PARCHMENT GLUTAMINE INDEX (PQI): A NOVEL METHOD TO ESTIMATE GLUTAMINE DEAMIDATION LEVELS IN PARCHMENT COLLAGEN OBTAINED FROM LOW-QUALITY MALDI-TOF DATA

Bharath Nair[1,*], Ismael Rodríguez Palomo[2], Bo Markussen[3], Carsten Wiuf[3], and Matthew Collins[1,2]

[1]GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
[2]McDonald Institute for Archaeological Research, University of Cambridge, The United Kingdom
[3]Department of Mathematical Sciences, University of Copenhagen, Denmark

*Corresponding Author

Bharath Nair
GLOBE Institute
University of Copenhagen
Denmark
email: bharath@palaeome.org

Keywords: ZooMS; MALDI-TOF; biocodicology; deamidation; collagen; parchment; PQI

Sun 13th Mar, 2022, 14:06

Contents

1 Abstract **1**

2 Introduction **2**

3 Materials **3**

4 Methods **4**

4.1 Selection of peptides 4

4.2 Pre-processing of raw data 4

4.3 Estimation of deamidation level of peptides 5

4.3.1 Theory 5

4.3.2 Weighted least square and linear regression 6

4.3.3 Deamidation fractions 8

4.4 Estimation of the overall deamidation level 9

4.4.1 PQI Model 9

4.5 Analysis workflow 11

5 Results and Discussion **12**

5.1 Relative rates of deamidation 12

5.2 Applications of PQI 15

6 Conclusion **15**

List of Figures

1 **Summary of the workflow developed as an `R` package `MALDIpqi`.** `MALDIpqi` consists of three steps, a) pre-processing of MALDI-TOF mass spectra, b) estimation of q of selected peptides, and c) prediction of PQI. 22

2 **Illustration of overlapping isotopic distribution (up to 6 peaks) of deamidated and non-deamidated fraction of peptides used in the model before (a), and after preprocessing (b).** We show the baseline estimated on the smoothened spectra and the peaks that are detected after the preprocessing step 3). Note the variation in intensity and the difference in the signal to noise ratio of peaks for each peptide. Whereas there is a clear distinction of individual peaks (hence, high signal to noise ratio) for COL1α1 508-519, peak distinction becomes complex for COL1α1 9-42 or COL1α2 756-789 due to the noisy spectra (hence, less signal to noise ratio). Five randomly selected samples are shown here. 23

3 **Diagnostic plots for PQI model fitting for COL1α1 508-519 indicating a valid statistical model.** a) Residual plot between fitted $\log(q)$ values and Pearson residuals, and b) quantile-quantile plot of residuals wherein the grey line indicates the 1-1 line. Both diagnostic plots indicates that the PQI model is valid except slightly too heavy tails in the normal distribution. 200 randomly selected data points are shown in each plot. 24

4 **Residual plots between fitted values and Pearson residuals** for a) COL1α1 508-519, b) COL1α1 270-291, c) COL1α1 375-396, d) COL1α1 934-963, e) COL1α2 756-789, f) COL1α2 535-567, g) COL1α1 9-42 (5 Pro $\rightarrow$ Hyp), and h) COL1α1 9-42 (7 Pro $\rightarrow$ Hyp), explores the model fitting quality. Random distribution of standardised residuals around 0 within ± 2 suggests that the proposed linear mixed effect model fits well. However, there are a few badly modelled q values for some of the peptides. 200 randomly selected data points are shown in each plot. 24

5 **Normal quantile-quantile plots of Pearson residuals for the model fit** for a) COL1α1 508-519, b) COL1α1 270-291, c) COL1α1 375-396, d) COL1α1 934-963, e) COL1α2 756-789, f) COL1α2 535-567, g) COL1α1 9-42 (5 Pro $\rightarrow$Hyp), and h) COL1α1 9-42 (7 Pro $\rightarrow$Hyp). Except a few deviations from the inserted 1-1 lines, in particularly at the tails for some of the peptides, the quantile-quantile plots indicates normality of the residuals as proposed by the PQI model. 200 randomly selected data points are shown in each plot. 25

6 **Plots depicting applications of PQI:** a) Comparison of PQI against different species used for the production of parchment depicting that manuscripts made out of calfskin were of better quality than the ones made with sheepskin or goatskin. b) Comparison of PQI against different typology the parchments were used for. Biblical manuscripts were written on calfskin, having the highest PQI, which is in accordance with the findings in (Ruffini-Ronzani et al., 2021). Sheepskin was commonly used to produce grammar and theology texts with an intermediate deamidation index. c) Comparison of PQI against production period for parchment locally produced in Orval scriptorium (bottom panel) and for imported parchments (top panel) starting from 9th century until 17th century. The timeline is organised by thirds of a century (early, mid, and late). Orval scriptorium was founded in the early 12th century. The use of calfskin to produce parchments remained constant during the "golden age"(first third and second third of the 13th century) of the scriptorium. d) Comparison of PQI against the thickness indices of codicological units. The thickness was determined depending on the number of folios in the codicological unit (Thickness index; 1 = less than 10 folia, 2 = 11-100 folia, 3 = 101-200 folia, 4 = greater than 200 folia.). (Icons of calf, goat, and sheep created with BioRender.com) 26

List of Tables

1 **List of peptides used in the analysis, using the Brown et al. (2020) nomenclature for ZooMS peptides.** Sequences for each mass were inferred from Mascot analysis of parchment datasets (SF and JW personal communication). Masses consistent with typical hydroxylation patterns of collagen except where indicated by "?". 21

2 **Summary of the dataset** 21

3 **Summary of the extent of deamidation in peptides and Restricted Maximum Likelihood estimates for fixed effects and relative rates of deamidation.** Herein, q is the extent of deamidation in the peptides (from weighted least square linear regression) and $exp(\hat{\theta})$ is the fixed effect estimates (from the PQI model). 21

ABSTRACT

Parchment was used as a writing material in the Middle Ages and was made using animal skins by liming them with $Ca(OH)_2$. During liming, collagen peptides containing Glutamine (Q) undergo deamidation resulting in a mass shift of 0.984 Da. Assessing the extent of deamidation can inform us about parchment production patterns and quality. In this study, we propose a simple three-step workflow, developed as an `R` package called `MALDIpqi()`, to estimate deamidation in parchment derived collagen using low-resolution MALDI spectra. After pre-processing raw spectra, we used weighted least-squares linear regression to estimate Q deamidation levels from the convoluted isotopic envelope for seven collagen-peptide markers. Finally, we employed a linear mixed effect model to predict the overall deamidation level of a parchment sample termed Parchment Glutamine Index (PQI). To test the robustness of the workflow, we applied `MALDIpqi()` to previously published ZooMS data generated from almost an entire library of the Cistercian monastery at Orval Abbey, Belgium. In addition to reliably predicting PQI, we observed interesting patterns pertaining to parchment production. `MALDIpqi()` holds excellent potential for biocodicological and other archaeological studies involving collagen, such as bone, but we also foresee its application in the food and biomedical industry.

INTRODUCTION

Glutaminyl and asparaginyl residues are molecular clocks which deamidate with predetermined half-lives [Robinson et al., 2006]. Glutamine (Gln, Q) deamidation occurs via two mechanisms, 1) direct hydrolysis, and 2) through the formation of a cyclic imide intermediate. Irrespective of the mechanism, instability of reaction intermediates results in slower deamidation rates of Gln. Because the deamidation rates of glutaminyl residues are slower than asparaginyl residues [Robinson et al., 2004, Wright, 1991], it has been advanced that Gln deamidation could be a better tool at our disposal to investigate chemical processes such as assessing the quality of skins in the food and leather industries [Maffia et al., 2004] and the age of fossils [van Doorn et al., 2012, Wilson et al., 2012], although in the latter case [Schroeter and Cleland, 2016] argue that Gln deamidation is an indicator of preservational quality and environmental conditions rather than age (and authenticity) of ancient proteins. Here, we use Gln deamidation to assess variability in parchment production.

Mass spectrometry is well suited to detecting sites of deamidation, which increases the molecular weight of deamidated peptide molecules by 0.984 Dalton (Da). This is easily detected and localised in the sequence by a mass shift in MS2 spectra. However, in MS1, the similarity in mass gain with that of a neutron (1.007 Da) means that the deamidated and non-deamidated isotopic envelopes overlap.

We explore a mathematical approach to derive the level of deamidation in glutamine from MS1 data from peptide mass fingerprinting of collagen. We use an isotopic envelope deconvolution method (similar to the approach used by Wilson et al., 2012) to estimate the extent of glutamine deamidation in selected tryptic peptides but then integrate the individual deamidation estimates to derive an overall index for a given sample. In order to develop the method we have used published MALDI spectra of parchment (e.g. [Fiddyment et al., 2015]) which we have then applied to a newly released data set from Orval Abbey [Ruffini-Ronzani et al., 2021].

Parchment is the dehaired and limed skin of an animal [Reed, 1972, Ryder, 1964]. Liming

is typically the first stage in parchment and leather preparation, it loosens the hairs from the hides, swells the collagen and saponifies some of the skin lipids prior. Gln deamidation occurs when the skins are soaked in lime, a solution of calcium hydroxide $Ca(OH)_2$, which at ambient temperature has an average pH of about 12.4 and is used in different strengths during the parchment making process. The alkaline environment results in direct side chain hydrolysis of the amide group on asparagines and glutamines. A longer exposure (or higher concentration and/or temperature) of lime results in an increase in the extent of deamidation. If not controlled correctly, an excessive exposure to lime can compromise the integrity of skin and to weaken it to such a degree that is no longer usable. By measuring the level of deamidation present in different samples we can start to assess the different production qualities from different regions and time and correlate this to prices and availability of parchment obtained from historic records. Consequently, the extent to which these skins are limed can be interrogated through the measurement of the level of glutamine deamidation.

By assessing the relative rates of deamidation of different tryptic peptides we derived a single value (with associated errors) which we term the Parchment Glutamine Index (PQI). Samples which retain the most intact glutaminyl residues have the highest PQI values; as deamidation increases, PQI falls. In MALDI-TOF mass spectrometry baseline noise in the spectra [Kolibal and Howard, 2006, Krutchinsky and Chait, 2002] results in a distortion of the relative intensity of the peaks across an isotope envelope which in turn affects estimates of deamidation based on the deconvolution of the envelope. Consequently values greater than 1 (ie. no deamidation) are possible due to noisy baselines, while values close to 0 are never observed in parchment, as this would follow complete gelatinisation.

MATERIALS

In order to establish the model, we used available published ZooMS [Buckley et al., 2009] data to establish correlations between the rates of deamidation of different tryptic peptides. We then test our model using data generated from almost the entire library of the Cisterian monastery at Orval Abbey, Belgium [Ruffini-Ronzani et al., 2021]. Explanation of the data

generated can be found in the data article [Bethencourt et al., 2022] and the ZooMS data is uploaded in Zenodo (https://doi.org/10.5281/zenodo.5648106).

METHODS

4.1 Selection of peptides

In order to assess the overall PQI we used as many peptides as possible and selected them based upon the following criteria:

1. They contain at least one glutamine.
2. They are consistently and reliably detected in the MALDI-TOF MS analysis.
3. They are present in all three species used to make parchment (calf, sheep and goat). All these peptides have the same mass in the different species, except m/z 3033 and 3093, which are the equivalent peptides for calf/sheep and goat, respectively.

A final list consisting of eight peptides was compiled (Table 1), of which a maximum of seven can be detected in any one sample due to the equivalence of peptides m/z 3033 and 3093), and used to run the subsequent analysis.

4.2 Pre-processing of raw data

We performed pre-processing of the spectra using the `R` [R Core Team, 2021] package `MALDIquant` [Gibb and Strimmer, 2012]:

- The Savitzky-Golay-filter [Savitzky and Golay, 1964] smoothed the spectra and reduced small, highly frequent noise. This allows for better subsequent baseline and noise estimation and peak maxima determination. We used a moving half-window size of 8, following the recommendation by [Bromba and Ziegler, 1981] of keeping it smaller than the full width at half maximum of the peaks.
- We estimated the baseline (and then subtracted) using the Statistics-sensitive Non-linear Iterative Peak-clipping algorithm (SNIP) [Ryan et al., 1988] implemented in `MALDIquant`; the iterations parameter of the algorithm is set to 20.

- We estimated noise using the SuperSmoother [Friedman, 1984] method. Peaks are detected if they are a maximum with a half-window size of 20 and are above a signal to noise ratio of 1.5.

- Finally, we extracted isotopic-like distributions for each of the selected peptides by finding the canonical m/z value and 5 following peaks (if detected) at the isotopic distance of 1 Da, allowing for a small tolerance deviation of $1.5 \cdot 10^{-4} \cdot mass$ units.

Figure 2 shows the spectra before and after pre-processing for five randomly selected samples.

4.3 Estimation of deamidation level of peptides

Deamidation of a peptide consisting of glutamine (Q) at a single site results in a mass shift of approximately +0.984 Da so that the first peak of the isotope distribution for the deamidated peptide coincides with the second peak of the isotope distribution for the non-deamidated peptide (at the resolution of our data). For a peptide with k possible deamidation sites, each additional deamidation results in a further +0.984 Da mass shift leading to k overlapping isotope distributions. The level of deamidation of a peptide can be estimated by deconvoluting the two overlapping isotopic distributions.

4.3.1 Theory

In order to explain the method, we focus on one peptide and assume there are $m+1$ isotopic peaks of the peptide available with isotope distribution

$$I_i, \quad i = 0, \ldots, m, \quad I_0 + I_1 + \ldots + I_m = 1.$$

We followed the method as described in Wilson et al. [2012] to calculate theoretical isotopic distributions for the peptides. For convenience, we put $I_i = 0$ for $i < 0$. During deamidation, we expect a shift in the isotope distribution

$$P_i = \beta_0 I_i + \beta_1 I_{i-1} + \ldots + \beta_k I_{i-k}, \quad i = 0, \ldots, m,$$

where $1 \leq k < m$ and $\beta_\ell \geq 0$, $\ell = 0, \ldots, k$, is the probability that ℓ positions are deamidated, such that $\beta_0 + \beta_1 + \ldots + \beta_k = 1$. It holds that $P_0 + P_1 + \ldots + P_m = 1$. In the current study, we take $k = 1$.

4.3.2 Weighted least square and linear regression

We developed a general theory assuming $m+1$ measurements, one for each isotope, replicated n times. However, to estimate the overall deamidation level from multiple peptides and replicates simultaneously (see Section 4.4) we obtained estimates of the deamidation level for each of the 3 replicates separately, that is, we apply the theory below with $n = 1$.

Notation for observed intensities of each isotopic peak and replicate:

$$x_{ij}, \quad i = 0, \ldots, m \text{ (isotopic peaks)}, \quad j = 1, \ldots, n \text{ (replicates)}.$$

There might be missing values and/or missing replicates.

The measurements are proportional to P_i, $i = 0, \ldots, m$, hence in particular the measurements do not sum to one. In general, consider the linear model

$$x_{ij} = \gamma_0 I_i + \gamma_1 I_{i-1} + \ldots + \gamma_k I_{i-k} + \epsilon_{ij},$$

where $\gamma_\ell \geq 0$, $\ell = 0, \ldots, k$, are parameters, and ϵ_{ij} is (unobserved) noise. The deamidation fractions are obtained as

$$\beta_\ell = \frac{\gamma_\ell}{\gamma_0 + \ldots + \gamma_k}.$$

We avoided assuming a specific noise structure (for example, normal distributed noise) and used weighted least square to estimate the unknown parameters,

$$\hat{\gamma} = (\widehat{\gamma}_0, \ldots, \widehat{\gamma}_k) = \text{argmin}_{\gamma,c} \sum_{j=1}^{n} \sum_{i=0}^{m} w_{ij}(x_{ij} - c_j X_i \gamma)^2, \tag{1}$$

where $\gamma = (\gamma_1, \ldots, \gamma_k)^\top$ is a column vector and $c = (c_1, \ldots, c_n)$ is a row vector. Here c_j is a scaling factor for the j's replicate with $c_1 = 1$. The idea being that replicates show the same

trend but might vary in signal intensity, hence scaling is required to adjust the parameters. Furthermore, w_{ij} is a weight for the x_{ij}'s data point, and X_i is the ith row of

$$X = \begin{pmatrix} I_0 & 0 & 0 & \dots & 0 \\ I_1 & I_0 & 0 & \dots & 0 \\ I_2 & I_1 & I_0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ I_k & I_{k-1} & I_{k-2} & \dots & I_0 \\ I_{k+1} & I_k & I_{k-1} & \dots & I_1 \\ \vdots & \vdots & \vdots & & \vdots \\ I_m & I_{m-1} & I_{m-2} & \dots & I_{m-k} \end{pmatrix}.$$

In total there are $(k+1)+(n-1) = k+n$ parameters and $n(m+1)$ measurements, assuming none are missing. If measurements are missing the corresponding terms in Equation (1) are omitted.

The estimates can be obtained by weighted linear regression with design matrix X and diagonal weight matrix

$$W_j = \mathrm{diag}(w_{0j}, w_{1j}, \dots, w_{mj}).$$

Let X_j^o be the matrix X with the rows corresponding to the missing measurements of replicate j omitted (upper index o for omitted), let W_j^o be the matrix W_j with the rows and columns corresponding to the missing measurements of replicate j omitted, and let x_j^o be the column vector with missing measurements or replicate j omitted.

Then the estimates can be obtained iteratively by

$$\widehat{\gamma}^{\,i+1} = \left(\sum_{j=1}^{n} (\widehat{c}_j^{\,i})^2 (X_j^o)^T W_j^o X_j^o \right)^{-1} \sum_{j=1}^{n} \widehat{c}_j^{\,i} (X_j^o)^T W_j^o x_j^o, \tag{2}$$

and

$$z_j^{\,i} = \frac{(W_j^o x_j^o)^T X_j^o \widehat{\gamma}^{\,i}}{(W_j^o X_j^o \widehat{\gamma}^{\,i})^T X_j^o \widehat{\gamma}^{\,i}}, \qquad \widehat{c}_j^{\,i} = \frac{z_j^{\,i}}{z_1^{\,i} + \ldots + z_n^{\,i}}, \tag{3}$$

where $\widehat{\gamma}^i$ and $\widehat{c}^i$ are the ith iterated estimates of the column vectors γ and $c = (c_1, \ldots, c_n)$. One continues until the difference between consecutive estimates is small with initial estimate $\widehat{c}^1 = (1, \ldots, 1)$.

Equation (2) is the standard weighted least square estimate assuming the scaling factors are known. Equation (3) is an update of the scaling factors assuming the other parameters are known.

For use in the later stage of the workflow to estimate the overall deamidation level (see Section 4.4.1) we define the *Reliability measure*

$$\text{Reliability} = \sum_{j=1}^{n} \sum_{i=0}^{m} w_{ij} (x_{ij} - \hat{c}_j X_i \hat{\gamma})^2 \tag{4}$$

If there is only one replicate ($n = 1$) or one estimates γ separately for each replicate, then

$$\widehat{\gamma}^{(j)} = \left((X_j^o)^T W_j^o X_j^o \right)^{-1} (X_j^o)^T W_j^o x_j^o, \tag{5}$$

and there is no need for iteration.

The weights might be chosen in different ways. Here, we assume the noise term on measurements is additive, so at the same level for each measurement. In that case, one might apply

$$w_{ij} = \frac{1}{n_{ij}}, \quad i = 0, \ldots, m, \quad j = 1, \ldots, n,$$

where n_{ij} is the estimated background noise for that particular measurement. Within one replicate, n_{ij} is roughly of the same size for all $i = 0, \ldots, m$, but differs between replicates. Herein, we calculate the noise as explained in Section 4.2 using the SuperSmoother method [Friedman, 1984].

4.3.3 Deamidation fractions

By normalisation

$$\widehat{\beta}_\ell = \frac{\widehat{\gamma}_\ell}{\widehat{\gamma}_0 + \ldots + \widehat{\gamma}_k}$$

are estimates of the deamidation fractions corresponding to the parameters β_ℓ, $\ell = 0, \ldots, k$, in Section 4.3.1. In our application of the theory we take $k = 1$, and let q be the (least square estimated) fraction of the remaining, undeamidated Q, i.e. $q = \widehat{\beta}_0 = \frac{\widehat{\gamma_0}}{\widehat{\gamma_0}+\widehat{\gamma_1}}$. Herein, we define the fraction of undeamidated Q (q) as the extent of deamidation with $q = 1$ meaning no deamidation where as $q = 0$ referring to complete deamidation. Histograms of the extent of deamidation (q) of eight peptides are shown in Supplementary figure 1.

4.4 Estimation of the overall deamidation level

We propose a linear mixed effect model, henceforth called the Parchment Glutamine Index (PQI) model, that integrates the estimated values of q and their analytical reliability with which the deamidation level is estimated. The Parchment Glutamine Index (PQI) model thus predicts an overall level of deamidation in a sample and an associated error of prediction.

4.4.1 PQI Model

The (PQI) model is a linear mixed effect model (LMM) that considers log-transformed q values as response variable with individual Peptide as the fixed effects, and Sample and Replicate as the random effects. As a result, the LMM fits the response variable at three different levels, namely, i) peptide, ii) sample, and iii) replicate. Herein, we use log-transformed q as the response variable to reflect the underlying kinetics of the loss of intact glutamine residues which follows pseudo-first order kinetics. Hence, the PQI model predicts the log-transformed deamidation level of a sample from the deamidation level of its individual peptides.

To simplify, we change the notation and structure of the data with respect to the previous section. Herein, the dataset is structured with $\log(q)$ values, *Reliability* estimates (see Section 4.3.2), factors that identifies Peptide (P, with n_P levels), Sample (S, with n_S levels) and Replicate (R, with n_R levels). A summary of the dataset is given in Table 2.

Let $t = 1, \ldots, u$ denote the observation index, where u is the number of rows, and $u = n_S \cdot n_R \cdot n_P$. The statistical model that we will use is the linear mixed effects model given by

$$\log(q_t) = \theta(P_t) + Y(S_t) + Z(S_t, R_t) + \epsilon_t \tag{6}$$

where,

- $\theta(P_1), \ldots, \theta(P_{n_P})$ are the fixed effects of the each peptides,
- $Y(S) \sim \mathcal{N}(0, \sigma_S^2)$ and $Z(S, R) \sim \mathcal{N}(0, \sigma_R^2)$ are the random components from sample, and replicate respectively, and
- $\epsilon_t \sim \mathcal{N}\big(0, (\text{Reliability}_t)^{2\mu} \cdot \sigma_{P_t}^2\big)$ are the residuals.

The variance parameters for random effects are σ_S^2 (sample, S) and σ_R^2 (replicate, R), and for the fixed effects of each peptide are $\sigma_{P_1}^2, \ldots, \sigma_{P_{n_P}}^2$. Furthermore, we scaled the residual variances by *Reliability* values generated from weighted square linear regression, see Eq. (4), to some power 2μ, and the PQI model estimates μ.

The aim is to predict $Y(s)$ given the observations of $\log(q_t)$ for indices t with $S_t = s$. $Y(s)$ is the random effect in the proposed linear mixed effect model that gives us the overall level of deamidation in a given sample, termed as PQI. To formalize this we define

$$J_s = \big\{t = 1, \ldots, u : S_t = s\big\},$$

so that J_s are the set of observation indices belonging to sample s. Additionally, we will use the following notation as given below:

- $|J_s|$ is the size of J_s,
- $M^\top$ is the transpose of a matrix M,
- 1_v is the column vector of length v consisting of 1's,
- $\delta_{x=y}$ is the Dirac delta taking the value 1 when $x = y$ and 0 otherwise,
- $\text{diag}(w)$ is the diagonal matrix with the vector w in the diagonal.

Using this notation we have

$$\begin{pmatrix} Y(s) \\ \{\log(q_t)\}_{t\in J_s} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \{\theta(P_t)\}_{t\in J_s} \end{pmatrix}, \begin{pmatrix} \sigma_S^2 & \sigma_S^2 \cdot 1_{|J_s|}^\top \\ \sigma_S^2 \cdot 1_{|J_s|} & \Xi \end{pmatrix} \right)$$

with $\Xi = \mathrm{Var}\big(\{\log(q_t)\}_{t\in J_s}\big) \in \mathbb{R}^{|J_s|\times|J_s|}$ given by

$$\Xi = \sigma_S^2 \cdot 1_{|J_s|} 1_{|J_s|}^\top + \sigma_R^2 \cdot \left\{\delta_{R_p=R_q}\right\}_{p,q\in J_s} + \mathrm{diag}\Big(\left\{(\mathrm{Reliability}_t)^{2\mu} \cdot \sigma_P^2\right\}_{t\in J_s}\Big)$$

In particular, if we have observations of 3 replicates for all 7 peptides, then $\Xi \in \mathbb{R}^{21\times 21}$ and it is given by

$$\sigma_S^2 \cdot 1_{21} 1_{21}^\top + \sigma_R^2 \cdot 1_3 1_3^\top \otimes \mathrm{diag}(1_7) + \mathrm{diag}\Big(\left\{(\mathrm{Reliability}_t)^{2\mu} \cdot \sigma_P^2\right\}_{t\in J_s}\Big)$$

From the above joint normal distribution it follows by standard formulae that the conditional mean and the conditional variance of $Y(s)$ are given by

$$\begin{aligned} \mathrm{E}\big[Y(s)|\{\log(q_t)\}_{t\in J_s}\big] &= \sigma_S^2 \cdot 1_{|J_s|}^\top \Xi^{-1}\big(\{\log(q_t) - \theta(P_t)\}_{t\in J_s}\big), \\ \mathrm{Var}\big[Y(s)|\{\log(q_t)\}_{t\in J_s}\big] &= \sigma_S^2 - \sigma_S^4 \cdot 1_{|J_s|}^\top \Xi^{-1} 1_{|J_s|} \end{aligned}$$

Note that $\mathrm{E}\big[Y(s)|\{\log(q_t)\}_{t\in J_s}\big]$ is the prediction of $Y(s)$, and $\mathrm{Var}\big[Y(s)|\{\log(q_t)\}_{t\in J_s}\big]$ is the associated prediction variance.

4.5 Analysis workflow

We performed all the computations in the statistical programming language `R` [R Core Team, 2021] using the following packages: `nlme`[Pinheiro et al., 2021], `dplyr`[Wickham et al., 2021], and `ggplot2`[Wickham, 2016]. The prediction of $Y(s)$ can be extracted directly from the `lme`-object using the function `nlme::ranef()`. However, the computation of the prediction variance requires implementation of the matrix formula. We developed an `R` package `MALDIpqi` for the whole workflow consisting of pre-processing of raw spectra, estimation

of deamidation rates using weighted least squares linear regression, and applying linear mixed effect model to estimate the overall deamidation index of parchment, available at https://github.com/ismaRP/MALDIpqi.

RESULTS AND DISCUSSION

We applied the workflow starting from pre-processing of raw data followed by estimation of deamidation levels of individual peptides and finally predicting the overall sample deamidation level, termed as PQI.

We let q denote the fraction of remaining non-deamidated Q (see Section 4.3.3) in the peptide under consideration. We estimated q values for selected eight peptides using weighted least squares linear regression on the isotopic distribution as obtained from MALDI-TOF spectra. Table 3 shows the first and third quartile of estimated q values to give an overview of deamidation levels in the peptides.

5.1 Relative rates of deamidation

Assuming the deamidation level over time follows first-order kinetics (N. E. Robinson & Robinson, 2004), then denoting the amount of non-deamidated Q of a particular peptide at time t by $[Q]_t$, we have

$$[Q]_t = [Q]_0 e^{-kt}, \tag{7}$$

and hence

$$q = \frac{[Q]_t}{[Q]_0} = e^{-kt} \tag{8}$$

where, $[Q]_0$ is the amount of Q at time 0, k is the deamidation rate constant, and t is the age of the sample.

Let k_1 and k_2 denote the deamidation rate constants of Peptide 1 and Peptide 2, respectively, from a particular sample. Similarly, let q_1 and q_2 denote the deamidation fractions of Peptide 1 and Peptide 2, respectively. Then the ratio of the deamidation rate constant of

Peptide 2 to that of Peptide 1 can be expressed as

$$\frac{k_2}{k_1} = \frac{\log(q_2)}{\log(q_1)}. \tag{9}$$

Considering Peptide 1 to deamidate slowly, we obtain a relative deamidation rate profile for each sample that might be compared across samples. Using this ratio overcomes the need to establish the absolute rate of deamidation. A limitation of this approach is when the levels of deamidation are very high, the true extent may be obscured due to the correspondingly high influence of noise in the spectra.

Among the eight peptides the rates (see Table 3) are compared relative to COL1α1 508-519 (m/z 1105.58, VQG) which deamidates the most slowly. The only peptide which does not have a Glycine (Gly) C-terminal to the Gln is peptide COL1α1 376-396 (m/z 2040.97, GQD), and this is the most rapidly deamidated. This rapid deamidation explains the clustering of fitted values towards the left of the residual plot for this peptide, as shown in Figure 4. COL1α1 934-963 (m/z 2689.25), contains two glutamine residues both oriented in the same plane (PQGFQG), but even their combined rate is nevertheless slower than COL1α1 376-396. Curiously m/z 3084.42 (identified by Mascot [Perkins et al., 1999] as COL1α1 9-42) has a rate of deamidation which is one third that of m/z 3116.40, which was interpreted as the same peptide but with only two less oxygen atoms. The most probable explanation is that one of these peptides may have been misidentified, as it seems unlikely that additional oxidation/hydroxylations would have such a significant effect on the rate of deamidation.

The log-transformed q values were then transferred to the PQI model, which fits the deamidation at peptide level and predicts the sample level deamidation. We used the `lme` function in `R` from the package `nlme` to fit the linear mixed effect model using restricted maximum likelihood (REML). PQI model estimates for sample level variance(σ_S^2) is 0.01 and replicate level variance (σ_R^2) is $4.09 * 10^{-11}$ with $\mu = -0.06$. The back-transformed peptide level fixed effect estimates $\exp(\hat{\theta})$ and the relative levels of deamidation are given in Table 3.

We validated the model fitting using residual plots (residuals vs. fitted values) and nor-

mal quantile-quantile plots of the Pearson residuals (residuals standardized by their estimated standard deviation). Residual plots are the most common diagnostic tool to assess the constant variation of residuals [Pinheiro and Bates, 2000]. Diagnostic plots for PQI model fitting for the slowest deamidating peptide, COL1α1 508-519, are shown in Figure 3 showing a valid statistical model except slightly too heavy tails in the normal distribution. Pearson residuals for almost all samples are randomly distributed around 0 with magnitudes ranging between ± 2, and without any concerning patterns as depicted in Figure 4.

Normal quantile-quantile plots compare quantiles of Pearson residuals to quantiles of standard normal distribution. Linearity of the quantile-quantile plot implies that residuals are normally distributed as proposed in the Parchment Glutamine Index (PQI) model. With the exception of a few data points on both tails of the quantile-quantile plots, the model fits the deamidation well (Figure 5). The few data points that do not fall onto the quantile-quantile line for peptides COL1α2 756-789 and COL1α2 535-567 (see Figure 5) is the result of a low signal to noise ratio that affects the correct estimation of q values from the MALDI-TOF spectra.

The PQI model predicts the sample level $\log(q)$ value and we therefore argue that the $\exp(\log(q))$ value depicts the overall extent of deamidation in a sample, termed as the Parchment Glutamine Index (PQI). From the samples considered in the analysis, PQI predicted from the model ranges from 0.47 to 1.26 with 54% of the values above 1, although theoretically the full PQI range is from 0 to 1. A low value of PQI implies more liming and hence low quality of parchment while a value of 1 indicates no deamidation. The model generates some PQI values greater than 1 due to the problem of accurate baseline correction. A histogram of predicted PQI values are shown in Supplementary figure 2.

From the PQI model, we estimated peptide level fixed effects and sample level random effects. Whilst the fixed effect is the mean $\log(q)$ of each peptide, the random effect $Y(s)$ is the predicted overall deamidation level in a sample, PQI (see Section 4.4.1). A few q values for the peptides COL1α2 756-789 and COL1α2 535-567 were not fitted well in the PQI model implying inaccurate estimates from spectral peaks with low signal to noise ratio. Relative to

COL1α1 508-519, COL1α1 375-396 displayed higher rates of deamidation where as COL1α1 270-291 displayed lowest rate (see Table 3).

5.2 Applications of PQI

As an illustration of the application of the PQI model we explore levels of deamidation from a collection of manuscripts from the library at Orval Abbey, Belgium[Ruffini-Ronzani et al., 2021] by comparing PQI values with species, thickness, typology, and production period as shown in Figure 6.

PQI vs species: PQI varies with species, parchment produced from calf has higher PQI values, than those produced from sheep or goat (Figure 6a). Goat skin parchments had the lowest PQI, suggesting that they were the most aggressively limed. We also observe the highest PQI values in calfskin used for Bible, and we speculate that this would probably have been perceived as of the best quality. Law and science texts tended to use the lowest quality parchment, although within each group of texts there was considerable variation (Figure 6b).

PQI vs parchment thickness: Sheep showed the widest range of values, and goat had the lowest PQI values (most deamidated). Estimated thickness suggests that the (small number of) very finest parchment (Thickness index 1) are not of the best quality, an unexpected finding which should be explored further. There is nevertheless a gradual fall in PQI in the next three thickness groups as might be expected, with the greatest levels of deamidation in the coarsest membranes as shown in Figure 6d.

PQI vs time: A temporal comparison of parchment production from the 9th century until the 17th century reveals the highest PQI values occurred during the "golden age" of the Orval scriptorium (first half of the 13th century), presumably before the disastrous fire of 1252 (see Figure 6c).

CONCLUSION

The PQI model allows us to reliably estimate the quality of parchment production by deriving an index which combines the extent of deamidation of seven tryptic peptide markers

from MALDI-TOF analysis (also known as ZooMS). It uses a three step workflow, the preprocessing of spectra for optimal assessment of each mass envelope, estimating deamidation levels in peptides using weighted least square linear regression, and finally, predicting the overall deamidation level in a sample using a linear mixed effect model. Each step is coded in `R` as a package `MALDIpqi()`, enabling high throughput analysis of large datasets.

We applied the workflow to 3714 MALDI-TOF spectra from parchments in the library of the Orval Abbey and were able to observe a number of patterns. There is a large variation in PQI between membranes but some patterns are evident. Coarser membranes are more heavily limed than thinner folia, and calfskin is more gently processed than sheep and goatskin. Both of these would be anticipated based upon our knowledge of parchment production, although we were surprised by the low PQI values of goatskin, which is typically less fatty than sheepskin and therefore does not require such long exposure to saponify and hence remove lipids. More subtle observations are also apparent at Orval Abbey; texts acquired after the fire of 1252 are on average worse than those acquired during the so-called golden age which preceded it.

In addition to this biocodicological application of PQI, livestock collagen is widely used in the food industry and biomedicine. Therefore the developed three step workflow offers a simple method to assess levels of Gln deamidation of processed collagen.

ACKNOWLEDGEMENTS

BN is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 801199. At the time of producing this work, IR and MC were funded by the European Union's EU Framework Programme for Research and Innovation Horizon 2020 under Grant Agreement No. 787282 (B2C), and IR is currently the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 956410. BM is funded by the University of Copenhagen through the Data Science Laboratory. MC is also supported by Danish National Research Foundation (DNRF128) and CW by the Independent Research Fund Denmark (grant number: 8021-00360B) and the University of Copenhagen through the Data+ initiative. We thank Julie Wilson (JW) and Sarah Fiddyment (SF) for their helpful comments and advice.

REFERENCES

J. H. Bethencourt, I. R. Palomo, S. Hickinbotham, B. Nair, S. Soncin, M. Dieu, M. Collins, and O. Deparis. Data from "a biocodicological analysis of the medieval library and archive from orval abbey, belgium". *J. Open Archaeol. Data*, 10(0), Feb. 2022.

M. U. A. Bromba and H. Ziegler. Application hints for Savitzky-Golay digital smoothing filters. *Anal. Chem.*, 53(11):1583–1586, Sept. 1981.

M. Buckley, M. Collins, J. Thomas-Oates, and J. C. Wilson. Species identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 23(23):3843–3854, 2009. doi: https://doi.org/10.1002/rcm.4316. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/rcm.4316.

S. Fiddyment, B. Holsinger, C. Ruzzier, A. Devine, A. Binois, U. Albarella, R. Fischer, E. Nichols, A. Curtis, E. Cheese, M. D. Teasdale, C. Checkley-Scott, S. J. Milner, K. M. Rudy, E. J. Johnson, J. Vnouček, M. Garrison, S. McGrory, D. G. Bradley, and M. J. Collins. Animal origin of 13th-century uterine vellum revealed using noninvasive peptide fingerprinting. *Proc. Natl. Acad. Sci. U. S. A.*, 112(49):15066–15071, Dec. 2015.

J. H. Friedman. A variable span scatterplot smoother. *Stanford University Technical Report*, No 5, 1984.

S. Gibb and K. Strimmer. MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, 28(17):2270–2271, Sept. 2012.

J. Kolibal and D. Howard. MALDI-TOF baseline drift removal using stochastic bernstein approximation. *EURASIP J. Adv. Signal Process.*, 2006(1):1–9, Dec. 2006.

A. N. Krutchinsky and B. T. Chait. On the nature of the chemical noise in MALDI mass spectra. *J. Am. Soc. Mass Spectrom.*, 13(2):129–134, Feb. 2002.

G. J. Maffia, M. A. Seltzer, P. H. Cooke, and E. M. Brown. Collagen processing. *Journal of the American Leather Chemists Association*, 99, 2004.

D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, Dec. 1999.

J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2021. URL https://CRAN.R-project.org/package=nlme. R package version 3.1-153.

J. C. Pinheiro and D. M. Bates, editors. *Linear Mixed-Effects Models: Basic Concepts and Examples*, pages 3–56. Springer New York, New York, NY, 2000. ISBN 978-1-4419-0318-1. doi: 10.1007/978-1-4419-0318-1_1. URL https://doi.org/10.1007/978-1-4419-0318-1_1.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

R. Reed. *Ancient skins, parchments and leathers*. Seminar Press, London; New York, 1972.

N. E. Robinson, Z. W. Robinson, and others. Structure-dependent nonenzymatic deamidation of glutaminyl and asparaginyl pentapeptides. *J. Peptide Res.*, 2004.

N. E. Robinson, V. Zabrouskov, J. Zhang, K. J. Lampi, and A. B. Robinson. Measurement of deamidation of intact proteins by isotopic envelope and mass defect with ion cyclotron resonance fourier transform mass spectrometry. *Rapid Commun. Mass Spectrom.*, 20(23): 3535–3541, 2006.

N. Ruffini-Ronzani, J.-F. Nieus, S. Soncin, S. Hickinbotham, M. Dieu, J. Bouhy, C. Charles, C. Ruzzier, T. Falmagne, X. Hermand, M. J. Collins, and O. Deparis. A biocodicological analysis of the medieval library and archive from orval abbey, belgium. *R Soc Open Sci*, 8 (6):210210, June 2021.

C. G. Ryan, E. Clayton, W. L. Griffin, S. H. Sie, and D. R. Cousens. SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nucl. Instrum. Methods Phys. Res. B*, 34(3):396–402, Sept. 1988.

M. L. Ryder. Parchment—its history, manufacture and composition. *J. Soc. Arch.*, 2(9): 391–399, Jan. 1964.

A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36(8):1627–1639, July 1964.

E. R. Schroeter and T. P. Cleland. Glutamine deamidation: an indicator of antiquity, or preservational quality? *Rapid Commun. Mass Spectrom.*, 30(2):251–255, Jan. 2016.

N. L. van Doorn, J. Wilson, H. Hollund, M. Soressi, and M. J. Collins. Site-specific deamidation of glutamine: a new marker of bone collagen deterioration. *Rapid Commun. Mass Spectrom.*, 26(19):2319–2327, Oct. 2012.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.7.

J. Wilson, N. L. van Doorn, and M. J. Collins. Assessing the extent of bone degradation using glutamine deamidation in collagen. *Anal. Chem.*, 84(21):9041–9048, Nov. 2012.

H. T. Wright. Nonenzymatic deamidation of asparaginyl and glutaminyl residues in proteins. *Crit. Rev. Biochem. Mol. Biol.*, 26(1):1–52, 1991.

Peptide	m/z	nQ	Sequence	*Bos taurus* 🐄 *Ovis aries* 🐑 *Capra hircus* 🐐
COL1α1 508-519	1105.58	1	GV**Q**GPPGPAGPR (1 Hyp)	🐄 🐑 🐐
COL1α1 270-291	2019.95	1	GEPGPTGI**Q**GPPGPAGEEGKR (2 Hyp)	🐄 🐑 🐐
COL1α1 375-396	2040.97	1	TGPPGPAG**Q**DGRPGPPGPPGAR (3 Hyp)	🐄 🐑 🐐
COL1α1 934-963	2689.25	2	GFSGL**Q**GPPGPPGSPGE**Q**GPSGASGPAGPR (2 Hyp)	🐄 🐑 🐐
COL1α2 756-789	3033.50	1	GPSGEPGTAGPPGTPGP**Q**GLLGAPGFLGLPGSR (5 Hyp)	🐄 🐑
COL1α2 535-567	3093.48	1	GPSGEPGTAGPPGTPGP**Q**GFLGPPGFLGLPGSR (5 Hyp)	🐐
COL1α1 9-42	3084.42	2	GLPGPPGAPGP**Q**GF**Q**GPPGEPGEPGASGPMGPR (5 Hyp)	🐄 🐑 🐐
COL1α1 9-42	3116.40	2	GLPGPPGAPGP**Q**GF**Q**GPPGEPGEPGASGPMGPR (7 Hyp?)	🐄 🐑 🐐

Table 1: List of peptides used in the analysis, using the Brown et al. (2020) nomenclature for ZooMS peptides. Sequences for each mass were inferred from Mascot analysis of parchment datasets (SF and JW personal communication). Masses consistent with typical hydroxylation patterns of collagen except where indicated by "?".

Variable	Usage	Type	Range
Sample (S)	Random effect	Categorical	Levels: $n_S = 3714$
Technical replicate (R)	Random effect	Categorical	Levels: $n_R = 3$
Peptides (P)	Fixed effect	Categorical	Levels: $n_P = 8$
log(q)	Response variable	Continuous	[-5.50:0.84]
Reliability	Weight	Continuous	[0:436634.4]

Table 2: Summary of the dataset

Peptide	1st quartile of q	Median q	3rd quartile of q	$\exp(\hat{\theta})$	Relative rates of deamidation
COL1α1 508-519	1.01	1.08	1.15	1.07	1.00
COL1α1 270-291	0.93	1.06	1.14	1.03	6.56
COL1α1 375-396	0.32	0.49	0.78	0.46	15.38
COL1α1 934-963	0.82	1.00	1.18	0.98	5.71
COL1α2 756-789	0.72	0.89	1.08	0.85	9.76
COL1α2 535-567	0.64	0.78	0.93	0.74	10.34
COL1α1 9-42	0.81	0.92	0.96	0.91	9.66
COL1α1 9-42$*$	0.63	0.79	0.94	0.74	9.41

Table 3: Summary of the extent of deamidation in peptides and Restricted Maximum Likelihood estimates for fixed effects and relative rates of deamidation. Herein, q is the extent of deamidation in the peptides (from weighted least square linear regression) and $exp(\hat{\theta})$ is the fixed effect estimates (from the PQI model).

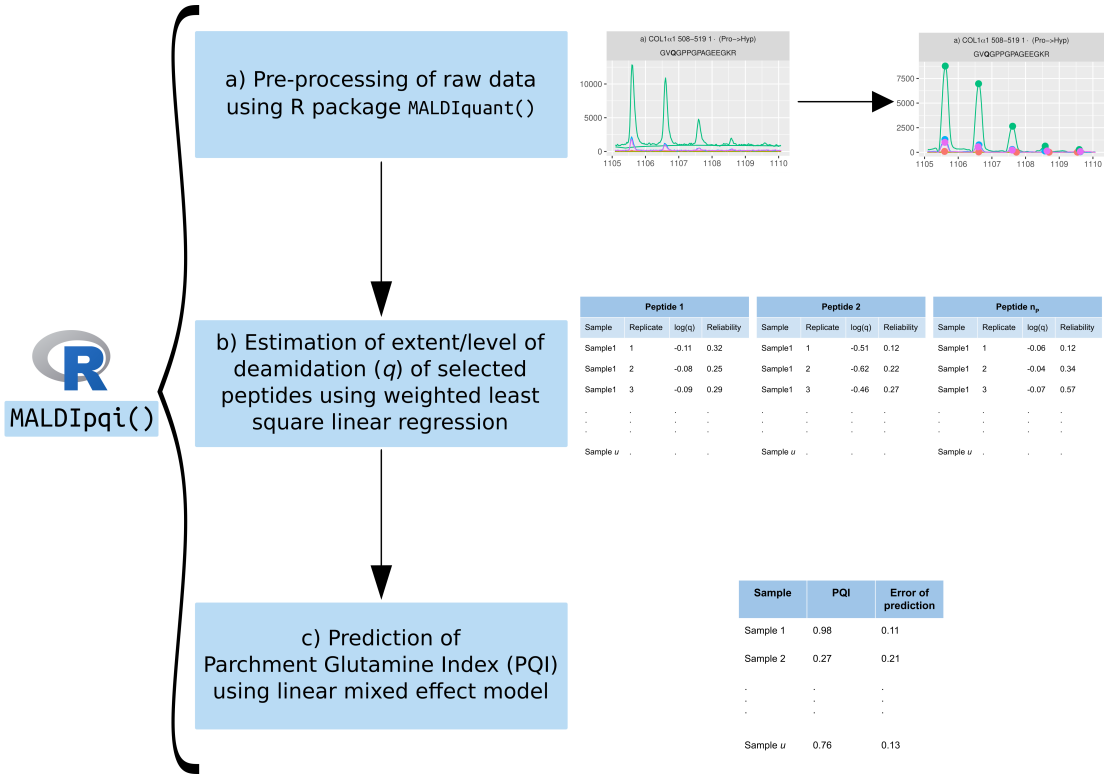

Figure 1: Summary of the workflow developed as an R package `MALDIpqi`. `MALDIpqi` consists of three steps, a) pre-processing of MALDI-TOF mass spectra, b) estimation of q of selected peptides, and c) prediction of PQI.

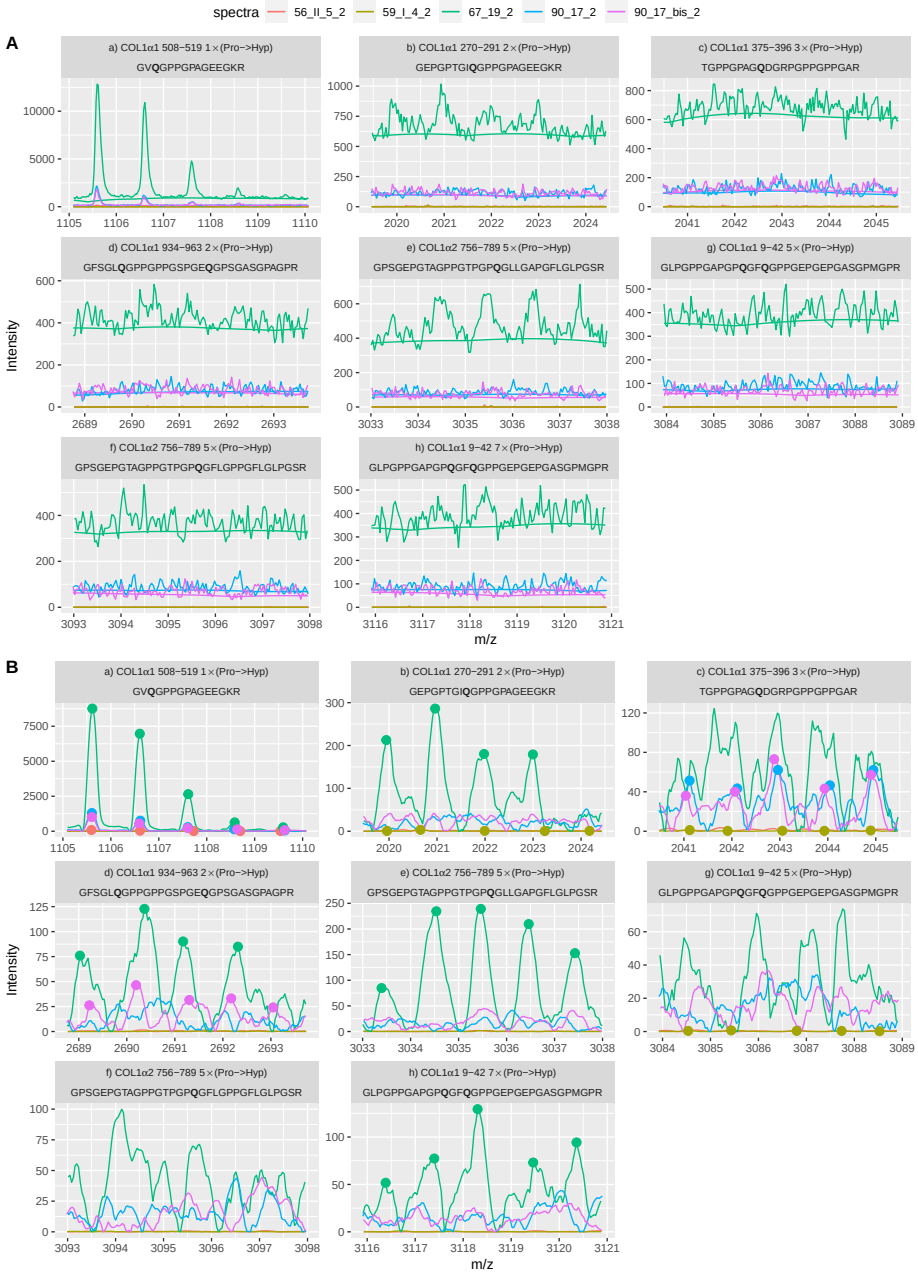

Figure 2: Illustration of overlapping isotopic distribution (up to 6 peaks) of deamidated and non-deamidated fraction of peptides used in the model before (a), and after preprocessing (b). We show the baseline estimated on the smoothened spectra and the peaks that are detected after the preprocessing step 3). Note the variation in intensity and the difference in the signal to noise ratio of peaks for each peptide. Whereas there is a clear distinction of individual peaks (hence, high signal to noise ratio) for COL1α1 508-519, peak distinction becomes complex for COL1α1 9-42 or COL1α2 756-789 due to the noisy spectra (hence, less signal to noise ratio). Five randomly selected samples are shown here.

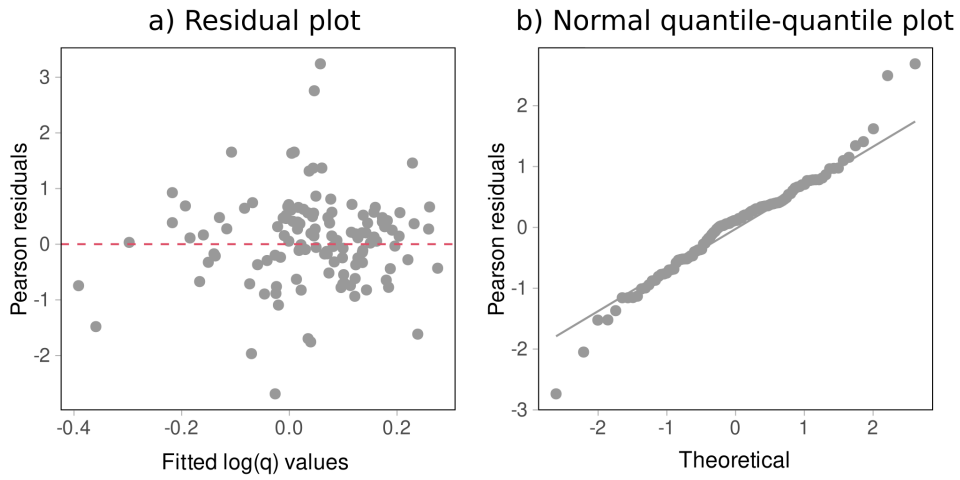

Figure 3: Diagnostic plots for PQI model fitting for COL1α1 508-519 indicating a valid statistical model. a) Residual plot between fitted log(q) values and Pearson residuals, and b) quantile-quantile plot of residuals wherein the grey line indicates the 1-1 line. Both diagnostic plots indicates that the PQI model is valid except slightly too heavy tails in the normal distribution. 200 randomly selected data points are shown in each plot.

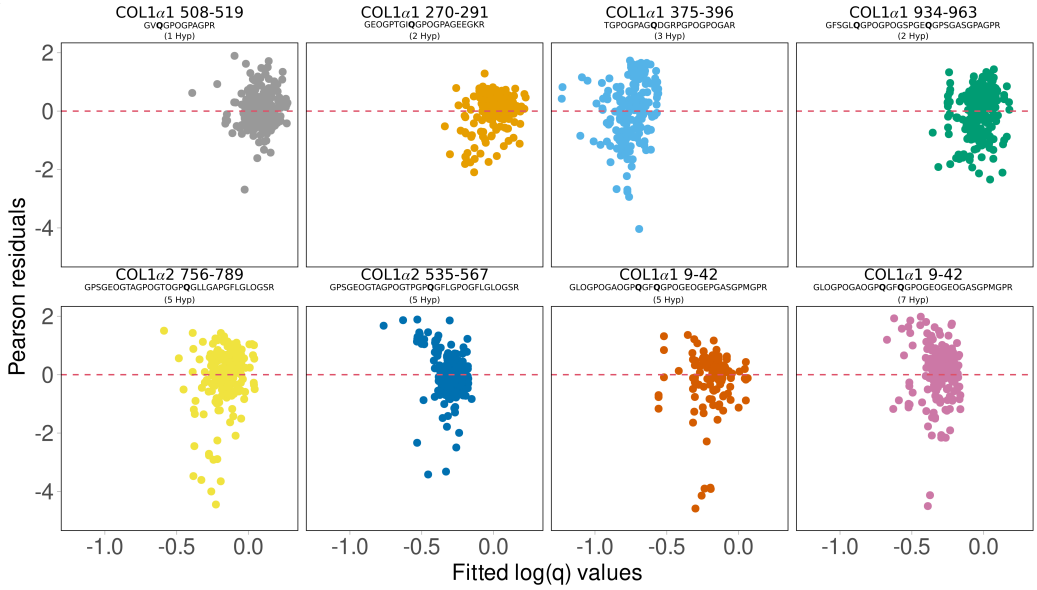

Figure 4: Residual plots between fitted values and Pearson residuals for a) COL1α1 508-519, b) COL1α1 270-291, c) COL1α1 375-396, d) COL1α1 934-963, e) COL1α2 756-789, f) COL1α2 535-567, g) COL1α1 9-42 (5 Pro $\rightarrow$ Hyp), and h) COL1α1 9-42 (7 Pro $\rightarrow$ Hyp), explores the model fitting quality. Random distribution of standardised residuals around 0 within ± 2 suggests that the proposed linear mixed effect model fits well. However, there are a few badly modelled q values for some of the peptides. 200 randomly selected data points are shown in each plot.

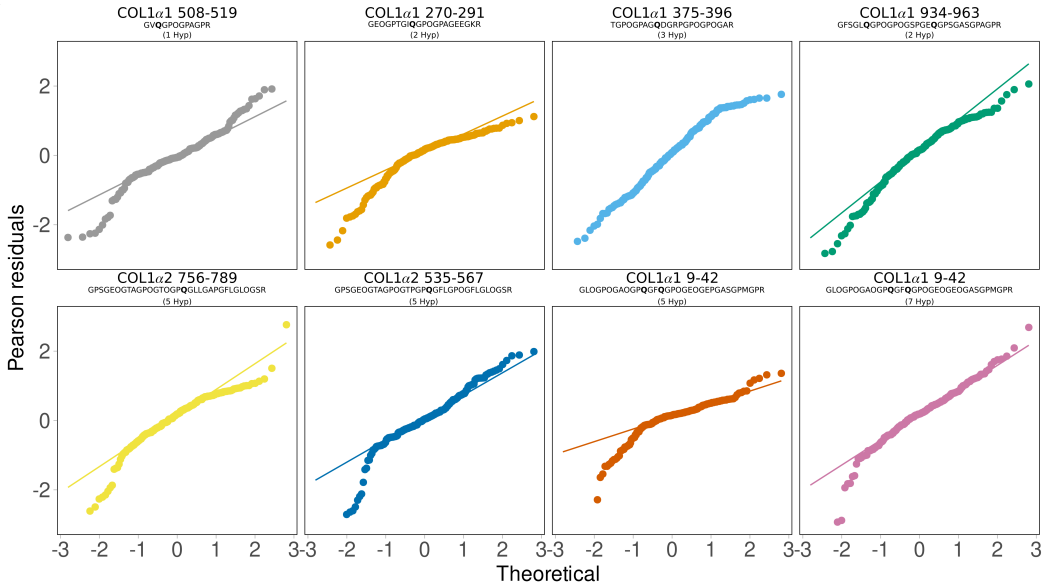

Figure 5: Normal quantile-quantile plots of Pearson residuals for the model fit for a) COL1α1 508-519, b) COL1α1 270-291, c) COL1α1 375-396, d) COL1α1 934-963, e) COL1α2 756-789, f) COL1α2 535-567, g) COL1α1 9-42 (5 Pro $\rightarrow$Hyp), and h) COL1α1 9-42 (7 Pro $\rightarrow$Hyp). Except a few deviations from the inserted 1-1 lines, in particularly at the tails for some of the peptides, the quantile-quantile plots indicates normality of the residuals as proposed by the PQI model. 200 randomly selected data points are shown in each plot.

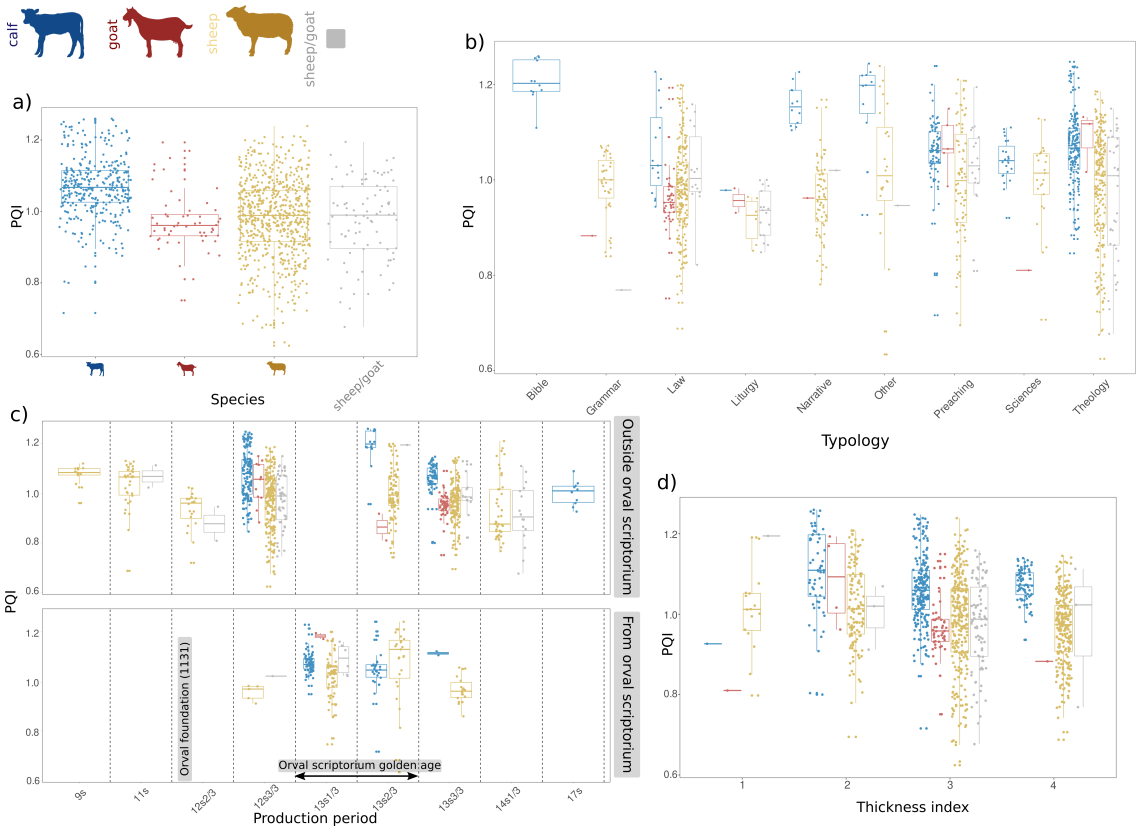

Figure 6: Plots depicting applications of PQI: a) Comparison of PQI against different species used for the production of parchment depicting that manuscripts made out of calfskin were of better quality than the ones made with sheepskin or goatskin. b) Comparison of PQI against different typology the parchments were used for. Biblical manuscripts were written on calfskin, having the highest PQI, which is in accordance with the findings in (Ruffini-Ronzani et al., 2021). Sheepskin was commonly used to produce grammar and theology texts with an intermediate deamidation index. c) Comparison of PQI against production period for parchment locally produced in Orval scriptorium (bottom panel) and for imported parchments (top panel) starting from 9th century until 17th century. The timeline is organised by thirds of a century (early, mid, and late). Orval scriptorium was founded in the early 12th century. The use of calfskin to produce parchments remained constant during the "golden age"(first third and second third of the 13th century) of the scriptorium. d) Comparison of PQI against the thickness indices of codicological units. The thickness was determined depending on the number of folios in the codicological unit (Thickness index; 1 = less than 10 folia, 2 = 11-100 folia, 3 = 101-200 folia, 4 = greater than 200 folia.). (Icons of calf, goat, and sheep created with BioRender.com)