



HAL
open science

Raman spectroscopy and multivariate analysis for the rapid discrimination between native-like and non-native states in freeze-dried protein formulations

S. Pieters, Y. Vender Heyden, J.M. Roger, M. d'Hondt, L. Hansen, B. Palagos, B. de Spiegeleer, J.P. Remon, C. Vervaet, T. de Beer

► **To cite this version:**

S. Pieters, Y. Vender Heyden, J.M. Roger, M. d'Hondt, L. Hansen, et al.. Raman spectroscopy and multivariate analysis for the rapid discrimination between native-like and non-native states in freeze-dried protein formulations. *European Journal of Pharmaceutics and Biopharmaceutics*, 2013, 85 (2), p. 263 - p. 271. 10.1016/j.ejpb.2013.03.035 . hal-00996051

HAL Id: hal-00996051

<https://hal.science/hal-00996051>

Submitted on 26 May 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Raman spectroscopy and multivariate analysis for the rapid discrimination between native-like and non-native states in freeze-dried protein formulations

Sigrid Pieters¹, Yvan Vander Heyden^{1*}, Jean-Michel Roger², Matthias D'Hondt³, Laurent Hansen⁴, Bernard Palagos², Bart De Spiegeleer³, Jean-Paul Remon⁵, Chris Vervaet⁵, Thomas De Beer⁴

¹ Department of Analytical Chemistry and Pharmaceutical Technology, Center for Pharmaceutical Research, Vrije Universiteit Brussel - VUB, Laarbeeklaan 103, B-1090 Brussels, Belgium

² Irstea, UMR1201, Rue J.F. Breton 361, 34191 Montpellier, France

³ Laboratory of Drug Quality & Registration (DruQuaR), Universiteit Gent, Harelbekestraat 72, B-9000 Ghent, Belgium

⁴ Laboratory of Pharmaceutical Process Analytical Technology, Universiteit Gent, Harelbekestraat 72, B-9000 Ghent, Belgium

⁵ Laboratory of Pharmaceutical Technology, Department of Pharmaceutics, Universiteit Gent, Harelbekestraat 72, B-9000 Ghent, Belgium

* Corresponding author: Tel.: +32 2 477 47 34

Fax: +32 2 477 47 35

Email: yvanvdh@vub.ac.be (Y. Vander Heyden)

Abstract

This study investigates whether Raman spectroscopy combined with multivariate analysis (MVA) enables a rapid and direct differentiation between two classes of conformational states, i.e. native-like and non-native proteins, in freeze-dried formulations. A data set comprising of 99 spectra, both from native-like and various types of non-native freeze-dried protein formulations, was obtained by freeze-drying lactate dehydrogenase (LDH) as model protein under various conditions. Changes in the secondary structure in the solid freeze-dried proteins were determined through visual interpretation of the blank corrected second derivative amide I band in the ATR-FTIR spectra (further called FTIR spectra) and served as an independent reference to assign class labels. Exploratory analysis and supervised classification, using Principal Components Analysis (PCA) and Partial Least Squares - Linear Discriminant Analysis (PLS-LDA), respectively, revealed that Raman spectroscopy is with 95% accuracy able to correctly discriminate between native-like and non-native states in the tested freeze-dried LDH formulations. Backbone (i.e. amide III) and side chain sensitive spectral regions proved important for making the discrimination between both classes. As discrimination was not influenced by the spectral signals from the tested excipients, there was no need for blank corrections. The Raman model may allow direct and automated analysis of the investigated quality attribute, opening possibilities for a real-time and in-line quality indication as a future step. However, the sensitivity of the method should be further investigated and where possible improved.

1. Introduction

The FDA's process analytical technology (PAT) initiative [1] is one of the concepts encouraging the (bio-) pharmaceutical industry to invest in analytical methods that provide ultra-fast (i.e. real-time) analysis results, while being information-rich and robust in the presence of interferents. Most of them rely on spectroscopic techniques combined with multivariate analysis (MVA) [2-4]. As they usually lack the conventional delay inherent to the standard analytical techniques for sampling, sample preparation, analysis, data extraction and interpretation, not only the laboratory work, time and costs may be reduced, but they may also allow decision making in real-time. This is essential for using these techniques as *in-situ* diagnostic tools during manufacturing processes, when the quality attribute of interest has to be assessed within a few seconds to make important decisions that ensure overall product and process quality.

The 3-D folded structure of a native protein is required to properly exert its biological action. The most important bonds that make up the secondary to quaternary structure are hydrophobic interactions (e.g. between the side chains of the amino acids) and hydrogen bonds (e.g. for constituting the secondary structural elements such as α -helix and β -sheets). However, these structures are fragile, since exposure to even mildly denaturing conditions may make them lose partially their natively-folded structure. These conditions can be numerous, and freeze-drying is well known to be one of them [5-7]. When the protein does not refold back to its native state after reconstitution, its activity might be reduced or absent. Partial unfolding can also result in the exposure of buried hydrophobic amino acid residues. This may favour aggregation when reconstituting the freeze-dried product with typically an up-concentrated protein fraction [8]. To avoid excessive unfolding and subsequent exposure of the hydrophobic sites it is thus necessary to take preventive actions, conserving the native conformation of the proteins as much as possible during the whole manufacturing process and

storage. The potential risk for reduced or absent activity of the non-native proteins, and immune responses of the aggregates, makes them important considerations in formulation and process development of protein drug products.

Among the many available techniques that may provide information on a protein's 3D-structural conformation (e.g. fluorescence spectroscopy, far-UV circular dichroism, NMR, FTIR...), Raman spectroscopy may have the potential to become an ultra-fast diagnostic tool in detecting in real-time protein conformational changes during the freeze-drying process. Whereas other techniques require the protein to be in solution, and/or are invasive (i.e. FTIR would affect the freeze-drying process) Raman spectroscopy is a non-invasive technique which is practically transparent for water and ice, and requires minimal or no sample preparation. Near-infrared (NIR) spectroscopy can also detect protein unfolding during freeze-drying, however due to its strong water and ice signals it could only be used for monitoring a sample after sublimation in the illuminated sample spot [9].

Infrared (IR) and Raman spectroscopy have been extensively used in protein structural characterization, especially for determining changes in the secondary and/or tertiary structure of proteins in different matrices. Identification of specific bands associated with the secondary structure has been achieved using homopolypeptides and proteins with known α -helix and β -sheet elements [10]. As these methods require a thorough peak-by-peak investigation to obtain the necessary information from the spectra of the protein formulations, they usually demand blank corrections, resolution enhancement techniques and expert knowledge. Visual interpretation of the amide I FTIR spectral band ($1700\text{-}1600\text{ cm}^{-1}$) is by far the most established technique for determining secondary structure of proteins in the solid state [10]. As the original amide I spectrum is usually featureless and can be overlapped by the signals of other components, the visualization of the underlying secondary structure components is hampered. Therefore, prior to visual inspection blank corrections and decomposition methods,

such as Fourier deconvolution or second derivation of the spectra, have become popular to delineate overlapping secondary structures in the original spectrum [10]. As delineating is more complex in the blank corrected Raman spectra, the interpretation is frequently based on changes in peak shape rather than on individual band assignments [11-13]. Usually, the focus is on the visual interpretation of the well-known amide I and amide III bands.

Although aforementioned methods are very informative, they have the drawback of being time consuming and unpractical for a rapid quality assessment, which is essential for in-line monitoring of the protein conformational state during freeze-drying processes. The proposed MVA method uses the Raman spectrum more efficiently, by extracting the relevant information directly from the sample spectrum, without the need for blank corrections, resolution enhancement or subjective interpretations. Multivariate calibration methods can cope with non-selectivity in the obtained spectral signals, provided the interferent signals are sufficiently incorporated into the calibration and their shape is not completely identical to the signal of the analyte [14]. Additionally, MVA may actually point out these spectroscopic marker bands differentiating between native and non-native proteins, thus offering insights into the biochemical origin of the discrimination. Mathematical approaches may also allow detecting more subtle, but consistent, spectral changes that are visually undetectable [3].

In the present paper we investigate whether Raman spectroscopy combined with MVA enables the discrimination between native-like and non-native conformational states of proteins in freeze-dried formulations in a rapid (i.e. a few seconds) and direct way. The aims are: (1) to investigate whether Raman spectroscopy is able to rapidly and accurately differentiate between native-like and non-native freeze-dried proteins, (2) to develop and validate a calibration model to classify freeze-dried protein samples, (3) to study the influence of formulation interferents (i.e. excipient signals) on the discrimination, (4) to examine which part(s) of the Raman spectra are necessary to make this discrimination, and whether they can

be linked to sound biochemical knowledge. Given its known instability during freeze-drying, lactate dehydrogenase (LDH) was selected as a model protein [15].

2. Methods and materials

2.1. Materials

L-Lactic Dehydrogenase (LDH) from rabbit muscle - Type II in ammonium sulfate was obtained from Sigma-Aldrich (Saint Louis, MO, USA). Prior to freeze-drying, 2 mL type I vials (Nipro, Authon-du-Perche, France) were filled with 720 μ l formulation containing approximately 9 mg/ml LDH. Table 1 shows the different LDH formulations that were prepared in this study. Formulations 1-12 were used in the design of experiments (DoE) and contained a 40 mM buffer pH 7.4, either sodium phosphate (Sigma Aldrich, Steinheim, Germany) (PHOS), L-histidine (HIS) or TRIS (Merck, Darmstadt, Germany). Some formulations contained a lyoprotectant, i.e. 5% sucrose (Suedzucker, Mannheim, Germany) and/or a cryoprotectant, i.e. 1% PEG 4000 (Fagron, Barsbüttel, Germany). Besides the DoE formulations, others (formulations 13-19) contained NaSCN, NaCl or ethanol (Sigma Aldrich, Steinheim, Germany), or were formulated in water.

2.2. Freeze-drying

Freeze-drying was performed using an Amsco FINN-AQUA GT4 freeze-dryer (GEA, Koeln, Germany). Eight batches were produced with different cooling methods and secondary drying temperatures (Table 2). The cooling methods included: (1) slow cooling at a rate of 0.2°C/min (SL), (2) fast cooling with pre-cooled shelf at -45°C (FA), (3) cooling at 1°C/min and after freezing an annealing step of 10h at -7°C (ANN), (4) cooling at 1°C/min with cooling stop for 60 min at -10°C (CH). Freezing was always performed for 2h at -45°C, primary drying for 20h at -20°C and 100 μ bar, and secondary drying for 6h at 100 μ bar with temperature ramps of 0.2°C/min in between.

The full factorial experimental design led to 96 freeze-dried samples (8 runs x 12 formulations) (Tables 1-2). Seven additional samples (formulations 13-19) were added to runs 1 and 2 to produce more extremely denatured LDH. This resulted in a total of 110 freeze-dried samples that were analyzed with both FTIR and Raman spectroscopy. After reconstituting the samples, they were analyzed using the LDH activity assay. However, since the amide I bands in the FTIR spectra of some freeze-dried samples could not be interpreted, those were omitted and the data set for developing and validating the model comprised 99 freeze-dried samples.

2.3. LDH activity assay

The enzymatic activity of LDH after reconstitution of the freeze-dried samples in 720 μ l water was determined with a UV-2401PC spectrophotometer (Shimadzu, Kyoto, Japan) by monitoring the rate of absorbance decrease per minute of NADH at 340 nm as a result of the reduction of pyruvate to lactate at 37°C. The Sigma-Aldrich (Steinheim, Germany) experimental procedure was followed [16], using all their necessary chemicals (i.e. NADH, sodium pyruvate and BSA). The remaining activity of the reconstituted freeze-dried samples was compared to that of native (not freeze-dried) LDH, and expressed as % LDH activity recovery after freeze-drying. Each sample was measured in triplicate and averaged. The assay had a coefficient of variation of 4.1% for repeatability and 11.0% for day-to-day intermediate precision.

2.4. Spectroscopic conditions and spectral pre-processing

2.4.1. Raman spectroscopy

A Raman Rxn1 spectrometer (Kaiser Optical systems, Ann Arbor, MI), equipped with an air-cooled CCD detector (back-illuminated deep depletion design), was used in combination with a fiber-optic non-contact MR probe. The Invictus NIR diode laser used a wavelength of 785

nm. Spectra were recorded with a resolution of 4 cm^{-1} and an exposure time of 40 s, using a laser power of 400 mW. For MVA, the Raman spectra of the freeze-dried solids were used without any pre-processing. For visual comparison of the Raman spectra, the corresponding blank spectrum of each formulation was subtracted and the spectra were scaled and Savitzky-Golay smoothed.

2.4.2. FTIR spectroscopy

FTIR spectroscopy functioned as a reference technique to evaluate potential changes in protein secondary structure in the freeze-dried solid. FTIR spectroscopy measurements were conducted on a Varian Scimitar 800 FTIR spectrometer, equipped with a Specac Golden Date diamond ATR module (Varian Scimitar, Middelburg, The Netherlands). The spectra were recorded from 4000 to 500 cm^{-1} in ATR mode at a resolution of 2 cm^{-1} . Each measurement was the average of 100 scans. The FTIR spectra of the same freeze-dried samples and the corresponding blanks were first background (air) corrected and converted to absorbance format. The resulting FTIR absorbance spectra were subsequently corrected for the spectral contributions of the blank formulations. Finally, the second derivative with 17-point Savitzky-Golay smoothing was performed and the FTIR spectra were area normalized over the amide I region ($1700\text{-}1600\text{ cm}^{-1}$) for direct comparison. Individual peak positions were identified and assigned according to the literature [10].

2.5. Data handling and MVA

All data handling was performed with Matlab 7.1 (The Mathworks, Natick, MA) and the PLS toolbox 6.2. (Eigenvector Research, Wenatchee, WA). For PLS-LDA and orthogonal projections in-house Matlab routines were used. Statistical data analysis was performed with Minitab 16 (Minitab, PA).

2.5.1. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised MVA technique that allows seeing trends and structures in data sets. It reduces the number of variables to a new set of latent variables (LV's) that describe the data variability more efficiently. These LV's, called principal components, are linear combinations of the original ones. They are defined in such a way that they explain most the (remaining) variability in the data and are orthogonal to each other. They can be decomposed into loadings and scores. The loading vector gives qualitative information about the samples by describing what type of information characterizes them, i.e. which variables are important. The associated weighted averages of the original variables are the scores, providing quantitative information, i.e. the amount of each loading vector in each sample [14, 17].

2.5.2. Partial Least Squares - Linear Discriminant Analysis (PLS-LDA)

PLS-LDA is a supervised classification algorithm that combines PLS for dimension reduction with LDA for obtaining a classification outcome. LDA seeks an optimal low-dimensional space in which the projected data points are linearly well-separated into different classes [18]. It aims to maximize the ratio of between-class variance to within-class variance. The Mahalanobis distance, taking the shape of the considered cluster into account, is used to calculate the probability that a new object belongs to a certain class. As the number of variables (i.e. wavenumbers) in the Raman and FTIR spectral data is much higher than the number of objects (i.e. samples), LDA cannot directly be applied to the spectra. Therefore, PLS first compresses the data dimensionality [19]. PLS is performed on a set of dummy responses (0 and 1) encoded to represent the 2 classes, and maximizes the co-variance between the measured signals (e.g. spectra) and the response [17].

For building the PLS-LDA models, the data set of 99 samples was randomly split into a training (75 freeze-dried LDH formulations) and a test set (24 freeze-dried LDH

formulations for validation of the model), but it was made sure that all possible variation in \mathbf{y} (i.e. containing native-like, mildly stressed (i.e. mainly altered ratio in α -helix and β -sheet structures, further called α/β ratio) and more strongly stressed (i.e. with the clear presence of non-native structures) LDH with % activity preservations ranging from <10% to >100%) was covered by both data sets. Both training and test set contained formulations with similar excipients originating from the 8 freeze-drying runs. The optimal model complexity was determined with cross validation (CV) on the training set. A permutation test was also part of the validation procedure [20]. This can assess whether the classification of the samples in the two designed groups performs significantly better than any other random classification in two arbitrary groups. This test should exclude the possibility of chance correlation in the model. For every permutation test, the class labels were randomly assigned to different samples. The permutation test was repeated a hundred times to obtain a distribution of classification expected to be non-significant.

2.5.3. Orthogonal projections to study the influence of the excipients

Protein formulations usually require excipients, such as buffers (here histidine, phosphate or TRIS), lyo- and/or cryo-protectants (here sucrose and PEG), because of their delicate physical stability. As we directly use the unprocessed Raman spectra of a freeze-dried formulation (without the usual blank subtraction), we have to investigate the influence of the excipient signals on the discriminating power of the model. Therefore we applied an orthogonal projection approach [21-24]. The contributions within the variable space of the calibration matrix \mathbf{X} (containing the spectra of the protein formulations) are originating from the pure protein signals (\mathbf{X}_p), interferences such as the excipient signals (\mathbf{X}_b) and other undefined spectral variance (ϵ) (eq. 1).

$$\mathbf{X} = \mathbf{X}_p + \mathbf{X}_b + \varepsilon \quad (1)$$

In practice, the interference space \mathbf{X}_b was estimated via an external data set containing only the Raman spectra of the freeze-dried blanks (\mathbf{B}) of all possible formulations used in the data set (Table 1 and Fig. 1) [21]. The column vectors of the so obtained matrix \mathbf{B} form a basis of the detrimental subspace. To remove those variations in \mathbf{X} originating purely from excipients signals, an orthogonal projector to \mathbf{B} , i.e. P_B^\perp , can be calculated. This corrects \mathbf{X} as follows (eq. 2) to obtain \mathbf{X}^* .

$$\mathbf{X}^* = \mathbf{X}P_B^\perp = \mathbf{X}(\mathbf{I} - \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T) \quad (2)$$

where \mathbf{I} is the identity matrix.

Thus, the algorithm finds the dimensions of the spectral space which are affected by the studied excipients and projects the spectra orthogonal to this variation. After the orthogonalization procedure, it can be assessed whether the discrimination and the classification between freeze-dried formulations with native-like and non-native proteins is affected, this way revealing the influence of the removed factor [22].

3. Results and discussion

3.1. Generation of a relevant data set

Whether and to which extent a protein will lose its native conformation depends upon the environmental conditions it is exposed to, and can be very erratic. Therefore, a relevant data set to study whether there can be discriminated between native-like and non-native states of proteins in freeze-dried formulations should cover sufficient native-like samples, as well as a distribution of denatured types (i.e. from mildly to more strongly stressed) obtained by freeze-drying. Therefore, the freezing method, secondary drying temperature and formulation composition were varied to yield different relevant examples of native-like (further called

NL) and non-native or denatured (further called DEN) LDH in freeze-dried formulations (Tables 1-2). An ANOVA was performed on the % LDH activity recovery of the 96 DoE samples, to investigate the influence of 5 process and formulation parameters on the preservation of the native protein conformation after reconstitution of the freeze-dried product (Table 3). Three factors, i.e. freezing method, buffer type and sucrose addition, showed statistically significant effects ($\alpha=0.05$) on the % LDH activity recovery. Slow cooling showed a negative effect (i.e. it showed the lowest activities) on the % LDH activity recovery compared to the other tested cooling methods. The use of phosphate buffer showed a negative effect compared to the others, and the addition of 5% sucrose as a protectant to the formulation positively affected the LDH activity preservation.

3.2. FTIR as a reference technique to assign class labels

Since FTIR easily allows analysis in the solid freeze-dried state, is well-known and frequently used, and is complementary to Raman spectroscopy, this technique was chosen as a reference for obtaining information on the protein conformation in the freeze-dried samples. To assign class membership labels (i.e. NL and DEN) to the samples, a two-step approach was used. First, we visually compared the experimental versus the native reference amide I FTIR spectrum to assign each sample to a class. Next, we performed a PCA on the spectra of the complete labeled sample set to independently confirm the correctness of the initial class labeling.

Changes in the distribution of secondary structures of freeze-dried proteins are most commonly assessed by visual comparison of the second derivative of the amide I band (1700-1600 cm^{-1}) in the blank corrected FTIR spectra of the sample with a native reference spectrum [10,15,25,26,27,28,29,30]. Freeze-dried formulations were designated as NL if the amide I spectrum was similar (in secondary structural elements distribution) to that of the native LDH before freeze-drying (Fig. 2). The amide I region of native LDH indicated primarily alpha

helix structures (near 1658 cm^{-1}), to a lesser extent beta-sheets (near 1636 cm^{-1}), and also a very small portion of intermolecular beta-sheets (near 1620 cm^{-1}), which is consistent with the literature [15]. The small frequency and intensity shifts of the main bands in their freeze-dried versus liquid form can be related to the altered hydration state of the protein molecules [10, 31]. Since it was not possible to preserve the native state in the freeze-dried product without sucrose, the NL formulations contained sucrose. Membership to the NL class was defined by two criteria: (1) no clear changes in the distribution of the native secondary structural elements, and (2) no clear formation of non-native structures, compared to the native state spectrum.

Freeze-dried samples showing signs of not preserving the fully native structure were designated as DEN. As protein denaturation in the freeze-dried product was diverse, some examples of the amide I FTIR spectra of DEN proteins are shown in Fig. 3. They are compared to the average of 5 native-like freeze-dried LDH formulations (bold dotted black line). Non-native LDH in the freeze-dried state was the result of an altered α/β ratio (i.e. clear changes in the peak intensities near 1658 and 1636 cm^{-1}), or a decrease of these structures with the formation of non-native structures (i.e. clearly appearing peaks near 1620 cm^{-1} and in the region $1660\text{-}1680\text{ cm}^{-1}$) [10]. The samples where the former phenomenon was mainly present can be seen as mildly stressed since they still comprise largely of native secondary structures. Some examples are shown in Fig. 3A. The remaining LDH activity after reconstitution of these samples ranged between 90.3 and 108.9%, which may reflect their stronger tendency to refold back to the native conformation during reconstitution. Hence, they may provide a lower quality risk. Some of them had sucrose in their formulation, which probably helped preventing more excessive unfolding. Several DEN freeze-dried samples also showed clearly the presence of non-native structures in their amide I FTIR spectrum. Some examples with various % LDH activity recoveries in their reconstituted product are shown in

Fig. 3B. Signals of non-native structures appeared in the amide I FTIR spectra around 1660-1680 cm^{-1} (turns and different unordered structures) and around 1620 cm^{-1} (intermolecular β -sheets). Protein unfolding was also indicated by band broadening in these regions, where clear absorbance signals at particular wavenumbers are lost due to a fading of structural homogeneity among the protein population [32]. These samples will provide most likely the highest quality risk at the time of reconstitution, since the refolding of the protein to its native state may be more hampered or impossible [8]. This was also seen by their generally lower % LDH activity recovery after reconstitution.

The data set of 99 samples that was used for the exploratory analysis, model development and model validation, comprised 31 NL and 68 DEN samples. Overall, the 99 freeze-dried formulations comprised a wide distribution of % LDH activity recoveries (Fig. 4). It can be seen that the reconstituted NL samples always had high % LDH activity recoveries (ranging from 88.1% to 109.3%), while those of DEN samples varied between 0.4% and 109.6%. The latter is the consequence of having different kinds of protein denaturation in the freeze-dried state, i.e. reversible and irreversible denaturation upon reconstitution. Both NL and DEN groups contained formulations with similar excipients (except samples formulated with NaSCN, NaCl, ethanol and pure water that were always DEN), originating from the 8 freeze-drying runs.

To independently confirm the visual assignment of the class labels to the samples, a PCA was performed on the full sample set constituting the same labeled amide I FTIR spectra as used for the visual interpretation. As spectral changes due to altered protein conformation largely exist in this set of samples, this type of systematic variation must be represented in the primary PC's. The separation among the scores of the NL and DEN samples in the PC1-PC2 scores plot (Fig. 5) was an independent justification for the correctness of the class labelling.

3.3. Raman spectroscopy: Exploratory analysis and classification

To explore trends in the Raman spectra related to the conformational state of the LDH, PCA was performed on the Raman spectra of the data set of 99 samples. No pre-processing of the spectra was applied since different methods (i.e. standard normal variates, normalization, multiplicative scatter correction, Savitsky-Golay smoothing, derivatives) did not improve the discrimination performance (data not shown). Different parts of the spectra within the 800-1800 cm^{-1} range were assessed, since the range below 800 cm^{-1} is known to be less sensitive for protein conformation [33]. The PC2-PC6 score plot of the Raman spectra in the range 1033-1725 cm^{-1} showed for the majority of the samples a discrimination between both NL and DEN states, but the scores of 5 DEN samples could not be differentiated from the NL group (Fig. 6).

The PLS-LDA model built from Raman spectra required 6 LV's and showed 5% (4 samples) misclassifications in CV, while 1 sample of the 24 test set samples was classified wrongly. The discrimination performance of the PLS-LDA model before and after orthogonal projection of the matrix of blank spectra (**B**) to the 75 spectra of the calibration matrix (**X**) was unchanged (Table 4), which indicates that the model is able to predict the protein's conformational state in the presence of this variation, while not being influenced by this variation. Hence, the signals of the tested excipients are not a confounding factor towards discrimination, neither positively nor negatively. Fig. 7 also shows that one is still able to discriminate for most of the samples between both NL and DEN states in a PCA score plot. This finding is of importance from a practical point of view. It implies that all the necessary information can be extracted from a single Raman spectrum of the sample, and that no corresponding blanks (at least for the tested excipients) should be freeze-dried nor measured, while a blank correction is required for visual interpretation of Raman (and FTIR) spectra.

The wrongly classified samples were always mildly denatured LDH (see their amide I FTIR spectra in Fig. 3A). These samples were assigned as DEN by the reference technique, but were predicted by the model as NL samples. In these samples, it was mainly the α/β ratio that was slightly changed, and almost no non-native structures were formed in the freeze-dried product. Their remaining LDH activities after reconstitution ranged between 90.3% and 108.9%. This is similar with those from NL samples, which implies that the denaturation is reversible upon reconstitution. These results suggest that the Raman model has lower sensitivity in detecting mildly denatured freeze-dried samples than the reference technique. It is indeed expected that when there is a continuum of states from NL to various kinds of DEN there will be a limit for detecting first signs of protein denaturation. These results suggest that this limit is higher for the Raman model than for the visual interpretation of the FTIR spectra.

When comparing the number of misclassifications for the Raman model (in green) to the predictions of the models generated by the random permutations of the class labels, we noticed a clear distinction between the permutation distribution and the original classification (Fig. 8). When comparing the original classification with all the permutations a p value of <0.001 could be obtained at the 0.01 significance level, both for CV and self prediction. This confirms that the classification is not based on spurious correlations but rather on consistent differences in spectral features between NL and DEN samples. It was studied whether these correlations could be linked to biochemical knowledge.

3.4. Spectral regions responsible for discrimination

From Fig. 9 it was seen that the loading vector of PC6 and the discriminant vector of the PLS-LDA model had very similar shapes. PC6 described only 0.03% of the variation in the spectra, which means that there is a lot of other variation in the Raman spectra, not responsible for the discrimination between both classes (e.g. light scattering, physical and environmental effects, instrumental effects...) [23, 34]. Yet the loading vector of PC6 or the

discriminant vector can be informative for studying the most important variables responsible for the discrimination. They indicated that the NL and DEN samples have maximum differences in the following regions of the Raman spectrum: (i) near 1070 cm^{-1} (C-N stretch), (ii) near 1130 cm^{-1} (NH_3 deformation), (iii) $1200\text{-}1320\text{ cm}^{-1}$ (amide III), and (iv) $1320\text{-}1500\text{ cm}^{-1}$ (mainly CH_n non-stretching with possible participations of CN and CC stretching). Surprisingly, the amide I band ($1600\text{-}1700\text{ cm}^{-1}$) was only slightly indicated. As the indicated bands comprised both backbone (i.e. amide III) and side chain sensitive regions, we tried to link them with biochemical knowledge. Therefore, we also visually inspected the blank corrected Raman spectra and the corresponding amide I FTIR spectra of some relevant NL and DEN samples (Fig. 10). When visually comparing the Raman spectra of those samples, we see that major spectral differences between NL and DEN samples are manifested in similar spectral regions as indicated by the loading and discriminant vectors. To study more in depth the influence of the different regions (linked to certain biochemical knowledge) of the Raman spectrum, different PLS-LDA models were built using each of these spectral regions separately or combined.

3.4.1. Backbone sensitive spectral regions

The amide I (peak with maximum near 1665 cm^{-1}) and III ($1200\text{-}1320\text{ cm}^{-1}$) Raman bands are well known to be related to the protein's backbone conformations (secondary structural elements) [35, 36]. Their shapes are the result of overlapping components corresponding to α -helices, β -sheets, turns and random structures. Therefore broadening of the peak width and/or differences in shape of the amide I and III bands suggest an altered distribution of secondary structures (native and/or non-native) [35-38]. However, among the PLS-LDA models built from different separate amide ranges (Table 5), the ones using the amide I and II bands showed a very high number of misclassifications, both for CV and for test set prediction. It confirms the knowledge that the amide II Raman band has little sensitivity for protein

secondary structure determination [35, 39]. This was also seen in Fig. 10A when visually inspecting the amide II band. Despite its very well known sensitivity for protein secondary structures, the poor performance of the PLS-LDA model built from the amide I band indicates that this band is little suitable for differentiation between NL and DEN in the studied formulations. This is also noticed when visually inspecting the amide I bands in the blank corrected Raman spectra (Fig. 10A). The amide I band of a DEN sample in a sucrose containing formulation resembles more that of a NL sample with sucrose than that of a DEN sample without sucrose. It is well known that co-lyophilized mixtures of sucrose and polymers (e.g. proteins) interact at a molecular level through hydrogen bonding, forming miscible amorphous molecular dispersions [40-42]. As there is evidence that the C=O stretching vibration is strongly influenced by the polymer-sucrose hydrogen bond [40], it is expected that this will notably influence the amide I band, typically dominated by signals of the backbone C=O stretching vibrations. Thus the strong influence of the extent of hydrogen bonding of the protein with the environment (i.e. hydration state) on the amide I Raman band is here possibly the reason for the modest discrimination between NL and DEN when only considering this band. Among the tested backbone sensitive regions, the amide III band clearly showed being the most suitable for building the model.

3.4.2. Side chain sensitive spectral regions

In Fig. 10A there were also differences between NL and DEN samples in the spectral regions related to some amino acid side chains, i.e. the bands 1320-1500 cm^{-1} , and near 1130 cm^{-1} and 1070 cm^{-1} [35, 43]. This matches again with the variables indicated by the loading and discriminant vectors. The performances of the PLS-LDA models built from these separate side chain sensitive regions are shown in Table 5, and could further improve when combining the different regions. Major changes in the CH_n non-stretching region (1320-1500 cm^{-1}) were linked to a change in 'exposure' of the hydrophobic amino acid residues to the outer surface,

i.e. buried versus exposed to the environment, as they were likewise assigned for heat-denatured proteins [35]. This can be considered as an indication of protein unfolding, since hydrophobic residues will be buried in a natively folded protein, yet a larger fraction of them will be exposed in the partially unfolded states. Wang et al. [44] also found drastic changes in the 1050-1200 cm^{-1} region of heat-denatured insulin and attributed this to protein unfolding.

3.4.3. Combined spectral regions

The classification performance of the PLS-LDA models built from fused backbone or side chain sensitive spectral regions surpassed those from single isolated bands (Table 5). Overall, the models considering the fused side chain sensitive regions and the full range (1033-1725 cm^{-1}) (Table 4) showed the highest accuracy. The latter model was preferred because it combines information from complementary sources, i.e. backbone (i.e. amide III) and side chain sensitive regions.

4. Conclusions

This study, performed on a data set containing NL and a wide distribution of DEN states, demonstrated that Raman spectroscopy is with 95% accuracy able to rapidly and directly discriminate between NL and DEN states of LDH in various freeze-dried formulations. Misclassifications were always the result of assigning mildly (and reversible) DEN samples (with minimal changes in secondary structure detected by the reference technique) to the NL class. Hence, the sensitivity to detect minimal signs of protein denaturation is lower than the chosen reference approach. Therefore, the sensitivity of the Raman model should be further investigated with appropriate data sets. For instance, further research should include (1) What is the limit for detecting first signs of protein denaturation by the Raman model?, (2) How can sensitivity be improved? and (3) What is an acceptable criterion for this limit considering the intended (i.e. future in-line) use of the model? Also, further testing of the model performance

on samples from novel batches is still required. This feasibility study suggests positive outlooks for Raman spectroscopy for the direct and automated analysis of the investigated quality attribute. However, as one analytical method can never fully characterize the conformational state of proteins, the proposed method rather aims giving a *quality indication*. The speed and automatic nature of the method advocate that a transfer of the Raman model to in-line and real-time diagnostic use during freeze-drying processes may be a possible future step.

Acknowledgements

SP is funded as an aspirant of the Research Foundation-Flanders (FWO). She would also like to thank the same institution and Irstea for the funding and hosting, respectively, of a research stay. MD and LH thank the Agency for Innovation by Science and Technology (IWT) for financial support.

References

- [1] www.fda.gov/downloads/Drugs/.../Guidances/ucm070305.pdf (accessed on 26/10/2012)
- [2] T. De Beer, A. Burggraeve, M. Fonteyne, L. Saerens, J.P. Remon, C. Vervaet, Near-infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes, *Int. J. Pharm.* 417 (2011) 32-47.
- [3] N.C. Dingari, G.L. Horowitz, J.W. Kang, R.R. Dasari, I. Barman, Raman spectroscopy provides a powerful diagnostic tool for accurate determination of albumin glycation, *PLoSOne* 7 (2012) 1-11.
- [4] B. Li, P.W. Ryan, B.H. Ray, K.J. Leister, N.M.S. Sirimuthu, A.G. Ryder, Rapid characterization and quality control of complex cell culture media solutions using Raman spectroscopy and chemometrics, *Biotechnol. Bioengin.* 107 (2010) 290-301.

- [5] J.J. Hill, E.Y. Shalaev, G. Zografi, Thermodynamic and dynamic factors involved in the stability of native protein structure in amorphous solids in relation to levels of hydration, *J. Pharm. Sci.* 94 (2005) 1636-1667.
- [6] J.C. Kasper, W. Friess, The freezing step in lyophilization: physico-chemical fundamentals, freezing methods and consequences on process performance and quality attributes of biopharmaceuticals, *Eur. J. Pharm. Biopharm.* 78, 2011, 248-263.
- [7] J.F. Carpenter, M.J. Pikal, B.S. Chang, T.W. Randolph, Rational design of stable lyophilized protein formulations: some practical advice, *Pharm. Res.* 14, 1997, 969-975.
- [8] S.K. Niazi, Handbook of biogeneric therapeutic proteins: regulatory, manufacturing, testing, and patent issues, Informa Healthcare, St Helier, Jersey, 2005, pp 265-289.
- [9] S. Pieters, T. De Beer, J.C. Kasper, D. Boulpaep, O. Waszkiewicz, M Goodarzi, C. Tistaert, W. Friess, J.P. Remon, C. Vervaet, Y. Vander Heyden, Near-infrared spectroscopy for in-line monitoring of protein unfolding and its interactions with lyoprotectants during freeze-drying, *Anal. Chem.* 84 (2012) 947-955.
- [10] A. Barth, C. Zscherp, What vibrations tell us about proteins, *Q. Rev. Biophys.* 35 (2002) 369-430.
- [11] S. Mangialardo, F. Piccirilli, A. Perucchi, P. Dore, P. Postorino, Raman analysis of insulin denaturation induced by high-pressure and thermal treatments, *J. Raman Spectrosc.* 43 (2012) 692-700.
- [12] A.A. Elkordy, R.T. Forbes, B.W. Barry, Study of protein conformational stability and integrity using calorimetry and FT-Raman spectroscopy correlated with enzymatic activity, *Eur. J. Pharm. Sci.* 33 (2008) 177-190.
- [13] R. Tuma, Raman spectroscopy of proteins: from peptides to large assemblies, *J. Raman Spectrosc.* 36 (2005) 307-319.

- [14] R. Bro, Multivariate calibration: What is in chemometrics for the analytical chemist?, *Anal. Chim. Acta* 500 (2003) 185-194.
- [15] A. Al-Hussein, H. Gieseler, The effect of mannitol crystallization in mannitol-sucrose systems on LDH stability during freeze-drying, *J. Pharm. Sci.* 101 (2012) 2534-2544.
- [16] www.sigmaaldrich.com/etc/medialib/docs/Sigma/Enzyme_Assay/I2500enz.Par.0001.File.tmp/I2500enz.pdf
- [17] B.G.M. Vandeginste, D.L. Massart, L.C.M. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of chemometrics and qualimetrics: Part B*, Elsevier Science, Amsterdam, 1998, pp. 87-160.
- [18] R.A. Fisher, The use of multiple measurement in taxonomic problems, *Ann. Eugen.* 7 (1936) 179-188.
- [19] M. Barker, W. Rayens, Partial least squares for discrimination, *J. Chemom.* 17 (2003) 166-173.
- [20] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Assessment of PLS-DA cross validation, *Metabolomics* 4 (2008) 81-89.
- [21] E.T.S. Skibsted, H.F.M. Boelens, J.A. Westerhuis, A.K. Smilde, N.W. Broad, D.R. Rees, D.T. Witte, Net analyte signal based statistical quality control, *Anal. Chem.* 77 (2005) 7103-7114.
- [22] C. L. Hansen, F. van den Berg, M. A. Rasmussen, S.B. Engelsen, S. Holroyd, Detecting variation in ultrafiltrated milk permeates - infrared spectroscopy signatures and external factor orthogonalization, *Chemom. Intell. Lab. Syst.* 104 (2010) 243-248.
- [23] R.C. Pinto, J. Trygg, J. Gottfries, Advantages of orthogonal inspection in chemometrics, *J. Chemom.* 26 (2011) 231-235.

- [24] J.C. Boulet, J.M. Roger, Pretreatments by means of orthogonal projections, *Chemom. Intell. Lab. Syst.* 117 (2012) 61-69.
- [25] J.T. Pelton, L.R. McLean, Spectroscopic methods for analysis of protein secondary structure, *Analytical Biochemistry* 277 (2000) 167-176.
- [26] P.I. Haris, F. Severcan, FTIR spectroscopic characterization of protein structure in aqueous and non-aqueous media, *J. Mol. Catal. B: Enzymatic* 7 (1999) 207-221.
- [27] J. Kong, S. Yu, Fourier transform infrared spectroscopic analysis of protein secondary structures, *Acta Biochim. Biophys. Sinica* 39 (2007) 549-559.
- [28] A. Dong, S.J. Pestrelski, S.D. Allison, J.F. Carpenter, Infrared spectroscopic studies on lyophilization- and temperature-induced protein aggregation, *J. Pharm. Sci.* 84 (1995) 415-424.
- [29] K. Izutsu, N. Aoyagi, S. Kojima, Protection of protein secondary structure by saccharides of different molecular weights during freeze-drying, *Chem. Pharm. Bull.* 52 (2004) 199-203.
- [30] M. Jackson, H.H. Mantsch, The use and misuse of FTIR spectroscopy in the determination of protein structure, *Critical Reviews in Biochemistry and Molecular Biology* 30 (1995) 95-120.
- [31] S.T.R. Walsh, R.P. Cheng, W.W. Wright, D.O.V. Alonso, V. Daggett, J.M. Vanderkooi, W.F. Degradó, The hydration of amides in helices: a comprehensive picture from molecular dynamics, IR, and NMR, *Protein Science* 12 (2003) 520-531.
- [32] S.D. Allison, T.D. Randolph, M.C. Manning, K. Middleton, A. Davis, J.F. Carpenter, Effects of drying methods and additives on structure and function of Actin: mechanisms of dehydration-induced damage and its inhibition, *Arch. Biochem. Biophys.* 358 (1998) 171-181.
- [33] A.M. Herrero, Raman spectroscopy for monitoring protein structure in muscle food systems, *Critical reviews in food science and nutrition* 48 (2008) 512-523.

- [34] K. Kjeldahl, R. Bro, Some common misunderstandings in chemometrics, *J. Chemom.* 24 (2010) 558-564.
- [35] A.M. Herrero, F. Jiménez-Colmenero, P. Carmona, Elucidation of structural changes in soy protein isolate upon heating by Raman spectroscopy, *International Journal of Food Science and Technology* 44 (2009) 711-717
- [36] N.C. Maiti, M.M. Apetri, M.G. Zagorski, P.R. Carey, V.E. Anderson, Raman spectroscopic characterization of secondary structure in natively unfolded proteins: α -synuclein, *J. Am. Chem. Soc.* 126 (2004) 2399-2408.
- [37] A. Hédoux, L. Paccou, S. Achir, Y. Guinet, *In situ* monitoring of proteins during lyophilization using micro-Raman spectroscopy: A description of structural changes induced by deshydration, *J. Pharm. Sci.* 101 (2012) 2316-2326.
- [38] N.C. Maiti, M.M. Apetri, M.G. Zagorski, P.R. Carey, V.E. Anderson, Raman spectroscopic characterization of secondary structure in natively unfolded proteins: α -Synuclein, *J. Am. Chem. Soc.* 126 (2004) 2399-2408.
- [39] R. Tantipolphan, T. Rades, N.J. Medlicott, Insights into the structure of proteins by vibrational spectroscopy, *Curr. Pharm. Anal.* 4 (2008) 53-68.
- [40] L.S. Taylor, G. Zografi, Sugar-polymer hydrogen bond interactions in lyophilized amorphous mixtures, *J. Pharm. Sci.* 87 (1998) 1615-1621.
- [41] T.W. Randolph, Phase separation of excipients during lyophilization: effects on protein stability, *J. Pharm. Sci.* 86 (1997), 1198-1203.
- [42] S.D. Allison, B. Chang, T.W. Randolph, J.F. Carpenter, Hydrogen bonding between sugar and protein is responsible for inhibition of dehydration induced protein unfolding, *Arch. Biochem. Biophys.* 365 (1999) 289-298.

[43] D. Ostrovskii, P. Jacobsson, B. Nystrom, O. Marstokk, H.B.M. Kopperud, Raman spectroscopic characterization of association and thermoreversible gelation in aqueous systems of poly(N-acetamidoacrylamide), *Macromolecules* 32 (1999) 5552-5560.

[44] M. Wang, M. Benford, N. Jing, G. Coté, J. Kameoka, Optofluidic device for ultra-sensitive detection of proteins using surface-enhanced Raman spectroscopy, *Microfluid Nanofluid* 6 (2009) 411-417.

Table 1. Different LDH formulations (appr. 9 mg/ml LDH) prepared in this study.

Formulation number	Excipient composition
1	HIS
2	HIS, sucrose
3	HIS, PEG
4	HIS, sucrose, PEG
5	PHOS
6	PHOS, sucrose
7	PHOS, PEG
8	PHOS, sucrose, PEG
9	TRIS
10	TRIS, sucrose
11	TRIS, PEG
12	TRIS, sucrose, PEG
13	50 mM NaSCN
14	150 mM NaSCN
15	300 mM NaSCN
16	600 mM NaSCN
17	0.9% NaCl
18	1.0% ethanol
19	water

Abbreviations: HIS= histidine buffer, PHOS = phosphate buffer, TRIS = TRIS buffer, all 40 mM at pH 7.4.

Table 2. Design of the freeze-drying runs performed.

Batch	Freezing method	Secondary drying temperature (°C)	Formulations*
1	SL	40	1-19
2	ANN	40	1-19
3	CH	5	1-12
4	ANN	5	1-12
5	FA	40	1-12
6	FA	5	1-12
7	CH	40	1-12
8	SL	5	1-12

Abbreviations: SL = slow cooling, FA = fast cooling, ANN = annealing, CH = cooling hold

* See table 1.

Table 3. Factor level space of the studied process and formulation variables, their mean response per factor level and their significance on the % LDH activity preservation.

Factor	Unit	Levels	Mean response	p value
Freezing method (A)	-	SL	78.89	0.003*
		FA	88.94	
		ANN	90.16	
		CH	89.28	
Secondary drying temperature (B)	°C	5	85.19	0.174
		40	88.44	
Buffer type (C)	-	HIS	91.34	<0.001*
		PHOS	79.63	
		TRIS	89.48	
Sucrose addition (D)	%	0	78.24	<0.001*
		5	95.40	
PEG addition (E)	%	0	86.68	0.908
		1	86.96	

*significant effect ($\alpha = 0.05$).

Abbreviations: SL = slow cooling, FA = fast cooling, ANN = annealing, CH = cooling hold, HIS= histidine buffer, PHOS = phosphate buffer, TRIS = TRIS buffer

Table 4. Performance of the PLS-LDA models built from Raman spectra.

Technique	Spectral region (cm⁻¹)	Number of PLS components	% misclassifications in CV	% misclassifications in test set
Raman	1033-1725	6	5	1/24
Raman blank corrected*	1033-1725	4	5	1/24

*after orthogonal projection to the blank matrix

Table 5. Performance of the PLS-LDA models built from smaller spectral (or fused) regions of the Raman spectra assigned to certain protein information.

Spectral region		Number of PLS components	% misclassifications in CV	% misclassifications in test set
<i>Band assignment</i>	<i>Wavenumber (cm⁻¹)</i>			
Amide I	1600-1700	5	33	7/24
Amide II	1500-1600	8	23	12/24
Amide III	1200-1320	6	7	2/24
Backbone sensitive regions fused*		6	5	2/24
CH _n non-stretch	1320-1500	5	7	2/24
C-N stretch + NH ₃ deformation	1033-1200	4	5	3/24
Side chain sensitive regions fused**		6	5	1/24

* fusion of the amide I, II, III regions

** fusion of the CH_n non-stretch, C-N stretch and NH₃ deformation regions

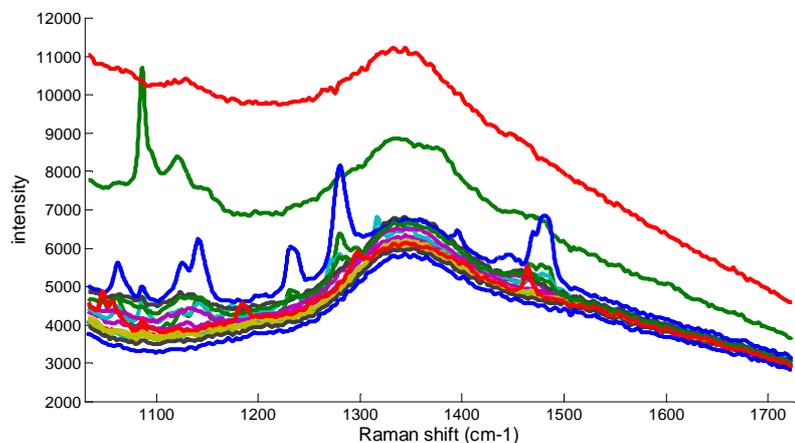


Fig. 1. Raman spectra of the blanks of the investigated formulations.

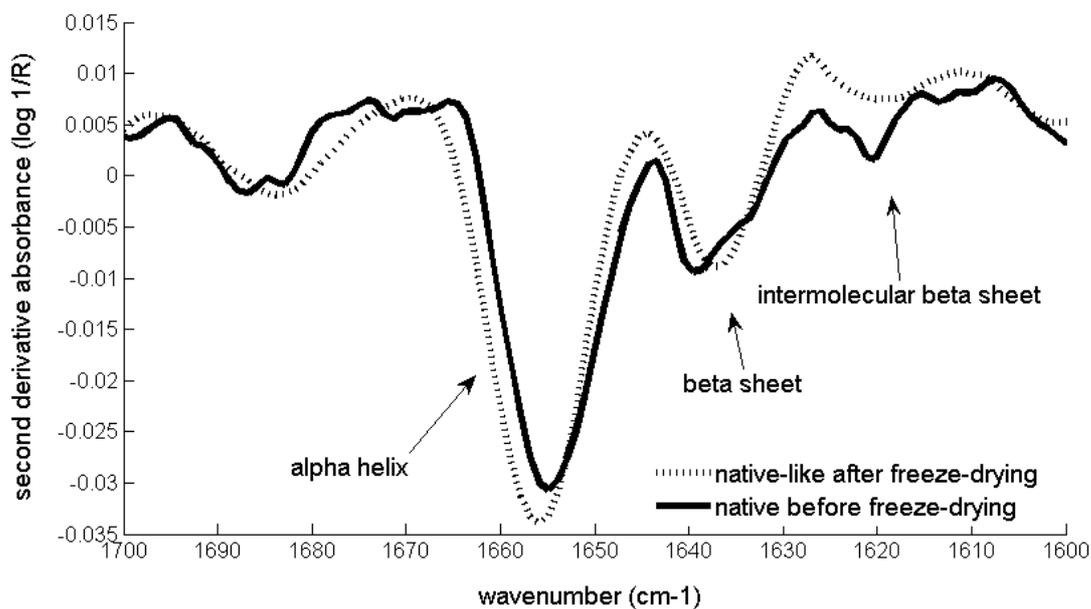


Fig. 2. Second derivative amide I FTIR spectrum of native LDH (in liquid) before freeze-drying (full line) and of native-like freeze-dried LDH (dotted line).

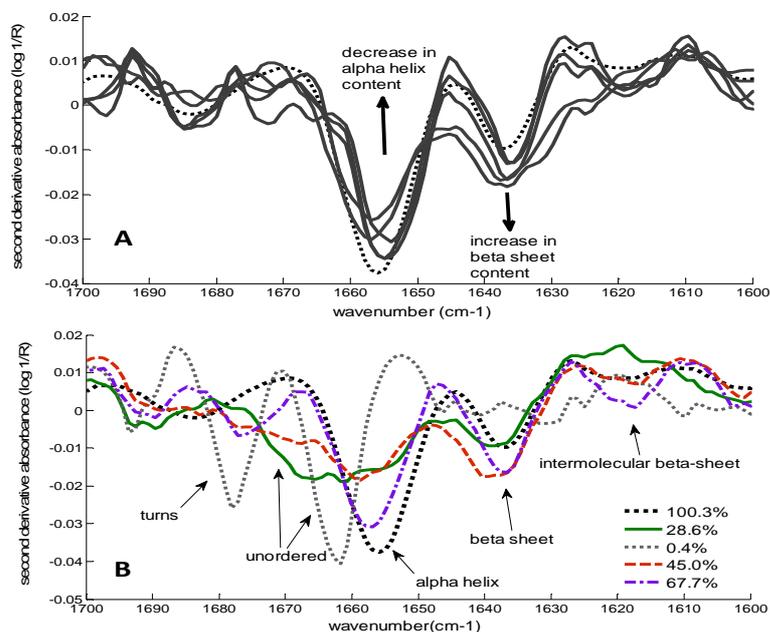


Fig. 3. Second derivative spectra of the amide I band (FTIR) of freeze-dried LDH formulations with different amounts of denaturation/unfolding. The black bold dotted line corresponds to native-like freeze-dried LDH. **(A)** Examples of mildly denatured freeze-dried LDH samples, having remaining LDH activities between 90.3% and 108.9% after reconstitution. **(B)** Examples of more denatured freeze-dried LDH samples, with formation of non-native secondary structures. Legend: % remaining LDH activity after reconstitution.

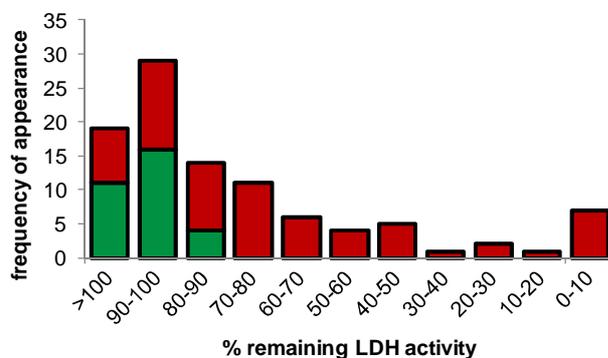


Fig. 4. Histogram of the distribution of the 99 freeze-dried LDH samples in the data set according to their % remaining LDH activity after reconstitution. The divided bars display the conformational state of these samples in the freeze-dried cake, i.e. NL (green) or DEN (red).

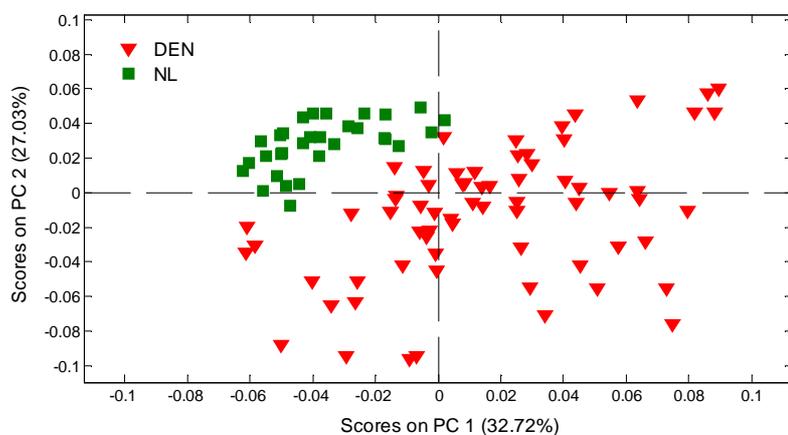


Fig. 5. PCA score plot (PC1-PC2) of the second derivative amide I FTIR spectra of the data set (99 samples) of freeze-dried LDH formulations. The spectra were mean centered before PCA. The spectra are independently labeled based on the visual interpretation of the amide I band (FTIR) as either NL (in green squares) or DEN (in red triangles).

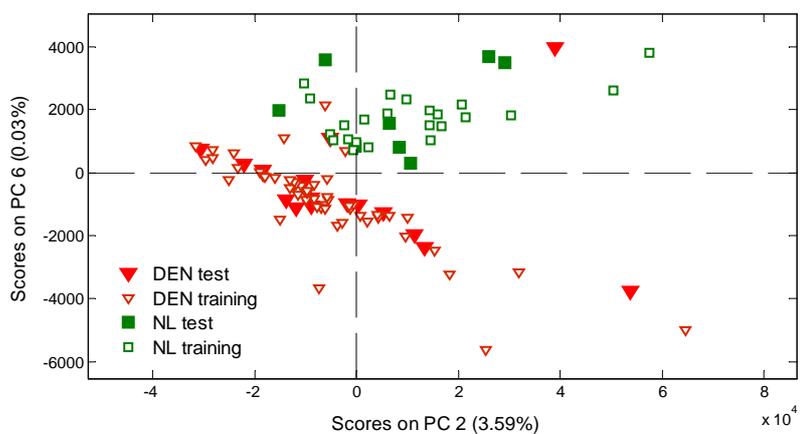


Fig. 6. PCA score plot (PC2-PC6) of the Raman spectra of the data set (99 samples) of the freeze-dried LDH formulations. The spectra were mean centered before PCA. The spectra are independently labeled based on the visual interpretation of the amide I band (FTIR) as either NL (in green squares) or DEN (in red triangles). The division of the samples in training and test set is visualized through filled and unfilled markers.

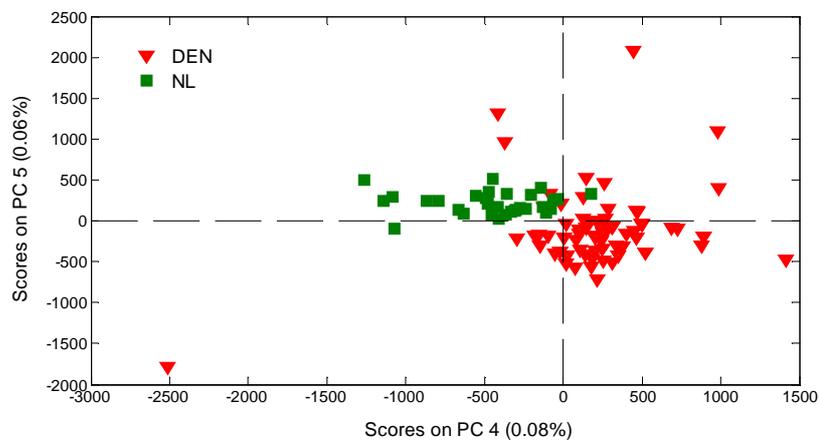


Fig. 7. PCA score plot of the Raman spectra after orthogonal projection to the blank matrix.

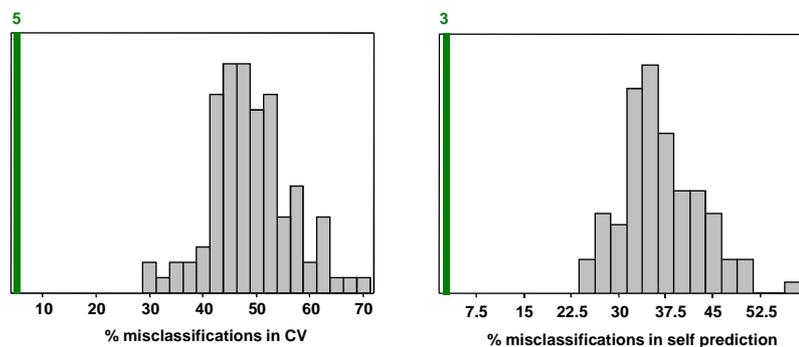


Fig. 8. Class prediction results (expressed as the % misclassifications) based on cross validation (left) and self prediction (right) of the training set with the original labeling (in green bold line) compared to the permuted data (grey histograms).

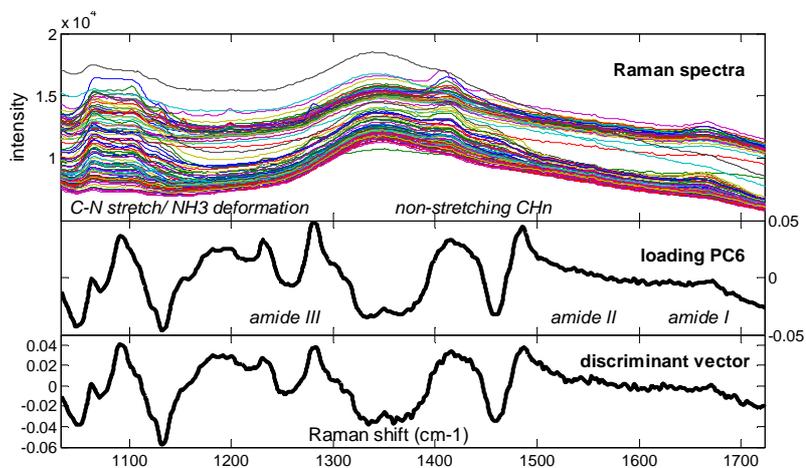


Fig. 9. Unprocessed Raman spectra of the data set (99 samples), the loading vector of the PC6 (0.03%) of the PCA, and the discriminant vector of the PLS-LDA model.

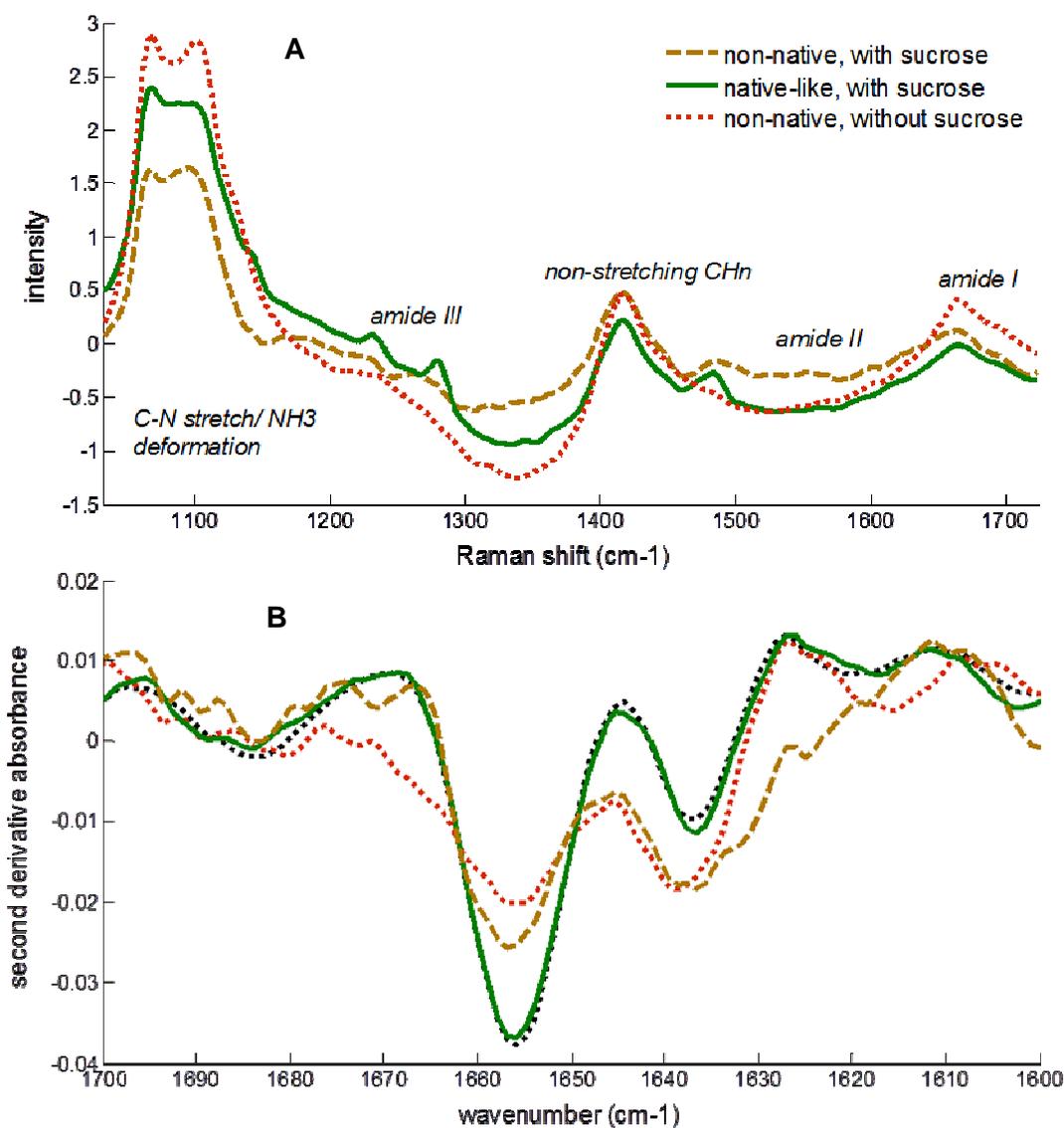


Fig. 10. Raman (A) and amide I FTIR (B) spectra of NL and DEN freeze-dried LDH samples. The Raman spectra were the result of the difference between the sample and corresponding blank spectra, and were subsequently Savitzky-Golay smoothed and normalized. The FTIR spectra were pre-processed as described in Methods and Materials. The black dotted line in (B) corresponds to the average spectrum of 5 NL freeze-dried LDH formulations.