Towards the discovery of exceptional local models: descriptive rules relating molecules and their odors

Guillaume Bosc[1], Mehdi Kaytoue[1], Marc Plantevit[1],
Fabien De Marchi[1], Moustafa Bensafi[2], Jean-François Boulicaut[1]

[1]LIRIS, Université Lyon 1, INSA-Lyon, France
[2]Centre de Recherche en Neurosciences de Lyon (CRNL), Université Lyon 1 Lyon, France

Definition and challenges of olfaction

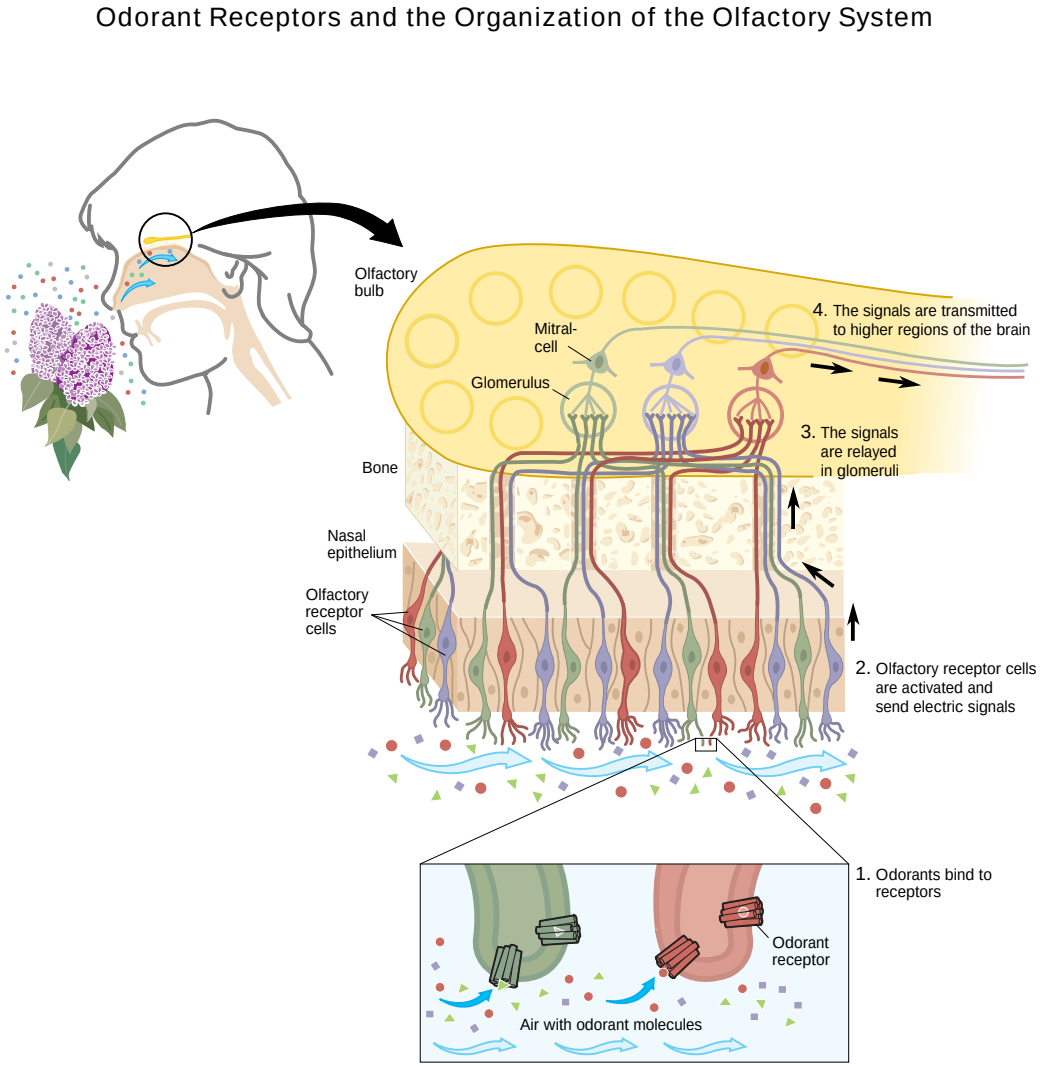
Olfaction, ...

- Ability to perceive odors
- Complex phenomenon from molecule to perception

C. Sezille et M. Bensafi

De la molécule au percept.
In *Biofutur*, 2013.

Definition and challenges of olfaction



http://www.nobelprize.org/nobel_prizes/medicine/laureates/2004/press.html

Definition and challenges of olfaction

Olfaction, ...

- Ability to perceive odors
- Complex phenomenon from molecule to perception

State of the art/challenges, ...

- Established links between physicochemical properties and olfactory qualities of molecules
- Difficulties to formulate/propose rules

C. Sezille et M. Bensafi
De la molécule au percept.
In *Biofutur*, 2013.

Definition and challenges of olfaction

Olfaction, ...

- Ability to perceive odors
- Complex phenomenon from molecule to perception

State of the art/challenges, ...

- Established links between physicochemical properties and olfactory qualities of molecules
- Difficulties to formulate/propose rules

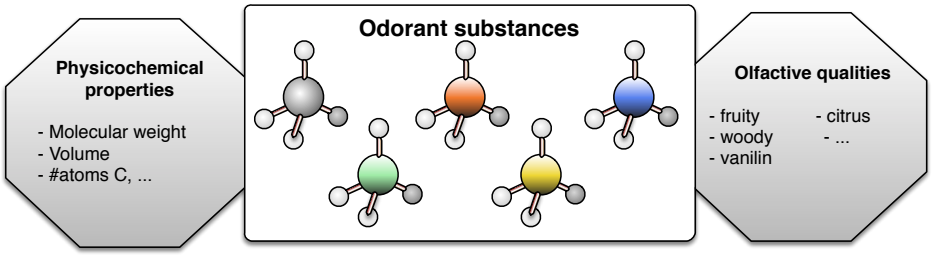
Interest, ...

- Fundamental neuroscience research
- Industry (agri-food industry , perfume industry, ...)
- Health (anosmia, ...)

C. Sezille et M. Bensafi
De la molécule au percept.
In *Biofutur*, 2013.

Problem setting

Given:

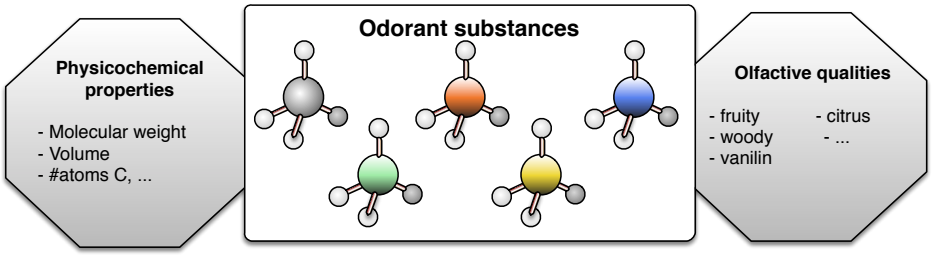


ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Problem setting

Given:



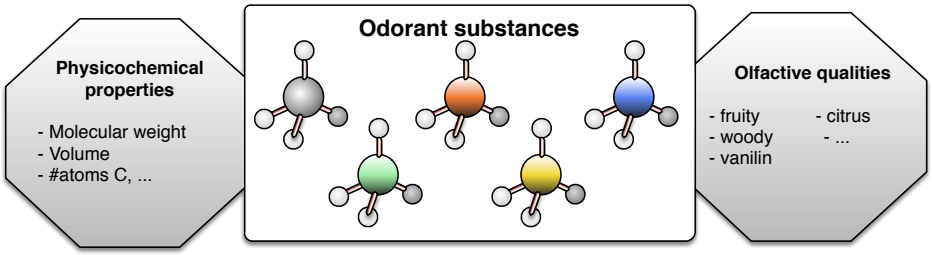
ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

⇝ How to **characterize** and **describe** the relationship between the **physicochemical properties** of a molecule and its **olfactory qualities** ?

Problem setting

Given:



ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

$\rightsquigarrow$ How to **characterize** and **describe** the relationship between the **physicochemical properties** of a molecule and its **olfactory qualities** ?

Find (e.g.):

- $\langle MW \leq 151.28, 23 \leq nAT \rangle \longrightarrow Honey$, with a high quality measure

P.K. Novak, N. Lavrač, and G.I. Web

Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining.
In *Journal of Machine Learning Research*, 2009.

Outline

Outline

Subgroup discovery 1/2

Task: *Find and describe subgroups of odorant molecules significantly different for* ***a*** *[1] (or* ***all*** *[2]) olfactory quality(ies).*

[1] S. Wrobel
An algorithm for multi-relational discovery of subgroups.
In *PKDD*, 1997.

[2] D. Leman, A. Feelders, and A. J. Knobbe
Exceptional model mining.
In *ECML/PKDD*, 2008.

Subgroup discovery 1/2

Task: *Find and describe subgroups of odorant molecules significantly different for* ***a*** *[1] (or* ***all*** *[2]) olfactory quality(ies).*

- **Subgroup:** described by a conjunction of attribute-values pairs (description) that covers a set of odorant molecules
  PHYSICOCHEMICAL DESCRIPTION $\longrightarrow$ OLFACTORY QUALITY(IES)

[1] S. Wrobel
An algorithm for multi-relational discovery of subgroups.
In *PKDD*, 1997.

[2] D. Leman, A. Feelders, and A. J. Knobbe
Exceptional model mining.
In *ECML/PKDD*, 2008.

Subgroup discovery 1/2

Task: *Find and describe subgroups of odorant molecules significantly different for* ***a*** *[1] (or* ***all*** *[2]) olfactory quality(ies).*

- **Subgroup:** described by a conjunction of attribute-values pairs (description) that covers a set of odorant molecules
  PHYSICOCHEMICAL DESCRIPTION $\longrightarrow$ OLFACTORY QUALITY(IES)
- **Interestingness:** quantifies the difference between the distribution of the labels of the projection of the subgroup and the entire dataset on the space of models (WRAcc or WKL)

[1] S. Wrobel
An algorithm for multi-relational discovery of subgroups.
In *PKDD*, 1997.

[2] D. Leman, A. Feelders, and A. J. Knobbe
Exceptional model mining.
In *ECML/PKDD*, 2008.

Subgroup discovery 1/2

Task: *Find and describe subgroups of odorant molecules significantly different for* ***a*** [1] *(or* ***all*** [2]*) olfactory quality(ies).*

- **Subgroup:** described by a conjunction of attribute-values pairs (description) that covers a set of odorant molecules
  PHYSICOCHEMICAL DESCRIPTION $\longrightarrow$ OLFACTORY QUALITY(IES)
- **Interestingness:** quantifies the difference between the distribution of the labels of the projection of the subgroup and the entire dataset on the space of models (WRAcc or WKL)
- **Algorithm:** heuristic approach (*beam-search*) to make search tractable

[1] S. Wrobel
An algorithm for multi-relational discovery of subgroups.
In *PKDD*, 1997.

[2] D. Leman, A. Feelders, and A. J. Knobbe
Exceptional model mining.
In *ECML/PKDD*, 2008.

Subgroup discovery 2/2

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Subgroup discovery 2/2

- 6 odorants $\{1, 24, 48, 60, 82, 1633\}$

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Subgroup discovery 2/2

- 6 odorants $\{1, 24, 48, 60, 82, 1633\}$
- 3 attributes of physicochemical properties $\{MW, nAT, nC\}$

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Subgroup discovery 2/2

- 6 odorants $\{1, 24, 48, 60, 82, 1633\}$
- 3 attributes of physicochemical properties $\{MW, nAT, nC\}$
- 1 class attribute *Qualities* with domain $\{Fruity, Vanilin, Honey\}$

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Subgroup discovery 2/2

- 6 odorants $\{1, 24, 48, 60, 82, 1633\}$
- 3 attributes of physicochemical properties $\{MW, nAT, nC\}$
- 1 class attribute *Qualities* with domain $\{Fruity, Vanilin, Honey\}$

- $d = \langle MW \leq 151.28, 23 \leq nAT \rangle$, $supp(d) = \{24, 48, 82, 1633\}$

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Subgroup discovery 2/2

- 6 odorants $\{1, 24, 48, 60, 82, 1633\}$
- 3 attributes of physicochemical properties $\{MW, nAT, nC\}$
- 1 class attribute *Qualities* with domain $\{Fruity, Vanilin, Honey\}$

- $d = \langle MW \leq 151.28, 23 \leq nAT \rangle$, $supp(d) = \{24, 48, 82, 1633\}$
  $WRAcc(d, Honey) = \frac{4}{6} \times (\frac{3}{4} - \frac{3}{6}) = 0.167$

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Subgroup discovery 2/2

- 6 odorants $\{1, 24, 48, 60, 82, 1633\}$
- 3 attributes of physicochemical properties $\{MW, nAT, nC\}$
- 1 class attribute *Qualities* with domain $\{Fruity, Vanilin, Honey\}$

- $d = \langle MW \leq 151.28, 23 \leq nAT \rangle$, $supp(d) = \{24, 48, 82, 1633\}$
  $WRAcc(d, Honey) = \frac{4}{6} \times (\frac{3}{4} - \frac{3}{6}) = 0.167$
  $WKL(d) = \frac{4}{6} \times ((\frac{2}{4} \log_2 \frac{3}{4}) + (\frac{3}{4} \log_2 \frac{3}{2}) + (\frac{3}{4} \log_2 \frac{3}{2})) = 0.45$

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

The drawbacks...

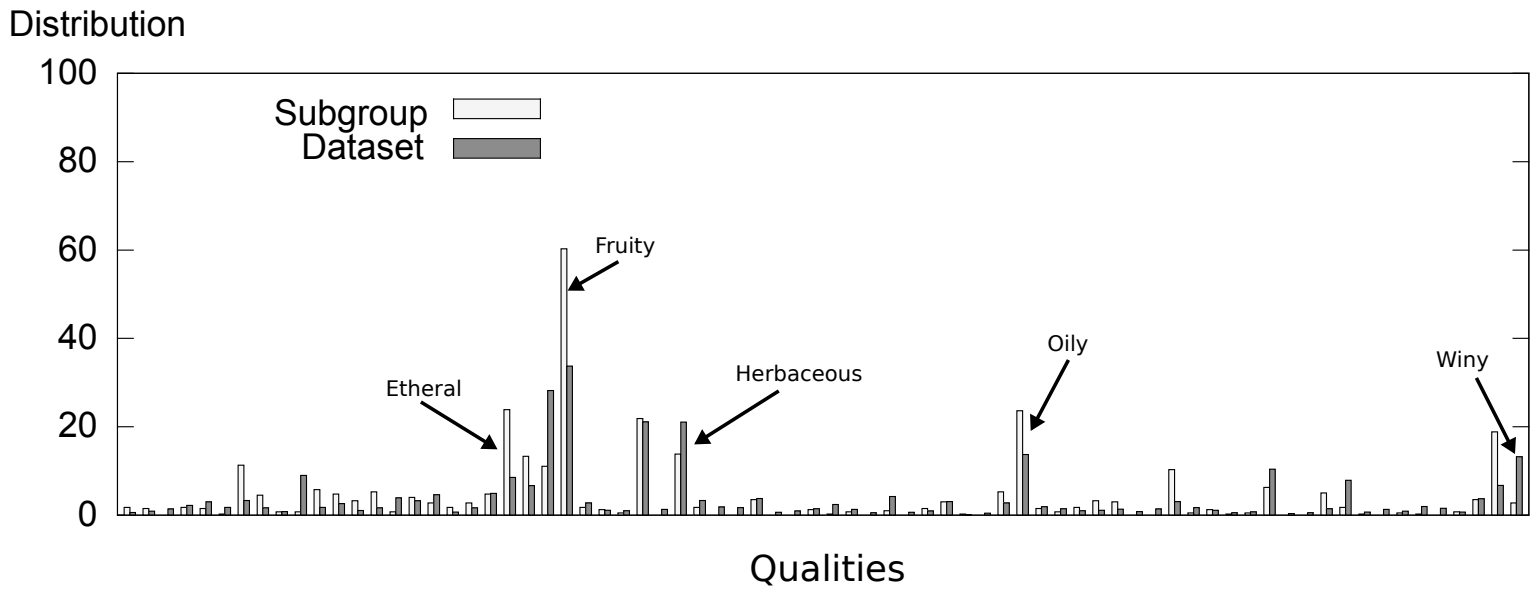
But these methods are unsuitable for olfaction:

- Characterize a single target not accurate enough
- Characterize all qualities together too coarse

The drawbacks...

But these methods are unsuitable for olfaction:

- Characterize a single target not accurate enough
- Characterize all qualities together too coarse



Outline

Exceptional local Model Mining 1/2

Principle: *Find and describe subgroups of objects (eg. odorants) significantly different for **a subset** of values of the class attribute (eg. olfactory qualities).*

Exceptional local Model Mining 1/2

Principle: *Find and describe subgroups of objects (eg. odorants) significantly different for **a subset** of values of the class attribute (eg. olfactory qualities).*

Quality measure: Given d a description and L a subset of values of the class attribute, we measure the quality of a local subgroup (d, L) with the $F_1 - score$:

$$F_1(d, L) = \frac{2 \times (P(d, L) \times R(d, L))}{P(d, L) + R(d, L)}$$

where $P(d, L) = \frac{E11}{E11+E10}$ is the precision
$R(d, L) = \frac{E11}{E11+E01}$ is the recall

with $E10 = |\{o \in \mathcal{O} | o \in supp(d), class(o) \cap L \neq L\}$,
$E11 = |\{o \in \mathcal{O} | o \in supp(d), class(o) \cap L = L\}$,
$E01 = |\{o \in \mathcal{O} | o \notin supp(d), class(o) \cap L = L\}$.

Exceptional local Model Mining 1/2

Principle: *Find and describe subgroups of objects (eg. odorants) significantly different for* ***a subset*** *of values of the class attribute (eg. olfactory qualities).*

Quality measure: Given d a description and L a subset of values of the class attribute, we measure the quality of a local subgroup (d, L) with the $F_1 - score$:

$$F_1(d, L) = \frac{2 \times (P(d, L) \times R(d, L))}{P(d, L) + R(d, L)}$$

Objective: The task of EIMM is to extract the top-k local subgroups (d, L) *wrt* the quality measure such that:

- $|supp(d)| \geq minSupp$
- $|d| \leq maxDesc$
- $|L| \leq maxLab$

Exceptional local Model Mining 2/2

Example:

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Exceptional local Model Mining 2/2

Example:

- $d = \langle MW \leq 151.28, 23 \leq nAT \rangle$

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Exceptional local Model Mining 2/2

Example:

- $d = \langle MW \leq 151.28, 23 \leq nAT \rangle$
- $L = \{Honey, Vanilin\}$

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Exceptional local Model Mining 2/2

Example:

- $d = \langle MW \leq 151.28, 23 \leq nAT \rangle$
- $L = \{Honey, Vanilin\}$
- $supp(d, L) = supp(d) = \{24, 48, 82, 1633\}$

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Exceptional local Model Mining 2/2

Example:

- $d = \langle MW \leq 151.28, 23 \leq nAT \rangle$
- $L = \{Honey, Vanilin\}$
- $supp(d, L) = supp(d) = \{24, 48, 82, 1633\}$

- $F_1(d, L) = \frac{2 \times P(d,L) \times R(d,L)}{P(d,L)+R(d,L)} = \frac{2 \times \frac{1}{2} \times 1}{\frac{1}{2}+1} = \frac{2}{3}$

ID	MW	nAT	nC	Qualities
1	150.19	21	11	{Fruity}
24	128.24	29	9	{Honey, Vanilin}
48	136.16	24	10	{Honey, Vanilin}
60	152.16	23	11	{Fruity}
82	151.28	27	12	{Fruity, Honey}
1633	142.22	27	10	{Fruity, Vanilin}

Toy dataset

Outline

Introducing...ElMMut

High-level overview of ElMMut:

Introducing...ElMMut

High-level overview of ElMMut:

- *Search space* based on a lattice structure of the local subgroups (partial order relation $\preceq$)

  eg. $\langle MW \leq 151.28, 23 \leq nAT, 10 \leq nC \rangle \preceq \langle MW \leq 151.28, 23 \leq nAT \rangle$

Introducing...ElMMut

High-level overview of ElMMut:

- *Search space* based on a lattice structure of the local subgroups (partial order relation $\preceq$)
- *Beam-search*: efficiently exploring search space top-down (from general to specific local subgroups)

Introducing...ElMMut

High-level overview of ElMMut:

- *Search space* based on a lattice structure of the local subgroups (partial order relation $\preceq$)
- *Beam-search*: efficiently exploring search space top-down (from general to specific local subgroups)
- *Pruning step* realized thanks to constraints

Introducing...ElMMut

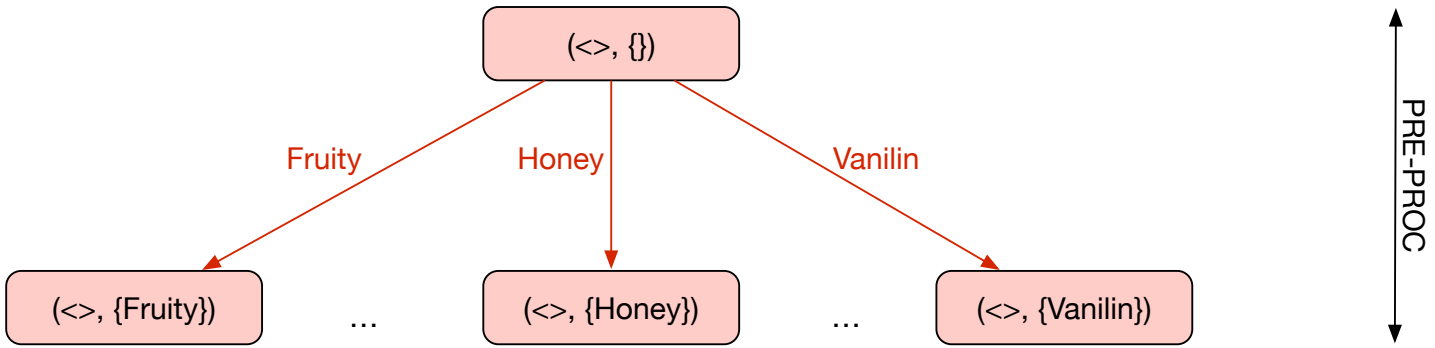
High-level overview of ElMMut:

- *Search space* based on a lattice structure of the local subgroups (partial order relation $\preceq$)
- *Beam-search*: efficiently exploring search space top-down (from general to specific local subgroups)
- *Pruning step* realized thanks to constraints
- *On-the-fly bucketing* to handle numerical attributes and find best cut points (improve the quality measure of local subgroups)

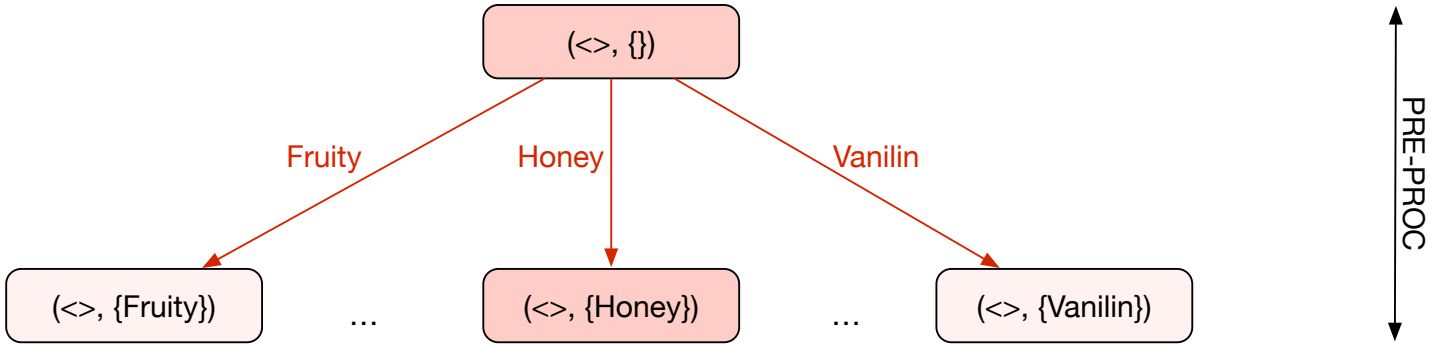
U.M. Fayyad, and K.B. Irani
Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning.
In *IJCAI*, 1993.

ElMMut example

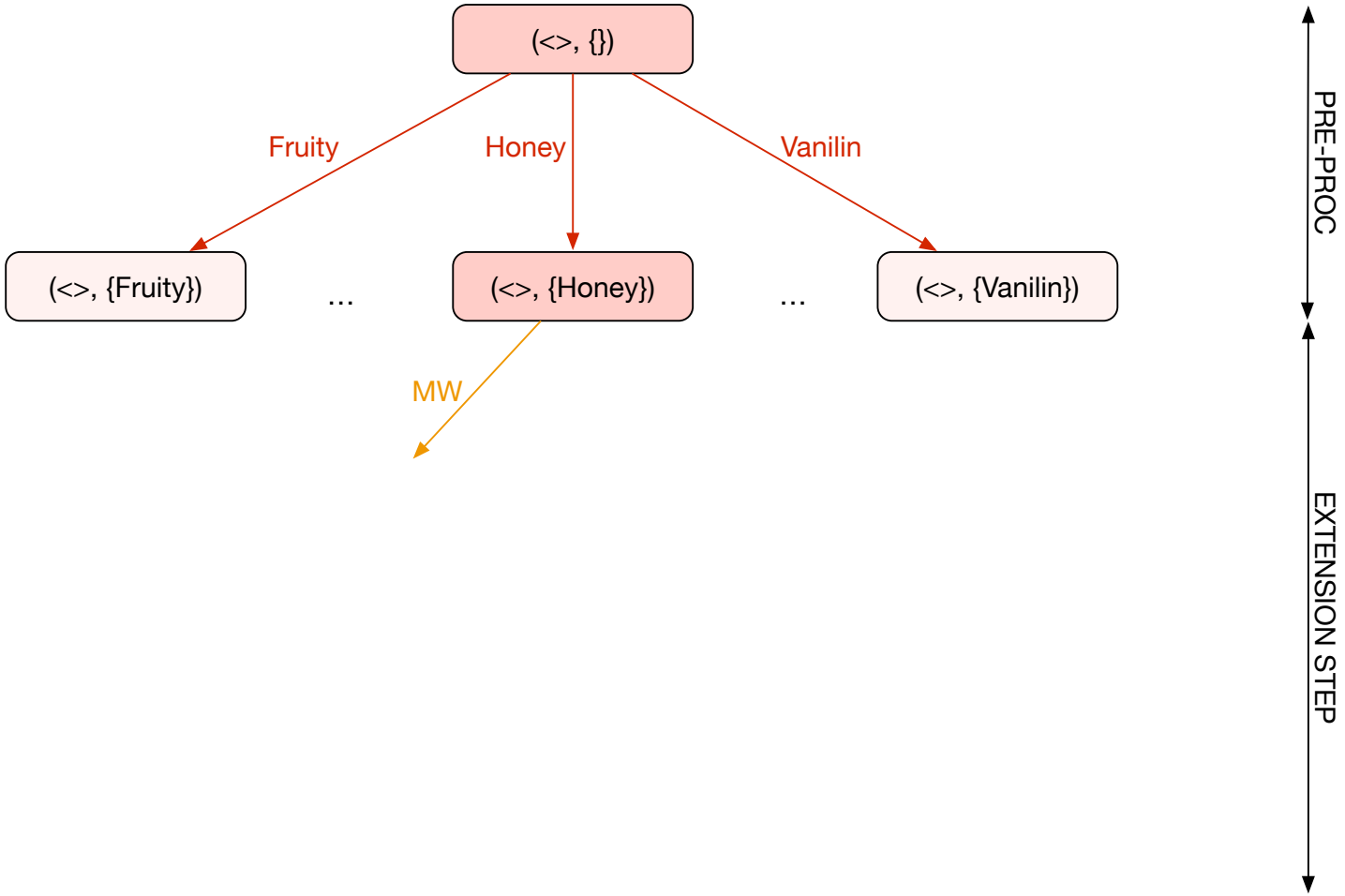
ElMMut example



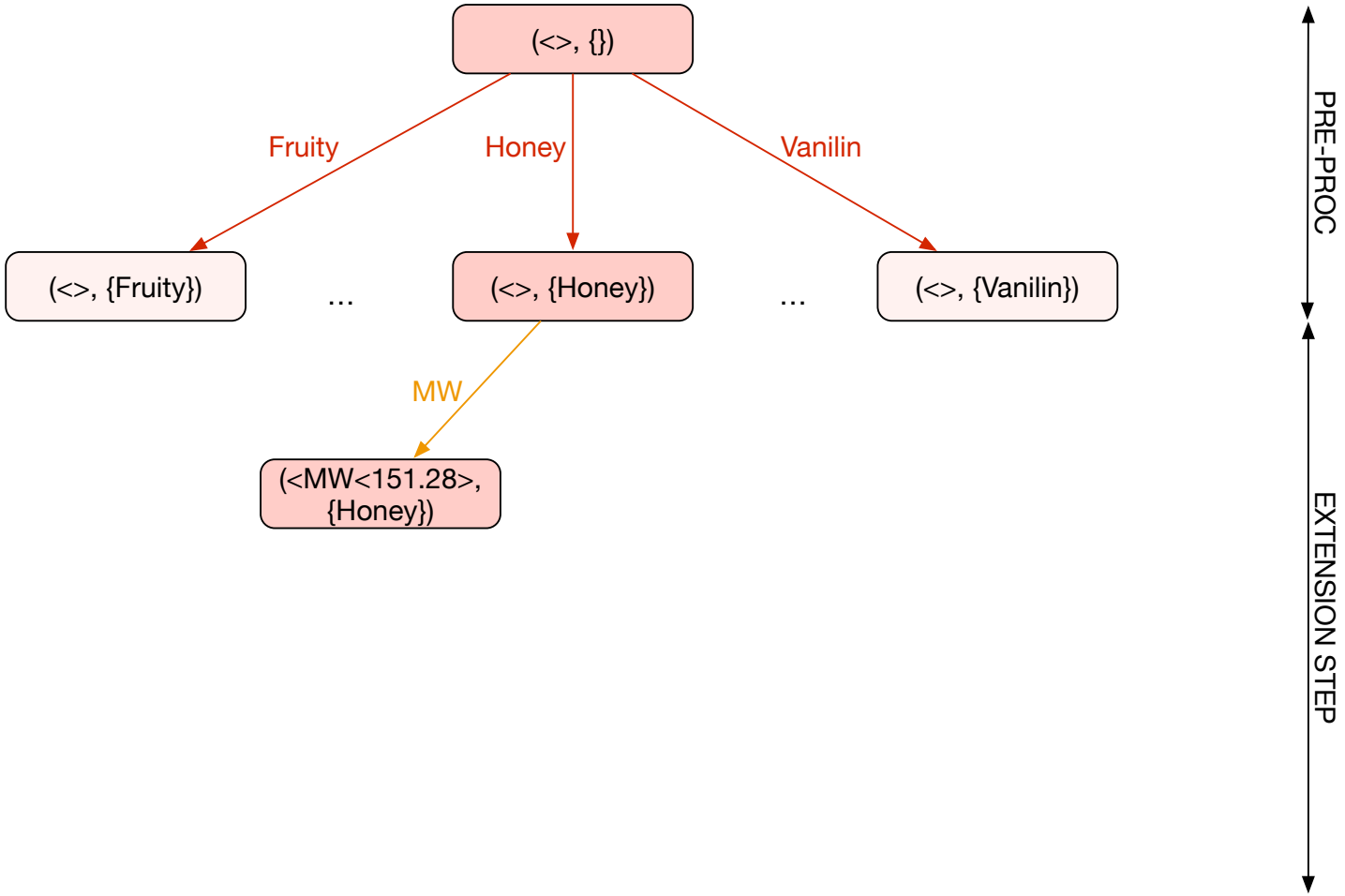
ElMMut example



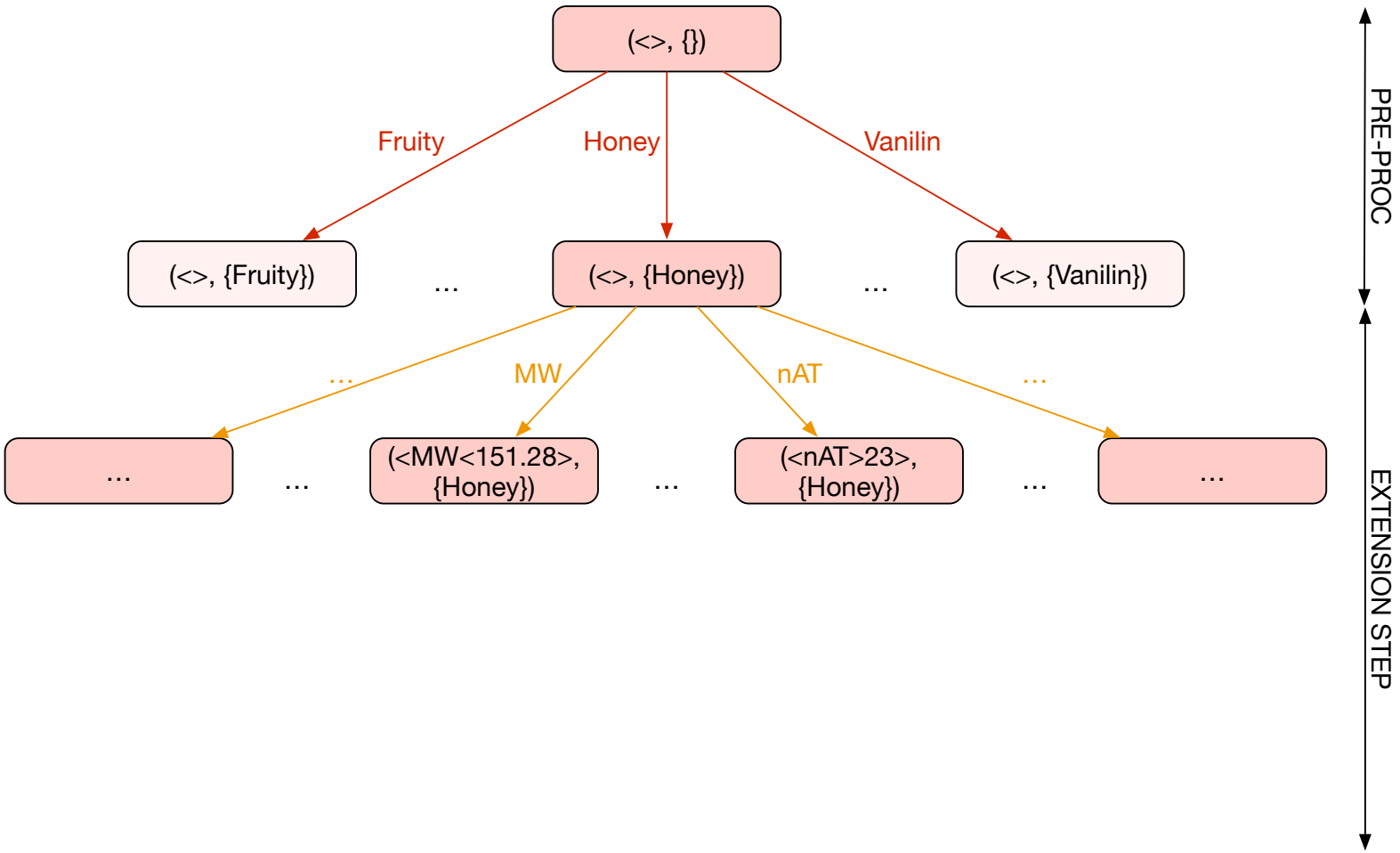
ElMMut example



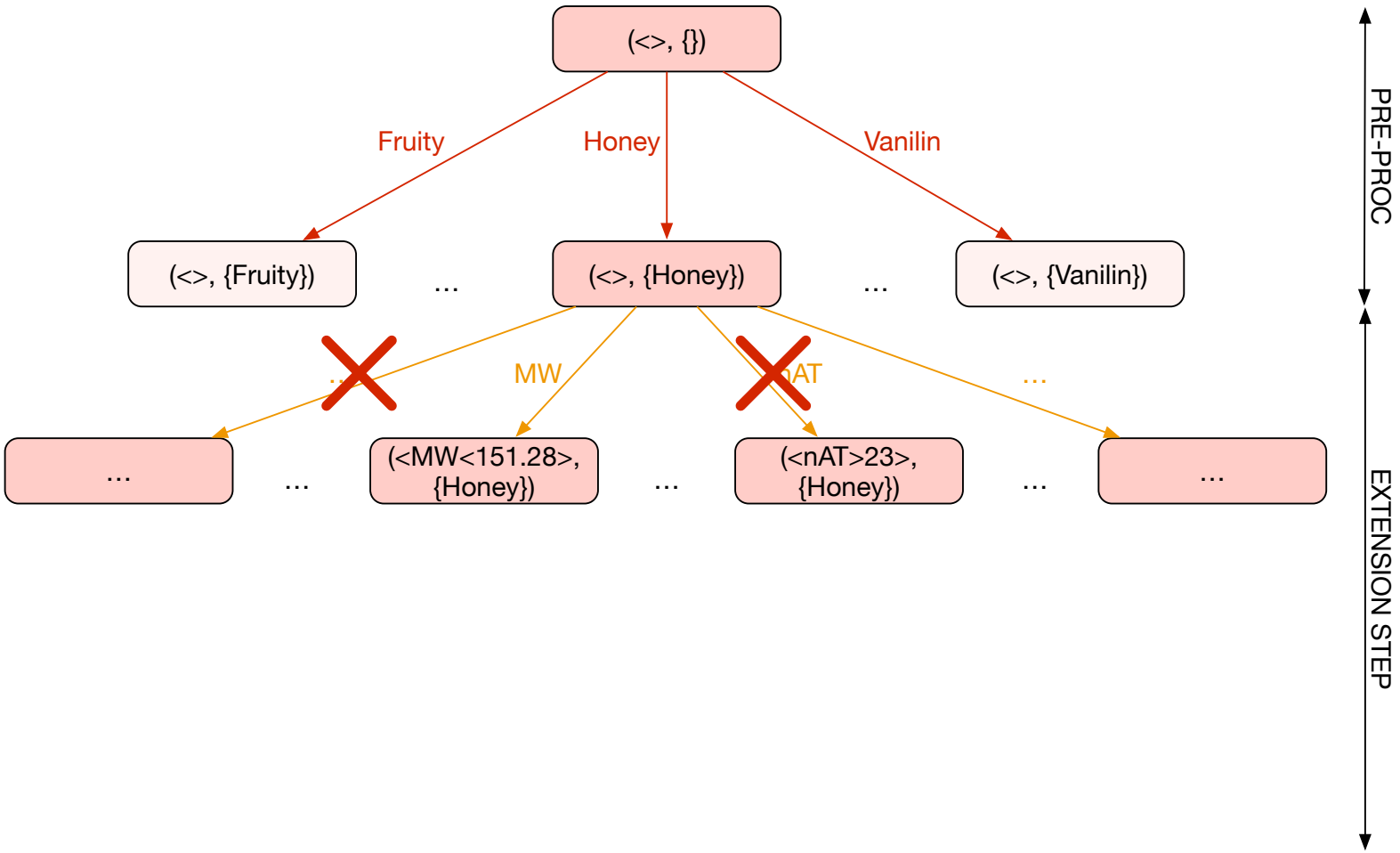
ElMMut example



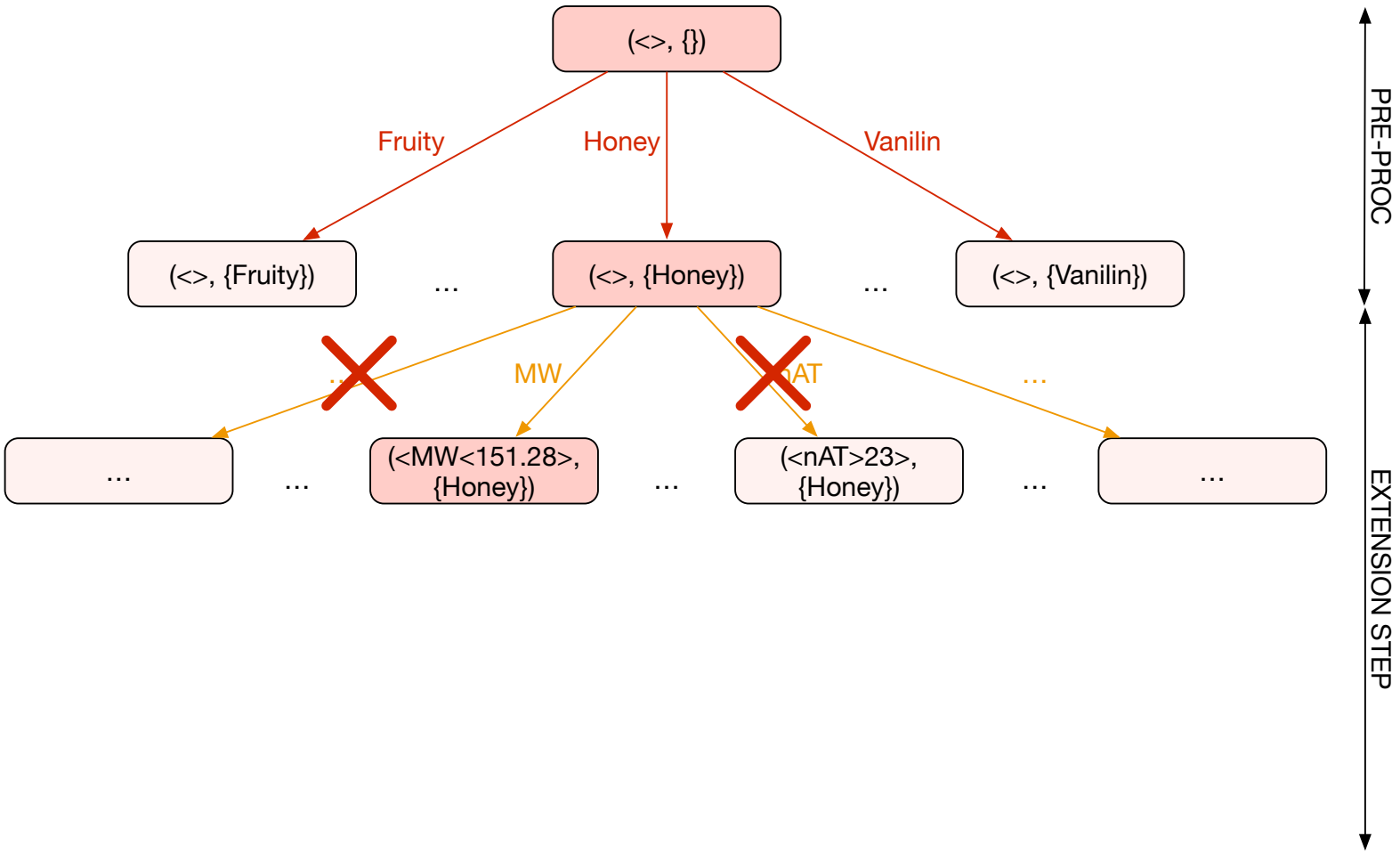
ElMMut example



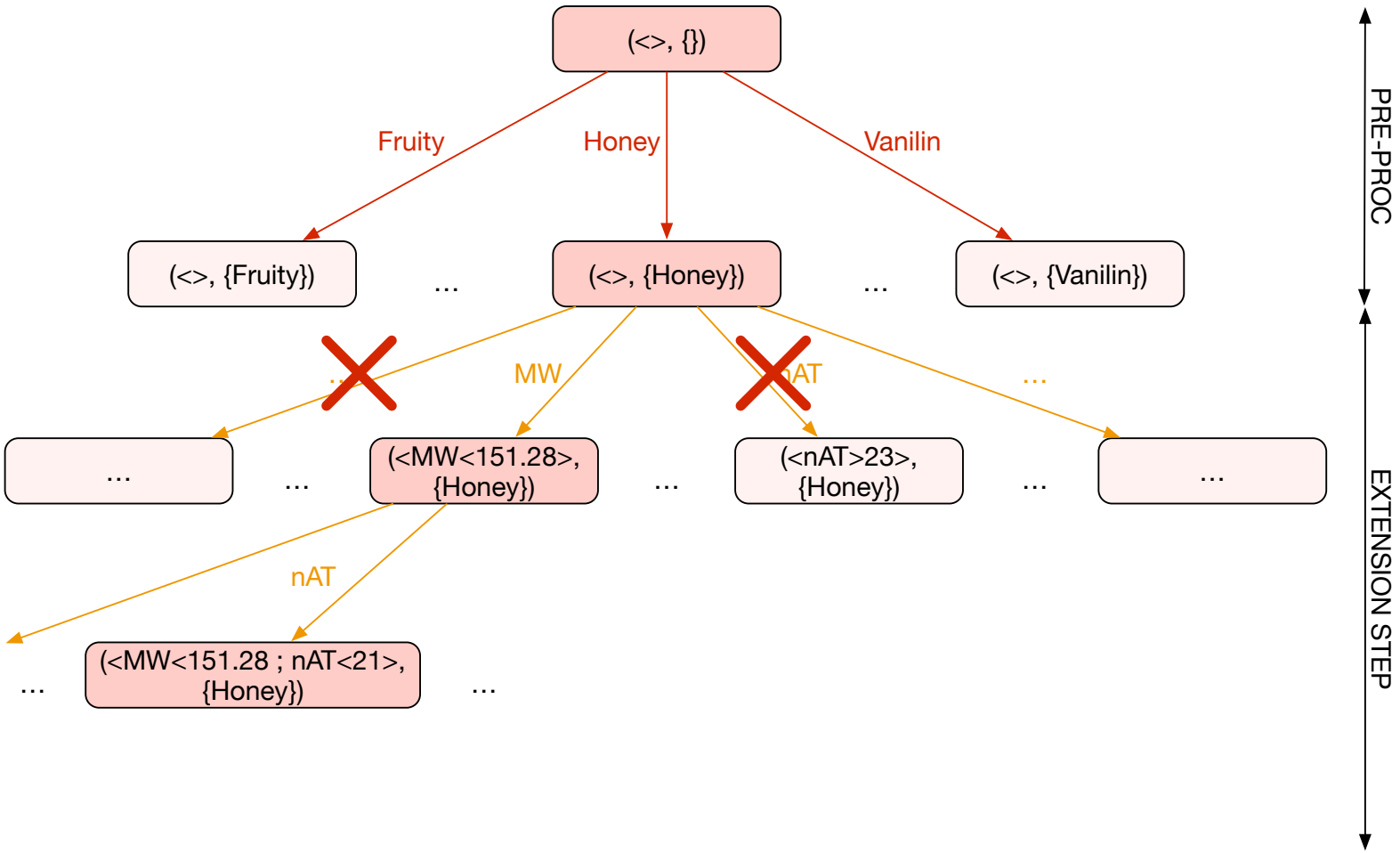
ElMMut example



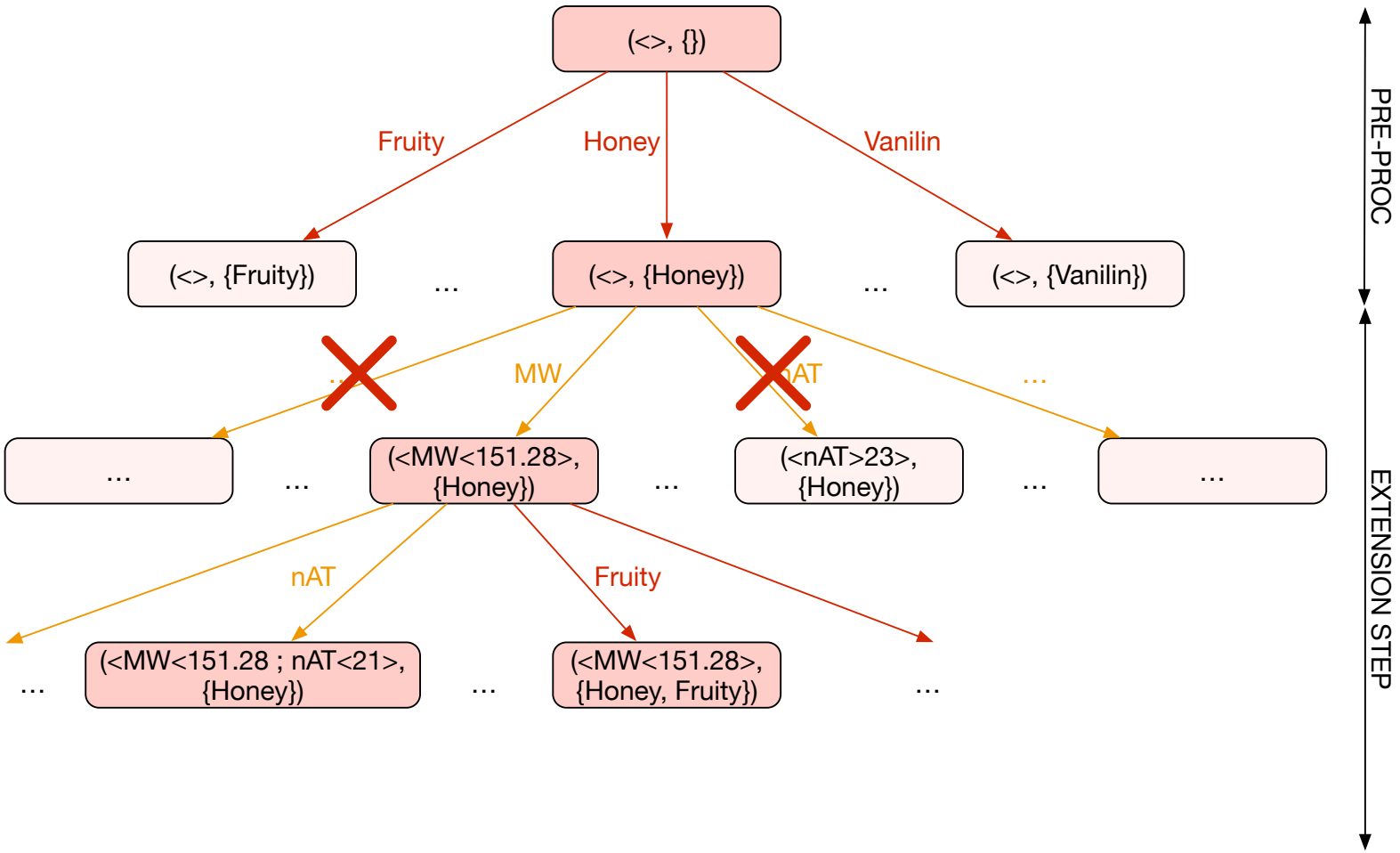
ElMMut example



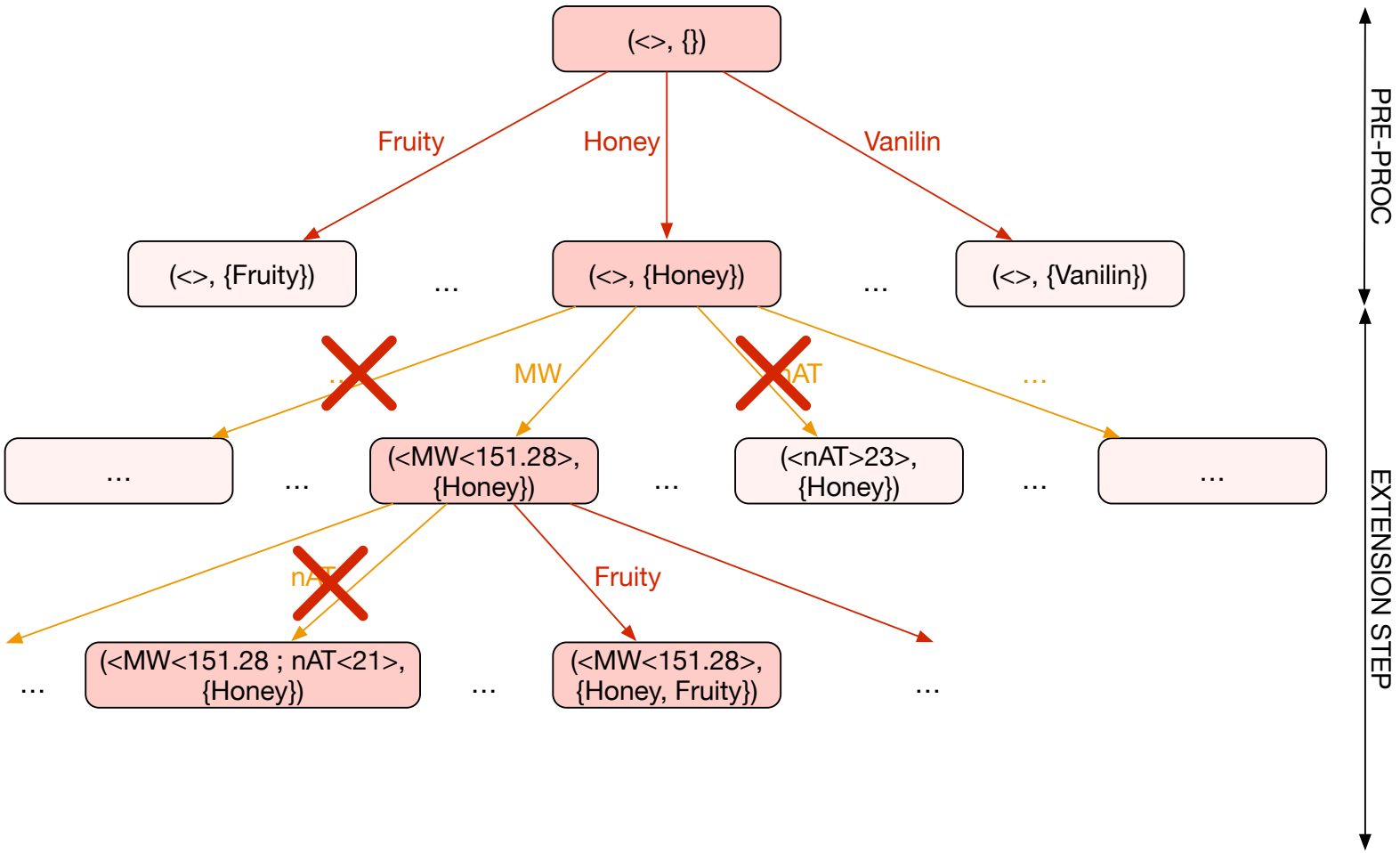
ElMMut example



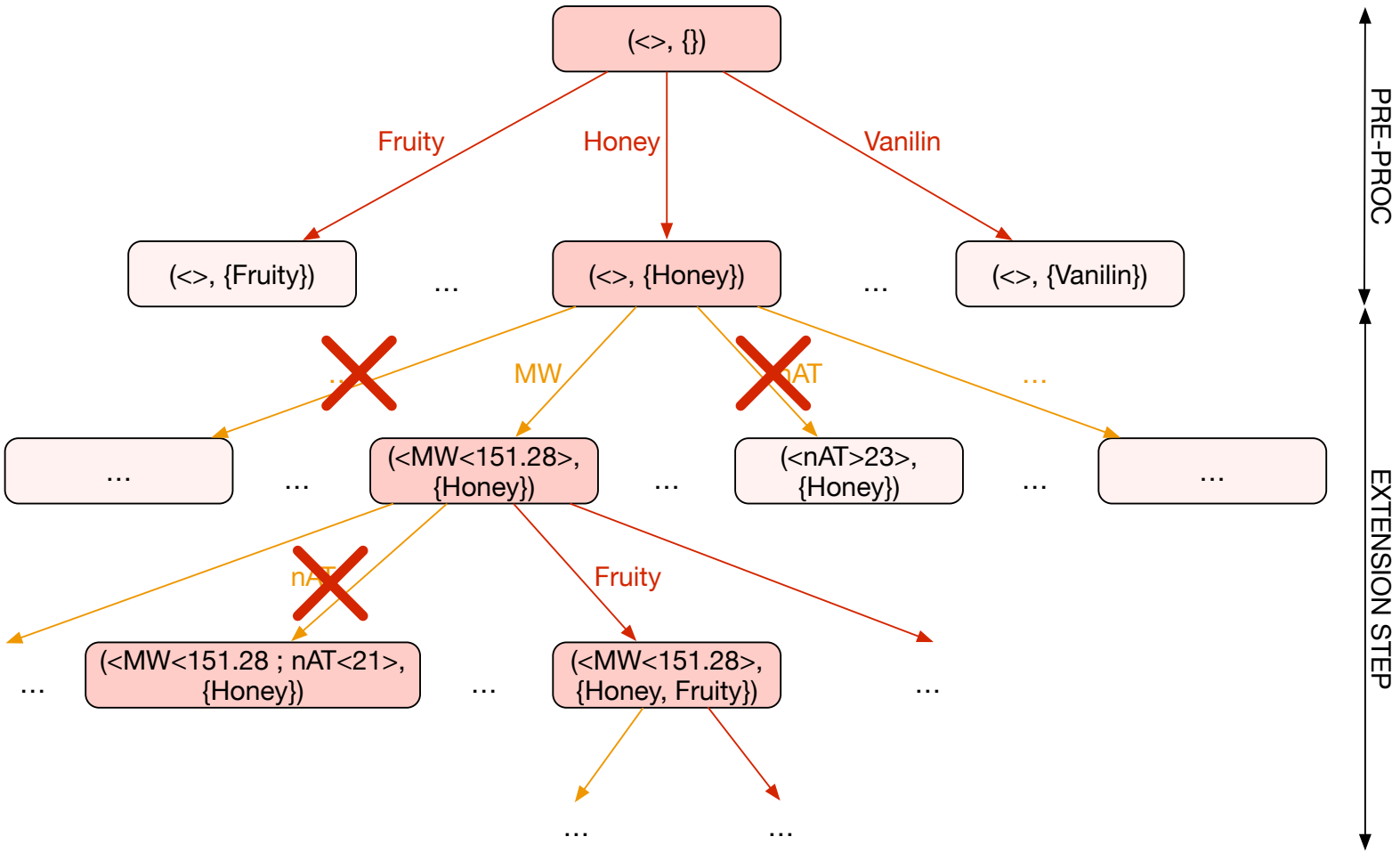
ElMMut example



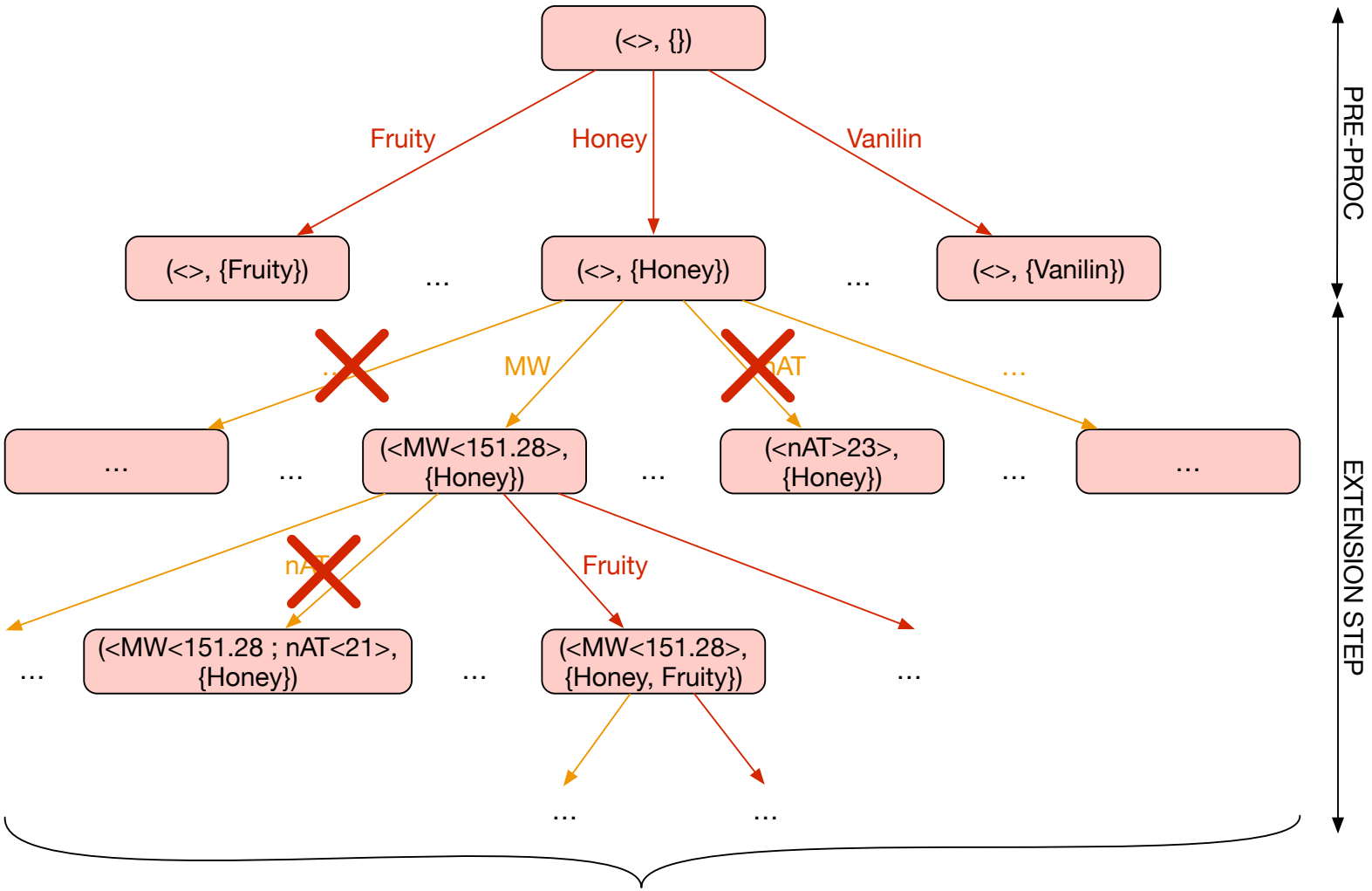
ElMMut example



ElMMut example



ElMMut example



Extract the top-k of explored local subgroups

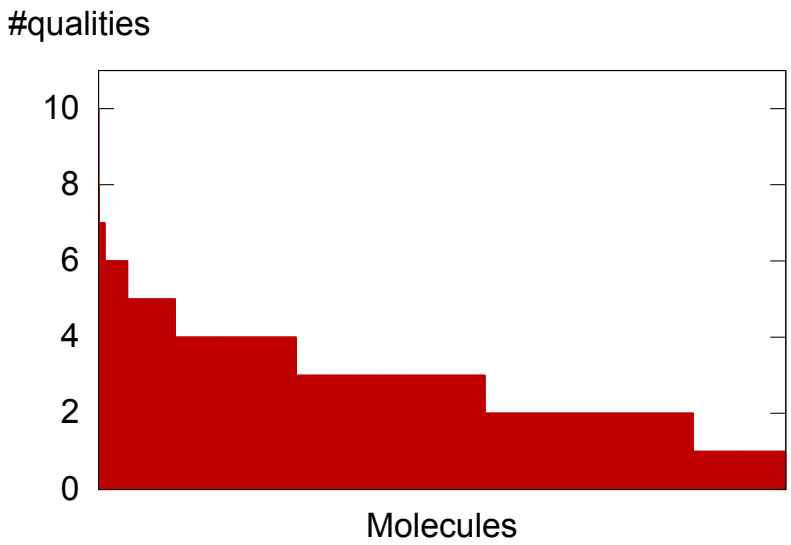
Outline

Establishment of datasets:

- 1 atlas (Arctander) provided by neuroscientists
- 2 datasets with different characteristics derived from atlas

	Dataset D1	Dataset D2
Atlas	Arctander	Arctander
Number of molecules	1689	1689
Number of physical properties	**43**	**243**
Number of olfactory qualities	74	74
Number of olfactory qualities per molecules	2.88	2.88

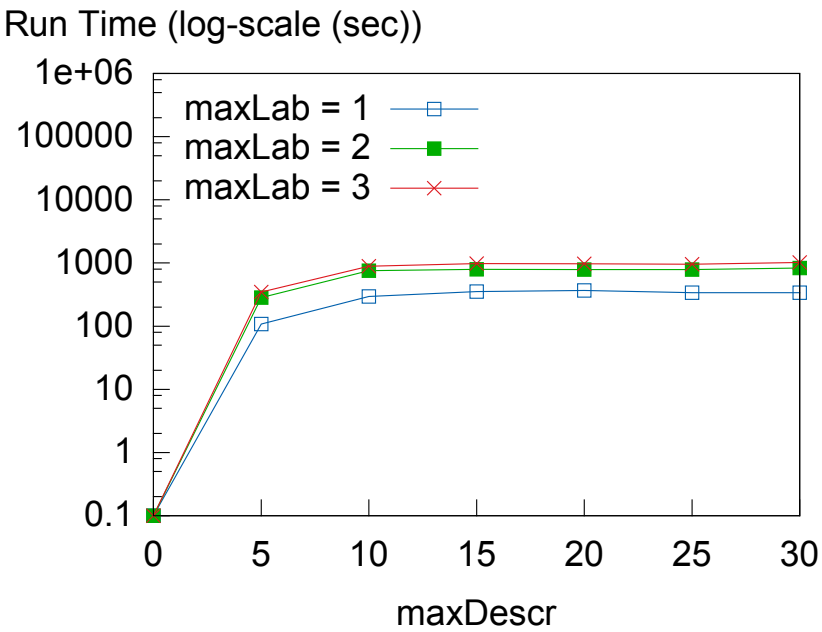
Characteristics of both datasets



S. Arctander
Perfume and flavor chemicals:(aroma chemicals).
In *Allured Publishing Corporation, Volume 2*, 1969.

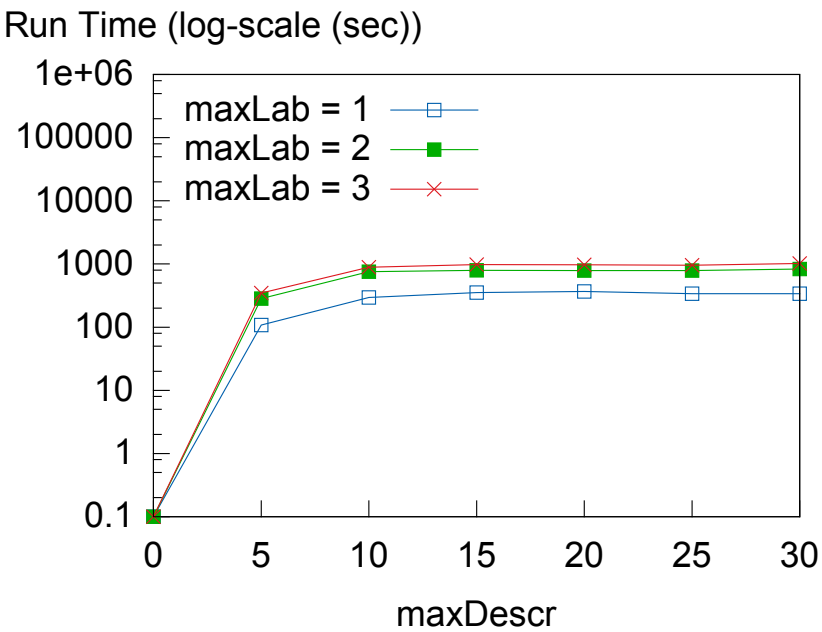
Quantitative results

- Dataset D_1, no optimization bucketing
- Key factors: *maxDescr* and *maxLab*
- *maxDescr* = 15
  maxLab = 2 or 3

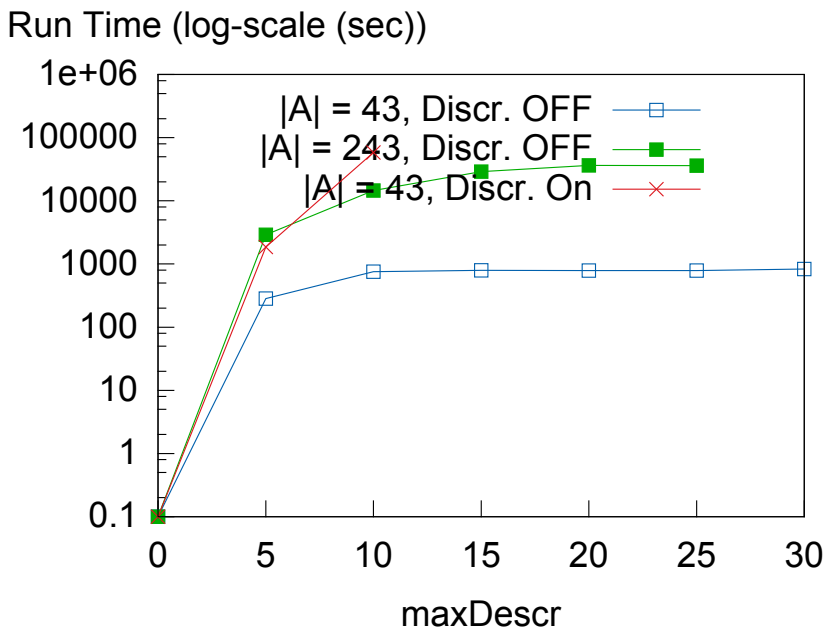


Quantitative results

- Dataset D_1, no optimization bucketing
- Key factors: *maxDescr* and *maxLab*
- *maxDescr* = 15
  maxLab = 2 or 3



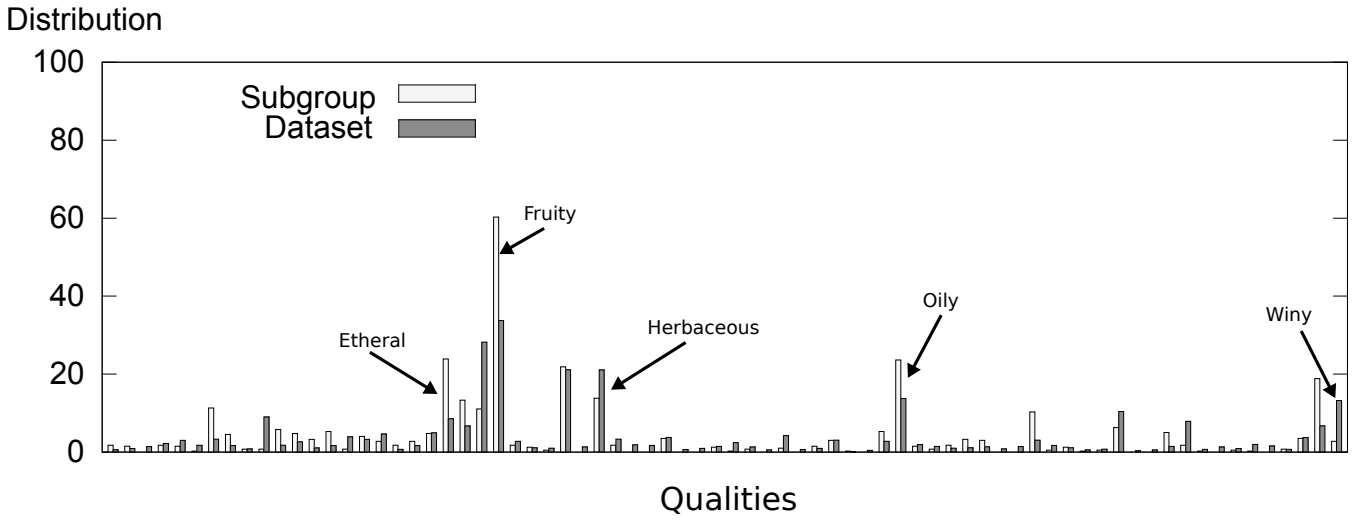
- Datasets D_1 and D_2
- Key factors: number of attributes and optimization bucketing



Qualitative results 1/2

Example of result existing method (EMM):

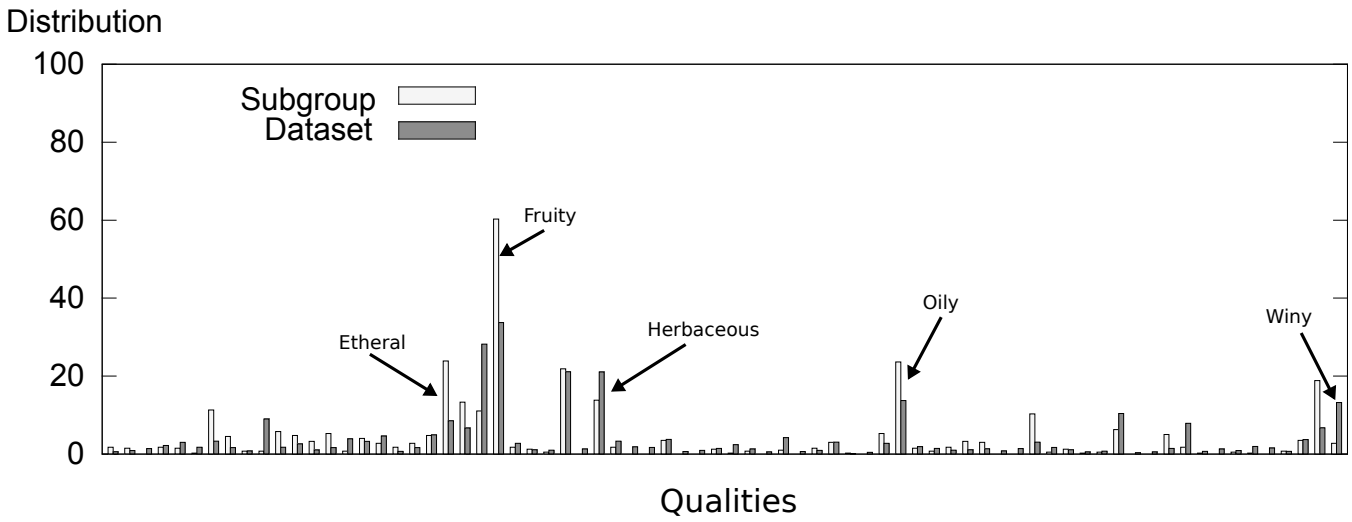
- Difficulties to interpret result
- Too many qualities are involved



Qualitative results 1/2

Example of result existing method (EMM):

- Difficulties to interpret result
- Too many qualities are involved



Example of result our method (EIMM):

- 74.6% of local subgroups involve 1 quality
- 22.9% of local subgroups involve 2 qualities
- 2.5% of local subgroups involve 3 qualities

Qualitative results 2/2

Top-5 local subgroups where:

- Dataset D_1
- Optimization bucketing
- $maxDescr = 10$, $maxLab = 2$, $minSupp = 30$

d	L	$\lvert supp(d) \rvert$	F_1
$\langle 0.116 < X\% < 0.314, 1.0 < nHet < 11.0, 5.159 < Sv < 8.792, 0.0 < nCIC < 0.0, 2.0 < nR03 < 8.0, 0.416 < Ui < 3.551, 4.0 < nArOH < 5.0, 1.0 < nCsp2 < 3.0, 12.0 < nCs < 47.0, 8.0 < nArCOOR < 25.0 \rangle$	{Fruity}	654	0.66
$\langle 134.19 < MW < 349.51, 14.0 < nCconj < 100.0, 4.76 < Sv < 8.277, 0.048 < X\% < 0.212, 22.0 < nCs < 49.0, 1.077 < Ui < 3.85, 18.0 < nAB < 49.0 \rangle$	{Floral}	740	0.55
$\langle 3.462 < Ui < 3.719, 30.0 < nCconj < 56.0, 40.0 < nAT < 57.0, 35.0 < nO < 50.0 \rangle$	{Musk}	32	0.5
$\langle 2.442 < TPSA(Tot) < 4.028, 4.74 < Sv < 6.095, 2.777 < Ui < 3.921, 0.208 < X\% < 0.31 \rangle$	{Oily}	213	0.44
$\langle 9.0 < nHet < 15.0, 6.095 < Sv < 8.258, 0.0 < Nr05 < 0.0, 2.749 < Ui < 3.517, 25.0 < nAB < 45.0, 2.279 < TPSA(Tot) < 3.334, 24.0 < nRCOOH < 34.0, 21.0 < nCconj < 51.0, 0.074 < X\% < 0.171 \rangle$	{Floral, Balsamic}	38	0.33

Outline

At the end...

The goals

- Characterize odors
- Study the relationship between physicochemical properties and olfactory qualities

At the end...

The goals

- Characterize odors
- Study the relationship between physicochemical properties and olfactory qualities

We propose a new problem setting and method

- Exceptional local Model Mining
- Generalization of existing approaches
- Introducing EIMMut

At the end...

The goals

- Characterize odors
- Study the relationship between physicochemical properties and olfactory qualities

We propose a new problem setting and method

- Exceptional local Model Mining
- Generalization of existing approaches
- Introducing EIMMut

Promising results

- Experiment on datasets provided by CRNL
  - Results considered interesting by expert
  - Theoretical avenues of improvements identified
- We invite you to explore further: ($\rightarrow$ *http://liris.cnrs.fr/olfamining*)

Perspectives

Extract *non-redundant* and a *statistically significant* (*p-value* test) local subgroups.

Perspectives

Extract *non-redundant* and a *statistically significant* (*p-value* test) local subgroups.

How to formalize and use a hierarchy over physicochemical attributes and/or olfactory qualities to improve our technique?

Perspectives

Extract *non-redundant* and a *statistically significant* (*p-value* test) local subgroups.

How to formalize and use a hierarchy over physicochemical attributes and/or olfactory qualities to improve our technique?

2D and 3D representation of molecules: how take into account these key parameters to improve the efficiency of our method ?

Thank you for your attention.

Any questions ?