A new diagram for amino acids:
User study comparing rainbow boxes to Venn/Euler diagram

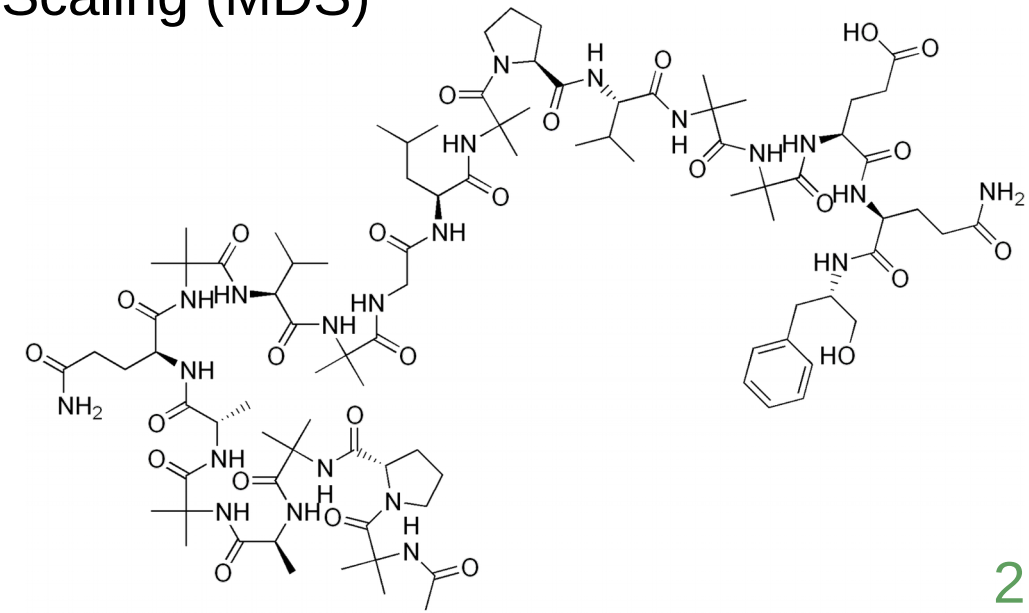
Jean-Baptiste Lamy

jean-baptiste.lamy @ univ-paris13.fr

LIMICS
Université Paris 13, Sorbonne Paris Cité, 93017 Bobigny
Sorbonne Universités, Paris
INSERM UMRS 1142

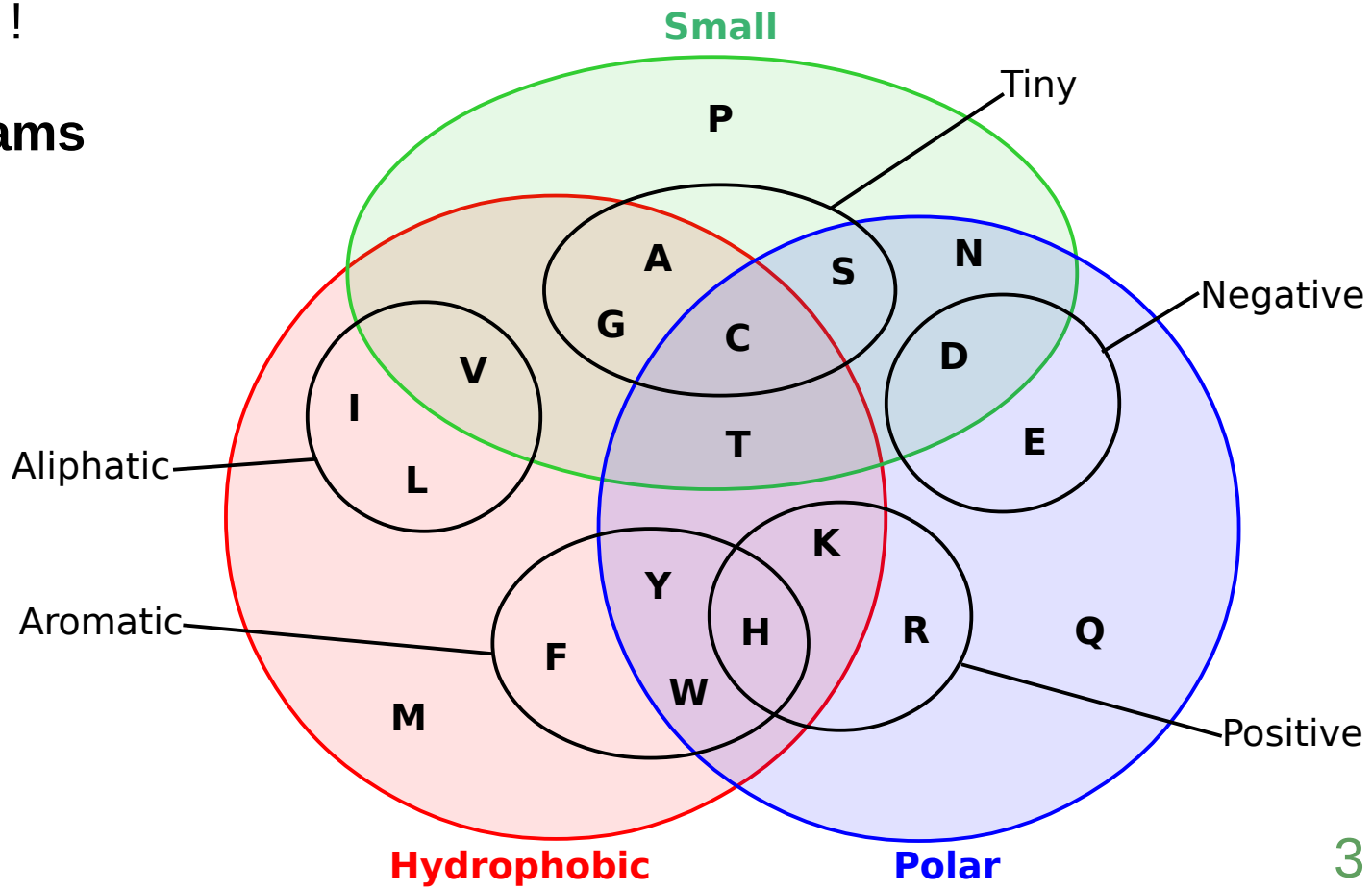
Introduction

- **The 20 amino acids: the building blocks for proteins**
- **Various groups of amino acids**
  - Size, Electrical charge, Chemical properties
  - Taylor’s classification in 8 overlapping groups
- **The 20 amino acids and the 8 groups are commonly visualized as an Euler diagram**
  - Produced *via* Multidimensional Scaling (MDS)

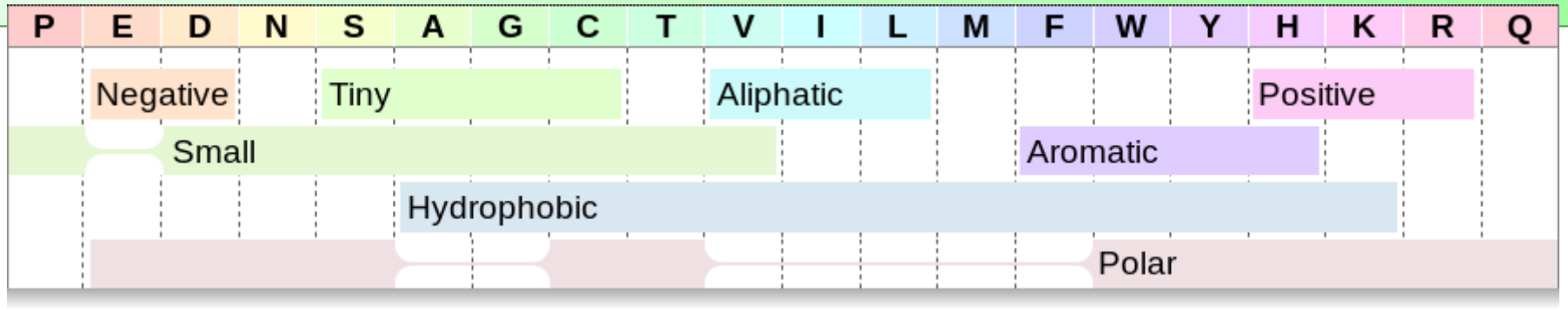


Introduction

- **The « Venn diagram of amino acid properties »**
  - Very well-known in biology and bioinformatics
  - Google : 364,000 results for “Venn|Euler diagram amino acid”
  - Actually an Euler diagram, not a Venn diagram !
- **But Euler/Venn diagrams are difficult to read**
  - More than 4 sets

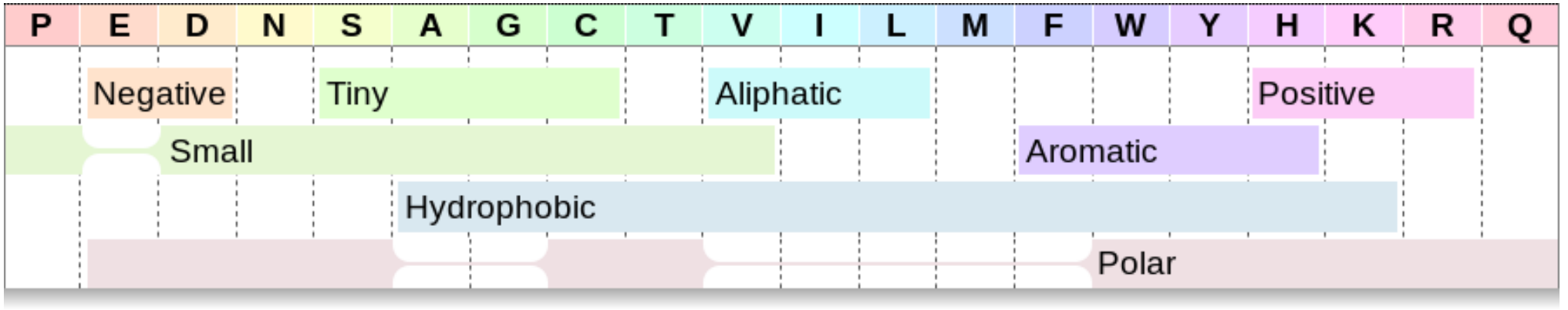


Introduction



- **Rainbow boxes : a recent technique for set visualization**
  - elements => columns
  - sets => rectangular boxes
  - color => one color per element
  - box color is the mean of its elements color
  - non continguous element in a set => box hole
  - elements are ordered so as to minimize the number of holes
  - box are stacked vertically by size
- **Objectives:**
  - Evaluating rainbow box *vs* Euler diagram on amino acids
  - Evaluating the possibility to enrich the diagram with new properties

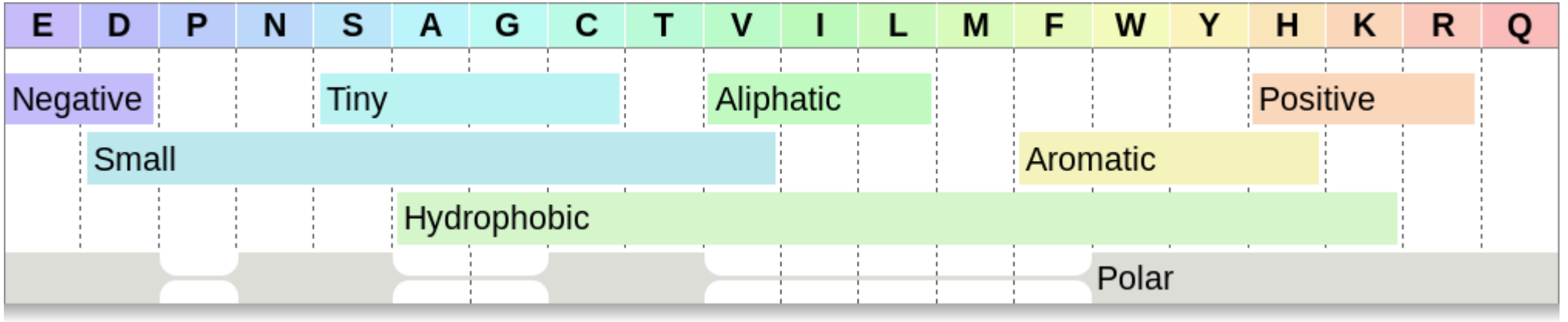
Improving the Rainbow Boxes diagram



New column order: the 3 holes are in the same box

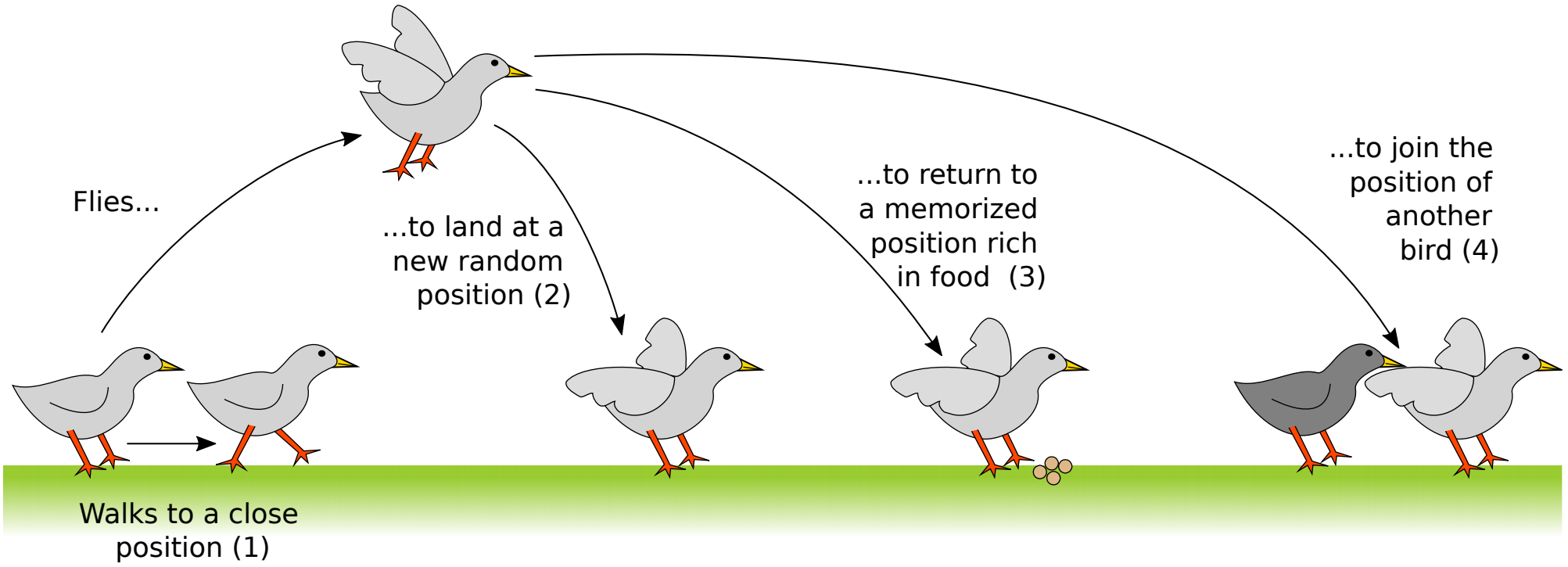
Verify that no column order with less than 3 holes exists

Better color contrast



AFB metaheuristic

Artificial Feeding Birds (AFB) [Springer]
Inspired by the behaviour of pigeons



- **Simple**
- **Performant**
- **Generic**

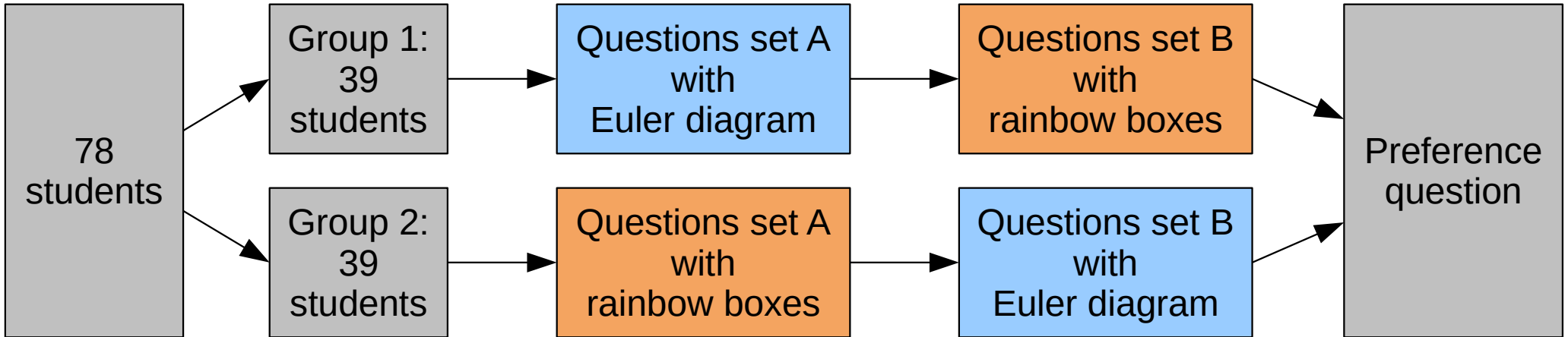
An optimisation problem = A triplet (*coût*(), *vol*(), *marche*())

Recruitment

- **78 biology students at University Paris 13**
  - 56 students in third year of bachelor degree (B3 / L3)
  - 22 students in first year of master degree (M1)
  - Majority of female students
  - 5 sessions
  - Single-blind
  - B3 students were « naive »
    - New to the Euler diagram of amino acids

Protocol

- **Cross-over protocol**
  - Each student tested both diagrams
  - Two groups of 39 students each (28 B3 + 11 M1)
  - Two sets of similar questions, A and B
- **Measurement**
  - Accuracy of responses (main criteria)
  - Response times
  - User preference (Euler diagram, rainbow boxes, no opinion)
- **Dynamic website**



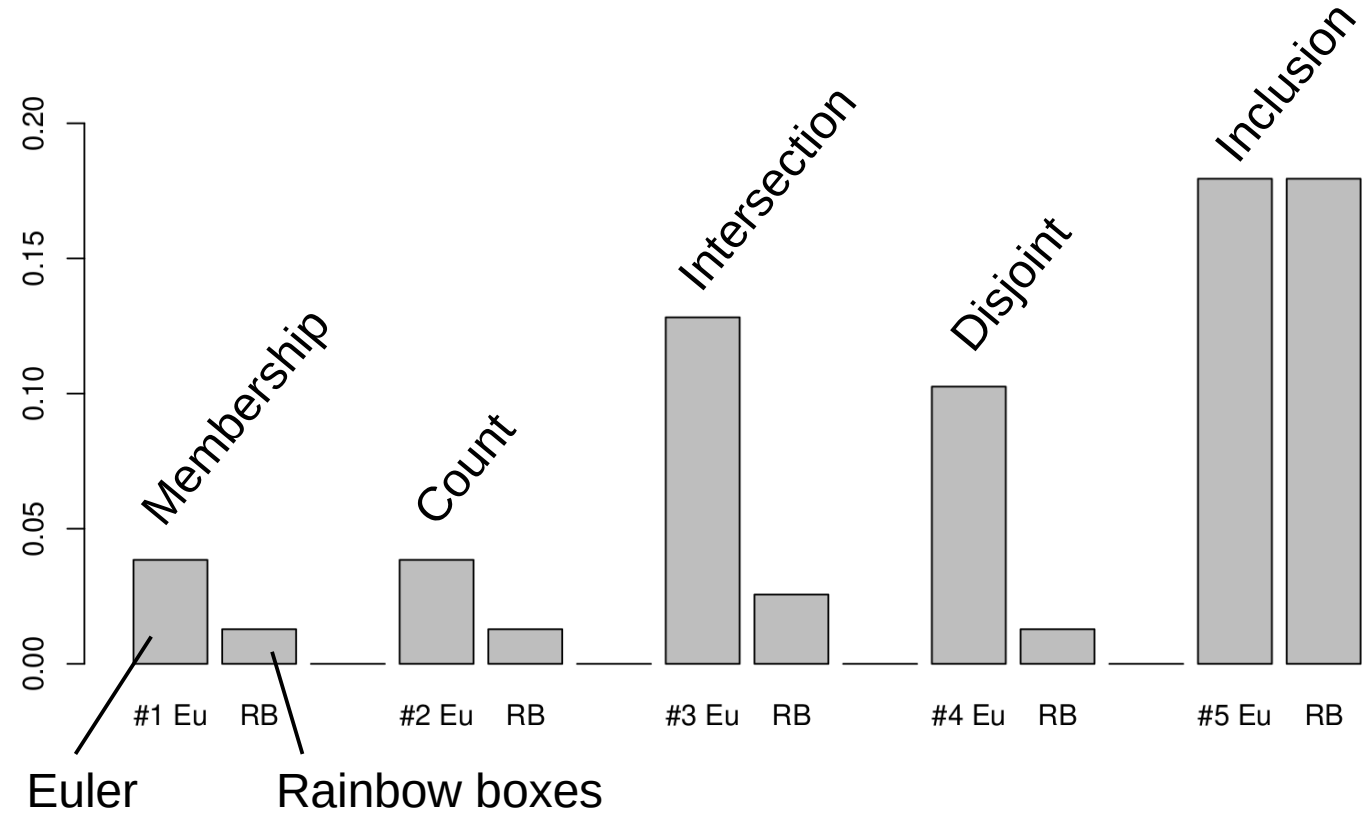
Questions

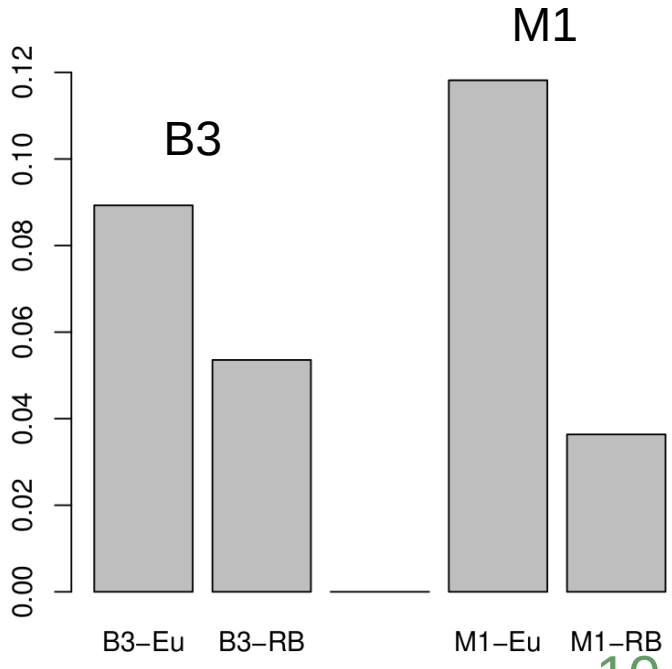
- **5 categories of questions**

	#	Category	#	Question	Right answer
Set A		(Warm-up)	1	Click on the small and aliphatic amino acid	Valine (V)
	1	Membership	2	Is Valin (V) polar?	No
	2	Count	5	How many tiny amino acids are there?	4
	3	Intersection	6	Click on the aromatic and positive amino acid	Histidine (H)
	4	Disjoint	3	Click on the tiny and positive amino acid	None
	5	Inclusion	4	Are all small amino acids polar ?	No
Set B		(Warm-up)	1	Is Cysteine (C) small?	Yes
	1	Membership	2	Is Tyrosine (Y) hydrophobic?	Yes
	2	Count	5	How many positive amino acids are there?	3
	3	Intersection	3	Click on the small and negative amino acid	Aspartate (D)
	4	Disjoint	4	Click on the aromatic and aliphatic amino acid	None
	5	Inclusion	6	Are all aromatic amino acids hydrophobic ?	Yes

Evaluation results: error rates

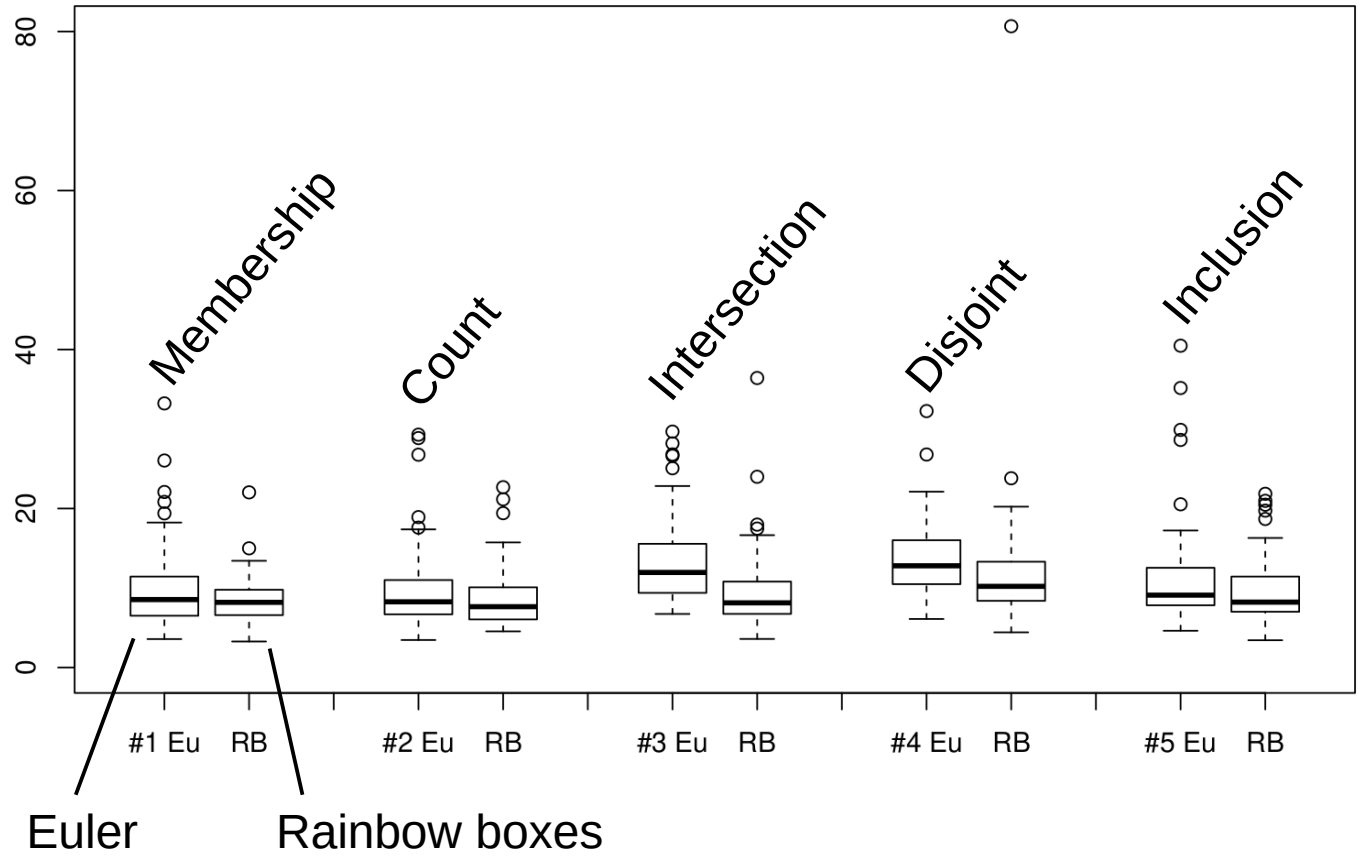
- 38 errors with the Euler diagram (error rate 9.7%, 95% CI: 6.8-12.7)
- 19 errors with rainbow boxes (error rate 4.9%, 2.8-7.0).
- **Significant difference, *p* value = 0.013** (Fisher exact test)
- Diagram and question category are the two significant factors (GLM)

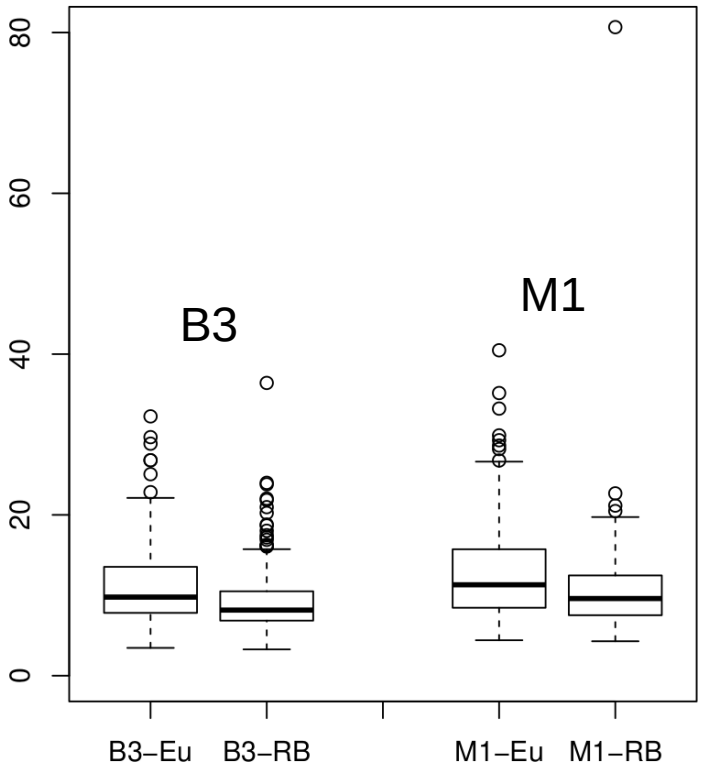




Evaluation results: response times

- 11.57 seconds with Euler diagram (11.02-12.12)
- 9.63 seconds with rainbow boxes (9.10-10.16)
- **Significant difference, *p* value < 10^{-6}** (Welch 2 samples T test on log(time))



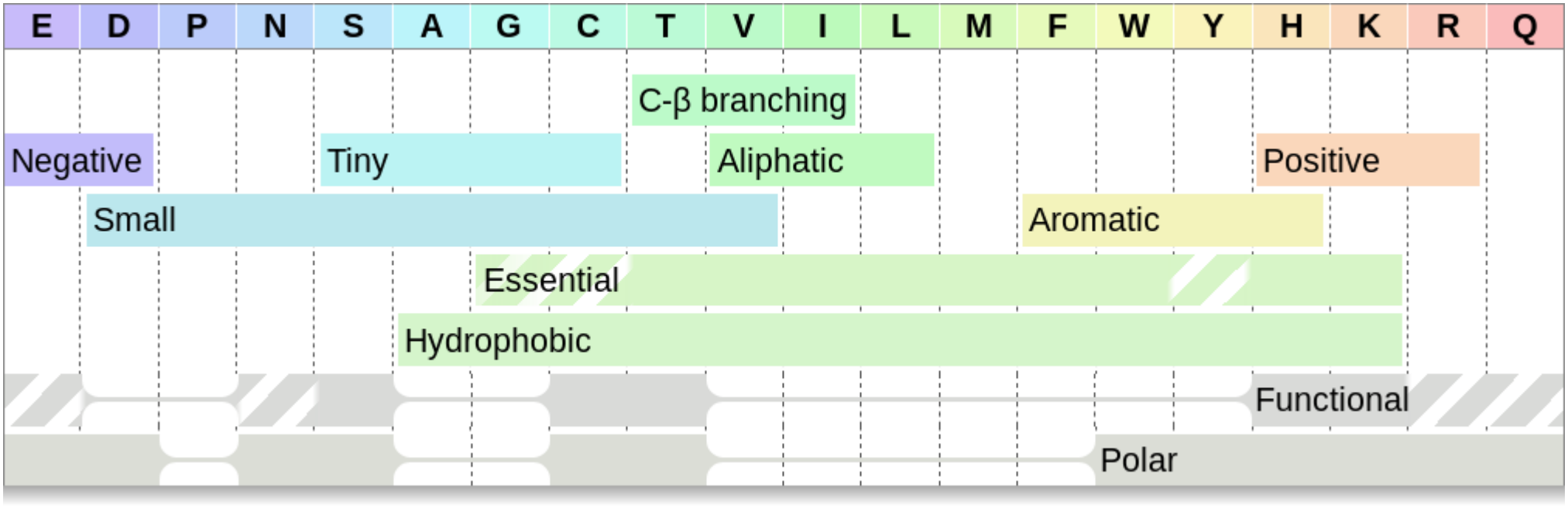


Evaluation results: user preference

- **User preference**
  - 19 students preferred the Euler diagram
  - 51 preferred rainbow boxes
  - 8 indicated no opinion
- **User comments**
  - Students commented on their preference for rainbow boxes or (more rarely) for Euler diagram
  - « Euler diagram is more complex at the beginning, but then “we deal with it” »
  - « Rainbow boxes were easier to read because the boxes were rectangular while the closed areas in the Euler diagram were rounded »

Enriching the diagram

- **3 new properties**
  - **Essential** amino acids: not synthesized by Human cells
  - **C-β branching**: have 2 non-hydrogen substituents attached to their Cβ carbon
  - **Functional**: frequently involved in interaction with non-protein compounds
- **Hatches**
- **Colors highlight similarities**



Discussion

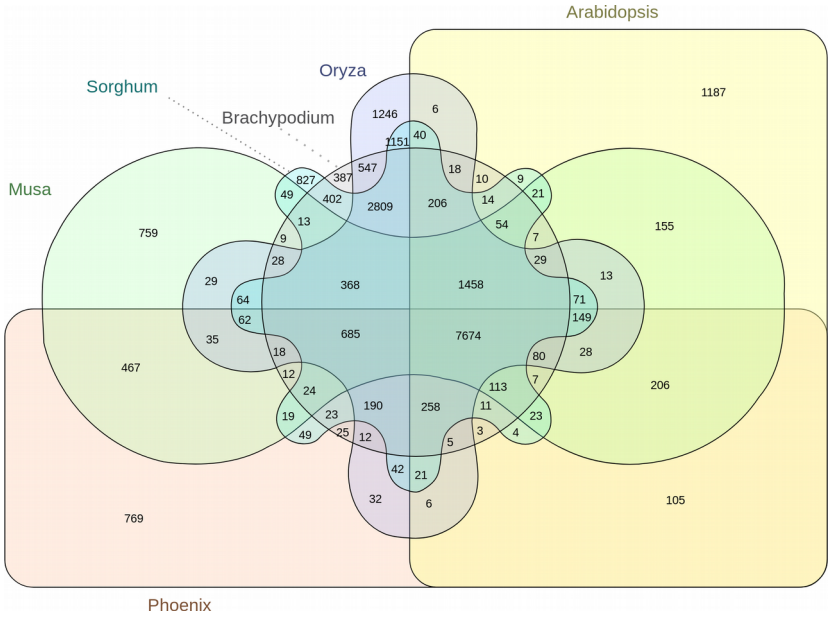
- **Rainbow boxes performed significantly better than Euler diagram, in terms of:**
  - Error rates (efficacy)
  - Response times (efficiency)
  - User preference (satisfaction)

  3 components, possibly independent
- **Rainbow boxes diagram can be enriched with new properties**
  - Euler diagram is already too complex for enrichment
- **Rainbow boxes are easier to produce automatically**
  - Up to 100 elements and 250 sets
  - Euler diagrams are chalenging above 6 sets
    - Up to 60 sets [Simonetto]
- **The new diagram could be used in textbooks and courses**
  - In complement of, or instead of, the Euler diagram

Perspectives

- **Evaluation software => training software**
- **Euler and Venn diagrams are commonly used in biology and medicine**
  - Comparison of gene/protein lists
  - Proportional Venn diagrams showing disease repartition

=> Rainbow boxes might be an alternative



References

[iV2016] : Lamy JB, Berthelot H, Favre M. Rainbow boxes: a technique for visualizing overlapping sets and an application to the comparison of drugs properties. International Conference Information Visualisation (iV) 2016;:253-260

[JVLC] : Lamy JB, Berthelot H, Favre M, Ugon A, Duclos C, Venot A. Using visual analytics for presenting comparative information on new drugs. J Biomed Inform 2017;71:58-69

[Taylor] : Taylor WM. The classification of amino acid conservation. J Theor Biol 1986;119(2):205-218

[Simonetto] : Simonetto P, Auber D, Archambault D. Fully automatic visualisation of overlapping sets. Computer Graphics Forum 2009;28(3):967-974