Journal of Pharmaceutical and Biomedical Analysis 149 (2018) 46–56

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

# Plasma biomarker discovery for early chronic kidney disease diagnosis based on chemometric approaches using LC-QTOF targeted metabolomics data

S. Benito[a], A. Sánchez-Ortega[b], N. Unceta[a], J.J. Jansen[c], G. Postma[c], F. Andrade[d], L. Aldámiz-Echevarria[d], L.M.C. Buydens[c], M.A. Goicolea[a], R.J. Barrio[a,*]

[a] Department of Analytical Chemistry, University of the Basque Country (UPV/EHU), Faculty of Pharmacy, Paseo de la Universidad 7, 01006 Vitoria-Gasteiz, Spain
[b] Central Service of Analysis (SGiker), University of the Basque Country (UPV/EHU), Laskaray Ikergunea, Miguel de Unamuno 3, 01006 Vitoria-Gasteiz, Spain
[c] Radboud University, Institute for Molecules and Materials (Analytical Chemistry-Chemometrics), P.O. Box 9010, 6500 GL Nijmegen, The Netherlands
[d] Group of Metabolism, BioCruces Health Research Institute, CIBER de Enfermedades Raras (CIBERER), Plaza de Cruces 12, 48903 Barakaldo, Spain

ARTICLE INFO

*Article history:*
Received 26 April 2017
Received in revised form 10 October 2017
Accepted 28 October 2017
Available online 29 October 2017

*Keywords:*
Metabolomics
Biomarker
Chronic kidney disease
Citrulline
S-adenosylmethionine
Symmetric dimethylarginine

ABSTRACT

Chronic kidney disease (CKD) is a progressive pathological condition in which renal function deteriorates in time. The first diagnosis of CKD is often carried out in general care attention by general practitioners by means of serum creatinine (CNN) levels. However, it lacks sensitivity and thus, there is a need for new robust biomarkers to allow the detection of kidney damage particularly in early stages. Multivariate data analysis of plasma concentrations obtained from LC-QTOF targeted metabolomics method may reveal metabolites suspicious of being either up-regulated or down-regulated from urea cycle, arginine methylation and arginine-creatine metabolic pathways in CKD pediatrics and controls. The results show that citrulline (CIT), symmetric dimethylarginine (SDMA) and S-adenosylmethionine (SAM) are interesting biomarkers to support diagnosis by CNN: early CKD samples and controls were classified with an increase in classification accuracy of 18% when using these 4 metabolites compared to CNN alone. These metabolites together allow classification of the samples into a definite stage of the disease with an accuracy of 74%, being the 90% of the misclassifications one level above or below the CKD stage set by the nephrologists. Finally, sex-related, age-related and treatment-related effects were studied, to evaluate whether changes in metabolite concentration could be attributable to these factors, and to correct them in case a new equation is developed with these potential biomarkers for the diagnosis and monitoring of pediatric CKD.



## 1. Introduction

Chronic kidney disease (CKD) is a major worldwide public health problem, affecting both children and adults, in which kidney function declines progressively. In clinical practice, different equations based on creatinine (CNN) concentration are used to estimate the glomerular filtration rate (GFR). This value reflects kidney function and is useful for the diagnosis of CKD and for assigning the stage or degree of the disease. Detection of CKD is considered a priority for primary care, because early treatment of CKD and its complications may delay or prevent the development of end-stage renal disease (ESRD) [1]. In clinical practice, one of the most commonly used equation in pediatrics to diagnose the disease for the first time is the Schwartz formula which is based on height, serum CNN concentration and $k$ coefficient [2], as shown in Eq. (1).

$$GFR\ (\mathrm{mL\,min^{-1}/1.73\,m^2}) = k \times Height\ (\mathrm{cm})/Serum\ creatinine\ (\mathrm{mg\,dL^{-1}}) \quad (1)$$

$k$ coefficient has changed along the years, and is 0.45 for first year term infants, 0.55 for children and adolescent girls and 0.7 for adolescent boys currently.

Even if CNN is the classic biomarker used for the assessment of renal function in primary care attention, it has several drawbacks. Indeed, it lacks sensitivity and often reveals kidney damage when an important nephronic loss has already occurred. For that reason, in several early CKD patients a proper diagnosis by the general practitioner (GP) using the available CNN-based screening blood or urine tests is not possible until the disease progresses or more specific tests like abdominal computed tomography scan, abdominal

* Corresponding author.
*E-mail address:* r.barrio@ehu.es (R.J. Barrio).

https://doi.org/10.1016/j.jpba.2017.10.036
0731-7085/

ecography, kidney histopathology and immunohistochemistry or renal scintigraphy are carried out by nephrologists [3].

It would be ideal for GPs to be able to carry out better screening including more sensitive biomarkers for CKD in addition to CNN, as it is estimated that the majority of the population visits their GP within a 3-year period and can be subjected to screening [1]. Screening tests can be performed using either urine or blood biofluids. The utility of urinalysis is at times overestimated due to the inaccuracies in quantitatively collection of urine [4]. For that reason, blood analysis is preferred in children for diagnostic purposes. Thus, there is a need for new biomarkers (in addition to CNN) to be included in the equations used by GP in screening blood tests. Indeed, an earlier diagnosis of CKD, a better approximation to the CKD stage defined by nephrologists, monitoring of the progression of the disease and evaluation of the response to therapy are required. The early detection of CKD and the approximation of the CKD stage due to the implementation of new biomarkers in the screening tests carried out by GP would allow the early referral to the nephrologist, often leading to a better outcome of the patient.

CKD is associated with alterations in multiple metabolic pathways [5]. Arginine-creatine metabolic pathway, arginine methylation and the urea cycle were suspicious to be affected in pediatric patients with CKD and thus, some metabolites from these metabolic pathways were expected to be increased or decreased in comparison to control pediatric patients. It has to be taken into account that depending on the metabolic pathways, differences in fold change concentration of metabolites can be lower or higher. For instance, the concentration of metabolites in the central metabolism is relatively constant. Concentrations of metabolites present in secondary metabolism-related pathways may differ more in concentration, depending on environmental conditions. Indeed, all biological systems are easily perturbed by a number of intra-individual or inter-individual experimental or environmental factors, such as age, diet, growth phase, media, nutrients, pH, sex, and temperature, which should be taken into consideration. This is known as induced biological variation [6]. Central metabolic pathways include glucolysis, the pentose phosphate pathway, the tricarboxylic acid cycle, anaplerotic reactions and biosynthetic pathways of fatty acids and amino acids, and those reactions not included in central metabolic pathways are considered intermediate or secondary metabolic pathways [7]. The urea cycle is considered a central metabolic pathway, as it is part of arginine biosynthesis metabolic pathway, whereas arginine-creatine metabolic pathway and arginine methylation are not included in the central metabolic pathway, and thus are considered secondary metabolic pathways.

Metabolomics aims at studying the dynamic changes, interactions and responses to stimuli of metabolites in different metabolic pathways [8]. The feasibility of metabolomics for biomarker discovery is supported by the assumption that metabolites play an important role in biological systems and that diseases cause disruption of biochemical pathways [9]. It has to be taken into account that each biofluid contains a large number of metabolites with concentration levels that can vary by orders of magnitude. However, from a biological point of view, metabolites present in high concentrations are not necessarily more important than those at low concentrations [6].

For all these reasons, in addition to the careful planning of experiments and analytical measurements, statistical and chemometric pre-processing are essential. Indeed, chemometrics, defined as the art of extracting chemically relevant information from data produced in chemical experiments, is indispensable to obtain consistent information and discard irrelevant information [10].

The aim of this work has been to perform data analysis on the plasma concentrations of 16 metabolites from arginine-creatine, urea cycle and arginine methylation metabolic pathways in thirty-two patients at different stages of CKD and twenty-four control patients not suffering from CKD in order to find new potential biomarkers. These metabolites were selected because they were suspicious of being altered in pediatrics with CKD and were measured by means of a recently developed ion-pairing LC-QTOF-MS methodology targeted at these metabolites [11]. To turn these measurements into scores, we used complementary chemometric tools to extract the diagnostically relevant information that remain unseen with the naked eye. We have also identified the potential biomarkers for early diagnosis of CKD and checked whether these metabolites could be affected by age, sex or treatment received.

## 2. Material and methods

### 2.1. Chemicals and reagents

Acetonitrile used for both standard preparation and mobile phase was supplied by Scharlau (Sentmenat, Spain). In addition, LC–MS grade ammonium formate eluent additive from Fluka Analytical, Sigma-Aldrich (Steinheim, Germany) and perfluoroheptanoic acid (PFHA) from Acros Organics (New Jersey, USA) completed the mobile phase. Standard preparation required the use of LC–MS grade methanol from Scharlau (Sentmenat, Spain), ultrahigh purity water obtained from pretreated tap water by means of Elix reverse osmosis followed by a Milli-Q system from Millipore (Bedford, MA, USA) and chlorhydric acid, obtained from Merck (Darmstadt, Germany) as well.

Amino acid and amino acid derivative standards were supplied by different manufacturers. L-Cysteine (CYS), creatine (CTN), betaine (BET) and reduced glutathione (GSH) were supplied by TCI (Tokyo, Japan). L-Methionine (MET), L-arginine (ARG), glycine (GLY), L-homocysteine (HCYS), *S*-adenosyl-L-homocysteine (SAH), NG,NG′-dimethyl-L-arginine di(*p*-hydroxyazobenzene-p′-sulfonate) (SDMA), NG,NG-dimethylarginine dihydrochloride (ADMA), *S*-adenosyl-L-methionine (SAM) and citrulline (CIT) were provided by Sigma-Aldrich (Steinheim, Germany). Fluorochem (Hadfield, UK) provided *N,N*-dimethylglycine (DMG) and homoarginine (HARG) and creatinine (CNN) was pursached from Alfa Aesar (Karlsruhe, Germany). Finally, creatine-$d_3$ $H_2O$ from CDN Isotopes (Quebec, Canada) and SDMA-$d_6$, and creatinine-$d_3$ (methyl-$d_3$) from Toronto Research Chemicals, TRC Canada (North York, Canada), and glutathione-$^{13}C_2{}^{15}N$ from Cambridge Isotope Laboratories (Andover, MA, USA) were used as internal standards. Dithiothreitol obtained from Fisher Scientific (Pittsburgh, PA, USA) was used as thiol reductant agent.

### 2.2. LC/QTOF method

Chromatographic analysis of plasma samples was carried out on an Agilent 1200 Series HPLC system coupled to Agilent 6530 Series hybrid quadrupole time-of-flight mass spectrometer (Agilent Technologies, Santa Clara, CA, USA). Agilent Jet Stream ESI was used as an ion source. Chromatographic separation and mass spectrometry are further described in a previous work [11]. This validated ion-pairing LC/QTOF methodology was used to quantify 16 metabolites from arginine-creatine metabolic pathway, arginine methylation and urea cycle in these plasma samples: ADMA, ARG, BET, CIT, CYS, CTN, CNN, DMG, GSH, GLY, HARG, HCYS, MET, SAH, SAM and SDMA. It has to be noted that for the aminothiols (CYS, GSH and HCYS) total concentrations were quantified by means of a reduction process in sample treatment with dithiothreitol. Creatine-$d_3$ was used to correct the signal of GLY, CIT, DMG, BET, CYS and CTN, whereas gluthatione-$^{13}C_2{}^{15}N$ corrected GSH signal. Similarly, creatinine-$d_3$ adjusted for the concentration of HCYS, ARG, MET, CNN and HARG, and SDMA-$d_6$ was used for ADMA, SDMA, SAM and SAH adjustment.

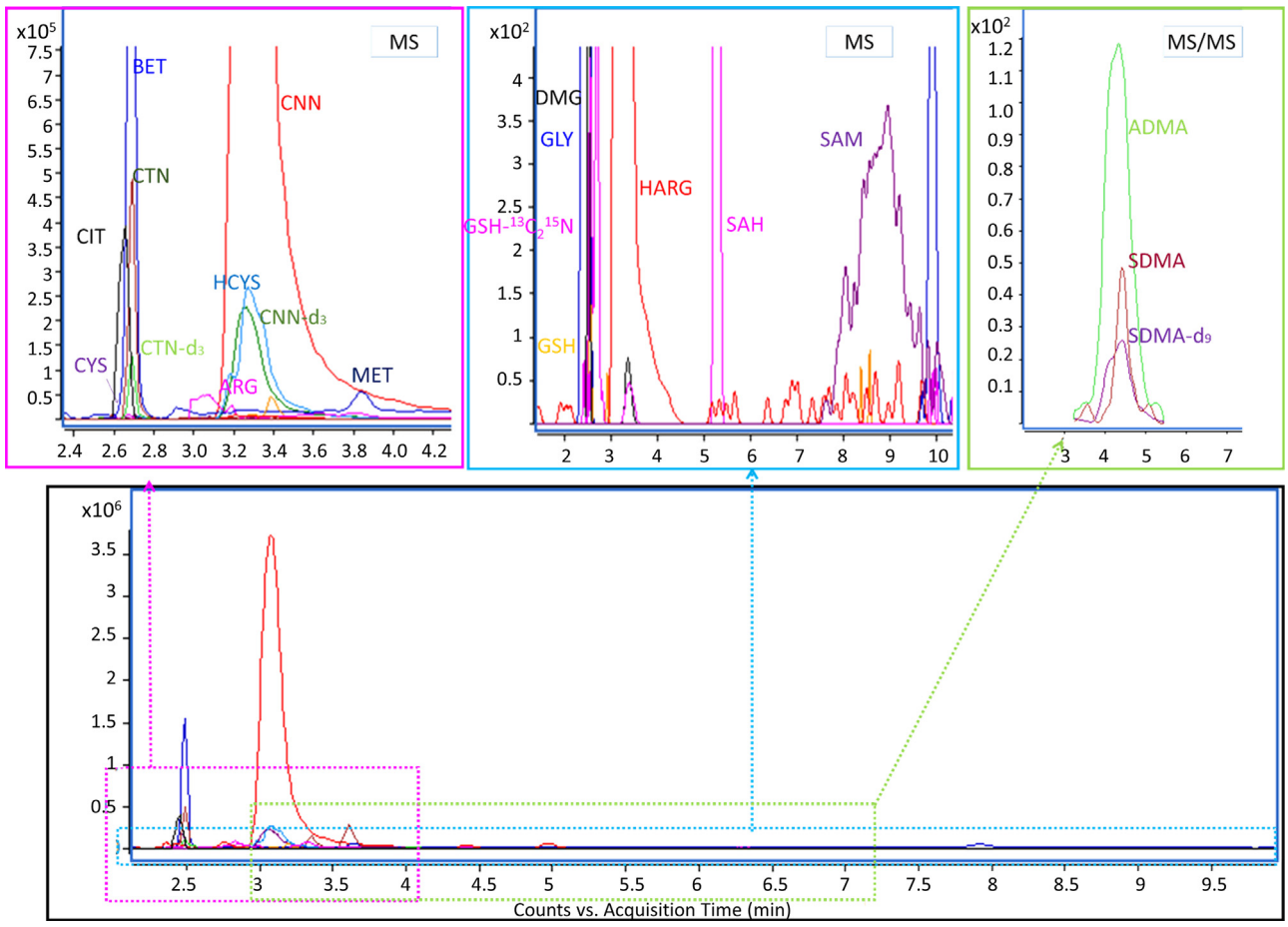



**Fig. 1.** Chromatogram of a plasma sample containing extracted ion chromatograms in MS/MS mode for ADMA and SDMA, and in MS mode for the rest of the analytes.

All the metabolites except for ADMA and SDMA were acquired in MS mode, with a scan rate of 2.2 spectra $s^{-1}$ in 50–1000 *m/z* mass range. Taking into account that ADMA and SDMA are structural isomer compounds and elute at the same retention time, MS/MS mode with a scan rate of 3 spectra $s^{-1}$ and a mass range from 30 to 1000 *m/z* was used with a collision energy of 10 V, which allows the detection of different ion products for each of the analytes. The protonated form [M+H$^+$] of the analytes was detected for all the analytes in MS mode, and the specific transitions 203.1503 > 46.0657 *m/z* and 203.1503 > 172.1086 *m/z* were monitorized for ADMA and SDMA, respectively in MS/MS mode (Table 1). Fig. 1 shows an example chromatogram of a plasma sample containing both MS and MS/MS extracted ion chromatograms.

Regarding EU's decision 2002/657/CE proposed for the collection of identification points, confirmation of the presence of the analytes in samples is possible with this methodology, as retention time of each analyte and the accurate mass measurement of the *m/z* with an error less than 5 ppm are used [12].

**Table 1**
Retention times, accurate *m/z* rations in MS and MS/MS modes and the internal standard (IS) used for signal correction for each examined analyte.

| Compound | RT (min) | *m/z* | IS |
|---|---|---|---|
| GLY | 2.40 | 76.0393 | Creatine-$d_3$ |
| CIT | 2.43 | 176.1030 | Creatine-$d_3$ |
| DMG | 2.46 | 104.0706 | Creatine-$d_3$ |
| BET | 2.50 | 118.0865 | Creatine-$d_3$ |
| CYS | 2.52 | 122.0270 | Creatine-$d_3$ |
| CTN | 2.54 | 132.0768 | Creatine-$d_3$ |
| GSH | 2.59 | 308.0911 | Glutathione-$^{13}C_2{}^{15}N$ |
| HCYS | 2.71 | 136.0427 | Creatine-$d_3$ |
| ARG | 3.08 | 175.1193 | Creatinine-$d_3$ |
| MET | 3.11 | 150.0583 | Creatinine-$d_3$ |
| CNN | 3.23 | 114.0667 | Creatinine-$d_3$ |
| HARG | 3.53 | 189.1346 | Creatinine-$d_3$ |
| ADMA | 4.58 | 203.1503 > 46.0657 | SDMA-$d_6$ |
| SDMA | 4.74 | 203.1503 > 172.1086 | SDMA-$d_6$ |
| SAH | 5.37 | 385.1289 | SDMA-$d_6$ |
| SAM | 8.52 | 399.1445 | SDMA-$d_6$ |
| Glutathione-$^{13}C_2{}^{15}N$ | 2.59 | 311.0954 | – |
| Creatine-$d_3$ | 2.54 | 135.0956 | – |
| Creatinine-$d_3$ | 3.23 | 117.0850 | – |
| SDMA-$d_6$ | 4.74 | 209.1879 > 175.1274 | – |

## 2.3. Study samples

Thirty-two patients suffering from chronic kidney disease aged 3–17 years were recruited for this study according to the following criteria: subjects were followed up in Cruces University Hospital and clinically stable at the time of the study. The exclusion criteria include suffering from hepatopathy, anuria and/or insulin-dependent diabetes mellitus. The glomerular filtration rate (GFR) for each patient was calculated in the Pediatric Nephrology Service at Cruces University Hospital using Schwartz formula. Then, patients were classified by nephrologists according to glomerular filtration rate (GFR) and nephrological criteria based on complementary tests into stages CKD2 (60–89 mL/min/1.73 $m^2$), CKD3 (30–59 mL/min/1.73 $m^2$), CKD4 (15–29 mL/min/1.73 $m^2$) and CKD5 (<15 mL/min/1.73 $m^2$) of the disease [13]. In addition to the CKD stage, information regarding age, sex, and whether they received renal replacement therapy (dialysis or transplant) or not was available as well. For comparison purposes, 24 patients not suffering

**Table 2**
Characteristics of the patients involved in the study.

| CKD STAGE | Characteristics of the population | | | Number of patients (n) |
|---|---|---|---|---|
| | SEX (M/F) | AGE (2–12 y/13–18 y) | TREATMENT Not treated/Dialyzed/Transplanted | |
| CONTROL | 18/6 | 15/9 | 24/0/0 | 24 |
| CKD2 | 8/6 | 10/4 | 9/0/5 | 14 |
| CKD3 | 5/1 | 2/4 | 4/0/2 | 6 |
| CKD4 | 2/4 | 3/3 | 5/0/1 | 6 |
| CKD5 | 2/4 | 5/1 | 1/5/0 | 6 |

from chronic kidney disease aged 6–18 years were recruited for the study (Table 2).

Blood samples were withdrawn in the morning after an overnight fasting and were immediately cooled in an ice-water bath. Then, they were centrifuged at 1000*g* for 5 min at 4 °C. Samples were stored at −80 °C until sample treatment and analysis were carried out [11].

The study protocol was approved by Cruces Hospital Ethics Committee of Clinic Research and informed consent was given by patients' parents.

### 2.4. Sample treatment

First of all, samples were thawed and pooled plasma was made of both control and CKD samples to be used after quantifying it by standard additions to make a calibration curve in pooled plasma. In addition, 50 μL of plasma were placed in an Eppendorf tube to quantify each plasma samples in triplicate. Plasma samples were then spiked with 10 μL of a mix solution containing 15 μg mL$^{-1}$ of creatinine-$d_3$, creatine-$d_3$, and glutathione-$^{13}C_2{}^{15}N$ and 5 μg mL$^{-1}$ of SDMA-$d_6$. Then, 50 μL of dithiothreitol (77,000 mg L$^{-1}$) were added and samples were incubated for 15 min at room temperature to perform the reduction of aminothiol compounds. In addition, 150 μL of cold acetonitrile were added to the mixture for plasma protein precipitation and was vortexed before centrifuging it for 10 min at 13,000 rpm at 4 °C. The obtained supernatant was then transferred to chromatographic vials, evaporated in nitrogen stream and the residue was reconstituted in 200 μL of ammonium formate 5 mM. Finally, these vials were transferred to autosampler for the analysis by means of the LC-QTOF method.

### 2.5. Chemometric analysis

#### 2.5.1. Multivariate analysis

Principal Component Analysis (PCA) was carried out on logarithm transformed, autoscaled data obtained from the analysis of plasma samples with the developed LC-QTOF method. Matlab R2015a (Mathworks, Natick, Massachusetts, United States) software was used for performing the multivariate analysis. Outlying samples were defined according to Hierarchical Clustering with single, complete and average linkage and removed.

We compared several multivariate classification methods in their predictive ability for CKD, being Maximum Likelihood – Linear Discriminant Analysis (ML-LDA) [14], Fisher-Linear Discriminant Analysis (Fisher-LDA) [15], Quadratic Discriminant Analysis (QDA) [16] and *k* Nearest Neighbours (kNN) [17]. For that purpose, a random division of samples was done to obtain 80% of samples in a training set and 20% of the samples in a test set, and leave-one-out approach (LOO) was used on the training set to get the accuracy for each classification method. For kNN, it was necessary to select the optimum number of neighbours to be used, which was done by testing different *k* numbers of neighbours from 1 to 10 using LOO on the training set. The random division of the samples was repeated 50 times and the mean performances obtained for each classifier and each parameter using the leave-one-out approach on the training sets were used to compare the results.

Similarly, Partial Least Squares – Discriminant Analysis (PLS-DA) [18] was applied after optimizing the number of latent variables (LV), and Sparse Partial Least Squares – Discriminant Analysis (sPLS-DA) [19] after selecting the number of metabolites to be included. In the case of sPLS-DA the number of latent variables was selected according to the number of groups minus one, as it is advised for generating the most stable models.

Once the classification method and the optimal parameters had been selected, they were applied to the samples from the test set and this operation was repeated 50 times to obtain the average performance of the selected model. To judge whether the 15 amino acids and related compounds measured improve the performance and add to the CNN-based diagnosis, the optimized performance of the classifiers were compared to the diagnostic performance by CNN alone.

## 3. Results and discussion

### 3.1. Summary statistics

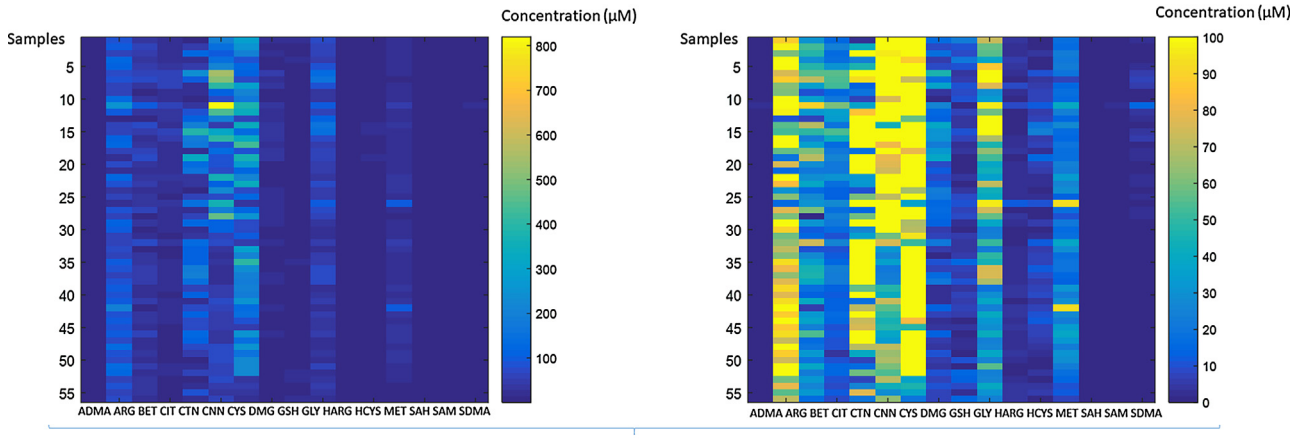
To have an approximate idea of the orders of magnitude of the concentrations of the 16 metabolites analyzed in plasma, Fig. 2 shows an image made using Matlab R2015a (Mathworks Natick, Massachusetts, United States) representing in a color scale the concentration of each analyte in each patient's sample. In addition, a table summarizing concentration levels of each amino acid in plasma from both control and CKD pediatrics is showed in the same figure.

This picture showed a big variation because even if almost all of the compounds were expected to be within the range from 0 to 100 μM, some metabolites like CNN had concentrations up to 800 μM[1]. After re-scaling the colors in the color bar of the image from 0 to 100 μM, again 2 groups were differentiated: one containing analytes from 30 to 100 μM -and higher- (ARG, BET, CYS, CIT, CTN, CNN, DMG, GLY, GSH, HARG, HCYS, and MET) and another one containing metabolites approximately from 0 to 2 μM (ADMA, SDMA, SAH and SAM).

### 3.2. Multivariate analysis

#### 3.2.1. Scaling and construction of PCA model

Multivariate data analysis tools take into consideration the intrinsic interdependency of the metabolite concentrations. As the concentrations of several metabolites were non-normally distributed across samples, logarithm transformation was carried out. Subsequently, autoscaling was performed by normalizing the concentration of each metabolite in each sample by subtracting mean metabolite concentration and dividing it by their standard deviation. This scaling method assumes that all metabolites are equally eligible to be important biomarkers, despite being present in higher or lower concentrations in plasma. One CKD sample was excluded for being an outlier according to Single Linkage, Complete Linkage




Concentration of the analytes of interest expressed as median (3rd – 97th interquartile range) in μM:

| Metabolite | Control | CKD |
|---|---|---|
| ADMA | 0.7 (0.2-1.1) | 0.9 (0.6-1.4) |
| ARG | 84.4 (71.8-93.0) | 76.3 (56.8-105.1) |
| BET | 29.0 (9.4-47.8) | 33.3 (6.8-75.1) |
| CYS | 191.5 (31.4-325.2) | 157.6 (49.5-386.3) |
| CIT | 12.6 (6.8-22.8) | 21.7 (12.6-55.4) |
| CTN | 76.3 (22.3-187.6) | 73.2 (8.4-331.0) |
| CNN | 45.1 (32.8-71.6) | 150.3 (58.3-560.5) |
| DMG | 6.8 (1.9-18.4) | 15.5 (3.9-38.8) |
| GLY | 32.0 (22.6-75.9) | 57.3 (24.0-155.9) |
| GSH | 6.2 (1.6-15.3) | 6.2 (1.3-21.8) |
| HARG | 2.6 (1.0-4.1) | 2.1 (0.6-9.0) |
| HCYS | 6.1 (0.6-11.1) | 5.4 (0.4-19.2) |
| MET | 22.8 (9.4-55.0) | 18.8 (9.4-38.9) |
| SAH | 0.021 (0.003-0.091) | 0.094 (0.016-0.830) |
| SAM | 0.055 (0.008-0.108) | 0.263 (0.100-0.394) |
| SDMA | 0.4 (0.1-0.7) | 1.7 (0.4-6.9) |

**Fig. 2.** Concentration of each amino acid per sample represented in a matrix according to a color scale (up) and a summary of the plasma concentrations for each metabolite per population group expressed as median and interquartile range in μM (down).

and Average Linkage Hierarchical Clustering and was also visible as an outlier when doing Principal Component Analysis (PCA) plot.

This data pre-treatment enabled a separation between control and CKD sample groups when doing PCA, as shown in the biplot in Fig. 3. Moreover, the relation between sample groups and metabolites can be obtained from the representation. SDMA, SAH, CNN, SAM, CIT, ADMA, GSH, DMG and GLY were found to be increased in CKD patients comparing with control samples. The first three principal components account for 58% of variance.

The information obtained from PCA was compared with the literature to check whether these up- or down-regulations could be explained by any known mechanism. SDMA and ADMA are thought to be increased due to the decreased excretion in renal patients. Moreover, the reduced activity of ADMA catabolism by dimethylarginine dimethylaminohydrolase (DDAH) would also be responsible for the increase of ADMA [20]. Regarding high CIT and GLY concentrations, low activity of the enzyme arginine/glycine amidinotransferase (AGAT) could be suggested [21]. Similarly SAH and SAM are thought to be increased in part due to impaired renal clearance, as kidneys play a major role in aminothiol metabolism [22]. Furthermore, it has been suggested that a reduction in the ability to metabolize and/or excrete SAH could be a primary event in CKD [23]. Several biomarkers of oxidative stress may be elevated in patients with CKD, such as glutathione [24], even if the nature of the oxidative stress in these patients still remains unclear. This impaired oxidative balance could be the result of increased reactive oxygen species (ROS) production and reduced clearance, or due to an ineffective antioxidant defense mechanism. Previous studies had already shown increased total GSH increased levels in adults suffering from CKD [25], but some other studies performed also in adults showed the opposite [26] or no difference between healthy and CKD patients [27]. Regarding DMG, it is a feedback inhibitor of betaine-homocysteine methyltransferase enzyme, which is normally excreted in urine or metabolized to sarcosine [28]. Thus, it is assumed that this increase could be related to accumulation, due to an impaired urinary excretion, in addition to any unknown factor that could also be involved.

The results of this unsupervised dimensionality reduction matched quite well with the results obtained using univariate analysis based on Student's *t* test for analytes with normal distribution and *U* test in accordance to Mann and Whitney for those with non-normal distribution after verifying the normal distribution of samples using Kolmogorov-Smirnov test. Indeed, the mean values of GLY, CIT, CNN, ADMA and SDMA were significantly increased in CKD patients when comparing with control individuals [11].

### 3.2.2. *Classifying between early CKD and control samples using all metabolites*

In order to find out whether any of these metabolites are interesting as potential biomarkers for the early diagnosis of CKD, a PCA model was constructed and a classification method was selected considering only early disease samples (those at CKD2 stage) and control samples (Fig. 4). Following these approaches, not only the separation between groups was complete except for one CKD2 sample which was placed with control samples, but also an improvement in the performance of the classification was gained when using all the metabolites including CNN, in comparison with the single use of CNN. Indeed, after the optimization of the classifiers, using all the metabolites with PLS-DA with 2 LV the percentage of success was 76% and with sPLS-DA with 1 LV and 4 metabolites was 67%. CIT, CNN, SAM, SDMA were the output metabolites for sPLS-DA variable reduction and classification method. On the other hand, QDA was selected as the best classification method when using only CNN with a performance of 71%.

Moreover, in order to have an idea of how well all non-CNN metabolites performed in comparison with CNN, classification was also executed using the 15 metabolites, and a performance of 59% was obtained using PLS-DA with 2 LV. Therefore, even if CNN alone

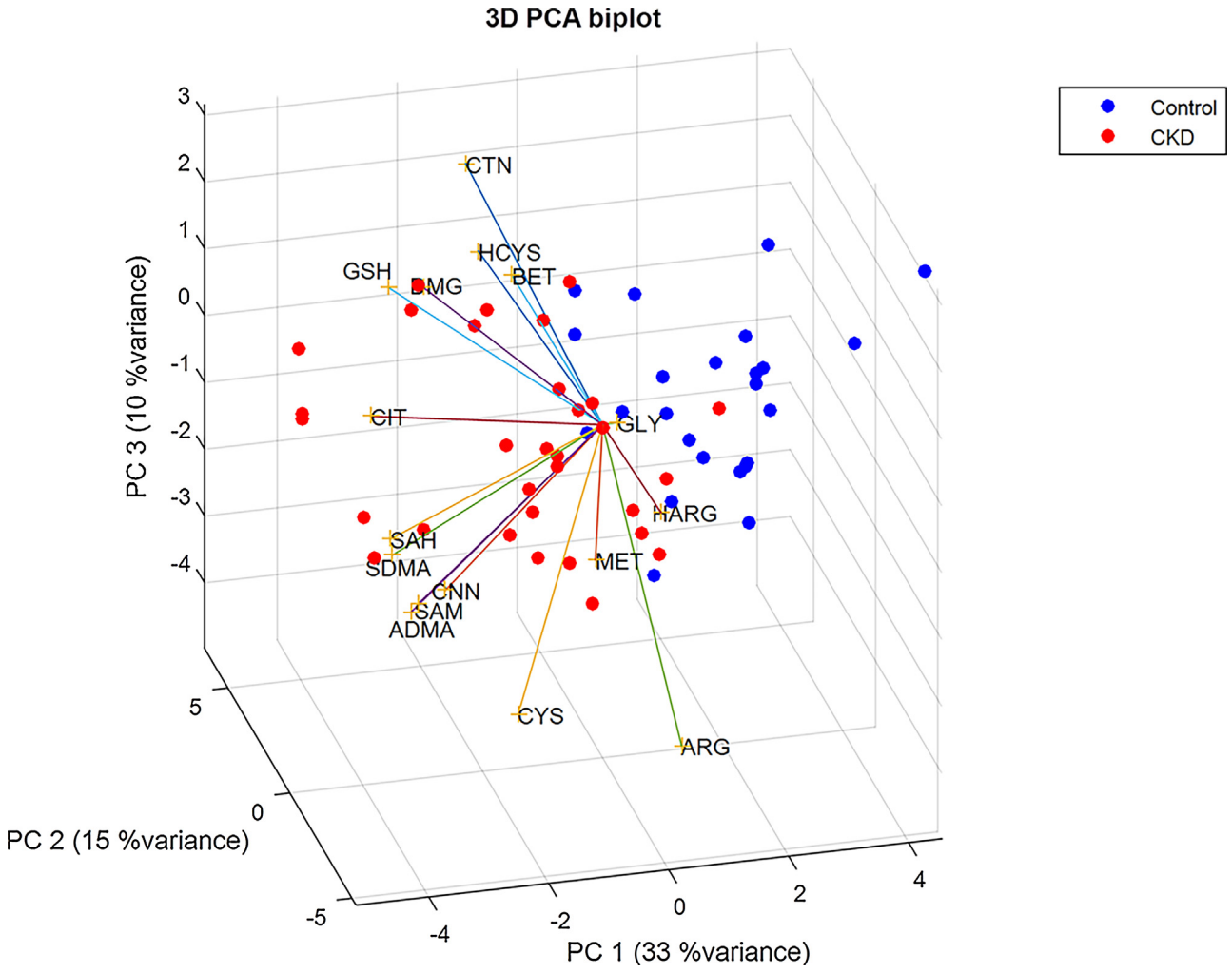
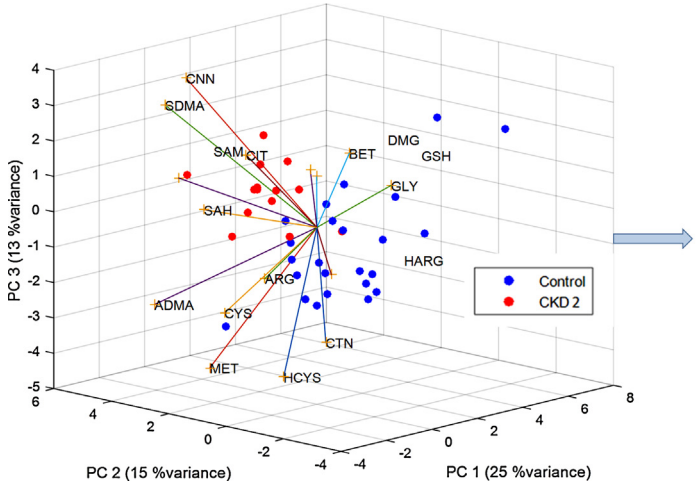



**Fig. 3.** PCA biplot containing the score plot for control and CKD samples and the loading plot with the 16 amino acids for the first 3 principal components.




**Sample classification**

| DATA | Optimal method | Performance |
|---|---|---|
| **Only creatinine** | QDA | 71 % |
| **All metabolites** | PLS-DA (LV = 2) | 76 % |
| **All except for CNN** | PLS-DA (LV=2) | 59 % |
| **SDMA, CNN, CIT and SAM** | PLS-DA (LV = 1) | 89 % |

**Fig. 4.** PCA biplot for control and early CKD samples (on the left) and the performances for different sample classification methods (on the right).

is better than the rest of the metabolites for prediction purposes, the rest of the metabolites could be useful biomarkers in combination with CNN.

As a consequence, CIT, SAM and SDMA and CNN were found to be the biologically relevant features according to the metabolite selection carried out with sPLS-DA, and it would match with the fact that CNN would be of special interest for classification purposes in comparison with the rest of the metabolites. Univariate statistical analysis was performed to verify if the two groups of samples were significantly different from each other for each of these 4 metabolites. The univariate statistical analysis based on the use of Student's *t*-test for normal distributed metabolites and *U* test in accordance to

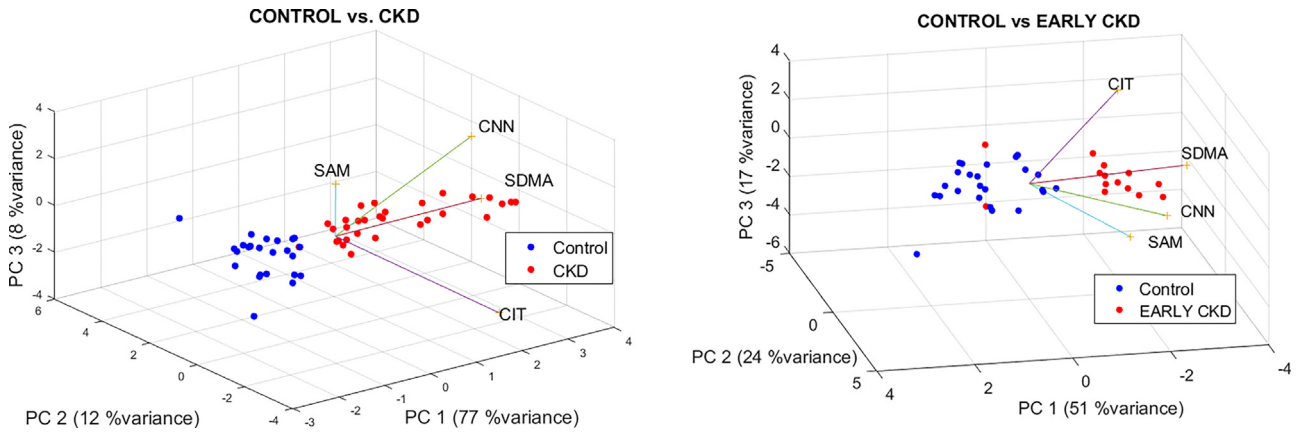



**Fig. 5.** PCA for CNN, CIT, SAH and SDMA data for control and CKD samples regardless of the stage of the disease (left) and for control and early CKD samples (right).

Mann-Whitney for non-normal ones showed p-values below 0.001 for CIT, SDMA and CNN. Despite the fact that SAM did not have a normal distribution and was not significant according to the tests ($p > 0.05$), median values and interquartile range showed slight differences between both populations, in addition to a correlation of SAM with SDMA ($r = 0.92$; $p < 0.05$), CIT ($r = 0.74$, $p < 0.01$) and CNN ($r = 0.77$, $p < 0.05$). Thus, the use of a higher population is suggested in order to clarify if there is any significant difference between both populations. Taking into account all of these evidences, it is worth using only these four metabolites as an input to check the accuracy of the different classifiers, described in the following sections.

### 3.2.3. *Classifying between early CKD and control samples using CIT, SAM, SDMA and CNN*

The most practical approach and the one of special interest would be to select the minimum number of metabolites capable of increasing prediction accuracy compared to CNN alone, and if possible also an increased one in comparison with using all the 16 metabolites. CNN, CIT, SAM and SDMA metabolites showed a similar efficiency for the early diagnosis of CKD in the output obtained from sPLS-DA in comparison with the prediction of CNN alone, as shown in the previous subsection. Furthermore, separation in the PCA model of CKD and control groups was almost complete (Fig. 5) and whether these metabolites used as an input could improve the accuracy of classification was checked as well. Once again, classifiers were optimized for these 4 metabolites and a performance of 89% was obtained using PLS-DA (LV = 1) for the classification of early CKD and control samples. This means an increase in performance of 18% when using CIT, SAM and SDMA in addition to CNN in comparison with using CNN only.

### 3.2.4. *Stage-independent CKD diagnosis using all metabolites*

Although the main objective of this work was to find a few potential biomarkers for early diagnosis, it is also interesting to verify if these 16 metabolites could also be useful regardless of the stage of CKD to be applicable in clinical practice. For that reason, the same process was repeated for all the CKD samples, regardless of the CKD stage using the 16 metabolites.

First of all, the various classification methods were compared and PLS-DA (LV = 1) with a classification accuracy of 84% resulted in the best outcome using all the metabolites including CNN, whereas QDA was again the best classification method using CNN alone to sort the samples as control or CKD (81% performance). In both cases performance was similar, so using all metabolites for classification would neither improve the accuracy in the diagnosis nor jeopardize the performance.

### 3.2.5. *Stage-independent CKD diagnosis using metabolites CIT, SAM, SDMA and CNN*

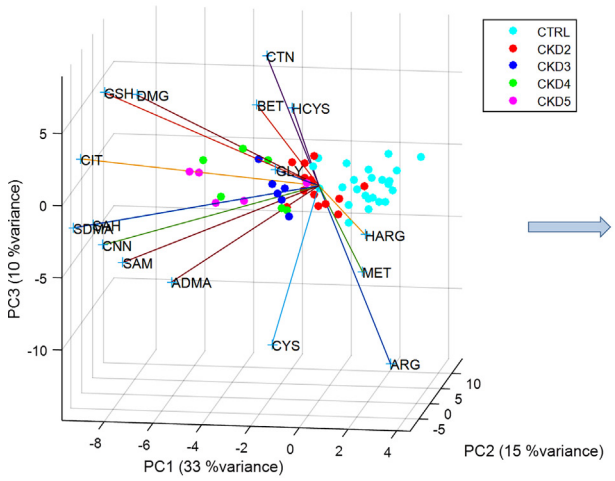
As CIT, SAM, SDMA and CNN provide the best results for the early diagnosis of CKD, the classification procedure was repeated using as input the concentration information for these metabolites only in all the samples, regardless of the CKD stage. In this case, PLS-DA (LV = 1) proved to be the best classifier to be used and once applied to the total amount of samples a performance of 91% was achieved, thus enabling an improvement of 10% in comparison with CNN alone.

Accordingly, metabolites CIT, SAM and SDMA might be of special interest to add to CNN, as they improve the accuracy of CNN diagnosis not only for early disease samples but also when using all the samples, regardless of the CKD stage. Moreover, SDMA ($r = 0.91$, $p < 0.001$), citrulline ($r = 0.84$, $p < 0.001$) and SAM ($r = 0.77$, $p < 0.05$) were found to be correlated with CNN.

### 3.2.6. *Classification performance for CKD stages*

Within the years different equations have been developed to obtain the GFR with the aim of diagnosing those patients suffering from CKD and to classify them according to the stage of the disease, such as: Schwartz, Bedside Schwartz equation, Modification of Diet in Renal Disease (MDRD), Counahan-Barratt and Cockroft-Gault [4]. For the development of new equations "gold standards" like inuline, iohexol, iothalamate, $^{51}$Cr-EDTA or $^{99m}$Tc-DTPA are commonly used to obtain a real GFR value and to adjust the equations to get a similar estimated GFR value based on the measurement of endogenous metabolites. Despite being this the ideal, as a preliminary study prior to the development of any new equation, metabolites were evaluated using PCA and different classification methods were used to determine the stage of the disease.

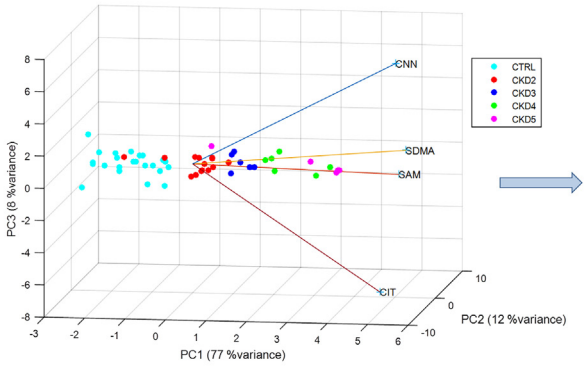
Concerning the differences in metabolite profile for different CKD stages, PCA model obtained from using all the metabolites showed that concentrations of some metabolites had increased with disease severity, such as SAH, SAM, SDMA, ADMA, CNN, CIT, GSH and DMG, as shown in Fig. 6. Moreover, there was a complete separation between the groups made of CTRL-CKD2 and CKD3 to CKD5 groups. Likewise, including only CIT, SAM, SDMA and CNN a good separation was achieved and even in the case in which CNN was not included, a gradation on CKD was observed (data not shown). To check whether all the metabolites or some of them could add, the accuracy of classification using the plasma concentrations of the proposed metabolites was compared with the use of CNN concentration only. Some increase in performance was found using all the 16 metabolites including CNN with PLS-DA and 2 LV (64%) and even a higher increase using CIT, SAM, SDMA and CNN with PLS-DA and 2 LVs (74%), in comparison with using only CNN




**ALL METABOLITES**

| PLS-DA (LV = 2) | CTRL pred | CKD2 pred | CKD3 pred | CKD4 pred | CKD5 pred |
|---|---|---|---|---|---|
| CTRL known | 3.5 | 0.2 | 0.0 | 0.0 | 0.0 |
| CKD2 known | 0.9 | 1.6 | 0.1 | 0.0 | 0.0 |
| CKD3 known | 0.0 | 0.7 | 1.1 | 0.6 | 0.4 |
| CKD4 known | 0.0 | 0.0 | 0.2 | 0.6 | 0.6 |
| CKD5 known | 0.0 | 0.0 | 0.0 | 0.1 | 0.3 |

**Performance: 64 %**




**4 METABOLITES: CIT, CNN, SAM, SDMA**

| PLS-DA (LV = 2) | CTRL pred | CKD2 pred | CKD3 pred | CKD4 pred | CKD5 Pred |
|---|---|---|---|---|---|
| CTRL known | 3.1 | 0.4 | 0.0 | 0.0 | 0.0 |
| CKD2 known | 0.6 | 1.9 | 0.2 | 0.0 | 0.0 |
| CKD3 known | 0.1 | 0.5 | 1.1 | 0.1 | 0.2 |
| CKD4 known | 0.0 | 0.0 | 0.2 | 1.1 | 0.3 |
| CKD5 known | 0.0 | 0.0 | 0.0 | 0.3 | 0.9 |

**Performance: 74 %**

**Fig. 6.** PCA for all the metabolites showing the gradation of the samples according to the metabolite concentration and the stage of the CKD. In addition, the average confusion matrix for the best classifier for all the metabolites in 50 different test sets is shown (up). Similarly PCA for the 4 selected metabolites and its corresponding confusion matrix (down) are shown.

with QDA classification (52%). Even if in general these accuracies do not appear high, closer inspection of the confusion matrix shows that almost all misclassifications occur one stage above or below in comparison with the CKD stage assessed by the nephrologists and less than 3% of the samples are misclassified 2 stages or more above or below (Fig. 6). From a clinical point of view, misclassification of the samples into a different stage could affect the action plan chosen for the patient. However, these recommended approaches for each stage are general and do not change substantially when classifying the samples one stage above or below.

It has to be taken into account that in our predictions classification performance using plasma CNN alone might be that low because age, gender, height and/or weight of pediatrics are commonly used in different equations to correct for some factors affecting serum CNN concentration for the GFR calculation and there is a chance that the classification into the different CKD stages could be improved by using these factors, instead of using CNN as a linear predictor (Eq. (1)). For instance, Schwartz equation uses height to correct for CNN production, as it is a function of muscle mass which is related with body mass and among the variables of body size tested by them, body length divided by CNN concentration provided the best correlation with GFR [29].

### 3.2.7. *Age-, sex- and treatment-related effects*

As mentioned before, it is necessary to check that the concentrations of these metabolites are not sex-related, age-related or treatment-related, in case these effects are present to be able to correct them in future equations, as it is done in Schwartz formula (Eq. (1)). For this purpose, PCA using CIT, CNN, SAM and SDMA was applied on the samples using two age groups ranging from 2 to 12 years old and from 13 to 18 years old, in agreement with the division made by Way et al., according to the normal GFR values shown for children [30]. Although there was no clear separation of the groups, some trend was observed, as shown in Fig. 7. Therefore, in order to find any relation between age and metabolite concentrations Partial Least Squares (PLS) regression method was carried out to predict the age according to the metabolite levels in plasma. No good prediction was obtained when all samples or only CKD samples were considered, whereas when only control samples were considered the prediction was much better, thus meaning that there is a positive correlation between age and metabolite concentrations in healthy pediatrics. The results were similar for either using all the metabolites or using only CNN, SDMA, SAM and CIT as variables.

Regarding sex, samples from both male and female groups were overlapped in the score plot of the PCA, thus showing that there is not any difference in metabolite concentrations according to sex. This is something that could be expected since the patients were pediatrics and no major differences in sex are presumed at these ages. In addition, the same operation was performed using only adolescent samples, that is, those aged between 13 and 18 years old. Nevertheless, no clear separation was observed neither using all the metabolites nor the 4 selected ones.

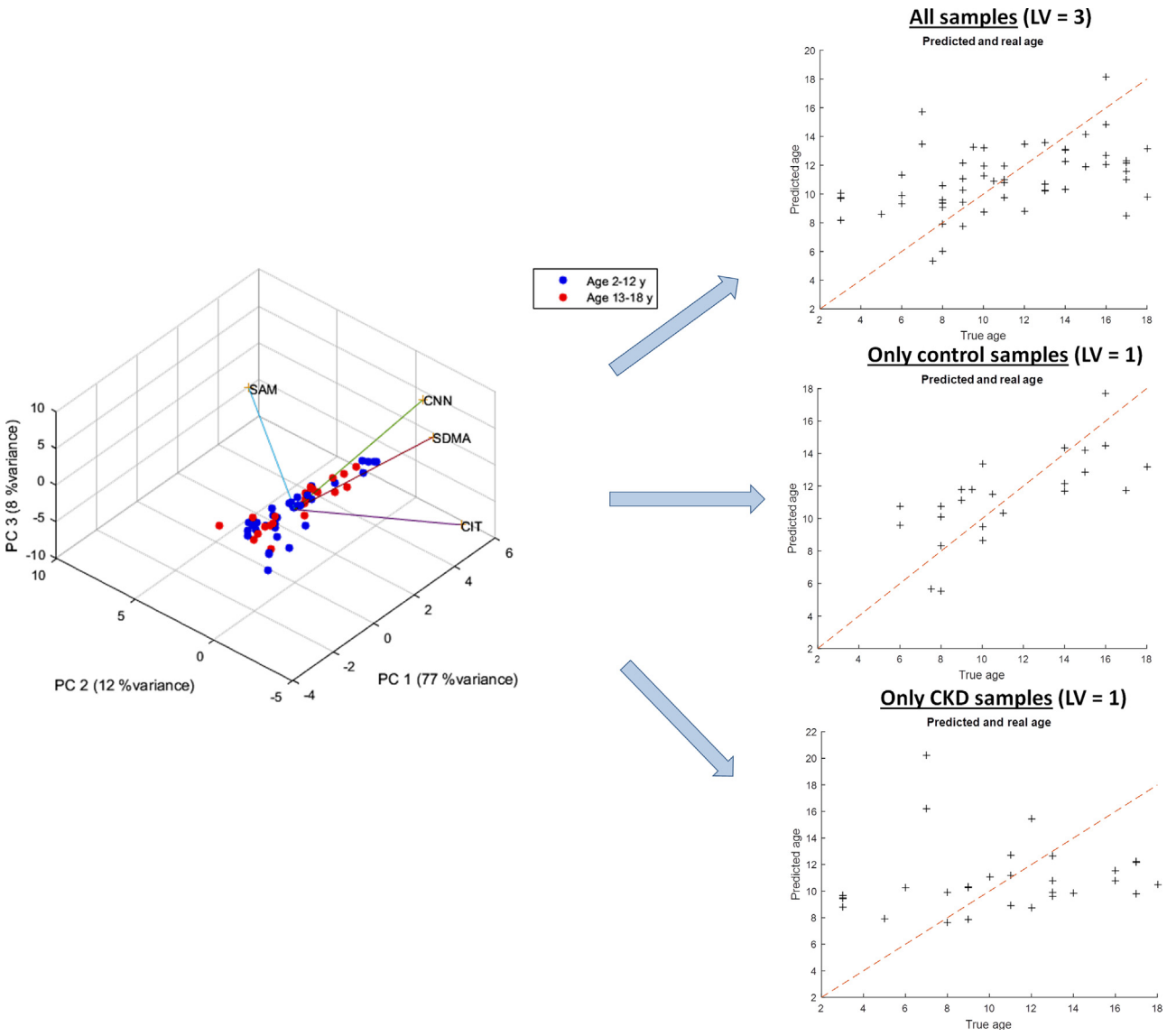



**Fig. 7.** On the left, PCA for all the samples with CIT, CNN, SAM and SDMA data, colored according to different age groups. On the right, a representation of a PLS regression to predict the age with all the samples, only control samples and only CKD samples.

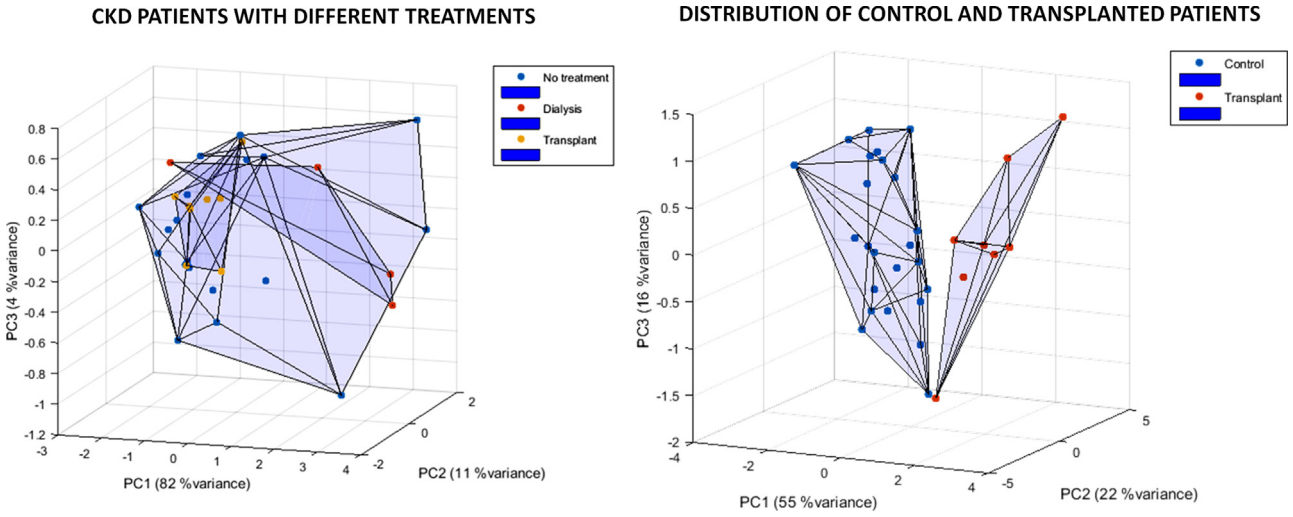



**Fig. 8.** On the left, PCA with convex hull showing different distributions of samples according to the concentrations of CIT, CNN, SAM and SDMA and the relation with the treatment received by CKD patient. On the right, sample distributions for control and transplanted patients.

Besides, to check if there could be some effect on metabolite concentration according to the treatment, PCA was carried out for untreated, transplanted and dialyzed CKD patients, but no separation was found between groups. However, transplanted patients showed a more homogeneous amino acid profile in comparison with not treated and dialyzed ones (Fig. 8). Therefore, it could be concluded that after receiving a transplant the amino acid profile tends to normalize and equalize. Moreover, if transplanted CKD patients are compared to control patients, the distribution of the samples is quite similar, thus implying that amino acid profile tends to equalize, even if both groups are still separated in PCA. This means that amino acid profiles from controls and transplanted patients have a similar distribution, but there are still some differences between both amino acid profiles.

This standardization in the metabolite concentration after transplant could be expected as patients were transplanted when other renal replacement therapy had failed and in all cases samples were collected at least a year after having received the transplant.

## 4. Conclusions

Three new metabolites, CIT, SAM and SDMA, have been proposed as potential biomarkers in addition to the commonly used CNN to be implemented since they enable a better diagnosis of early stages of CKD in pediatrics. Moreover, these metabolites showed greater improvement also for the diagnosis of the rest of the stages in CKD. In addition, some gradation in the concentration of these metabolites according to the CKD stage has also been found. Therefore, it is reasonable to think of including them in a new equation to be used in general practitioners' blood screening tests with some other commonly used parameters as cystatin C protein concentration, CNN concentration, age, height and weight, in order to be able to approximate the prediction of the disease performed by nephrologitsts and to foresee the stage in CKD patients. In order to comply with this purpose, it is suggested to collect a higher number of samples, to obtain GFR by using some contrast agent like iohexol, which is used intravenously for radiologic procedures even in the presence of renal disease, and which is not secreted, metabolized or reabsorbed by the kidneyfor future research. This would allow adjusting a new equation for GFR calculation including CIT, SDMA and SAM and comparing the performances obtained from this equation with other commonly used equations like Schwartz in pediatrics.

## Disclosures

No relevant conflicts of interest to declare.

## Acknowledgements

The authors thank for technical and human support provided by SGIker of UPV/EHU and European funding (ERDF and ESF) as well as the Division of Metabolism belonging to Cruces University Hospital (Barakaldo, Spain) for supplying real samples for this study. This work was funded by the Department of Industry, Innovation, Commerce and Tourism of the Basque Government (SAI 12/25 Project) and by the Basque Government, Research Groups of the Basque University System (Project No. IT3338-10). The Basque Government is also gratefully acknowledged for a predoctoral grant (PRE_2013_1_899) and for a mobility grant (EP_2016_1_0003) for Sandra Benito (Department of Education, Language Policy and Culture). This grant made possible doing an stay in Chemometrics group in the Analytical Chemistry Department at Radboud University, which is also thanked for the opportunity given. Dr. Ewa Szymanska is also gratefully acknowledged for kindly providing sPLSDA routine.

## References

[1] M. Rosenberg, R. Kalda, V. Kasiulevicius, M. Lember, Management of chronic kidney disease in primary health care: position paper of the European Forum for Primary Care, Qual. Prim. Care 16 (2008) 279–294.
[2] G.J. Schwartz, D.F. Work, Measurement and estimation of GFR in children and adolescents, Clin. J. Am. Soc. Nephrol. 4 (2009) 1832–1843.
[3] S.F. Simoneaux, L.A. Greenbaum, Diagnostic imaging, in: E.D. Avner, W.E. Harmon, P. Niaudet, N. Yoshikawa (Eds.), Pediatric Nephrology, Springer, Berlin, 2009, pp. 535–564.
[4] G.J. Schwartz, A. Munoz, M.F. Schneider, R.H. Mak, F. Kaskel, B.A. Warady, S.L. Furth, New equations to estimate GFR in children with CKD, J. Am. Soc. Nephrol. 20 (2009) 629–637.
[5] A.D. Slee, Exploring metabolic dysfunction in chronic kidney disease, Nutr. Metab. 9 (2012) 36.
[6] R.A. Van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genom. 7 (2006).
[7] C. Yang, A.D. Richardson, A. Osterman, J.W. Smith, Profiling of central metabolism in human cancer cells by two-dimensional NMR, GC–MS analysis, and isotopomer modeling, Metabolomics 4 (2008) 13–29.
[8] A. Kalivodova, K. Hron, P. Filzmoser, L. Najdekr, H. Janeckova, T. Adam, PLS-DA for compositional data with application to metabolomics, J. Chemom. 29 (2015) 21–28.
[9] M.S. Monteiro, M. Carvalho, M.L. Bastos, P. Guedes de Pinho, Metabolomics analysis for biomarker discovery: advances and challenges, Curr. Med. Chem. 20 (2013) 257–271.
[10] R. Madsen, T. Lundstedt, J. Trygg, Chemometrics in metabolomics. A review in human disease diagnosis, Anal. Chim. Acta 659 (2010) 23–33.
[11] S. Benito, A. Sanchez, N. Unceta, F. Andrade, L. Aldamiz-Echevarria, M.A. Goicolea, R.J. Barrio, LC-QTOF-MS-based targeted metabolomics of arginine-creatine metabolic pathway-related compounds in plasma: application to identify potential biomarkers in pediatric chronic kidney disease, Anal. Bioanal. Chem. 408 (2016) 747–760.
[12] 2002/657/EC. Commission Decision of 12 August 2002 implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results, Official Journal of the European Communities (2002).
[13] National Kidney Foundation, K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification, Am. J. Kidney Dis. 39 (2002) S1–S266.
[14] T. Hastie, R. Tibshirani, J. Friedman, Linear methods for classification, in: T. Hastie, R. Tibshirani, J. Friedman (Eds.), The Elements of Statistical Learning, Springer, New York, 2008, pp. 101–118.
[15] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[16] S. Geisser, Posterior odds for multivariate normal distributions, J. R. Stat. Soc. B Met. 26 (1964) 59–76.
[17] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE T. Inf. Theory IT 13 (1) (1967) 21–27.
[18] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[19] E. Szymanska, E. Brodrick, M. Williams, A.N. Davies, H.-J. van Manen, L.M.C. Buydens, Data size reduction strategy for the classification of breath and air samples using multicapillary column-ion mobility spectrometry, Anal. Chem. 87 (2015) 869–875.
[20] F. Mihout, N. Shweke, N. Bige, C. Jouanneau, J.-C. Dussaule, P. Ronco, C. Chatziantoniou, J.-J. Boffa, Asymmetric dimethylarginine (ADMA) induces chronic kidney disease through a mechanism involving collagen and TGF-B1 synthesis, J. Pathol. 223 (2011) 37–45.
[21] F. Andrade, J. Rodriguez-Soriano, J.A. Prieto, J. Elorz, M. Aguirre, G. Ariceta, S. Martin, P. Sanjurjo, L. Aldamiz-Echevarria, The arginine-creatine pathway is disturbed in children and adolescents with renal transplants, Pediatr. Res. 64 (2008) 218–222.
[22] A. Valli, J.J. Carrero, A.R. Qureshi, G. Garibotto, P. Barany, J. Axelsson, B. Lindholm, P. Stenvinkel, B. Anderstam, M.E. Suliman, Elevated serum levels of S-adenosylhomocysteine, but not homocysteine, are associated with cardiovascular disease in stage 5 chronic kidney disease patients, Clin. Chim. Acta 395 (2008) 106–110.
[23] K. Jabs, M.J. Koury, W.D. Dupont, C. Wagner, Relationship between plasma S-adenosylhomocysteine concentration and glomerular filtration rate in children, Metab. Clin. Exp. 55 (2006) 252–257.
[24] M. Romeu, R. Nogues, L. Marcas, V. Sanchez-Martos, M. Mulero, A. Martinez-Vea, J. Mallol, M. Giralt, Evaluation of oxidative stress biomarkers in patients with chronic renal failure: a case control study, BMC Res. Notes 3 (2010).
[25] A. Zinellu, S. Sotgia, G. Loriga, L. Deiana, A.E. Satta, C. Carru, Oxidative stress improvement is associated with increased levels of taurine in CKD patients undergoing lipid-lowering therapy, Amino Acids 43 (2012) 1499–1507.
[26] I. Ceballos-Picot, V. Witko-Sarsat, M. Merad-Boudia, A.T. Nguyen, M. Thevenin, M.C. Jaudon, J. Zingraff, C. Verger, P. Jungers, B. Descamps-Latscha, Glutathione antioxidant system as a marker of oxidative stress in chronic renal failure, Free Radic. Biol. Med. 21 (1996) 845–853.

[27] J. Himmelfarb, E. McMenamin, E. McMonagle, Plasma aminothiol oxidation in chronic hemodialysis patients, Kidney Int. 61 (2002) 705–716.
[28] D.O. McGregor, W.J. Dellow, M. Lever, P.M. George, R.A. Robson, S.T. Chambers, Dimethylglycine accumulates in uremia and predicts elevated plasma homocysteine concentrations, Kidney Int. 59 (2001) 2267–2272.
[29] G.J. Schwartz, G.B. Haycock, C.M. Edelmann Jr., A. Spitzer, A simple estimate of glomerular filtration rate in children derived from body length and plasma creatinine, Pediatrics 58 (1976) 259–263.
[30] A.F. Way, A.M. Bolonger, J.G. Gambertogli, Pharmacokinetics and drug dosing in children with decreased renal function, in: E.D. Avner, M.A. Holliday, T.M. Barratt (Eds.), Pediatric Nephrology, Williams&Williams, Berlin, 1994, p. 1306.