PEPA test: fast and powerful differential analysis from relative quantitative proteomics data using shared peptides

Laurent Jacob*,[1], Florence Combes[2], Thomas Burger[2]

[1]Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Lyon, France
[2]BIG-BGE (Université Grenoble Alpes, CNRS, CEA, INSERM), 38000 Grenoble, France
* laurent.jacob@univ-lyon1.fr

Abstract

We propose a new hypothesis test for the differential abundance of proteins in mass-spectrometry based relative quantification. An important feature of this type of high-throughput analyses is that it involves an enzymatic digestion of the sample proteins into peptides prior to identification and quantification. Due to numerous homology sequences, different proteins can lead to peptides with identical amino acid chains, so that their parent protein is ambiguous. These so-called shared peptides make the protein-level statistical analysis a challenge and are often not accounted for. In this article, we use a linear model describing peptide-protein relationships to build a likelihood ratio test of differential abundance for proteins. We show that the likelihood ratio statistic can be computed in linear time with the number of peptides. We also provide the asymptotic null distribution of a regularized version of our statistic. Experiments on both real and simulated datasets show that our procedures outperforms state-of-the-art methods. The procedures are available via the `pepa.test` function of the `DAPAR` Bioconductor R package.

Keywords: Likelihood ratio test; Differential analysis; Discovery proteomics; Shared peptides;

1 Introduction

Quantitative proteomics refers to the identification and quantification of the proteins present in a biological sample. This field has rapidly grown mature over the last decade, allowing for a refined understanding of a wide variety of biomolecular processes: phenotypes of new forms of life, such as giant viruses [17], host-pathogen cell interactions [15] or microbial infections [12]. As many other omics sciences, it is based on a large scale sequencing approach, that is bound to high throughput measurements whose statistical processing is a central issue. The most classically used measurement pipeline [4] is referred to as *relative bottom-up MS/MS quantification* [24]. The term *bottom-up* refers to the fact

that proteins are not directly identified: instead, they are first digested by an enzyme into smaller molecules called *peptides*, that are easier to analyze by MS/MS. The term *MS/MS* refers to the fact that two kinds of mass spectrometry (MS) measurements are alternatively performed. This instrumental pipeline is particularly useful in discovery proteomics, where the goal is to find a short-list of proteins that are significantly differently abundant. Several samples are collected under different biological conditions (*e.g.*, in healthy vs. disease, wild-type vs. mutant, etc.) and analyzed with the aforementioned pipeline, leading to a list of identified peptides and their intensity for each sample. Several methods have been proposed to detect differentially abundant proteins and can be divided in two main families: *peptide-based* and *aggregation-based* methods, also referred to as summarization-based in [9]. In the latter ones, peptide-level information is first aggregated at the protein level and proteins are then tested for differential abundance using these summaries [20, 19, 5, 6]. They mostly differ by the set of peptides that they aggregate and the test statistic they use after aggregation. Peptide-based models on the other hand do not rely on an aggregation step and build a test statistic using peptide intensities as a sampling unit [2, 1, 3, 10]. A more detailed discussion of the literature and preliminary comparisons among aggregation-based methods are provided in Section A of the Supplementary Material.

For both families of methods, deciding which proteins are differentially abundant from peptide level observations is made difficult by the presence of *shared peptides* (as opposed to *protein-specific* ones): due to the numerous homology sequences between different genes, some peptides can belong to several distinct proteins [16]. This problem has long been reported in the literature [13]. It affects all data produced by a bottom-up approach – including label-free as studied in this paper – as well as isobaric tag data, where similar aggregation problems were reported [11, 14]. According to [5], up to 50% of peptides can be shared in the proteome of complex organisms. We illustrate the extent of this phenomenon and its effect on the detection of differentially abundant proteins in Section B of the Supplementary Material, on a proteomic dataset obtained from the LC-MS/MS analysis of mouse liver samples.

To the best of our knowledge, the few solutions to the shared peptide problem available in the literature have hardly spread to proteomics platforms, as they are computationally expensive and do not scale to large proteomics datasets. [2] for example exploit a model akin to the one we use in this paper but include a factor representing the peptide-specific relationship between the measured peptide intensity and its actual abundance. This factor causes the negative log-likelihood to be non-convex in the set of parameters making its minimization non-trivial and possibly expensive – the authors restrict themselves to peptides shared by no more than two proteins. No algorithm or code is available for this method, to the best of our knowledge. More recently [1] have proposed AllP, which is also based on a similar model as our work but uses a log-normal model. The use of this distribution corresponds to a common assumption on observed peptide distribution [18]. Unfortunately, maximizing the corresponding likelihood is more computationally demanding than that of the normal distribution we use, as no closed form maximizer is available. As reported in its original article, a synthetic datasets with 100 proteins requires 3 days of computation and for such a dataset the algorithm does not converge in 18% of the cases.

In this context, our contribution is four fold:

1. We introduce a linear model which relates measured peptide MS intensities to latent protein abundances, and use it to build a PEptide based Protein differential Abundance (PEPA) likelihood ratio test which accounts for shared peptides.

2. Our linear model allows for faster estimation than existing log-linear models but still involves an $nq \times (p+q)$ design matrix, where n is the number of samples, q the number of peptides and p the number of proteins. Computation of our statistic can therefore be slow and require a large amount of memory in practice using naive least square implementations. We show that our likelihood ratio statistic can be computed in $\mathcal{O}(nq)$ nevertheless, making it compatible with proteomic platform throughput.

3. We empirically observe that regularized estimators of the variance parameter lead to more powerful tests. We show that under the null hypothesis of homogeneity, the regularized log-likelihood ratio statistic is still asymptotically χ^2 distributed up to some normalization.

4. We provide R code for all our methods in the `DAPAR` Bioconductor R package, so they can be routinely used by proteomic practitioners.

2 Methods

We consider a proteomics experiment measuring the intensity of q previously identified peptides that map onto a set of p known proteins. A biological sample consists of observed intensities for q peptides. Each peptide in turn can belong to several among p proteins, and the abundance of a peptide is the sum of the abundance of all proteins containing this peptide. Formally, if the proteins have respective abundance values $\theta_1, \ldots, \theta_p$ in a sample, then the abundance of peptide k in this sample should be $\sum_{j=1}^{p} x_{kj}\theta_j$, where $x_{kj} = 1$ if peptide k belongs to protein j, 0 otherwise.

2.1 Model

The observed intensities $\tilde{y}_k$ from an MS/MS experiment are typically modeled as samples from a log-normal distribution [2, 18, 1]:

$$\ln \tilde{y}_k | X, \theta, \alpha \sim \mathcal{N}\left(\ln \sum_{j=1}^{p} x_{kj}\theta_j + \alpha_k, \sigma^2\right), \tag{1}$$

where $\sigma^2 > 0$ is the variance of the distribution, α_k is a peptide-specific effect, $X \in \{0,1\}^{q \times p}$ is a binary matrix whose elements are the x_{kj} and $\theta \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}^q$ are vectors containing the protein abundances and peptide effects respectively.

The parameters of interest for differential analysis are the protein abundances $\theta_1, \ldots, \theta_p$. They are unobserved, and we want to test whether they change between two experimental conditions of interest. More precisely, we assume the $n = n_1 + n_2$ biological samples are measured under two different

experimental conditions (n_1 under the first condition, n_2 under the second) and we want to test

$$\mathbf{H}_0 : \theta^{(1)} = \theta^{(2)} \quad \text{vs.} \quad \mathbf{H}_1 : \theta_j^{(1)} \neq \theta_j^{(2)}, \theta_l^{(1)} = \theta_l^{(2)} \forall l \neq j, \tag{2}$$

where $\theta^{(1)}$ and $\theta^{(2)} \in \mathbb{R}^p$ are protein abundance vectors under the two conditions and j is the protein being tested for differential abundance.

The data we use to test (2) consist of $q \times n$ i.i.d. intensity measurements $\{\tilde{y}_k^i\}_{k=1,\dots,q}^{i=1,\dots,n}$. To make the analysis and computation easier, we make the approximation that:

$$\ln \tilde{y}_k | X, \theta, \alpha, \sigma^2 \sim \mathcal{N}\left(\sum_{j=1}^{p} x_{kj}\theta_j + \alpha_k, \sigma^2\right), \tag{3}$$

which amounts to replacing $\ln \sum_{j=1}^{p} x_{kj}\theta_j$ by its first order Taylor expansion, as discussed in Supplementary Material C. Consistently, we observe that the likelihood ratio statistic relying on (3) leads to good empirical performances to test $\mathbf{H}_0$, even on real data or simulated ones from log-normal distributions. Computing this statistic only requires solving linear problems, which is typically much faster than solving the non-linear problem associated with (1). It also makes the computation amenable to an even faster procedure which we introduce in Section 2.2. In the rest of this paper, we therefore assume that the $\tilde{y}_k^i$ are sampled from (3) and let y_k^i denote the log intensities $\ln \tilde{y}_k^i$.

The maximum likelihood (ML) estimators of $\beta = (\theta, \alpha)$ for observations $\{y_k^i\}_{k=1,\dots,q}^{i=1,\dots,n}$ from (3) under $\mathbf{H}_0$ and $\mathbf{H}_1$ are obtained by ordinary least square:

$$\hat{\beta}_k \in arg\,min_{\beta_k} \|y - \mathbf{X}_k\beta_k\|^2,\ k = 0, 1. \tag{4}$$

$\mathbf{X}_0$ is a vertical concatenation of n copies of the $\mathbb{R}^{q \times p+q}$ $(X\ I_q)$ matrix where I_q is the identity matrix in $\mathbb{R}^q$, and $\mathbf{X}_1$ is a vertical concatenation of n_1 copies of the $\mathbb{R}^{q \times p+q+1}$ $(X_{-j}\ x_j\ 0\ I_q)$ matrix and n_2 copies of the $\mathbb{R}^{q \times p+q+1}$ $(X_{-j}\ 0\ x_j\ I_q)$ matrix:

$$\mathbf{X}_0 = \begin{pmatrix} X & I_q \\ \vdots & \vdots \\ X & I_q \end{pmatrix} \in \mathbb{R}^{nq \times p+q}, \qquad \mathbf{X}_1 = \begin{pmatrix} X_{-j} & x_j & 0 & I_q \\ \vdots & \vdots & \vdots & \vdots \\ X_{-j} & 0 & x_j & I_q \end{pmatrix} \in \mathbb{R}^{nq \times p+q+1}. \tag{5}$$

X_{-j} is the X matrix without its j-th columns x_j. The matrix $y \in \mathbb{R}^{nq}$ contains all $\{y_k^i\}_{k=1,\dots,q}^{i=1,\dots,n}$, *i.e.*, the n_1 peptide intensity measurements under the first condition, followed by n_2 ones under the second. Finally, $\beta_0 = (\theta, \alpha) \in \mathbb{R}^{p+q}$ and $\beta_1 = (\theta_{-j}, \theta_j, \theta'_j, \alpha) \in \mathbb{R}^{p+q+1}$.

Considering σ as a fixed parameter, the ML estimator of σ^2 is

$$\hat{\sigma}_k^2 = (nq)^{-1} \|y - \mathbf{X}_k\hat{\beta}_k\|^2, k = 0, 1. \tag{6}$$

Using an inverse gamma prior $\sigma^2 \sim$ Inv-Gamma(α, β) for $\alpha \geq 0$, the maximum a posteriori (MAP) estimator of σ^2 is

$$\hat{\sigma}_k^2 = (3 + 2\alpha)^{-1} (nq)^{-1} \|y - \mathbf{X}_k\hat{\beta}_k\|^2 + s, k = 0, 1, \tag{7}$$

with $s = 2\beta$. This estimator is implicitly used in test statistics like SAM [21]. It amounts to regularizing the variance estimate and can lead to better power than t-tests to detect differential abundance when only few samples are available. To choose s in practice, we generalize the heuristic of [21]: we compute our statistic for all proteins across a grid of values of s and retain the one leading to the smallest coefficient of variation of the statistic across variance levels. The motivation of the heuristic is that the amplitude of the regularized statistic should not be determined by the variance of the residuals.

Individual effects such as our peptide effect α_k are commonly modeled as random variables and endowed with a prior distribution. In our experiments, using a fixed or random α_k made little difference so we opted for the fixed effect model, as it is amenable to fast computation using Proposition 1 (see below).

Finally, in practice X often involves disjoint sets of proteins with no peptide in common. When testing (2) for a protein j, we only use peptides whose observation affects the estimation of θ_j. Concretely, we identify connected components in the bipartite graph whose nodes are peptides and proteins with edges between each peptide and its parent proteins, and apply our procedure to each connected component separately. Section D of the Supplementary Material further discusses this point and introduces a heuristic procedure which does not separate connected components.

2.2 Computation of the test statistics

We consider the likelihood ratio statistic for (2) under model (3):

$$\lambda(\hat{\sigma}_0^2, \hat{\sigma}_1^2) = nq \left(\ln \hat{\sigma}_0^2 - \ln \hat{\sigma}_1^2 \right), \tag{8}$$

where $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ are obtained by solving either (6) for ML estimation or (7) for MAP estimation. Both estimators require solving the least square problem in (4) for $k = 0$ and $k = 1$. A naive implementation explicitly storing the $nq \times (p+q)$ and $nq \times (p+q+1)$ design matrices $\mathbf{X}_k$ would be slow and possibly run out of memory even for small n when dealing with connected components involving thousands of peptides. In the experiments on simulated data with 50% of shared peptides presented in Section 4.1, the largest connected component contained 993/1000 proteins and 4981/5000 peptides. Computing the test statistic (8) under $\mathbf{H}_0$ across only 2×3 samples took 4.8Gb of memory and 23 minutes on a 3Ghz i7 core with an implementation using the `lm` function of R. Computing the statistic (8) only requires to know the sum of the squared residuals $\min_\beta \|y - \mathbf{X}_k\beta\|^2$ and not necessarily $\hat{\beta}_k \in arg\,min_\beta \|y - \mathbf{X}_k\beta\|^2$. An implementation of the projection matrix $\mathbf{X}_0(\mathbf{X}_0^\top \mathbf{X}_0)^\dagger \mathbf{X}_0^\top y$ over the span of $\mathbf{X}_0$ relying on the singular value decomposition of $(X\ I_q)$ took 1.6Gb and 3 minutes. We now show how this sum of squared residuals can be computed in linear time with the size of y. On the same simulation, our approach allowed to compute (8) in 0.002 seconds with only a marginal memory usage.

Proposition 1 *Let* $\mathbf{X}_0$ *and* $\mathbf{X}_1$ *be defined as in* (5) *and* $y \in \mathbb{R}^{nq}$ *contain the* n

stacked $\{y_k^i\}_{k=1,\dots,q}^{i=1,\dots,n}$ *samples, then*

$$\min_\beta \|y - \mathbf{X}_0\beta\|^2 = \|y\|^2 - n\,\|\bar{y}\|^2 \tag{9}$$

$$\min_\beta \|y - \mathbf{X}_1\beta\|^2 = \|y\|^2 - n\,\|\bar{y}\|^2 - n^{-1}n_1n_2\left(\|x_j\|^{-1}x_j^\top\left(\bar{y}^{(1)} - \bar{y}^{(2)}\right)\right)^2, \tag{10}$$

where $\bar{y} = n^{-1}\sum_{i=1}^n y^i \in \mathbb{R}^q$ *is the average across the* n *samples and* $\bar{y}^{(l)} \in \mathbb{R}^q$ *is the average across the* n_l *samples under condition* $l \in \{1, 2\}$.

The proof is in Section E of the Supplementary Material. When X is a binary matrix, $\|x_j\|^{-1}x_j^\top\left(\bar{y}^{(1)} - \bar{y}^{(2)}\right) = q_j^{\frac{1}{2}}\left(\bar{y}_j^{(1)} - \bar{y}_j^{(2)}\right)$, where $\bar{y}_j^{(l)} \in \mathbb{R}$, $l = 1, 2$ is the average across samples under condition l of the log-intensities of all peptides belonging to protein j and $q_j = \|x_j\|^2$ is the number of peptides belonging to protein j. The additional term in $\hat{\sigma}_1$ can then be interpreted as the squared difference between the average log intensities across peptides between the two conditions.

An important consequence of Proposition 1 is that the log-likelihood ratio statistic (8) can be computed in $\mathcal{O}(nq)$ by computing averages of subsamples of y, without storing the $\mathbf{X}_k$ matrices or diagonalizing the $(\mathbf{X}_k^\top\mathbf{X}_k)$ matrices. Using Proposition 1 to compute the likelihood ratio statistic moves the computational bottleneck to the identification of connected components, as illustrated in Section 4.1. The heuristic we discuss in Section D of the Supplementary Material skips this identification step and retains all peptides for each tested protein.

2.3 Null distribution of λ

By Wilk's theorem [23], we know that λ converges in law to a χ_1^2 distribution as $n \to \infty$ and the y_{ik} are sampled i.i.d. under $\mathbf{H}_0$ (*i.e.*, $\theta^{(1)} = \theta^{(2)}$) and when using maximum likelihood estimators of $(\theta, \alpha, \sigma^2)$ from (4) and (6). This result provides asymptotic levels for our test, as rejecting $\mathbf{H}_0$ when $\lambda > \chi_{1,\alpha}^2$, where $\chi_{1,\alpha}^2$ is the $1-\alpha$ quantile of the χ_1^2 distribution, asymptotically leads to a false positive rate of α. The asymptotic is in n even though the number of sampling units is nq, as the size of the parameter α also increases with q.

When using a MAP estimator (7) for σ^2, Wilk's theorem does not hold anymore, and indeed we observed in our experiments that the null distribution of λ under $\mathbf{H}_0$ deviates from the χ_1^2 distribution. However, Proposition 2 shows that multiplying λ by a constant factor is enough to recover a correct asymptotic level.

Proposition 2 *Let* $\beta, \beta' \in \mathbb{R}^p$, $\sigma \in \mathbb{R}_+$, $x_i \in \mathbb{R}^p$, $y_i|x_i, \beta, \sigma^2 \overset{iid}{\sim} \mathcal{N}(x_i^\top\beta, \sigma^2), i = 1, \dots, n_1$, $y_i|x_i, \beta, \sigma^2 \overset{iid}{\sim} \mathcal{N}(x_i^\top\beta', \sigma^2), i = n_1 + 1, \dots, n = n_1 + n_2$, $\hat{\sigma}_0^2$ *and* $\hat{\sigma}_1^2$ *denote the maximum likelihood estimator of* σ^2 *under* $\mathbf{H}_0 : \beta = \beta'$ *and* $\mathbf{H}_1 : \beta_k = \beta'_k \forall k \neq j, \beta_j \neq \beta'_j$ *respectively,* $s \geq 0$ *and* $\lambda(\hat{\sigma}_0^2, \hat{\sigma}_1^2) = n\left(\ln\hat{\sigma}_1^2 - \ln\hat{\sigma}_0^2\right)$. *If* $\beta = \beta'$, *then*

$$\frac{\sigma^2 + s}{\sigma^2}\lambda(\hat{\sigma}_0^2 + s, \hat{\sigma}_1^2 + s) \to \chi_1^2.$$

The proof is in Section F of the Supplementary Material. In practice, σ^2 is unknown but we obtained satisfying test levels on both real and simulated data by using the maximum likelihood $\hat{\sigma}^2$ instead.

3 Experimental setting

We use both simulated datasets and datasets resulting from spike-in samples to evaluate the performance of our test procedure. Simulated datasets allow us to control key parameters such as the proportion of shared peptides, but the conclusions we draw only hold for data which behave like the simulations. On the other hand, because of the difficulties that are inherent to the wet-lab procedures (see Section G.2 of the Supplementary Material), it is not possible to prepare real samples with an indisputable ground truth and which contain an important proportion of shared peptides. As a result, spike-in and simulated datasets provide complementary views on the performance evaluation. The R codes for our experiments is available at https://github.com/ThomasBurger/pepa-validation.

3.1 Simulated datasets

We simulate peptide intensities for each of $n_1 = n_2 = 3$ samples under two biological conditions, for $q = 5000$ peptides belonging to $p = 1000$ proteins. We purposely use a generative model – described in Supplementary Material G.1 – which differs from the normal model (3) used in our testing procedure. This allows us to obtain more realistic data, and to assess the robustness of our method to deviations from the model.

3.2 Spike-in datasets

The true set of differentially expressed proteins is generally not known in real data, making it difficult to compare differential analysis methods. We resort to using spike-in samples, for which the true set of differentially abundant proteins is known. We use the two datasets described in [7], and which are available in the DAPARdata R package [22]. These two datasets contain 6 samples that were prepared using the equimolar human protein mixture Sigma UPS1, including 48 human proteins. Their differential abundance ratios are 2 and 2.5 for the first and second datasets, that are respectively referred to as Exp1_R2_pept and Exp1_R25_pept. A limitation of this dataset is that it contains few shared peptides contrarily to real human samples. To cope with this issue, we artificially add shared peptides in these two real datasets, by merging pairs of peptides. In Section 4, we report experiments with respectively 0, 120, 200 and 280 artificial shared peptides. These numbers are to be compared to the total numbers of human peptides in the datasets that are 211 (respectively 290) in Exp1_R2_pept (respectively Exp1_R25_pept). More details about the spike-in datasets are available in Section G.2 of the Supplementary Material.

3.3 Compared methods

We evaluate two versions of PEPA test, both using the likelihood ratio statistic (8): PEPA-ML relies on the ML estimator of the variance (6), and PEPA-

MAP on the MAP estimator (7) of the variance with an inverse-gamma prior. Comparing these two procedures provides insights on the respective interests of the likelihood ratio test itself and the variance regularization. These two methods are compared to several other reference methods.

The only other peptide-based model that accounts for all shared peptides is AllP [1], which cannot cope with large scale datasets that we consider in our experiments so we compare to methods that only account for protein-specific peptides. The first one, referred to as PeptideModel, performs a two-sample t-test for each protein, where each group is formed by pooling all protein-specific peptides across all biological samples in one condition. This corresponds to a likelihood ratio test using model (3) without its peptide effect, and restricted to protein-specific peptides, so the performance increment between PeptideModel and PEPA-ML quantifies the interest of accounting for shared peptide in the context of this particular model. We also include the latest peptide-based method MSqRob from [10]. Like PeptideModel, MSqRob relies on a linear model using the peptides as its sampling unit but introduces a few improvements: first, a ridge penalty on the estimated effects; second, an empirical Bayes estimator of the variance (both of which should help when few unique peptides are available); and finally, a robust loss function to deal with outliers. In the absence of shared peptides, it is therefore similar to PEPA-MAP but uses a different type of regularization of the variance, an additional regularization of the regression parameters and a robust loss.

We also consider two aggregation-based models: the first one, denoted AllSpec-SAM, performs the SAM-test proposed by [21] (with automatic tuning of the fudge factor parameter, as discussed by [8]) over each protein summarized by the sum of intensities of all of its protein-specific peptides. The second one, denoted Top3Spec-TT, performs a t-test over each protein summarized by the sum of intensities of its 3 most abundant protein-specific peptides. AllSpec-SAM is the most accurate aggregation-based model, that is the best combination of aggregation and test, as discussed in Sections A.2 and A.4 of the Supplementary Material. On the other hand, because of its simplicity, Top3Spec-TT remains one of the most used methods on proteomics platforms, and it provides baseline performances.

To compare the performances of the different methods, we construct precision-recall (PR) curves. In order to stabilize the results, the PR curves are averaged over 30 runs for simulated data, and 10 runs on the spike-in data when using the peptide merging procedure.

3.4 Additional assumptions made by PEPA

We evaluate all methods on their ability to test $\mathbf{H'}_0 : \theta_j^{(1)} = \theta_j^{(2)}$ versus $\mathbf{H'}_1 : \theta_j^{(1)} \neq \theta_j^{(2)}$ which do not make assumptions on the differential abundance of other proteins. All competitors are built upon $\mathbf{H'}_0$ and $\mathbf{H'}_1$. By contrast, PEPA needs to make assumptions on the differential abundance of all proteins in the same connected component since (i) it exploits shared peptides and (ii) it is based on a likelihood ratio statistic. The likelihood ratio statistic for each protein involves other proteins, whose differential abundance status is not specified by $\mathbf{H'}_0$ and $\mathbf{H'}_1$. A Wald statistic exploiting shared peptides would not have this issue, but would not be amenable to the acceleration offered by Proposition 1.

Specifically PEPA tests $\mathbf{H}_0$ versus $\mathbf{H}_1$, assuming that no other protein is differentially abundant. Alternatively, we could assume that all other proteins are differentially abundant or that $\theta^{(1)} - \theta^{(2)}$ is sparse and estimate its support by using an ℓ_0 or ℓ_1 constraint. We tried the latter option and obtained a very moderate improvement when using the true number of differentially abundant proteins, at the cost of a large computational overhead. We therefore rely on the $\mathbf{H}_0$ approximation for PEPA but keep in mind that is is misspecified as long as other proteins in the connected component are differentially abundant, which could inflate both type I and type II errors. We show that in our experiments – where 50 proteins are differentially abundant – PEPA is reasonably well calibrated and allows for better precision/recall trade-offs even against competitors which are not making this approximation. Other choices of hypothesis could be better suited in cases where a large proportion of proteins are expected to be differentially abundant.

4 Results

In this Section we present PR curves on simulated and spike-in datasets, calibration curves and a runtime performance comparison of the evaluated methods.

4.1 Performances on simulated datasets

The performances on simulated datasets with 0%, 5%, 20% and 50% of shared peptides are displayed on Figure 1. Additional figures representing 1%, 10%, 33% and 67% of shared peptides are provided in Section K of the Supplementary Material.

Peptide-based dominate aggregation-based methods In all settings, the baseline method Top3Spec-TT is by far the least accurate. Overall, as noticed by [9], aggregation-based methods (Top3Spec-TT and AllSpec-SAM, depicted by lighter or darker green dotted curves respectively) are less accurate than other methods, whether or not they exploit shared peptides.

Benefit of using shared peptides In the absence of shared peptides, both PeptideModel and MSqRob are very similar to PEPA-ML and PEPA-MAP, respectively, as discussed in Section 3.3. Accordingly, the upper left panel of Figure 1 shows that both families perform comparably in this regime, whether they use shared peptides (PEPA-ML and PEPA-MAP in lighter or darker solid red curves respectively) or not (PeptideModel and MSqRob depicted by lighter or darker dashed blue curves respectively). As shared peptides are introduced, and as their number increases, the number of protein-specific peptides available mechanically decreases, affecting all methods which do not exploit shared peptides. On the other hand, the performances of our methods accounting for shared peptides are generally unaffected and clearly dominate all other methods as soon as a large enough proportion is reached (5 to 20%).

Benefit of regularization The regularized versions of the compared methods (darker colors) always dominate their unregularized counterparts (lighter

colors). This is true regardless of the proportion of shared peptides. As the number of shared peptides increases, the unregularized versions of methods which do not exploit these peptides (peptide-based PeptideModel and aggregation-based Top3Spec-TT) face a more severe dropout of their performances than the corresponding regularized methods (peptide-based MSqRob and aggregation-based AllSpec-SAM), suggesting that regularization helps more as fewer protein-specific peptides become available. We observe the opposite behavior for our methods accounting for shared peptides: the benefit of the regularization introduced in PEPA-MAP versus PEPA-ML decreases as the proportion of shared peptides increases. Indeed, as this proportion increases, our likelihood ratio test does not discard the shared peptides and additionally gains more peptides for each protein. Accordingly the pink curve of PEPA-ML actually improves as the proportion of shared peptides increases because its sample size also increases and regularization becomes less useful.

Overall, PEPA-MAP provides the best performances on simulated data, despite the strong misspecification of the data generating model with respect to the regression model.

4.2 Performances on spike-in datasets

We build PR curves to compare all methods on Exp1_R2_pept (Figure 2) and Exp1_R25_pept datasets (Figure 3) with 0, 120, 200 and 280 shared peptides. Additional figures representing the cases with 40, 80, 160 and 240 shared peptides are provided in Section K of the Supplementary Material. Exp1_R2_pept and Exp1_R25_pept originally contain 10722 (resp. 10601) peptides, among which 211 (resp. 290) come from human proteins, so that the proportion of shared peptides is rather small: for either datasets, the proportion of introduced shared peptides is smaller than 2.65% (to be compared with a proportion up to 50% of shared peptides as recalled in the introduction).

We notice a large difference of performances and of behavior between the two datasets. Exp1_R25_pept derives from a series of LC-MS/MS experiments that did not undergo any malfunction, so that the data is of rather high quality. On Exp1_R2_pept it is not possible to diagnose the source of the noise which could range from the MS acquisition to the bioinformatic data processing, yet it is clear that on this dataset, the UPS1 protein is more difficult to detect. Both datasets correspond to a scenario that can be faced on proteomics platforms.

Peptide-based methods do not always dominate aggregation-based methods
The domination of peptide-based over aggregation-based models that was clearly illustrated on simulated dataset does not hold on our spike-in datasets. This is especially true on Exp1_R2_pept, where AllSpec-SAM outperforms all other methods (including ours) when there are no shared peptides. In the presence of shared peptides, it is either competitive with or dominated by our method. All other methods (PeptideModel, Top3Spec-TT and MSqRob) are dominated by AllSpec-SAM, PEPA-ML and PEPA-MAP regardless of the number of shared peptides. A possible explanation is that when peptide-level intensity values are unreliable, the aggregation process somehow regularize the resulting protein-level intensity values. The same conclusions hold on the Exp1_R25_pept dataset,

yet with dimer magnitude: when the number of shared peptides increases, MSqRob keeps up but remains dominated by aggregation-based methods.

Benefit of shared peptides and regularization PEPA-ML and PEPA-MAP outperform all other methods as soon as enough shared peptides are introduced. In all cases, the regularized methods AllSpec-SAM and PEPA-MAP outperform their unregularized counterparts Top3Spec-TT and PEPA-MAP. MSqRob dominates PeptideModel on Exp1_R25_pept, but is outperformed for small recall values on Exp1_R2_pept, suggesting that on this dataset MSqRob assigns the lowest p-values to a few non differentially abundant proteins.

Our methods handle proteins with no specific peptide In both experiments, some proteins are lost by methods that only rely on protein-specific peptides because as the number of shared peptides increases, these proteins end up with only shared peptides. On Figures 2 and 3, this leads to the noticeable dropouts on the lower end of the curves for these methods. This illustrates an important practical issue: methods that only rely on protein-specific peptides are unable to deal with some proteins. Accounting for shared peptides like we suggest not only improves our ability to detect differentially abundant proteins among those that are handled by classical methods, but also increases the proteome coverage.

To conclude, these experiments show that both our method accounting for shared peptides and its regularized version improve our ability to detect differentially abundant proteins on proteomics datasets. The more shared peptides, the more important the increment of performances, but even on datasets with no shared peptides the methods remain accurate and never strongly underperforms. Finally, the regularized version always performs better than the other, making PEPA-MAP a safe choice.

4.3 p-value calibration

The last point to evaluate is the quality of the calibration of the p-values provided by our tests. In particular, it is necessary to check that the correction we introduced in Proposition 2 for our PEPA-MAP statistic leads to correct asymptotic levels using a χ^2 distribution like with the PEPA-ML statistic. To visually assess this point, we compare the expected and actual test levels for the methods evaluated in Sections 4.1 and 4.2 except for Top3Spec-TT, to avoid cluttering our graphs. In addition to PEPA-MAP we include a corrected version PEPA-MAP-RW. RW stands for reweighted: all the regularized likelihood ratio statistics are multiplied by $\frac{\hat{\sigma}^2+s}{\hat{\sigma}^2}$ as suggested by Proposition 2. We compute the mean square residuals of our model for each protein and average these estimates across proteins to obtain $\hat{\sigma}^2$. PEPA-MAP-RW would behave exactly like PEPA-MAP in the PR curves of Sections 4.1 and 4.2 as multiplying the test statistic of all proteins by the same weight does not affect their order. The p-values for PEPA-ML, PEPA-MAP and PEPA-MAP-RW are computed by comparing the corresponding statistics to the quantiles of a χ^2_1 distribution.

Figure 4 is a (log-log) plot of the empirical proportion of false positives obtained as a function of the p-value threshold, *i.e.*, the proportion of non-differentially abundant proteins (y-axis) which are assigned a p-value lower than

the threshold (x-axis). If a test is correctly calibrated, a proportion α of non-differentially abundant proteins has p-value lower than α for all $\alpha \in [0, 1]$ and its calibration plot is the $y = x$ axis. Additional figures representing the various calibration plots we obtained during the experiments are provided in Section K of the Supplementary Material.

Calibration of PEPA-ML The left panel of Figure 4 shows the plots obtained on simulated data from Section 3.1 with no shared peptide. All methods except PEPA-MAP are reasonably well calibrated. The small deviation observed for PEPA-ML can be explained by a combination of two factors. First, the χ^2 distribution of PEPA-ML is an asymptotic result, and we only have $n_1 = n_2 = 3$ observations for each group in this case. Second, as discussed in Section 3.4, the null hypothesis that we are testing is $\theta = \theta'$, *i.e.*, that no protein is differentially abundant. This model is misspecified as soon as the protein j that we are testing is in the same connected component as a differentially abundant protein j', *i.e.* $\theta_{j'} \neq 0$, even if indeed $\theta_j = 0$. This may happen in our simulations, as our dataset contains 50 differentially abundant proteins. We observe nonetheless that using the same simulation setting with 10 samples per group instead of 3 leads to a perfectly calibrated PEPA-ML (not shown), suggesting that the main issue is the low sample size. The right panel of Figure 4 shows the plots obtained on the Exp1_R2_pept data. PEPA-ML is more severely decalibrated, leading to a false positive rate of 7.3% when thresholding at 0.01 when AllSpec-SAM leads to a 1.9% rate, PeptideModel to 0.4% and MSqRob to 4.3%. The deviation is likely caused by the low sample size, as the number of differentially abundant proteins in the dataset is very similar to the one used in our simulation – where we recover a correct calibration by increasing the sample size.

Calibration of PEPA-MAP As predicted by Proposition 2, the PEPA-MAP statistics are not χ^2_1 distributed under $\mathbf{H}_0$ which is illustrated by the strong deviation of the grey curve from the $y = x$ axis – the selected regularization parameter s is large. Weighting our regularized statistics by the factor obtained in Proposition 2 leads to a test with similar calibration as our unregularized PEPA-ML, *i.e.*, whose small deviation from the correct level can be explained by the low sample size. On the Exp1_R2_pept data the deviation of PEPA-MAP is milder because the selected s is smaller. It actually leads to more accurate levels than PEPA-ML (1.4% false positive rate when thresholding the p-values at 0.01) by partially compensating the deviation incurred by PEPA-ML (because of small sample size) in the opposite direction. This is of course artefactual and should not be considered a good property as there is no guarantee the same phenomenon will systematically happen on new data. The weighting scheme mostly corrects the deviation of the levels of PEPA-MAP from those of PEPA-ML, leading to a 4.4% false positive rate when thresholding the p-values at 0.01. The remaining difference is probably caused by the poor quality of our estimate of $\hat{\sigma}^2$, and the fact that the data may not be well represented by i.i.d samples from a distribution with a common variance.

Additional calibration plots with varying number of artificially added peptides for Exp1_R2_pept and for Exp1_R25_pept are displayed in Section K

of the Supplementary Material.

4.4 Runtime evaluation

Table 1 shows the average runtime of all evaluated methods across five runs of the simulation described in Section 3.1. All coefficients of variation are below 10%. We only show a single result denoted as PEPA for PEPA-ML and PEPA-MAP as the marginal cost of adding a fudge factor is small. For PEPA, we also show the average time spent at computing the test statistic. The rest of the time is used to identify the connected component of the peptide-protein graph, and is unaffected by the speedup obtained in Proposition 1. We do not show the execution time of PEPA without the speedup but each experiment takes more than ten hours.

Most methods which do not use shared peptides have a runtime close to one second: computing their statistic only involves simple operations such as averages over small numbers of observations. MSqRob however is two orders of magnitude slower as it involves more observations and an iterated reweighted least square procedure, but it remains fast enough to be applicable to proteomics datasets with hundreds of proteins and thousands of peptides. Our methods run in one minute, most of which is spent computing the connected components of the peptide-protein graph. The computation of our test statistic using Proposition 1 takes less than 10 seconds, even though it requires the residuals of a linear regression problem with an $nq \times (p+q)$ design matrix.

5 Discussion and conclusions

We have proposed a linear model that accounts for shared peptides in relative quantification proteomic experiments based on mass-spectrometry analysis of peptides. This model can be used to build likelihood ratio tests relying either on the maximum likelihood or MAP estimator of its variance parameter. We have also introduced a faster way to compute the test statistic, making it amenable to datasets with thousands of peptides and proteins. The faster form relies on the fact that the likelihood ratio statistics only requires the regression residuals as opposed to estimates of the regression parameters, in particular of protein abundances. Using our model to estimate abundances – a task which is out of the scope of this paper and which we did not include in our experiments – would require to actually estimate the regression parameters, *e.g.* using explicit formulas for $(\mathbf{X}_k^\top \mathbf{X}_k)^\dagger \mathbf{X}_k^\top y$.

Experiments on simulated and spike-in data confirm that the proposed tests have a clear advantage against existing methods to detect differentially abundant peptides in the presence of shared peptides. In the absence of shared peptides or when very few of them are present, our tests behave like existing methods, suggesting they can be safely used in all cases. We have also shown that asymptotic levels could be obtained when using the MAP estimator instead of the maximum likelihood, providing asymptotic levels for this version of our likelihood ratio test – which systematically outperforms the maximum likelihood version in our experiments. Our tests are implemented in the `pepa.test` function of the Bioconductor package `DAPAR`.

Our work could be extended in several ways. First the experimental design of some proteomic analyses may be more complex than the one accounted for in our evaluations. The fast version of our statistic introduced in Proposition 1 is derived for a model with a protein and peptide fixed effect only, as opposed to *e.g.* technical replicates. Proposition 1 could be generalized to more grouping factors with fixed effects. Alternatively, or for random effects, one can use model (3) with additional factors and (slower) out of the box implementations of mixed model to compute the likelihood ratio statistic. Another possible extension regards the misspecification described in Section 4.3: using a Wald test instead of likelihood ratio test, one could simply estimate all protein abundances jointly instead of relying on models in which all proteins but one are differentially abundant. Wald tests however would not benefit from the acceleration allowed by Proposition 1 as they require parameter estimates as opposed to just likelihoods.

Supplementary Materials

The reader is referred to the Supplementary Materials for additional comparisons between methods, a more detailed description of existing methods, additional experiments and plots, a discussion of our linear approximation of the log-normal model, a fast heuristic for our testing procedure, proofs of Propositions 1 and 2, as well as a description of our simulation protocol.

Acknowledgments

The authors thank Samuel Wieczorek for his help in integrating the PEPA test to the `DAPAR` package and Zaïd Harchaoui for insightful discussions.

Fundings

LJ was supported by the ANR under grant numbers ANR-14-CE23-0003-01 and ANR-17-CE23-0011-01 (MACARON and FAST-BIG project). FC and TB were supported by grants from the "Investissement d'Avenir Infrastructures Nationales en Biologie et Santé" program (ProFI project, ANR-10-INBS-08) and by the ANR (GRAL project, ANR-10-LABX-49-01).
Conflict of Interest: None declared.

References

[1] Mélisande Blein-Nicolas, Hao Xu, Dominique de Vienne, Christophe Giraud, Sylvie Huet, and Michel Zivy. Including shared peptides for estimating protein abundances: A significant improvement for quantitative proteomics. *Proteomics*, 12(18):2797–2801, 2012.

[2] Yury V Bukhman, Moyez Dharsee, Rob Ewing, Peter Chu, Thodoros Topaloglou, Thierry Le Bihan, Theo Goh, Henry Duewel, Ian I Stewart, Jacek R Wisniewski, et al. Design and analysis of quantitative differential

proteomics investigations using lc-ms technology. *Journal of Bioinformatics and Computational Biology*, 6(01):107–123, 2008.

[3] Meena Choi, Ching-Yun Chang, Timothy Clough, Daniel Broudy, Trevor Killeen, Brendan MacLean, and Olga Vitek. Msstats: an r package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, 30(17):2524–2526, 2014.

[4] Commentary. Ten years of methods, 2014.

[5] Banu Dost, Nuno Bandeira, Xiangqian Li, Zhouxin Shen, Steven P Briggs, and Vineet Bafna. Accurate mass spectrometry based protein quantification via shared peptides. *Journal of Computational Biology*, 19(4):337–348, 2012.

[6] Sarah Gerster, Taejoon Kwon, Christina Ludwig, Mariette Matondo, Christine Vogel, Edward M Marcotte, Ruedi Aebersold, and Peter Bühlmann. Statistical approach to protein quantification. *Molecular & cellular proteomics*, 13(2):666–677, 2014.

[7] Quentin Giai Gianetto, Florence Combes, Claire Ramus, Christophe Bruley, Yohann Couté, and Thomas Burger. Calibration plot for proteomics: A graphical tool to visually check the assumptions underlying fdr control in quantitative experiments. *Proteomics*, 16(1):29–32, 2016.

[8] Quentin Giai Gianetto, Yohann Couté, Christophe Bruley, and Thomas Burger. Uses and misuses of the fudge factor in quantitative discovery proteomics. *Proteomics*, 16(14):1955–1960, 2016.

[9] Ludger JE Goeminne, Andrea Argentini, Lennart Martens, and Lieven Clement. Summarization vs peptide-based models in label-free quantitative proteomics: performance, pitfalls, and data analysis guidelines. *Journal of proteome research*, 14(6):2457–2465, 2015.

[10] Ludger JE Goeminne, Kris Gevaert, and Lieven Clement. Peptide-level robust ridge regression improves estimation, sensitivity, and specificity in data-dependent quantitative label-free shotgun proteomics. *Molecular & Cellular Proteomics*, 15(2):657–668, 2016.

[11] Shelley M Herbrich, Robert N Cole, Keith P West Jr, Kerry Schulze, James D Yager, John D Groopman, Parul Christian, Lee Wu, Robert N OâĂŹMeally, Damon H May, et al. Statistical inference from multiple itraq experiments without using common reference standards. *Journal of proteome research*, 12(2):594–604, 2013.

[12] Elisabeth Hodille, Ludmila Alekseeva, Nadia Berkova, Asma Serrier, Cedric Badiou, Benoit Gilquin, Virginie Brun, François Vandenesch, David S Terman, and Gerard Lina. Staphylococcal enterotoxin o exhibits cell cycle modulating activity. *Frontiers in microbiology*, 7:441, 2016.

[13] Shuangshuang Jin, Donald S Daly, David L Springer, and John H Miller. The effects of shared peptides on protein quantitation in label-free proteomics by lc/ms/ms. *Journal of proteome research*, 7(01):164–169, 2007.

[14] Kai Kammers, Robert N. Cole, Calvin Tiengwe, and Ingo Ruczinski. Detecting significant changes in protein abundance. *EuPA Open Proteomics*, 7:11 – 19, 2015.

[15] Clémentine Le Roux, Gaëlle Huet, Alain Jauneau, Laurent Camborde, Dominique Trémousaygue, Alexandra Kraut, Binbin Zhou, Marie Levaillant, Hiroaki Adachi, Hirofumi Yoshioka, et al. A receptor pair with an integrated decoy converts pathogen disabling of transcription factors to immunity. *Cell*, 161(5):1074–1088, 2015.

[16] Alexey I Nesvizhskii and Ruedi Aebersold. Interpretation of shotgun proteomic data the protein inference problem. *Molecular & cellular proteomics*, 4(10):1419–1440, 2005.

[17] Nadège Philippe, Matthieu Legendre, Gabriel Doutre, Yohann Couté, Olivier Poirot, Magali Lescot, Defne Arslan, Virginie Seltzer, Lionel Bertaux, Christophe Bruley, et al. Pandoraviruses: amoeba viruses with genomes up to 2.5 mb reaching that of parasitic eukaryotes. *Science*, 341(6143):281–286, 2013.

[18] Katharina Podwojski, Martin Eisenacher, Michael Kohl, Michael Turewicz, Helmut E Meyer, Jörg Rahnenführer, and Christian Stephan. Peek a peak: a glance at statistics for quantitative label-free proteomics. *Expert review of proteomics*, 7(2):249–261, 2010.

[19] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011.

[20] Jeffrey C Silva, Marc V Gorenstein, Guo-Zhong Li, Johannes PC Vissers, and Scott J Geromanos. Absolute quantification of proteins by lcmse a virtue of parallel ms acquisition. *Molecular & Cellular Proteomics*, 5(1):144–156, 2006.

[21] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[22] Samuel Wieczorek, Florence Combes, Cosmin Lazar, Quentin Giai Gianetto, Laurent Gatto, Alexia Dorffer, Anne-Marie Hesse, Yohann CoutÃĺ, Myriam Ferro, Christophe Bruley, and Thomas Burger. Dapar & prostar: software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics*, 33(1):135–136, 2017.

[23] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 03 1938.

[24] Yaoyang Zhang, Bryan R Fonslow, Bing Shan, Moon-Chang Baek, and John R Yates III. Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews*, 113(4):2343–2394, 2013.

Shared peptides	Top3Spec-TT	AllSpec-SAM	PEPA (test)	MSqRob	PeptideModel
0	1.67	0.44	56.84 (1.85)	403.36	0.63
0.05	1.55	0.49	60.16 (1.76)	494.20	0.61
0.2	1.43	0.55	64.77 (5.61)	660.61	0.57
0.5	1.13	0.49	67.74 (5.93)	992.35	0.50

Table 1: Average execution time in seconds across five runs for the evaluated methods on simulated data with 0%, 5%, 20% and 50% of shared peptide.

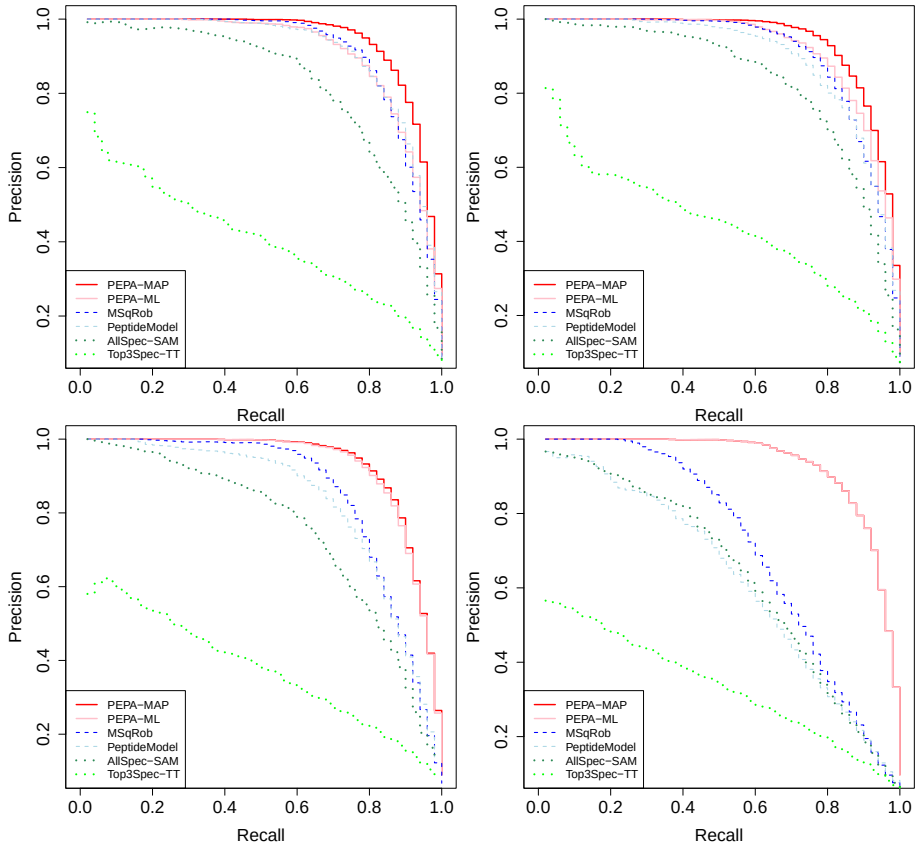

Figure 1: PR curve on simulated data with 0% (upper left), 5% (upper right), 20% (lower left) and 50% (lower right) of shared peptides.

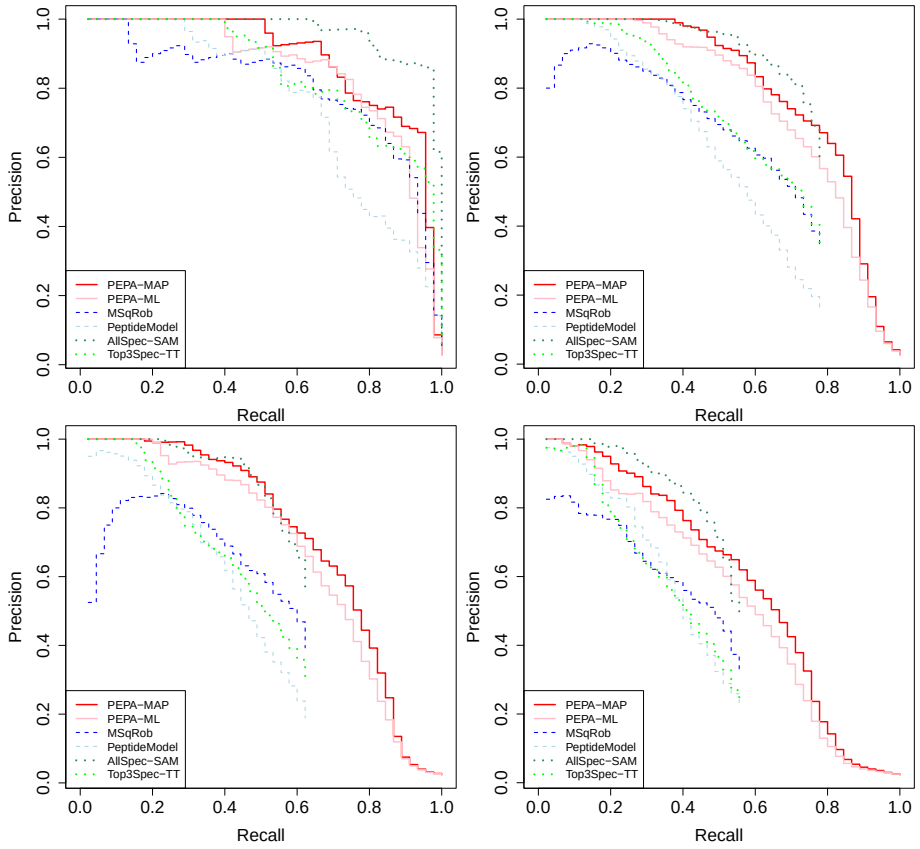

Figure 2: PR curve on Exp1_R2_pept data with 0 (upper left), 120 (upper right), 200 (lower left) and 280 (lower right) artificially added shared peptides.

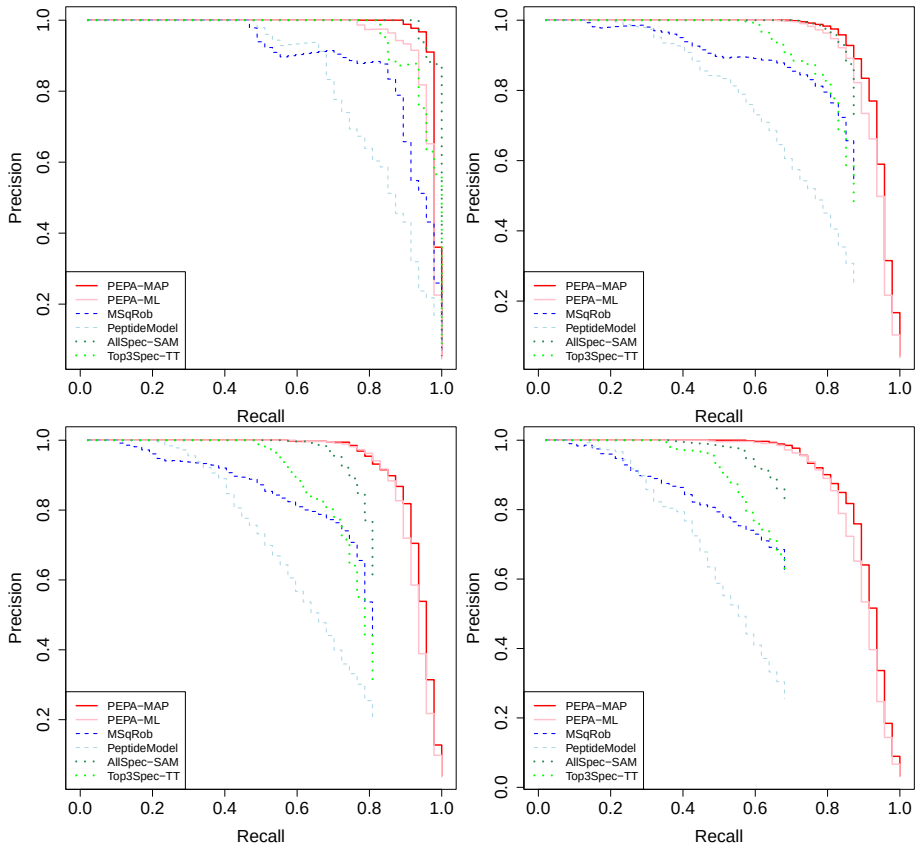

Figure 3: PR curve on Exp1_R25_pept data with 0 (upper left), 120 (upper right), 200 (lower left) and 280 (lower right) artificially added shared peptides.

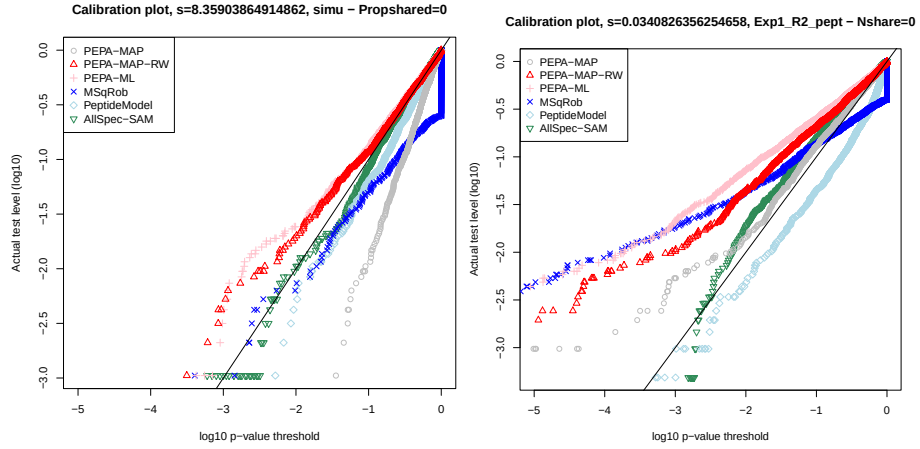

Figure 4: Calibration plots on simulated (left) and Exp1_R2_pept (right) data for all compared testing procedures.