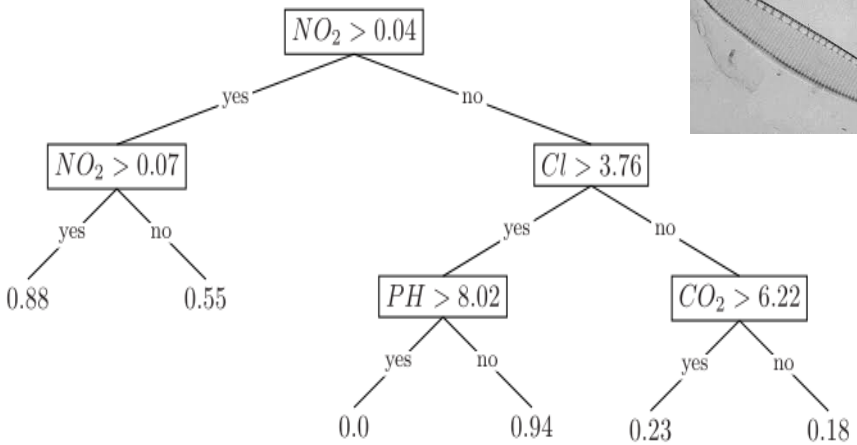
plecoptera
- *plecoptera_amphinemura*
  - *plecoptera_amphinemura_triangularis*
- *plecoptera_brachyptera*
  - ***plecoptera_brachyptera_risi***
  - *plecoptera_brachyptera_seticornis*

Slovenian rivers: Habitat models

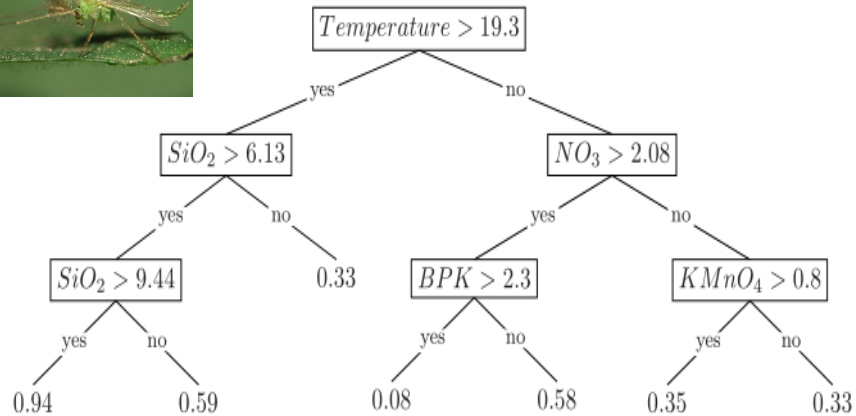


Bacillariophyta Cyclotella Comta

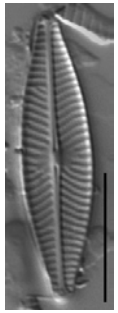


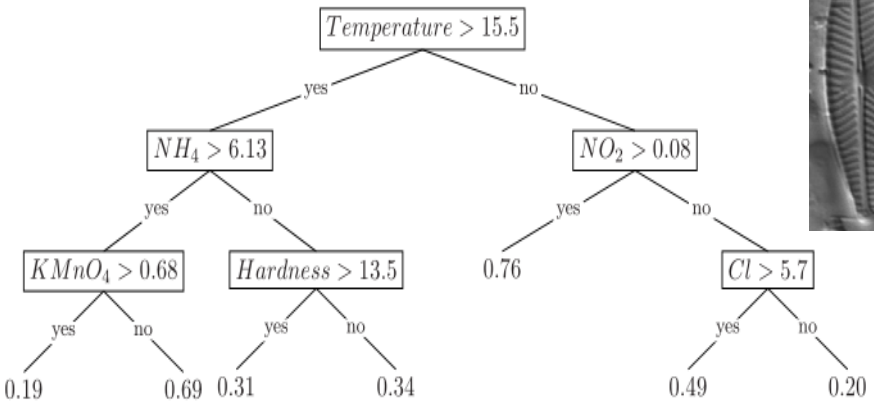


Bacillariophyta Nitzschia Palea



Diptera Chironomidae Zeleni

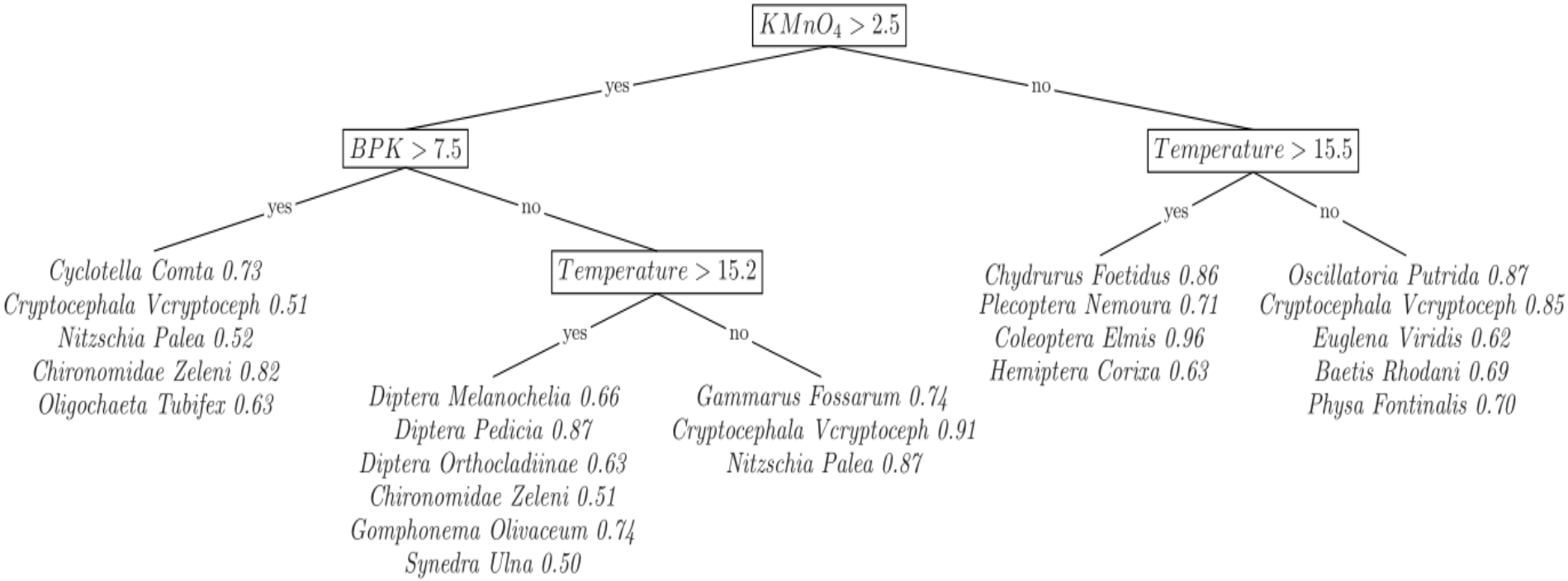




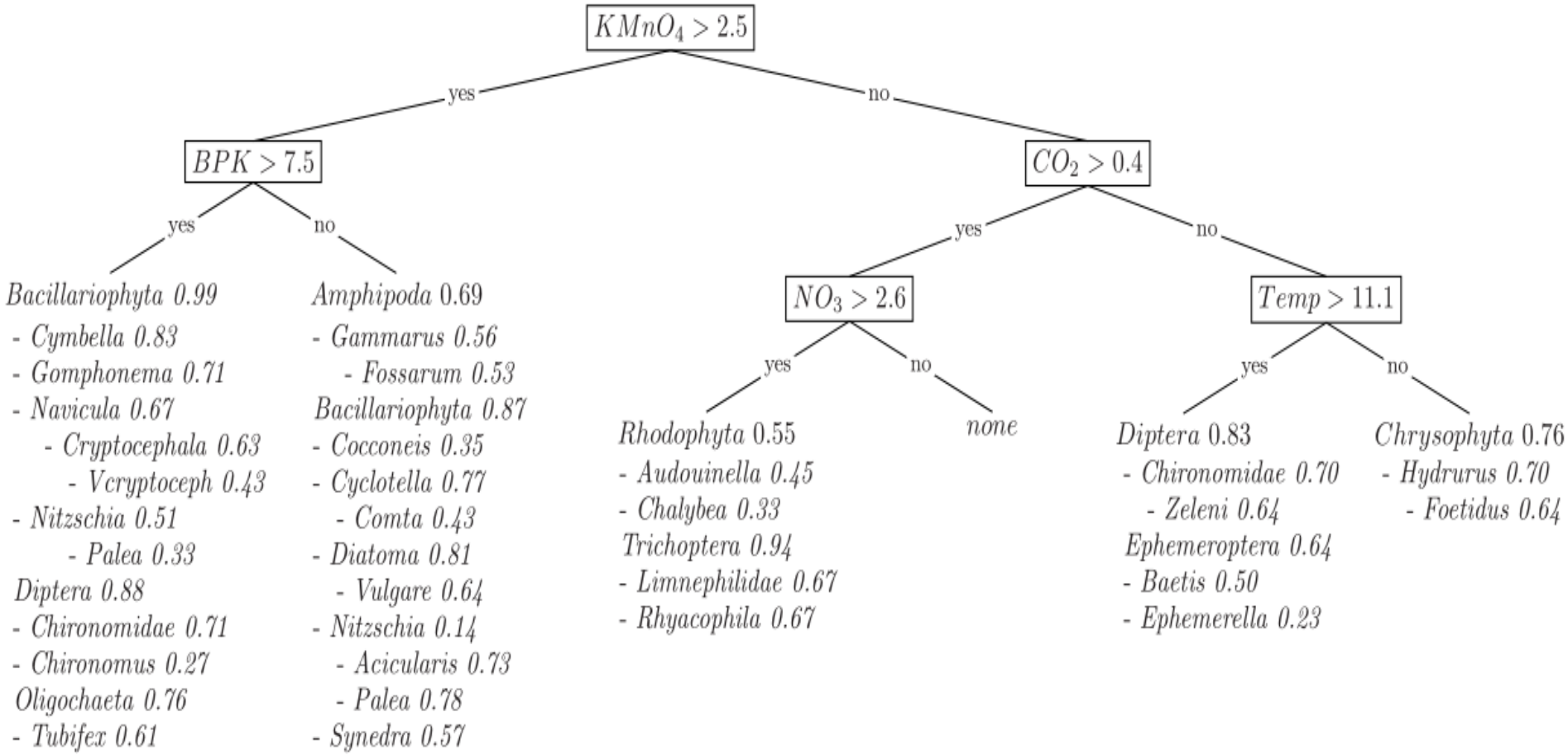
Bacillariophyta Navicula Cryptocephala Vcryptoceph

Slovenian rivers: Species comp.

- MLC: Multi-label classification tree



Slovenian rivers: Community struc.



Danish farms: Soil Microarthropods

- 1.944 soil samples
- 137 attributes/agricultural events and soil biological parameters
- 35 collembolan species
- data collected 1989-1993

Isotominae
- ***Isotominae_Isotoma***
  - *Isotominae_Isotoma_anglicana*
  - *Isotominae_Isotoma_notabilis*
  - *Isotominae_Isotoma_tigrina*

Lepidocyrtinae
- *Lepidocyrtinae_Lepidocyrtus*
  - *Lepidocyrtinae_Lepidocyrtus_cyaneus*
  - *Lepidocyrtinae_Lepidocyrtus_lanuginosus*
- *Lepidocyrtinae_Pseudosinella*
  - *Lepidocyrtinae_Pseudosinella_alba*
  - *Lepidocyrtinae_Pseudosinella_sexoculata*

Orchesellinae
- *Orchesellinae_Heteromurus*
  - *Orchesellinae_Heteromurus_nitidus*
- *Orchesellinae_Orchesella*
  - *Orchesellinae_Orchesella_cincta*
  - *Orchesellinae_Orchesella_villosa*

Sminthuridae
- *Sminthuridae_Smint*
- *Sminthuridae_Sminthurinus*
  - *Sminthuridae_Sminthurinus_aureus*
  - *Sminthuridae_Sminthurinus_elegans*
- *Sminthuridae_Sminthurus*
  - *Sminthuridae_Sminthurus_viridis*

Tomoceridae
- *Tomoceridae_Tomocerus*
  - *Tomoceridae_Tomocerus_flavescens*
  - *Tomoceridae_Tomocerus_minor*

Tullbergiidae
- *Tullbergiidae_Mesaphorura*

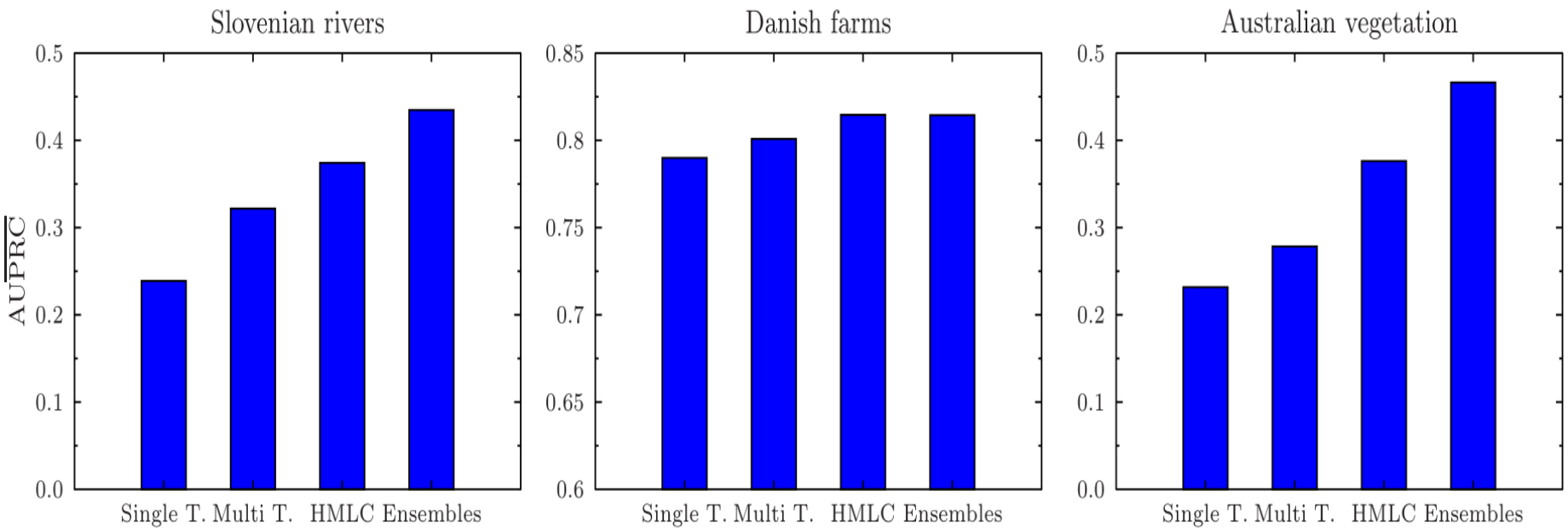
Victoria, Australia Vegetation

- 27.482 sites
- 81 env. attributes
- 3.173 species

DivisionConifer
- *DivisionConifer_callitris*
  - *DivisionConifer_callitris_endlicheri*
  - *DivisionConifer_callitris_glaucophylla*
  - *DivisionConifer_callitris_gracilis*
    - *DivisionConifer_callitris_gracilis_ssp~murrayensis*
  - *DivisionConifer_callitris_rhomboidea*
  - *DivisionConifer_callitris_verrucosa*

DivisionMonocotyledon
- *DivisionMonocotyledon_leucopogon*
  - *DivisionMonocotyledon_leucopogon_attenuatus*
  - *DivisionMonocotyledon_leucopogon_australis*
  - *DivisionMonocotyledon_leucopogon_clelandii*
  - *DivisionMonocotyledon_leucopogon_juniperinus*
  - *DivisionMonocotyledon_leucopogon_lanceolatus*
    - *DivisionMonocotyledon_leucopogon_lanceolatus_var~lanceolatus*
  - *DivisionMonocotyledon_leucopogon_maccraei*
  - ***DivisionMonocotyledon_leucopogon_microphyllus***
    - *DivisionMonocotyledon_leucopogon_microphyllus_var~pilibundus*
  - *DivisionMonocotyledon_leucopogon_montanus*
  - *DivisionMonocotyledon_leucopogon_neurophyllus*
  - *DivisionMonocotyledon_leucopogon_parviflorus*
  - *DivisionMonocotyledon_leucopogon_virgatus*
    - *DivisionMonocotyledon_leucopogon_virgatus_var~brevifolius*
    - *DivisionMonocotyledon_leucopogon_virgatus_var~virgatus*
  - *DivisionMonocotyledon_leucopogon_woodsii*
- *DivisionMonocotyledon_epacris*
  - *DivisionMonocotyledon_epacris_breviflora*
  - *DivisionMonocotyledon_epacris_celata*
  - *DivisionMonocotyledon_epacris_glacialis*
  - *DivisionMonocotyledon_epacris_gunnii*
  - ***DivisionMonocotyledon_epacris_impressa***
    - *DivisionMonocotyledon_epacris_impressa_var~grandiflora*
    - *DivisionMonocotyledon_epacris_impressa_var~impressa*

Community structure: Overall results



Dataset	Method	AUPRC	O_S	Learning time	Complexity
Slovenian rivers	Single-label	0.239	0.692	23.3	15,336
	HSC	0.309	0.591	10.2	25,035
	Multi-label	0.322	0.007	9.4	1
	HMC	**0.374**	0.132	0.6	37
Danish farms	Single-label	0.790	0.099	3.7	2605
	HSC	0.808	0.083	1.3	2873
	Multi-label	0.801	0.112	0.7	265
	HMC	**0.815**	0.065	0.4	259
Australian vegetation	Single-label	0.232	0.715	14,888.2	482,745
	HSC	0.306	0.591	76,023.2	648,970
	Multi-label	0.278	0.684	4639.5	23,699
	HMC	**0.376**	0.180	313.5	1279

Predicting Gene Functions

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

Predicting gene functions

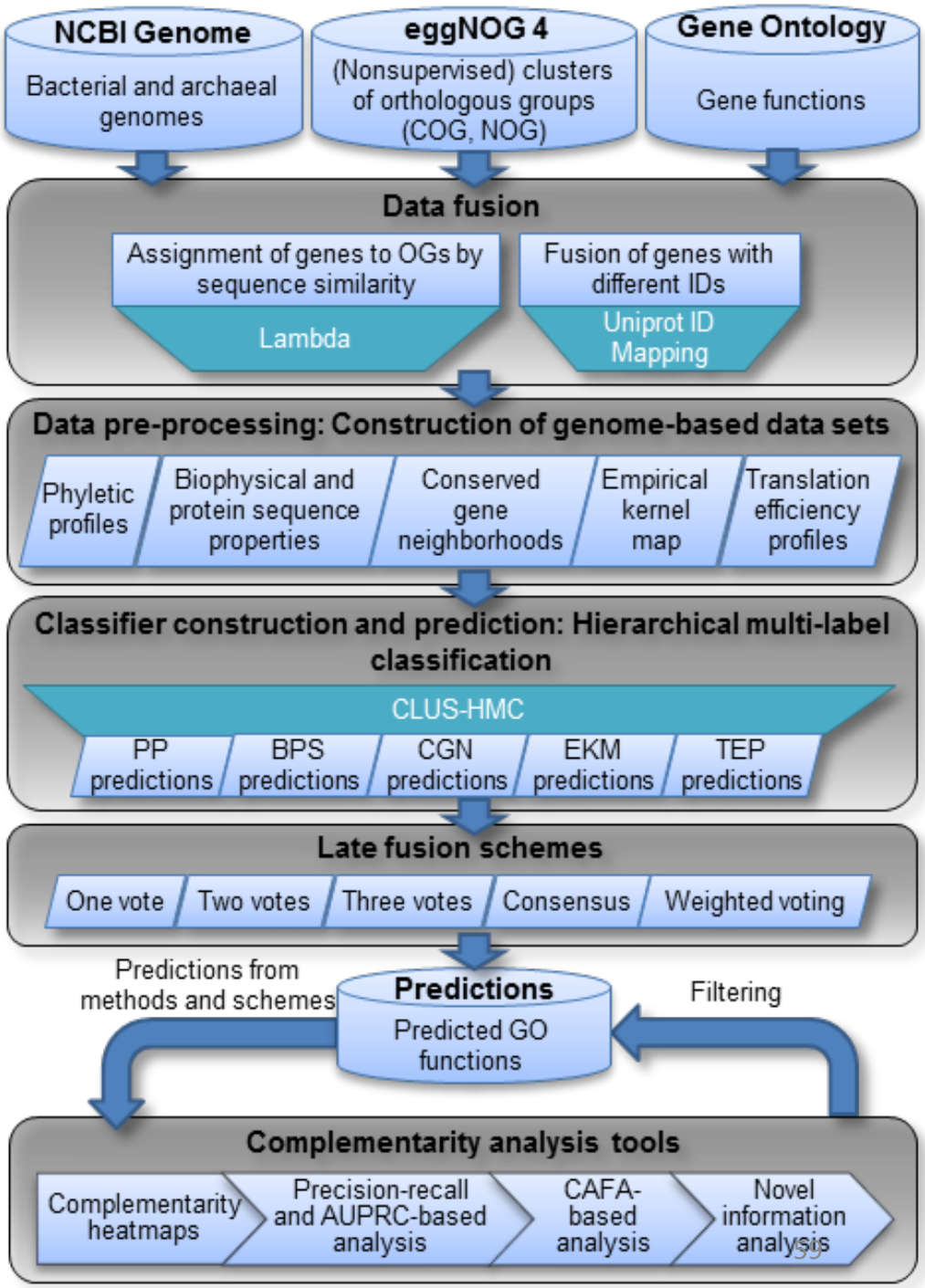
- In model organisms
  - Arabidopsis thaliana
  - Saccharomyces cerevisiae
  - Mus musculus
- In bacterial genomes
  - From different sets of features
  - Including phyletic profiles
- Using metagenome data

Predicting Gene Functions in Bacterial Genomes (RBI+JSI)

- The number of sequenced genomes and metagenomes continually increases and with them also the number of **genes with unknown biological functions**.
- Even in a well-known pathogen such as *Mycobacterium tuberculosis* **26% of genes are of unknown function**.
- Gene function prediction (GFP) is typically based on transfer of function by homology using sequence similarity.
- Recently, GFP methods based on different **feature sets** and **machine learning** algorithms received much attention.
- We examined **complementarity** between GFP methods **on a large scale** including 2071 microorganisms, 5 million genes and 4000 gene functions.

GFP Pipeline

- 5000 bacterial genomes
- 5 different feature sets
- Predictive models learned from each FS: Tree ensembles for HMLC
- Predictions combined
  - with late fusion
  - different voting schemes



Different Features Sets for GFP

Instances : 21,626 eggNOG 4 OGs

(a)	g_1	g_2	g_3	g_4	GO
OG_1	1	0	0	1	
OG_2	1	1	0	1	?
OG_3	0	1	0	1	
OG_4	1	0	1	1	?

Phyletic Profiles (PP) — Features: 2,071 microbial genomes — Class: 4,145 Gene Ontology (GO) functions

(b)	A Entropy	Cysteine Spaced Motif	hp2: AAAAC	GO
OG_1	-0.27	-0.21	0.28	
OG_2	-0.12	-0.19	0.09	?
OG_3	-0.15	-0.18	0.08	
OG_4	-0.77	-0.24	-1.11	?

Features: 1,170 biophysical and sequence derived attributes — Biophysical and Protein Sequence Properties (BPS)

(c)	g_1	g_2	OG_1	OG_2	GO
OG_1	0.71	0.53	1	0.71	
OG_2	0.48	0.25	0.71	1	?
OG_3	1.22	0.56	-0.27	0.44	
OG_4	0.66	0.56	0.34	-0.59	?

Features: 2,071 microbial genomes + 5,891 OGs that appear in at least 100 genomes — Translation Efficiency Profiles (TEP)

Instances : 21,626 eggNOG 4 OGs

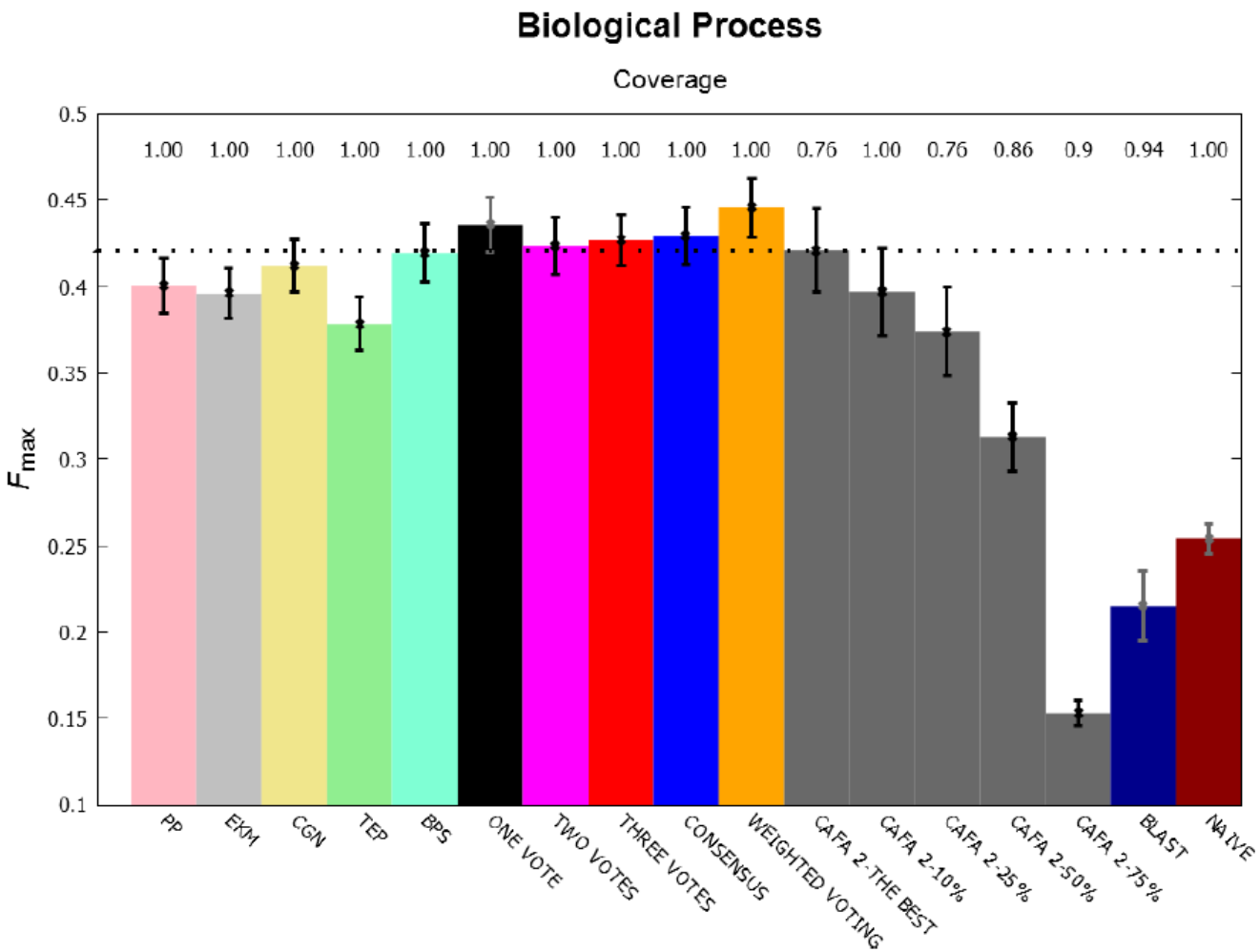
(d)	OG_1	OG_2	OG_3	OG_4	GO
OG_1	-33.22	19.96	13.88	11.21	
OG_2	19.96	-33.22	23.81	23.81	?
OG_3	13.88	23.81	-33.22	20.38	
OG_4	11.21	23.81	20.38	-33.22	?

Features: 5,891 OGs that appear in at least 100 genomes — Conserved Gene Neighborhoods (CGN)

(e)	OG_1	OG_2	OG_3	OG_4	GO
OG_1	0	0.24	6.64	6.64	
OG_2	0.24	0	-9.87	1.32	?
OG_3	6.64	-9.87	0	6.64	
OG_4	6.64	1.32	6.64	0	?

Features: 8,447 OGs from 6 model organism genomes — Empirical Kernel Map (EKM)

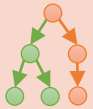
Gene Function Prediction: Predictive Performance



Metagenome Phyletic Profiles

Metagenome Phyletic Profiles (MPP)

Features: metagenomes

	m_1	m_2	m_3	m_4	GO
OG_1	0	10E-5	0	0	
OG_2	10E-6	0	10E-7	10E-9	?
OG_3	0.008	0.02	0	0.01	
OG_4	0	0	0.003	0	?

Feature values: sum of OG member genes abundances in metagenomes

Phyletic Profiles (PP)

Features: microbial genomes

	g_1	g_2	g_3	g_4	GO
OG_1	1	0	0	1	
OG_2	1	1	0	1	?
OG_3	0	1	0	1	
OG_4	1	0	1	1	?

Feature values: presence/absence of genes in genomes

MPP can predict hundreds of gene functions that would not be predicted using only PP

Multi-Target Prediction for Virtual Compound Screening

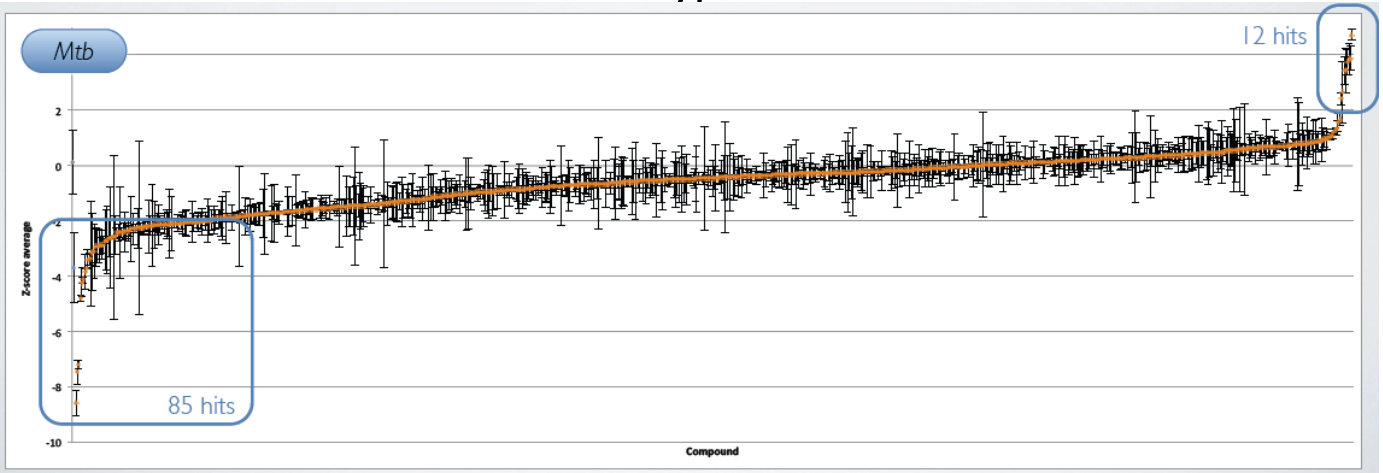
Interreg
ITALIA-SLOVENIJA
UNIONE EUROPEA
EVROPSKA UNIJA
TRAIN
Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj

Virtual compound screening

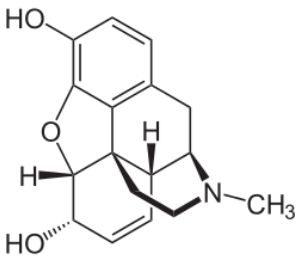
- Descriptive variables refer to compound structure
  - Functional groups
  - Fingerprints
  - Bulk properties
- May also describe the compound in terms of the proteins it targets (e.g. from PubChem)
  - Their functional annotations
  - Pathways they are involved in
  - Proteins that the targets interact with (and/or their functional annotations, pathways they are involved in)
- Target variables describe compound activity and toxicity

Host-targeted Drugs for MTB (Tuberculosis) and STM (Salmonella)

- Library of compounds
  - LOPAC library - Library Of Pharmacologically Active Compounds
    - 1260 compounds
  - Well-characterized compounds, many already applied in clinical practice for a range of conditions
- Flow cytometry (FACS) - measured reduction in bacterial load
  - MelJuSo cells infected with Mycobacterium tuberculosis at MOI 10 – Mtb
  - HeLa cells infected with Salmonella typhimurium at MOI 10 - Stm



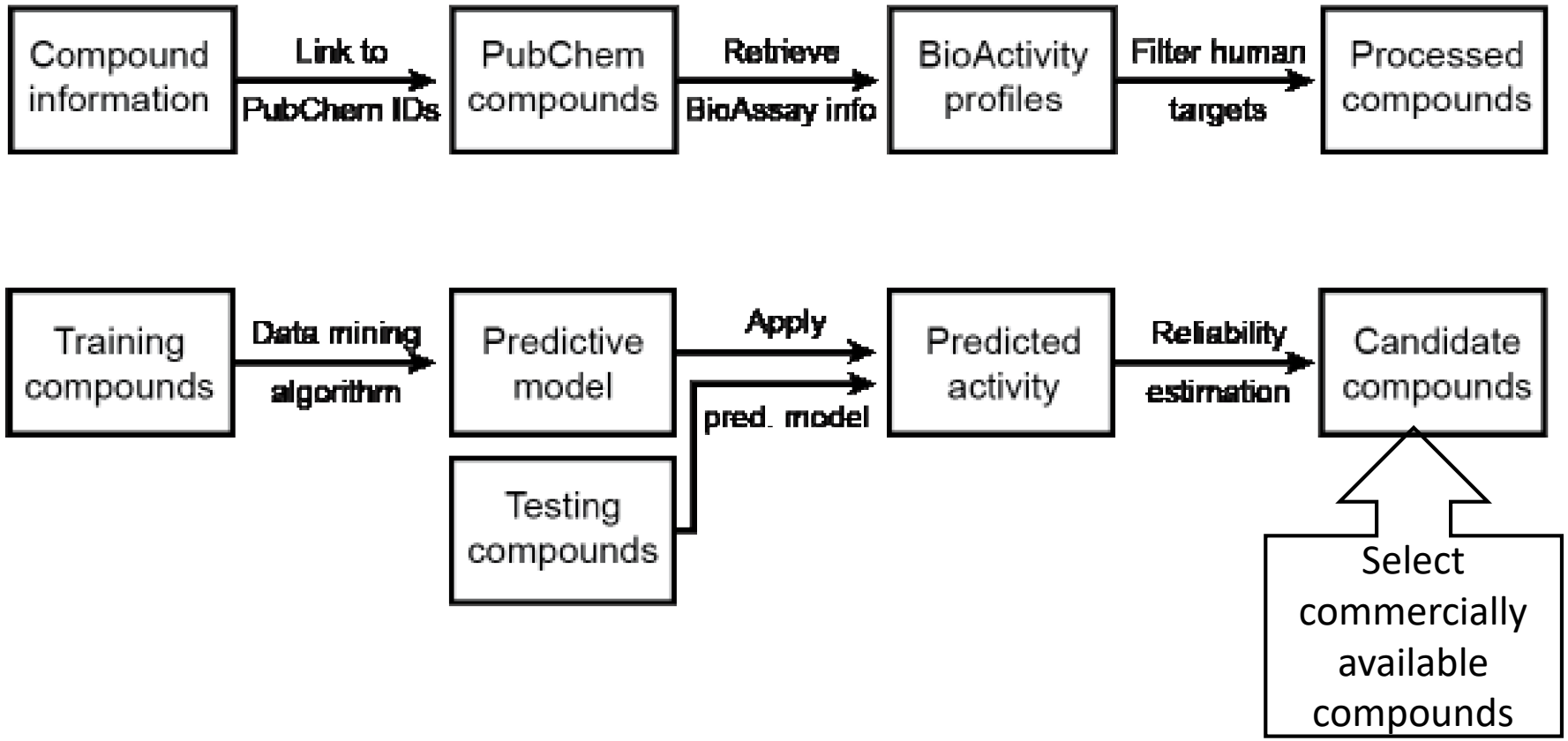
MTB&STM: Host-targeted Drugs



Morphine

- Given SDF files, find PubChemID
- PubChem repository
  - Retrieve the proteins that were found to be active in bio-assays with human cells
- Dataset
  - 964 compounds were found active on human protein targets
  - 711 distinct protein targets were identified
- Each compound is described with
  - the respective protein targets
  - functional annotations of the respective protein targets
  - functional annotations of both the respective protein targets and the proteins they interact with

MTB&STM: Host-targeted Drugs
The Data Analysis Workflow



MTB&STM: Host-targeted Drugs Results

- Greatly increased proportions of hit compounds
  - 5 out of 9 (55.6%) for Mtb and
  - LOPAC primary screen (90 out of 1260 (7.1%) for *Mtb*
- The *in silico* predictive model successfully identified active compounds *de novo*

Abbr.	*Compound name*	*Alternative name(s)*	*Primary sceen z-score*	*Rescreen z-score*	*Activity*
Mycobacterium tuberculosis					
SU	SU 6656	2,3-Dihydro-N,N-dimethyl-2-oxo-3-[(4,5,6,7-tetrahydro-1H-indol-2-yl)methylene]-1H-indole-5-sulfonamide	**-5.79**	**-10.51**	Src family kinase inhibitor
Q	Quinacrine dihydrochloride		**-5.25**	**-9.90**	MAO inhibitor
SB	SB 216763	3-(2,4-Dichlorophenyl)-4-(1-methyl-1H-indol-3-yl)-1H-pyrrole-2,5-dione	**-6.02**	**-8.29**	GSK-3 kinase inhibitor
G	GW5074	3-(3, 5-Dibromo-4-hydroxybenzylidine-5-iodo-1,3-dihydro-indol-2-one)	**-4.86**	**-6.98**	Raf1 kinase inhibitor
T494	Tyrphostin AG 494	N-Phenyl-3,4-dihydroxybenzylidenecyanoacetamide	**-3.83**	**-6.93**	EGFR kinase inhibitor
L	3',4'-Dichlorobenzamil hydrochloride	L-594,881	**-3.87**	-5.13	$Na^{+}/Ca2^{+}$ exchanger inhibitor
H	Haloperidol		**-3.77**	-2.96	D2/D1 dopamine receptor antagonist

Analyzing data from High-contents Screens

- Compounds described by fingerprints
- Generated by open-source chemoinformatics SW library RDkit
- The FCFP2 fingerprits were used (1024 features)
- Also considered profiles of targeted proteins
- These are the attributes

- Assays photographed under the microscope
- Features extracted from images
- These are then the targets

HTS: Modulating fibroblast to myofibroblast transition

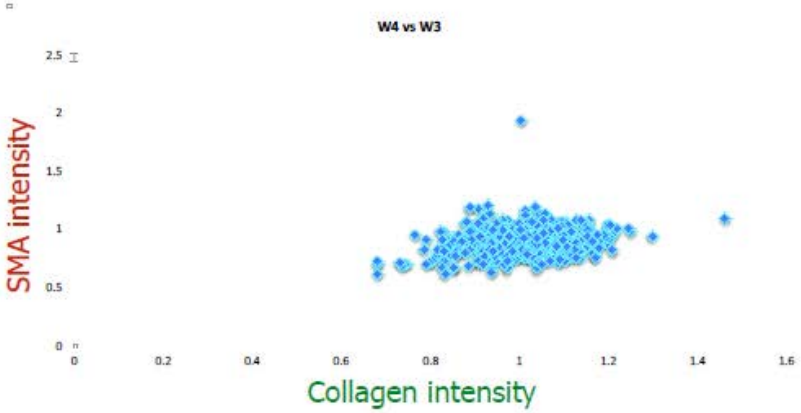
cardiac fibroblasts from a-SMA-RFP/
Coll α1(I)-EGFP mice

0h

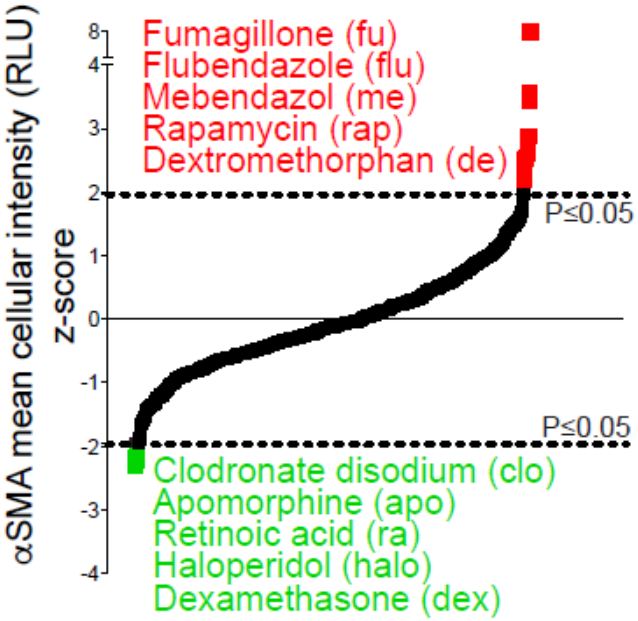
FDA-approved drugs (640 compounds)

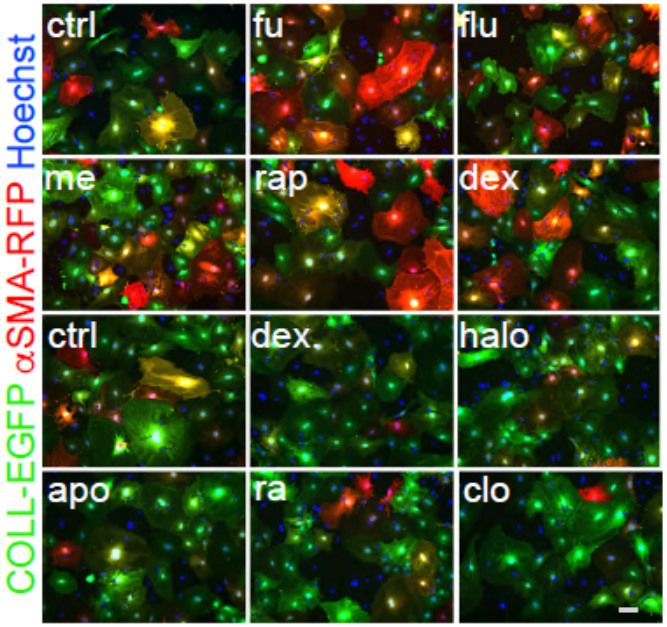
72h

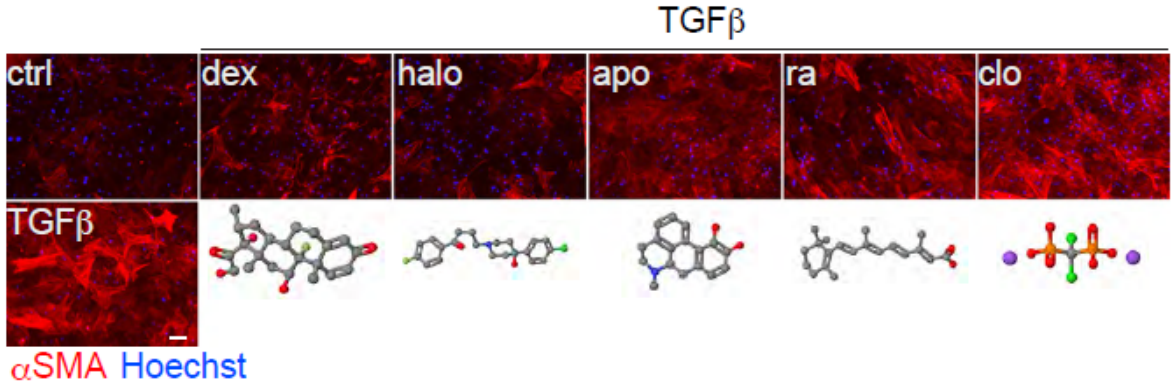
cell fixation, image acquisition and elaboration

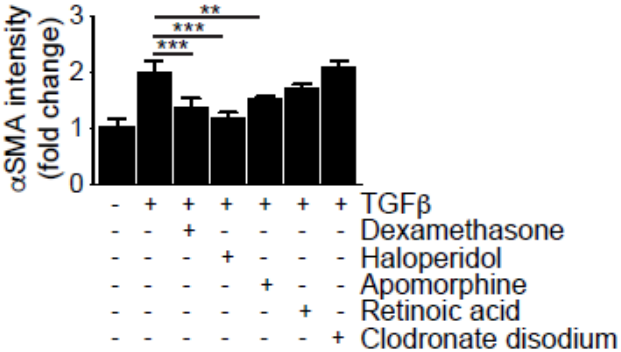


Hits in the HTS screen



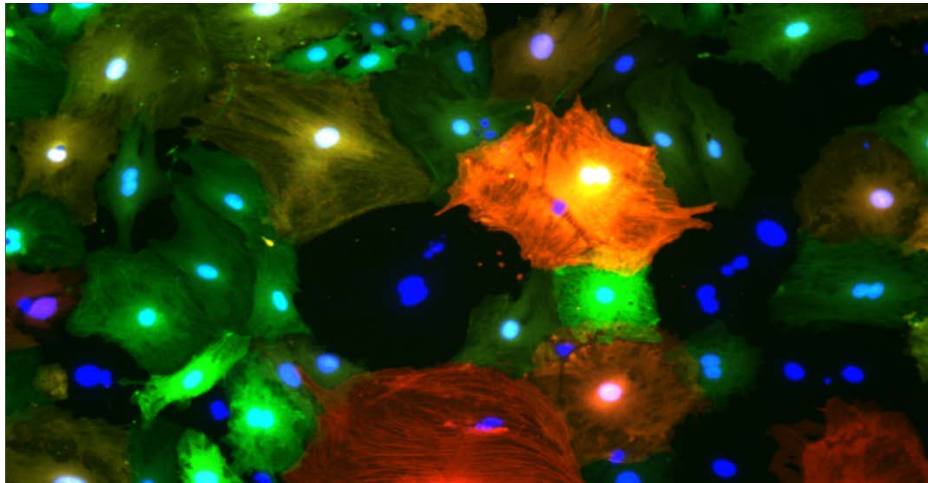






Reducing fibrosis in myocardial infarction

- High content screen using a library of 640 FDA approved drugs (ENZO)
- Identify drugs to reduce fibrosis in myocardial infarction
- Screen used murine cardiac fibroblasts which differentiate into myofibroblasts in culture, expressing increased alpha SMA-RFP and collagen-alpha1-EGFP
- Targets: Intensity of
  - alphaSMA
  - Collagen
- Attributes
  - Fingerprints



Testing the predictions

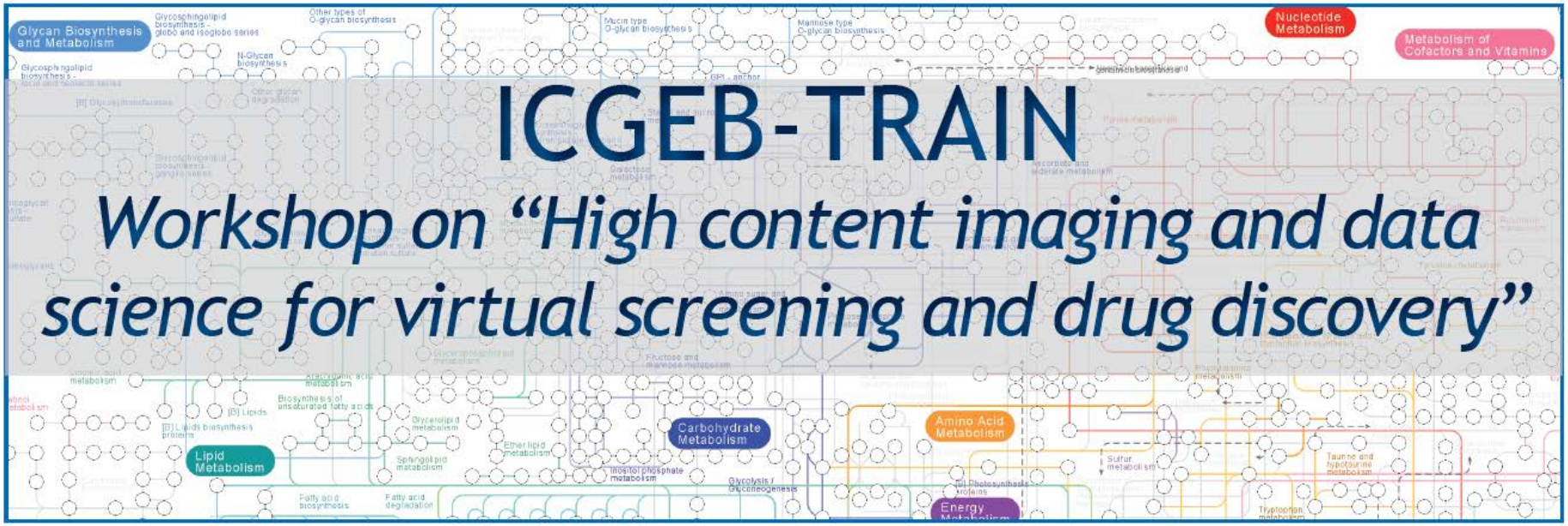
- Some domain-specific knowledge / constraints applied: Predicted compounds filtered for FDA approved drugs that are not corticosteroids
- SMILE strings used in Chemmine to identify substances with structural similarity to non commercial compounds with high predicted values
- Three related compounds identified which are described in literature to have an anti-fibrotic effect
- Four related compounds identified which were not previously described to have an anti-fibrotic effect
- Tested in the wet-lab and one works really well ☺

Spring school in Bled in May

Interreg
ITALIA-SLOVENIJA
TRAIN
UNIONE EUROPEA
EVROPSKA UNIJA

Progetto standard co-finanziato dal Fondo europeo di sviluppo regionale
Standardni projekt sofinancira Evropski sklad za regionalni razvoj

Institut "Jožef Stefan" Ljubljana, Slovenija



13-17 May 2019 | Bled, SLOVENIA

Conclusions

- Exciting new technology for mining big and complex data
- Can handle different aspects of complexity
  - Different types of structured outputs
  - Big data and data streams
  - Partially annotated data, network data
- Efficient, works fast!

  [What's the environmental footprint of deep learning?]
- Can produce accurate models
- Can produce understandable models