Etudes métabolomiques: analyse de données complexes Utilisation du Space Filling Design et des méthodes de sélections de variables

Pierre LANTERI & Yohann CLEMENT

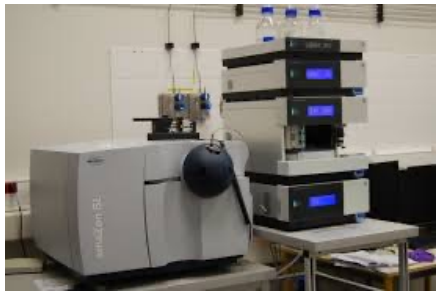
Institut des Sciences Analytiques (UMR 5280)
Université Claude Bernard lyon1 / CNRS
Villeurbanne, France

8 octobre 2021

Outils chimiométriques

Input

LC-MS



RMN

GC-MS

DOE

Stratégie expérimentale

Statistical analyses

Tests univariés
t-student
Correlation
False discovery rate
…

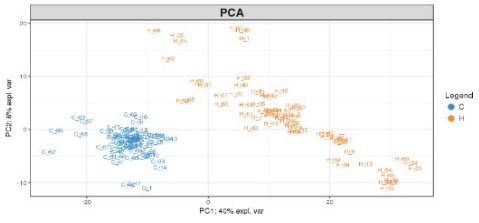
Méthodes descriptives
PCA
ICA
sPCA
GSPPCA
…

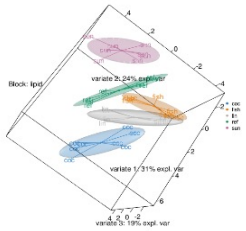
Méthodes supervisées
PLS
PLS-DA
SPLA-DA
OPLS
…

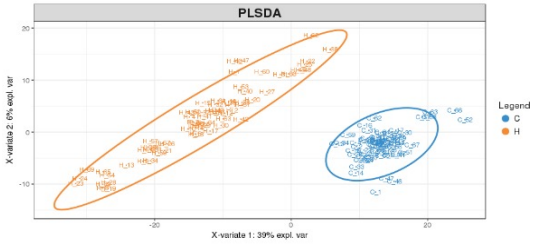
Méthodes Multi-blocks
Multiway PCA
AFM
PARAFAC
COMDIM
…

Graphics

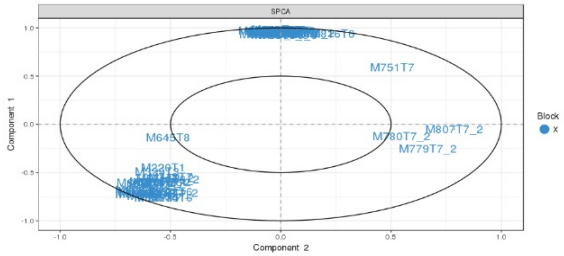
Scores Plot

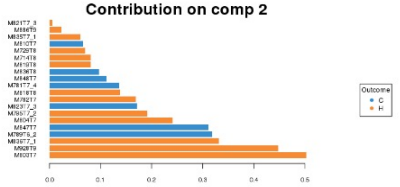


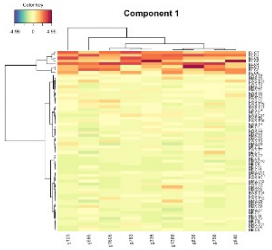


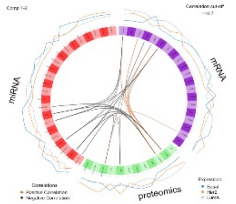


Loadings Plot









SPACE FILLING DESIGN

Théorie et rappels

Sélection des individus

Dans un espace multidimensionnel, la ***caractérisation*** d'un ***« remplissage »*** est ***complexe***. Elle implique de pouvoir répondre à deux questions :

- **1** – la ***distribution*** des points est-elle ***uniforme*** ou non ?
- **2** – l'***espace*** expérimental est-il correctement ***rempli*** ?

Sélection des individus

Pour qualifier l'***uniformité*** de la distribution, on examine l'écart entre une distribution de points donnée et une distribution de points uniforme : cet écart est quantifié par un ***critère*** appelé ***discrépance***.

La ***discrépance*** mesure ***l'écart*** entre une ***distribution*** de points ***uniforme*** et une distribution de points donnée ***dans un cube unité*** multidimensionnel. Ce critère rapporte le ***nombre de points*** contenus dans des pavés au ***volume*** de ces pavés.

pour ***deux*** ensembles d'un ***même nombre*** de points choisis dans un domaine expérimental, le ***meilleur*** sera celui dont la ***discrépance*** est ***la plus faible***

Sélection des individus

Pour qualifier la ***qualité du remplissage*** (la régularité des espacements) on examine la proximité d'une distribution de points donnée à celle d'une grille régulière. Cela fait appel à des ***critères*** utilisant les notions ***de*** « ***distances*** ».

Les ***critères*** les ***plus*** utilisés sont la distance euclidienne, le « recouvrement », le rapport des distances et les distances dites « ***Maximin*** » et « ***Minimax*** ».

Maximin : plan ***D*** qui ***maximise la plus petite*** distance de n'importe quel couple de points de ***D*** dans le domaine expérimental.

Minimax : plan qui ***minimise la plus grande*** distance entre tout point du domaine expérimental et ***D***.

Sélection des individus

Les techniques de remplissage:

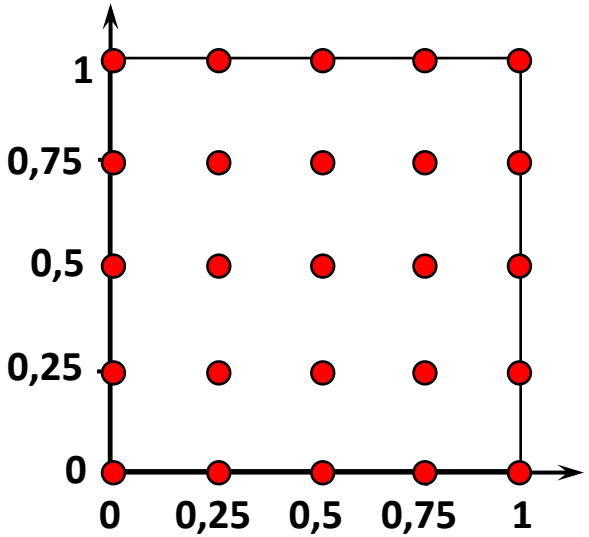
On peut répertorier ***6 méthodes*** principales de ***remplissage régulier*** d'un espace expérimental :

- Le ***maillage*** régulier (à partir de plans factoriels, Box-Behnken, Doelhert)
- L'échantillonage ***aléatoire***
- Les ***hypercubes latins***
- La répartition basée sur la ***notion de distances***
- La répartition basée sur la ***notion de discrépance***
- La méthode de ***Monte Carlo (avec chaînes de Markov)***

Sélection des individus

Maillage factoriel 5^2

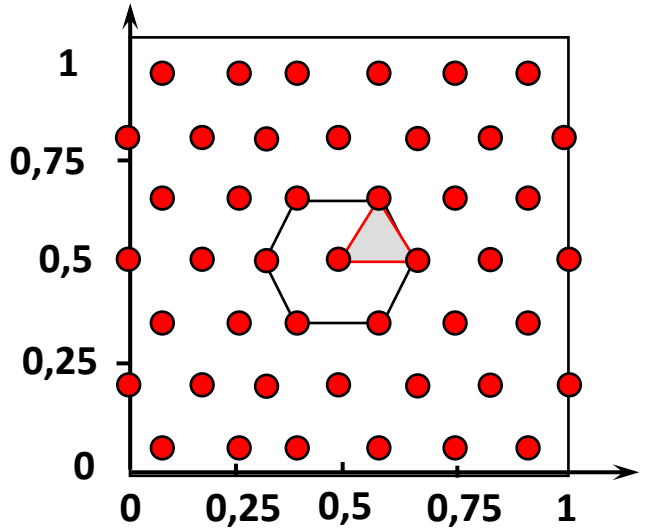
25 points



Maillage régulier:

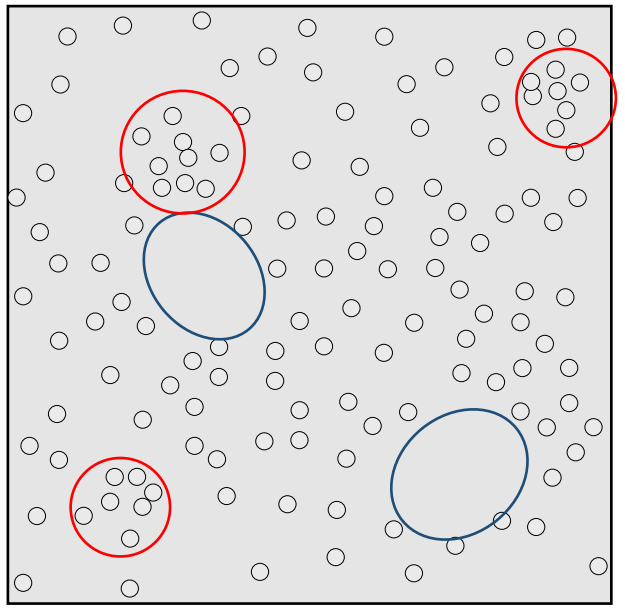
Maillage Doehlert

45 points



Echantillonnage aléatoire:

En général, la sélection obtenue présente à la fois des ***zones peu représentées*** et des ***agrégats de points*** :



Hypercubes latins:

Les ***hypercubes latins*** sont des tableaux ***orthogonaux***, qui appartiennent à la famille des ***plans à « marges uniformes »*** : cela signifie qu'une ***projection*** des points sur un axe ou sur un plan conduit à une ***répartition uniforme***

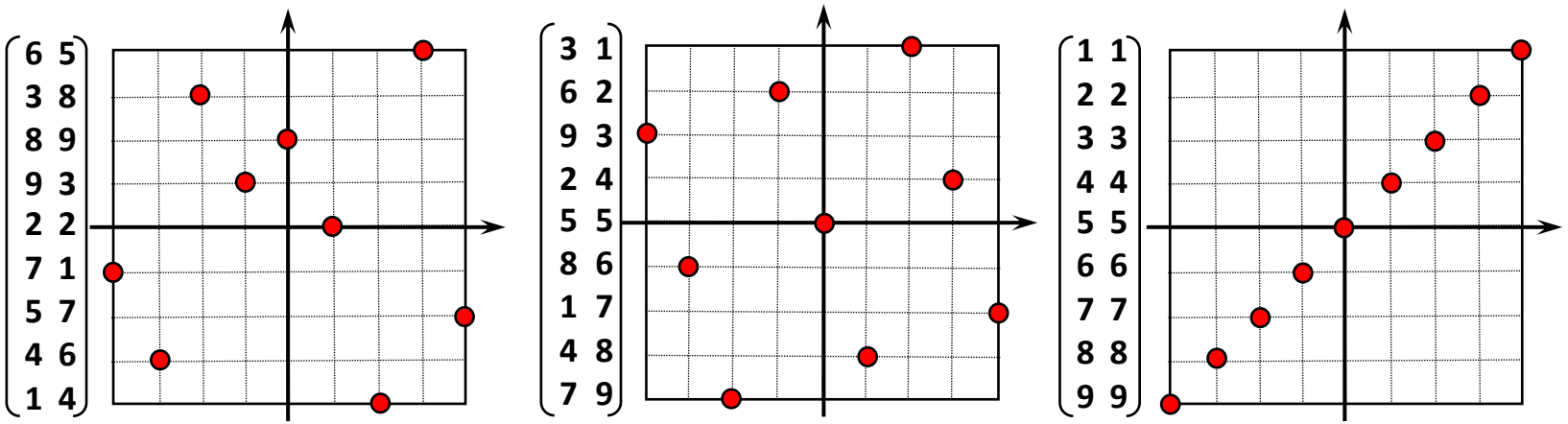
Un hypercube latin ***sélectionne*** alors **n** points parmi les $\mathbf{n^d}$ points d'une grille de façon à ce que les **n** ***niveaux*** des variables d'entrée soient ***testés une fois*** par les simulations

Ex. pour 3 facteurs et 10 niveaux

Exp	X_1	X_2	X_3
1	6	10	10
2	5	1	2
3	10	8	1
4	3	6	3
5	7	3	6
6	2	9	7
7	1	2	4
8	4	4	8
9	9	5	9
10	8	7	5

Sélection des individus

Exemple : ***comparons 3*** hypercubes latins de même taille.



3 critères de comparaison :
remplissage, indépendance, uniformité.

Sélection des individus

Selon la notion de distance

Il est possible d'***utiliser les distances*** pour calculer un critère de remplissage de l'espace permettant de ***choisir un sous ensemble*** de points à partir d'un grand ensemble de candidats.

Le principe de base de ces algorithmes consiste à déterminer ce sous-ensemble en ***améliorant un critère géométrique*** de remplissage de l'espace.

L'uniformité de la répartition des points peut être aussi assurée, en choisissant les points pour qu'ils soient simultanément à une ***distance minimale*** (***Dmin***) de chaque point déjà inclus dans le plan et ***aussi près que possible*** du centre de l'hypercube unité.

A titre d'illustration, nous avons choisi ***l'algorithme WSP*** (**Wooton, Sergent, Phan-Tan-Luu**) qui repose sur celui de *Kennard et Stone* (1969) et de *Wooton* (1975).

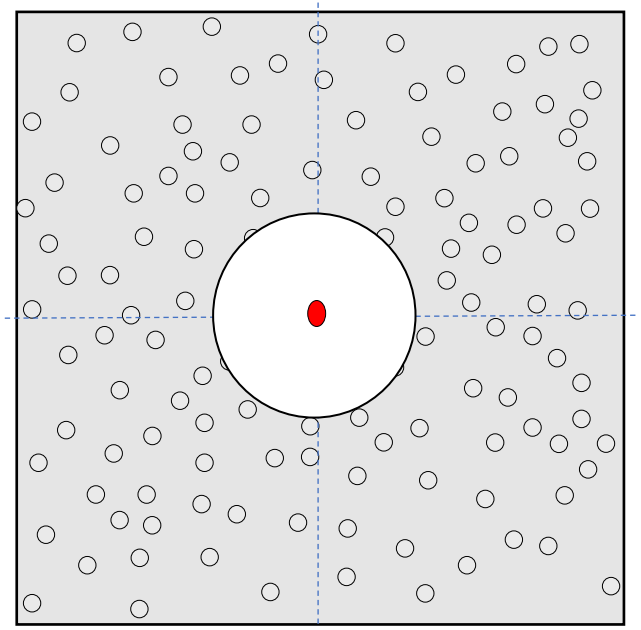
Sélection des individus

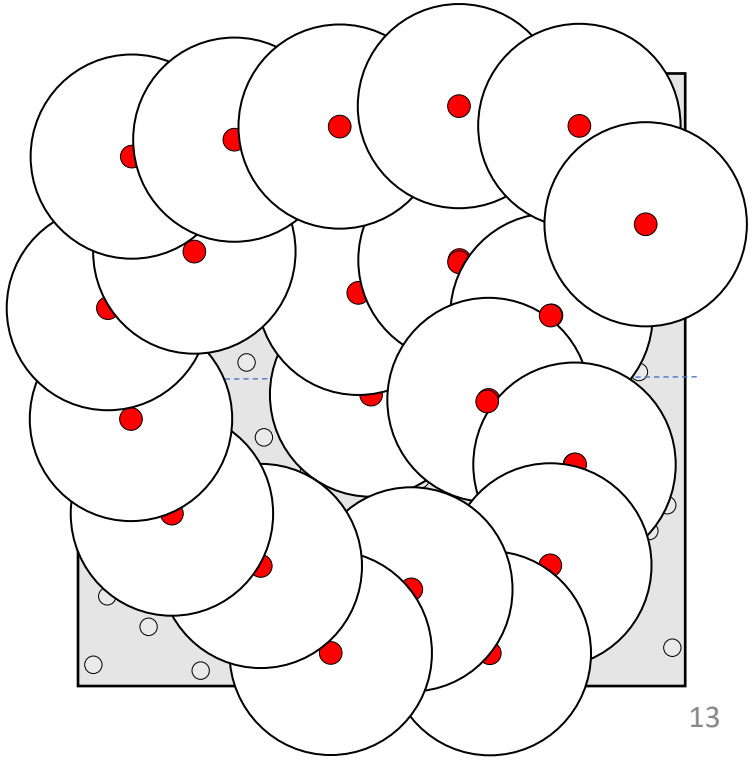
1 – ***Partir*** d'un ensemble de ***points candidats*** (générés ici aléatoirement) ;

2 – ***Choisir le point le plus près du centre*** du domaine comme ***point de départ*** ;

3 – ***Choisir le diamètre Dmin du cercle*** (de la sphère) à l'***intérieur*** duquel (de laquelle) tous les ***points*** seront ***éliminés*** ;

4 – ***réitérer le processus*** à partir du ***point le plus proche*** de ce cercle :

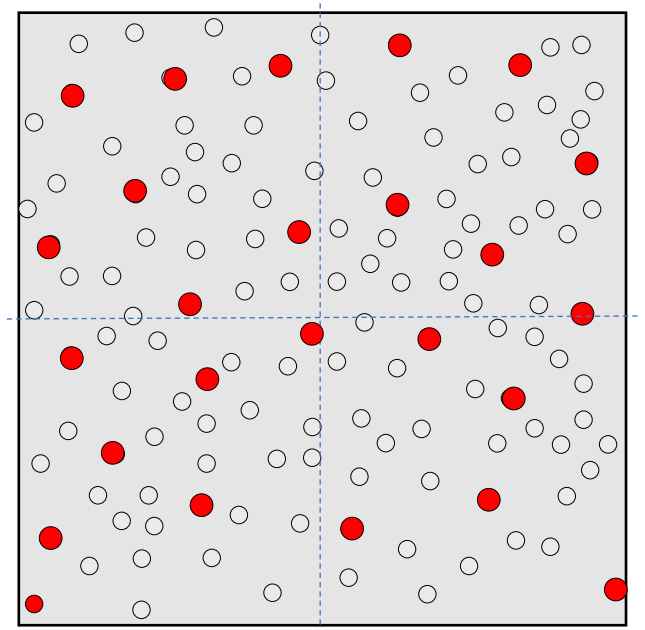




Sélection des individus

Sur cette simulation en deux dimensions, on constate qu'une fois le ***processus terminé*** on obtient une distribution de ***n points répartis uniformément*** dans l'hypercube,

où ***n*** est ≤ au nombre de points de la distribution initiale.



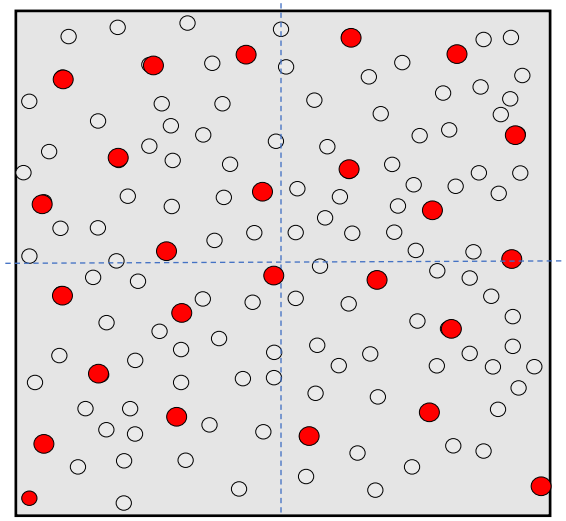
Sélection des individus

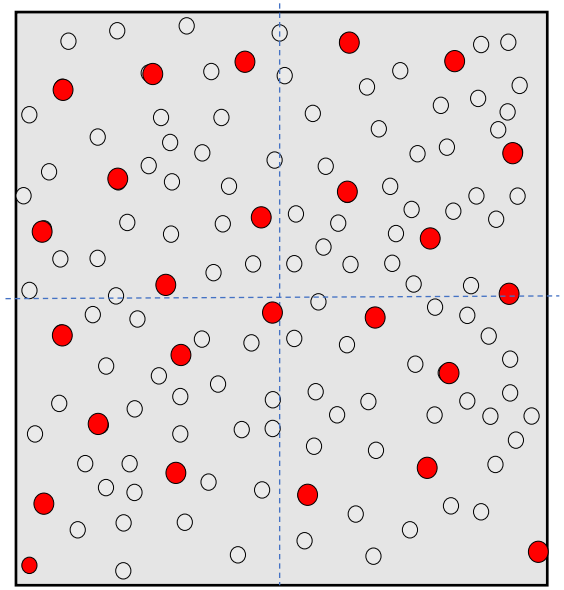
Il est évident que ***plus*** la distance ***Dmin est faible*** et plus le ***nombre de points*** sélectionnés est ***important***.

Une ***extension*** possible est de ***rechercher*** toutes les solutions ***en faisant varier*** la distance minimale.

Cela permet de choisir celle présentant un ***bon compromis entre*** la ***qualité*** du critère ***et*** le ***nombre de points***.

Le ***critère*** de choix généralement ***utilisé*** est le ***rapport entre la distance minimale et*** la ***distance maximale*** des plus proches voisins.

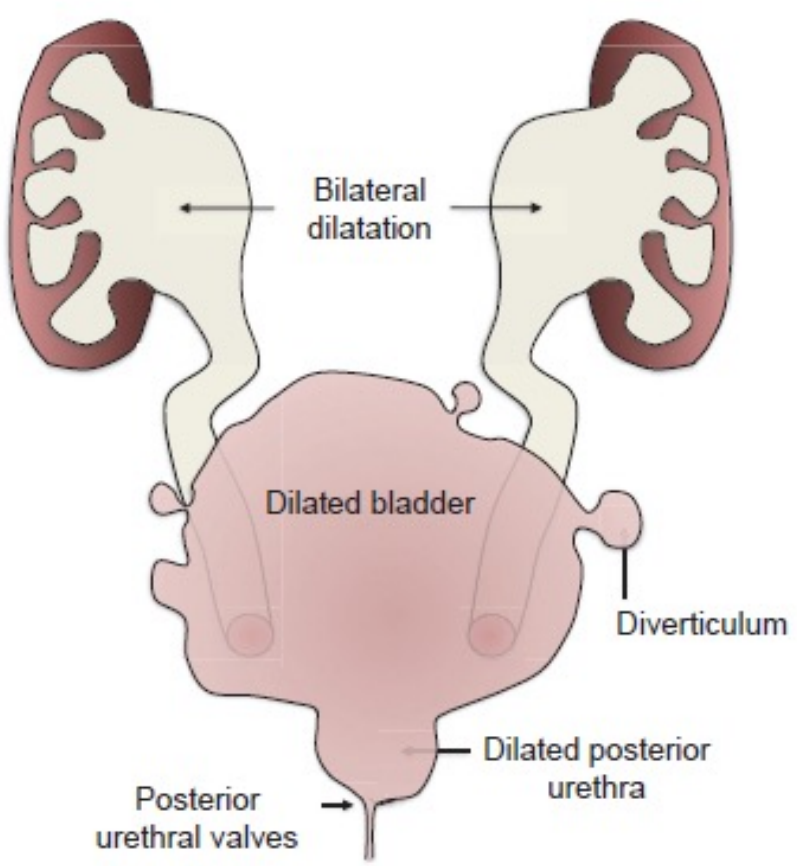




Applications:

Etude par RMN du profil métabolique urinaire chez le nouveau né dans la cas d'une obstruction de la jonction pyélo-urétérale

Le contexte



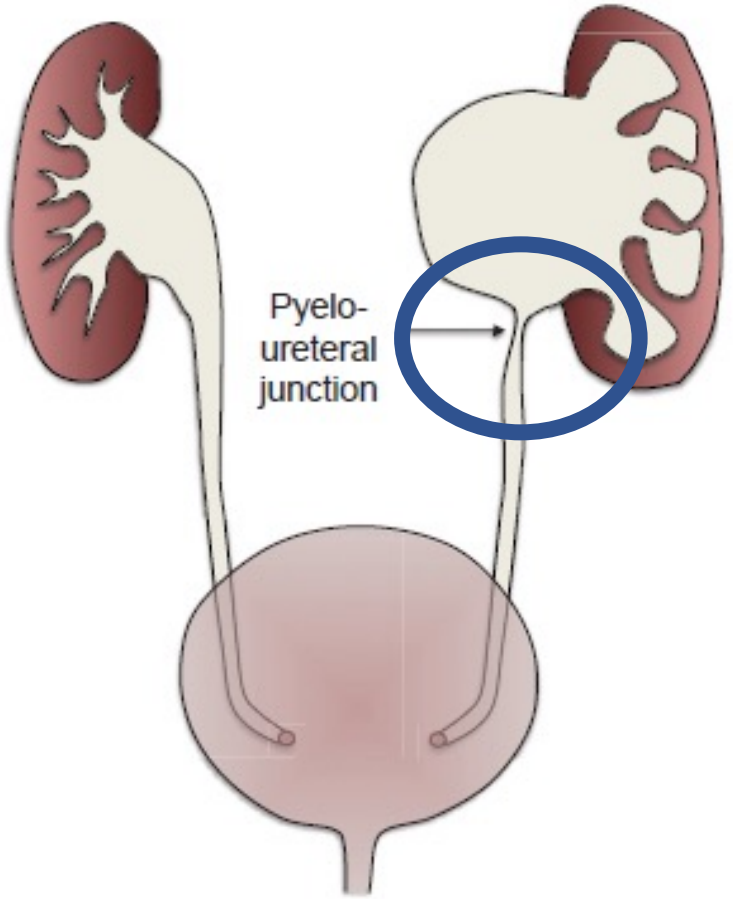


Figure :Schéma d'une anomalie de jonction pyélo-urétérale à droite
Le bassin rénal est dilaté mais l'uretère ne l'est pas.

Le contexte

140 urines de nouveaux nés ont été prélevées:

- 50 nouveaux nés avec symptômes:
  - 24 opérés
  - 26 avec dilatation transitoire
- 90 nouveaux nés témoins

L'ensemble des urines a été prélevé à l'hôpital avec un protocole identique

➢ *Analyse des échantillons par RMN ^{1}H*

1ère étude

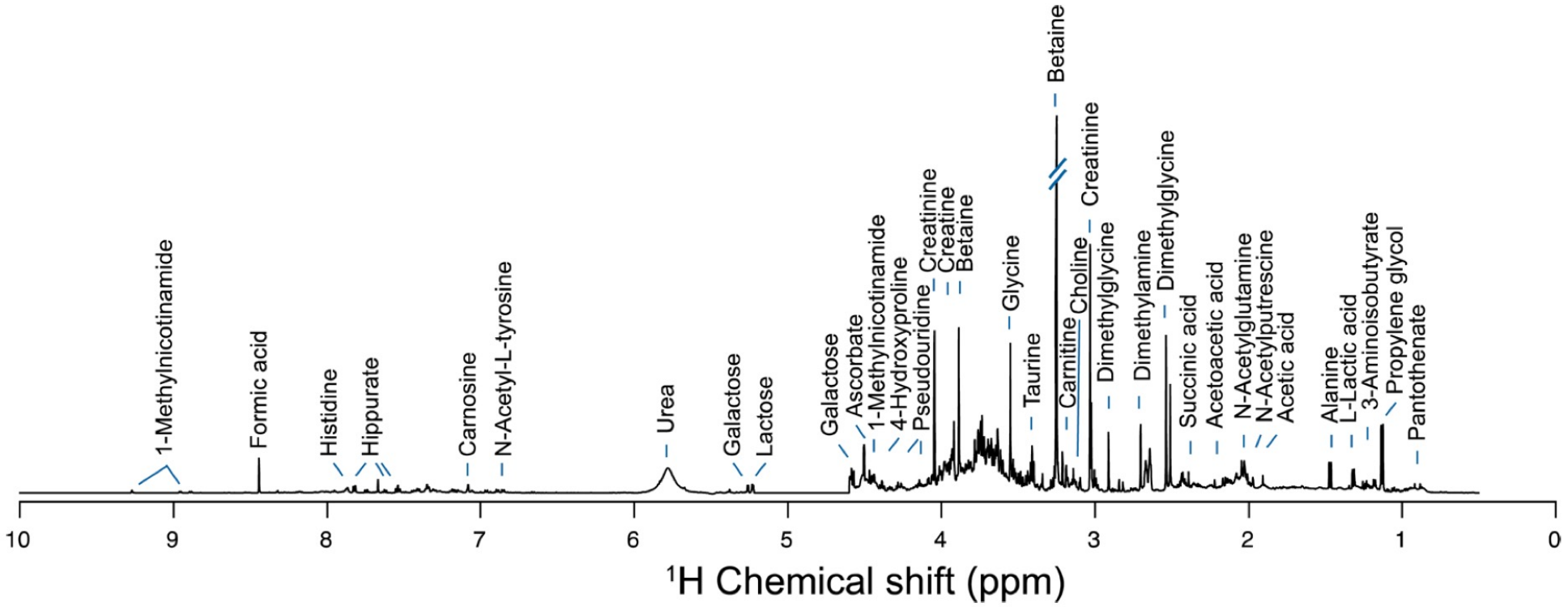


Figure 1. Mean 600 MHz ^{1}H NMR NOESY spectrum from all 90 newborns' urine samples represented with selected metabolites annotations.

Scalabre A. et al, **Evolution of Newborns' Urinary Metabolomic Profiles According to Age and Growth,** *J. Proteome Res.* 2017, 16, 10, 3732–3740

1ère étude

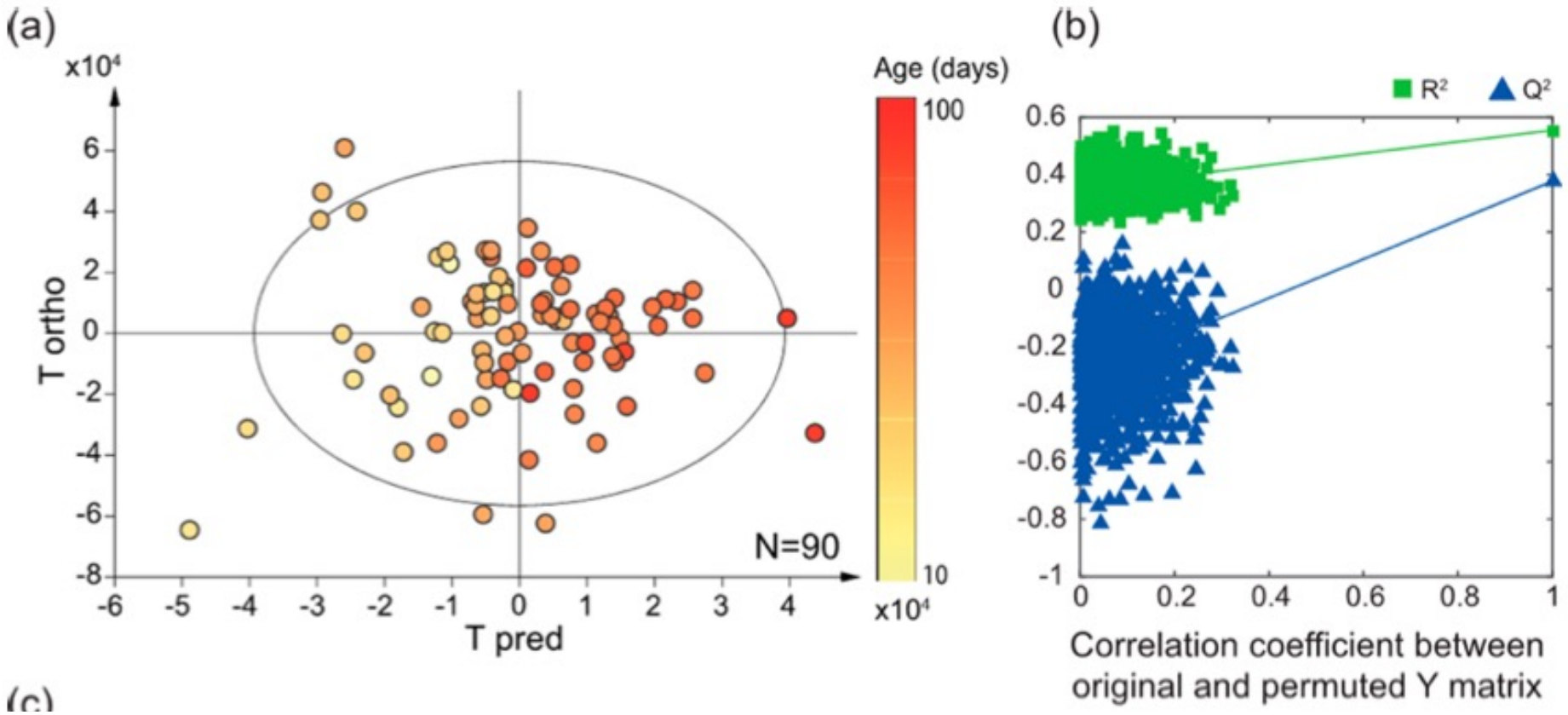


Figure 2. O-PLS model constructed from the 90 ^{1}H NMR spectra of newborns' urine for a regression based on age (1 + 1 components; R^2X = 0.235 ; R^2Y = 0.554; Q^2 = 0.376; *p*-value = 1.3 × 10^{-8} by CV-ANOVA). (a) Score plot; (b) model validation by resampling 1000 times under the null hypothesis; and (c) O-PLS loading plot after SRV analysis and Benjamini–Hochberg multiple testing correction. Statistically significant signals correspond to the colored spectral regions. Significant metabolite variations are summarized in Table 1.

1ère étude

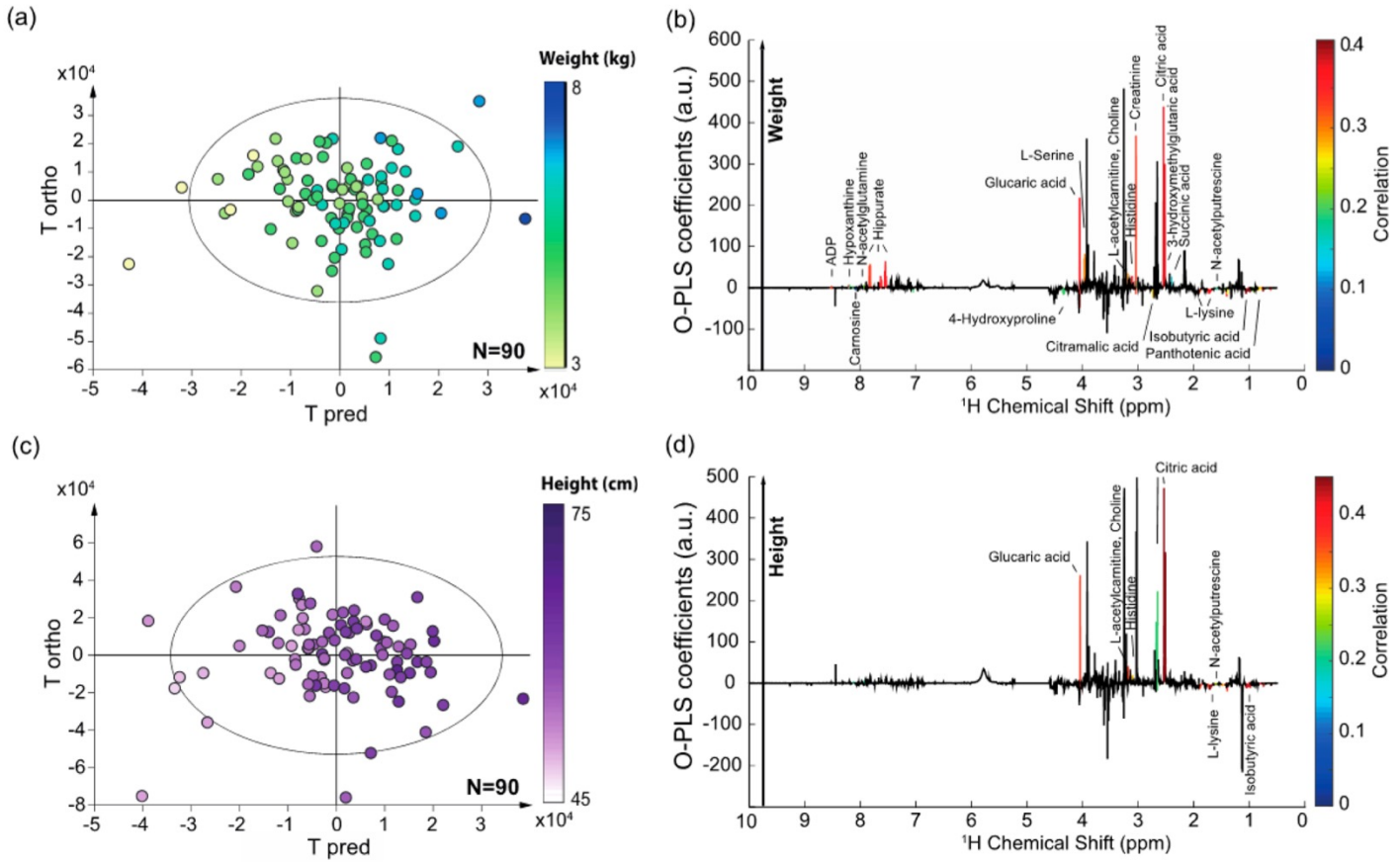


Figure 3. O-PLS regression model against newborns' weight, constructed from the set of 90 ^{1}H NMR spectra (1 + 1 components; $R^2X = 0.11$, $R^2Y = 0.588$, $Q^2 = 0.118$, p-value = 0.03 by CV-ANOVA): (a) score plot and (b) corresponding loadings plot after univariate analysis and Benjamini–Hochberg multiple testing correction. Statistically significant signals correspond to the colored spectral regions. Significant concentration variations correlated with weight are summarized in Table 1. O-PLS regression model against newborns' height (1 + 1 components; $R^2X = 0.196$, $R^2Y = 0.551$, $Q^2 = 0.341$, p-value = 5.1×10^{-7} by CV-ANOVA): (c) score plot, and (d) corresponding loadings plot after univariate analysis and Benjamini–Hochberg multiple testing correction. Significant concentration variations associated with infant height are reported in Table 1.

1ère étude

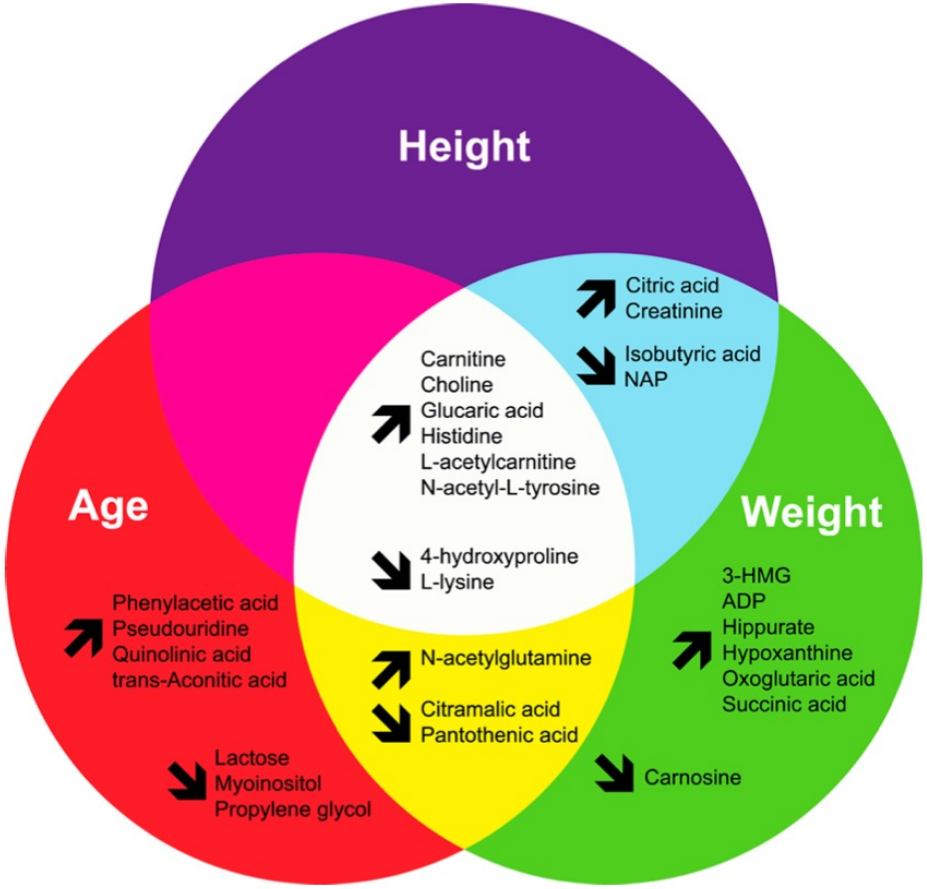
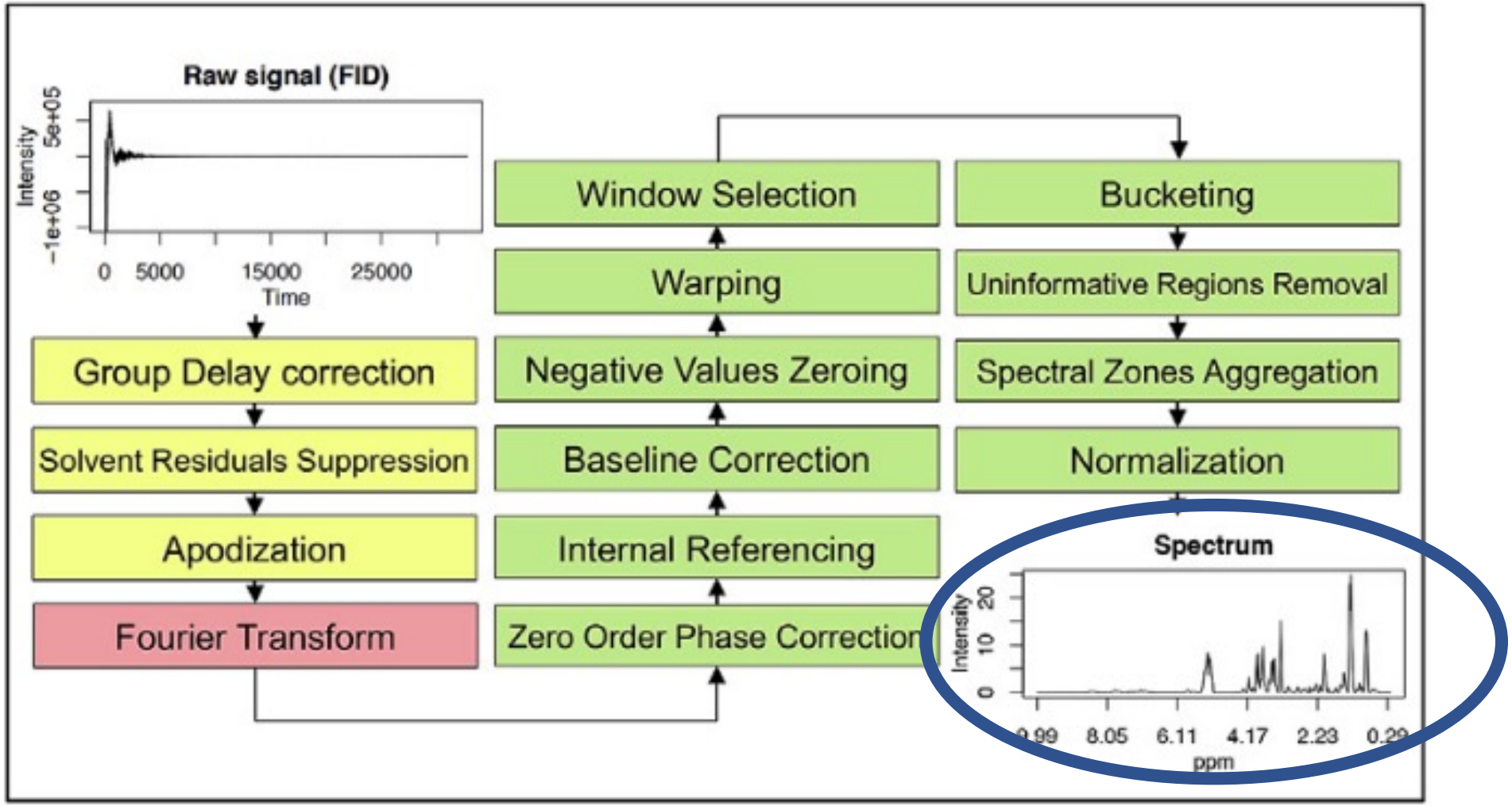


Figure 4. Venn diagram summarizing key metabolites significantly associated with age, weight, and/or height. Metabolites with at least one *q*-value < 0.05 are represented. NAP: *N*-acetylputrescine; 3 HMG: 3-hydroxymethylglutarate.

2ème étude

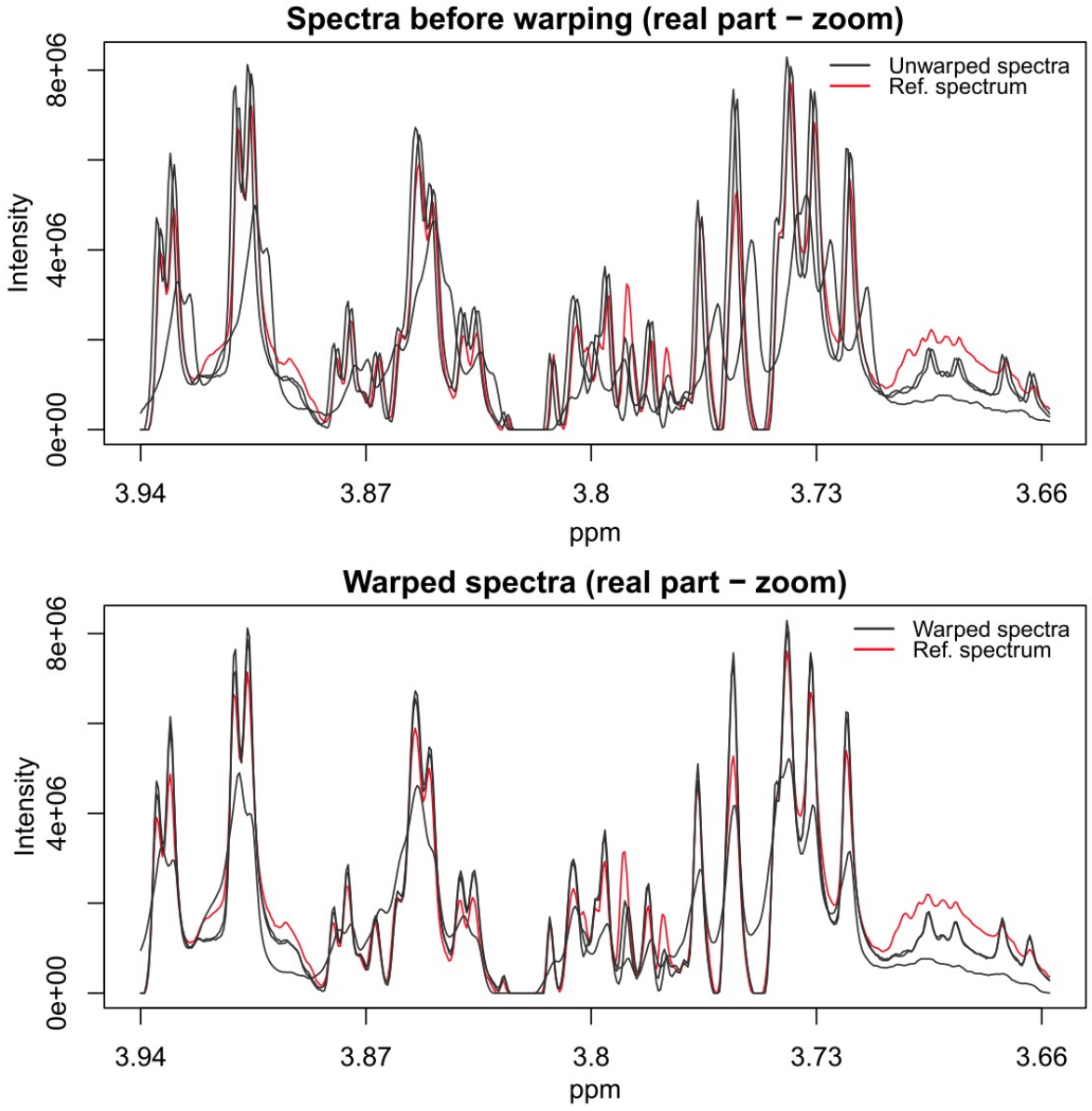
Prétraitement des données



Martin, M er al. (2018). PepsNMR for 1H NMR metabolomic data pre-processing. *Analytica Chimica Acta*, *1019*, 1–13.

Prétraitement des données

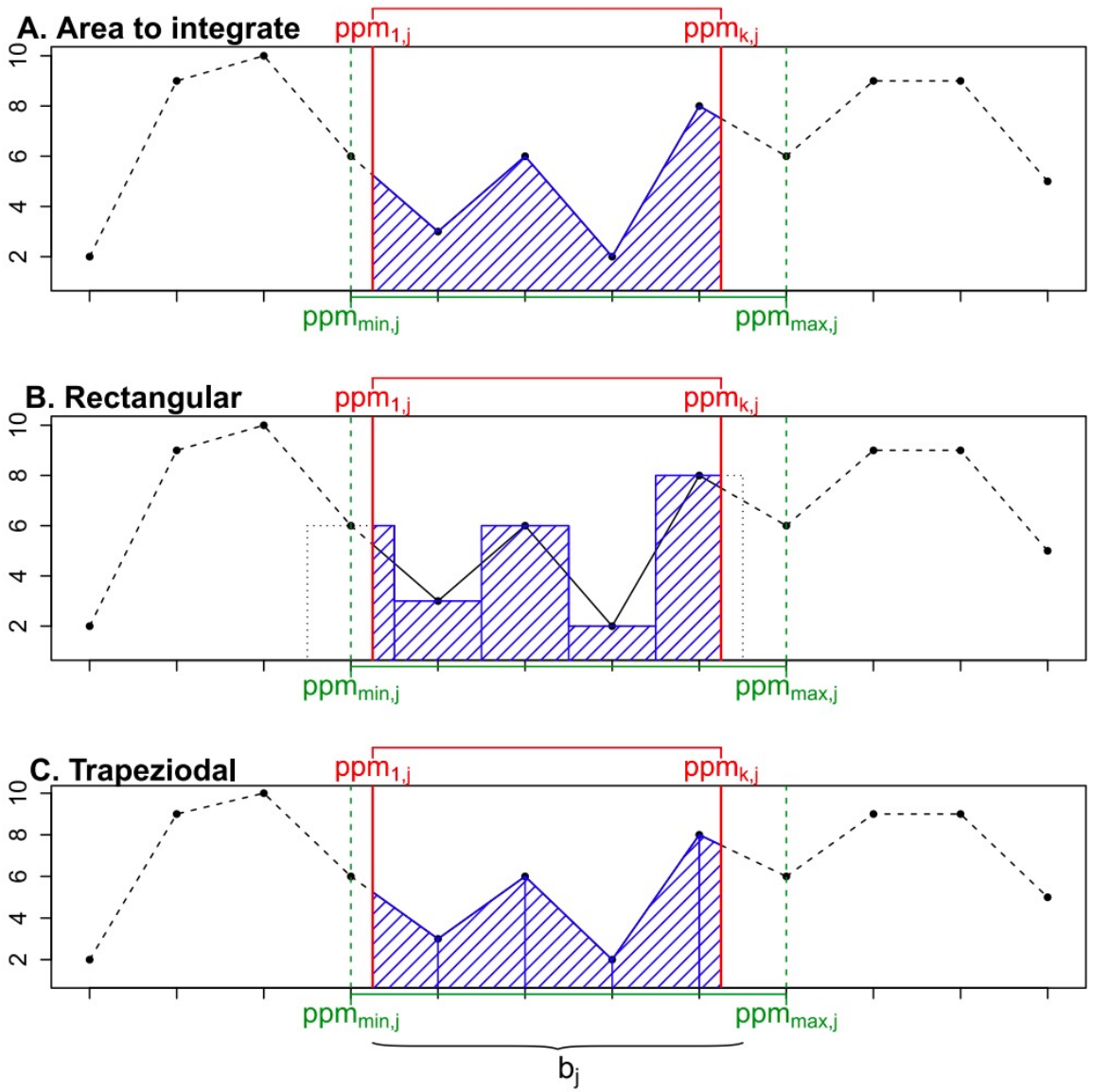
Alignement des spectres:



Prétraitement des données

Bucketing:

- Suppose que les spectres sont alignés
- Les décalages restant sont minimes
- intègre les m intensités spectrales originales dans n intervalles prédéfinis
- Réduction du nombre de variables finales



Sélection des individus

Dans le but d'obtenir une calibration et une validation robuste, une sélection des individus est effectuée par Space Filing Desing (SfD) sur les scores des ACP de chaque groupe

140 urines de nouveaux nés ont été prélevés et donc divisés en 2 groupes par SfD

- 25 nouveaux nés avec symptômes:
  - 12 opérés
  - 13 avec dilatation transitoire
- 45 nouveaux nés témoins

} Calibration

LA NORMALISATION

- Correction de la dérive des signaux (perte d'intensité, erreurs analytiques (bugs) , ….)
- Permet de comparer les spectres entre eux en corrigeant les effets d'échelles
- Pour les études en métabolomiques (urines par exemple) permet de réduire les effets de dilution
- Permet de corriger et comparer les spectres issus de différents lots d'analyses (différents jours par exemple)

Mizuno, H., Ueda, K., Kobayashi, Y., Tsuyama, N., Todoroki, K., Min, J. Z., & Toyo'oka, T. (2017). The great importance of normalization of LC–MS data for highly-accurate non-targeted metabolomics. *Biomedical Chromatography*, *31*(1). Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., Chen, Y., Xue, W., Li, X., & Zhu, F. (2017). NOREVA: Normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Research*, *45*(W1), W162–W170.
Karpievitch, Y. V, Taverner, T., Adkins, J. N., Callister, S. J., Anderson, G. A., Smith, R. D., & Dabney, A. R. (2009). *Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition*. *25*(19), 2573–2580.