Technical Report:
NITPICK: peak identification for mass spectrometry data

Bernhard Y. Renard^{‡,1,2}, Marc Kirchner^{‡,1,2}, Hanno Steen[3], Judith A. J. Steen[4], Fred A. Hamprecht^{*1,2}
[‡]These authors contributed equally

[1]Interdisciplinary Center for Scientific Computing, University of Heidelberg, Heidelberg, Germany [2]Department of Pathology, Children's Hospital Boston, Boston, MA, USA [3]Department of Pathology, Harvard Medical School and Children's Hospital Boston, Boston, MA, USA [4]Department of Neurobiology, Harvard Medical School and Department of Neurology, Children's Hospital Boston, Boston, MA, USA

Email: BYR - bernhard.renard@iwr.uni-heidelberg.de; MK - marc.kirchner@iwr.uni-heidelberg.de; HS - hanno.steen@childrens.harvard.edu; JAJS - judith.steen@childrens.harvard.edu; FAH*- fred.hamprecht@iwr.uni-heidelberg.de;

[*]Corresponding author

Abstract

Background: The reliable extraction of features from mass spectra is a fundamental step in the automated analysis of proteomic mass spectrometry (MS) experiments.

Results: This contribution proposes a sparse template regression approach to peak picking called NITPICK. NITPICK is a Non-greedy, Iterative Template-based peak PICKer that deconvolves complex overlapping isotope distributions in multicomponent mass spectra. NITPICK is based on *fractional averagine*, a novel extension to Senko's well-known averagine model, and on a modified version of sparse, non-negative least angle regression, for which a suitable, statistically motivated early stopping criterion has been derived. The strength of NITPICK is the deconvolution of overlapping mixture mass spectra.

Conclusions: Extensive comparative evaluation has been carried out and results are provided for simulated and real-world data sets. NITPICK outperforms pepex, to date the only alternate, publicly available, non-greedy feature extraction routine. NITPICK is available as software package for the R programming language and can be downloaded from http://hci.iwr.uni-heidelberg.de/mip/proteomics/.

Background

The reliable extraction of proteomic features from complex biological mixtures is of utmost interest for unraveling the intricate biomolecular interplay at the heart of many systems biology research questions. In this context, mass spectrometry (MS) has become a key technology which provides peptide and protein identification, modification characterization and quantification capabilities. In contrast to gene expression microarray technologies, MS analysis yields a direct view on the whole set of proteins (the proteome) present in the system under investigation and can thus contribute to a richer picture of protein interaction, real-time dynamics and their regulation [1]. MS contributes to clinical research and the diagnosis process [2], it is used to detect, grade and characterize cancer diseases [3], it serves as a general purpose tool for microorganism characterization [4, 5] and provides sensitive and specific means for pharmaceutical quality control.

MS experiments typically contain tens to thousands of spectra, each of which holds intensity information for tens to hundreds of thousands of mass channels. These data stem from a set of different mass analysis technologies, combining chemical separation procedures (chromatography), ionization methods (electrospray ionization, matrix-assisted laser desorption/ionization) and mass analyzers (time-of-flight, quadrupole, ion cyclotron motion). Despite physicochemical preprocessing and the availability of high mass resolution instruments, spectra which stem from complex biochemical mixtures (e.g. cell lysate, blood or serum) frequently exhibit overlapping isotope distributions of independent molecular species. Moreover, in many quantitative MS approaches, these mixtures are present by design and their manual unmixing, quantification and interpretation is tedious or infeasible.

As a consequence, the automated analysis and interpretation of multicomponent mass spectra is highly desirable. An (incomplete) set of challenges for MS feature extraction includes the sheer data set sizes, mixtures of isotope patterns, the presence of multiple charge states, chemical and detector noise, species-dependent ionization efficiencies, chemical reproducibility and deviations from detector linearity. Among all requirements that derive from these challenges, it is important to emphasize the crucial nature of the feature extraction step: as all subsequent analysis steps rely on the set of extracted features, meaningful biological conclusions are highly dependent on the adequacy and reliability of the feature extraction method.

Apart from few special alternate approaches [6, 7], all automated methods for feature extraction from isotope-resolved mass spectra compare the observed (experimental) spectral pattern to a set of precalculated theoretical isotope patterns. The calculation of isotope patterns is based on the estimation of average stoichiometries for a particular molecular mass (*averagine* [8] and related methods [9]) or on relative isotope abundance estimation [10] or on protein database-driven mean isotope distribution calculation [11]. The computation of isotope patterns is based on efficient implementations [12–14] of Yergey's original polynomial method [15, 16].

Comparison of theoretical and experimental isotope distributions is typically accomplished based on subtractive fitting and peak selection algorithms, attempting to sequentially detect the dominant components in a mixture spectrum. These subset selection methods attempt to determine a small set of basis functions capable of approximating the observed signal well. Facing the infeasibility of an exhaustive search over all possible subsets of explanatory basis functions, they apply greedy search strategies. Here, "greediness" refers to the fact that these approaches consistently overestimate individual feature contributions and are incapable of excluding a basis function once it has been included in the active set. Hence, although providing sparseness, they are not globally optimal. In the context of mixture modeling of mass spectra, these approaches amount to sequential isotope distribution template matching procedures [6, 8–11, 17–22]. Fitting is carried out via χ^2 distances [8, 20], least squares [9–11, 17, 21–23], weighted least squares [19], or cross-correlation [18, 24]. The automatic determination of the charge state associated with an isotope pattern present in an experimental spectrum is based on cross-correlation [19, 25] or on dot products in Fourier space [25, 26], exploiting the shift theorem of the Fourier transform. There are only few [27] non-greedy feature selection algorithms and mixture model approaches for MS data [28–31]. Among these, *Matching* [28] and Roussis' method [29] rely on manual preselection of contribution candidates. Sparse non-greedy procedures include *pepex* [30] and Du's method [31]. The *pepex* approach is suitable for single charge data and is based on a non-negative sparse regression scheme, with an approximate L_0-norm constraint. Du and Angeletti [31] perform data reduction prior to feature extraction and apply a sparseness-promoting variable selection scheme [32]. With the exception of Du's [31] and Kaur's [19] methods, none of the mentioned mixture model approaches provide support for the detection of a sparse set of *a priori* unknown peptide peaks under an arbitrary set of charge states. Du's method [31] and NITPICK overcome Kaur's greedy iterative weighted least squares fitting approach. In contrast to [31], NITPICK does not rely on a heuristic parameterization and is instead based on statistical model selection, making use of an algorithmically more efficient non-greedy sequential feature selection procedure with a statistically motivated termination criterion. NITPICK was designed to support the calculation of accurate monoisotopic peak lists from raw mass spectra and was specifically tailored to cases where the raw spectra stem from unknown, possibly overlapping experimental isotope patterns of multiple charge states.

The methods section details the mixture modeling approach, fractional averagine for improved stoichiometry estimation and data fitting, and our main contribution, a computationally efficient method for improved non-negative feature selection and the corresponding statistical complexity estimation approach in conjunction with the derivation of a lower bound for early termination. Comparative results on simulated and real-world data sets are given in the results and subsequently discussed. Eventually, we conclude and offer perspectives. Derivations of the formulas used in the main article are available in the appendix.

Methods

The NITPICK algorithm (cf. figure 1) models an observed mixture spectrum as a linear combination of theoretical isotope distribution patterns. Statistically, finding a sensible parameterization of this mixture model amounts to a constrained regression problem in which we seek to minimize the raw signal reconstruction error in a least-squares sense while adhering to a set of additional constraints. Such an approach requires reliable underlying isotope patterns, and we propose an improvement for the well-known averagine model to achieve this goal. We subsequently introduce NITPICK's iterative feature selection procedure, which employs a novel, non-greedy isotope distribution selection method and is based on a statistically motivated termination criterion, attempting to eliminate premature or late iteration termination.

Mixture model

We assume that observed spectra are available in a discrete (not necessarily equispaced) mass binning scheme defined by a mass vector $\boldsymbol{m} = (m_1, m_2, \ldots, m_N)^T$ and represent a raw multicomponent mass spectrum by a vector $\boldsymbol{s}$ of size $N \times 1$, where s_i corresponds to the abundance observed in the ith mass bin m_i. In practical applications, the vector $\boldsymbol{s}$ may also result from preprocessing steps such as relevant region detection [19] and may thus represent only a part of a complete raw spectrum. The basic assumption behind the mixture model approach is that $\boldsymbol{s}$ be a linear combination of mass spectrum abundances of K pure components $\boldsymbol{\phi}_i$,

$$\boldsymbol{s} = \sum_{k=1}^{K} c_i \boldsymbol{\phi}_i = \boldsymbol{\Phi}\boldsymbol{c}. \tag{1}$$

Each of the concentration coefficients $c_i, i = 1, \ldots, K$ is associated with a column $\boldsymbol{\phi}_i$ of the $N \times K$ model matrix $\boldsymbol{\Phi}$. We regard these columns as basis functions and their elements ϕ_{ji} correspond to the mass spectrum abundance expected in the jth mass bin m_j of the ith pure component $\boldsymbol{\phi}_i$.

For the estimation of the concentration vector $\boldsymbol{c}$, the model matrix $\boldsymbol{\Phi}$ has to be available, and in general this is not the case. One hence resorts to approximating the basis functions by a large set of theoretical isotope distributions (i.e. isotope abundance patterns) densely spread over the prespecified mass/charge binning scheme. Effectively, this recasts the original peak picking task into the framework of a feature (i.e. basis function) selection problem.

Model matrix calculation

Given an elemental stoichiometry, the corresponding theoretical isotope distribution is well-defined and can easily be calculated [12–15]. Hence, if a prespecified set of stoichiometries of potential pure components is available, the calculation of the respective set of theoretical isotope distributions (including chemical modifications and multiple charge states) is straightforward. These isotope distributions are subsequently convolved with instrument-specific, possibly mass-dependent peak shape functions, yielding the basis functions ϕ_i.

Fractional averagine

In many practical applications prior knowledge about potential components is not at hand. Thus, one needs to resort to expected average stoichiometry estimates as a best-effort approximation. In this case, the quality of the feature selection procedure is highly dependent on the quality of the stoichiometry model. We therefore extended the widely used *averagine* approach [8] to amend its discrete and discontinuous nature, gaining models without mass error and improved true isotope distribution reconstruction properties. *Fractional averagine (FA)* provides a real-valued element stoichiometry $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_5)^T$ according to the mapping $f : \mathbb{R} \to \mathbb{R}^5$ between a mass value and the number of element atoms in an averagine ($H_{7.75833}C_{4.9384}N_{1.35777}O_{1.4773}S_{0.0417}$) molecule. The calculation of the theoretical isotope distribution of $\boldsymbol{\rho}$ is based on the observation that isotope abundances follow a multinomial distribution [33], and that fractional numbers of trials in a multinomial can be modeled as linear interpolation between the probability functions of the multinomials parameterized with the surrounding integers (see appendix A). For computational ease, calculations are carried out in the realm of the corresponding moment generating function (MGF) [34] of the multinomial probability mass function. For the ith stoichiometry element, the MGF given ρ_i can be factorized according to

$$\begin{aligned}&M_x(t_1,\ldots,t_{k-1}|\rho_i)\\&=\left[p_1e^{t_1}+\cdots+p_{k-1}e^{t_{k-1}}+p_k\right]^{\lfloor\rho_i\rfloor+(\rho_i-\lfloor\rho_i\rfloor)}\\&=M_{x^1}(t_1,\ldots,t_{k-1}|\lfloor\rho_i\rfloor)M_{x^2}(t_1,\ldots,t_{k-1}|\left(\rho_i-\lfloor\rho_i\rfloor\right))\end{aligned}\tag{2}$$

where p_l is the probability of occurrence of the lth isotope ($\sum_{l=1}^{k}p_l=1$), $x=(x_1,\ldots,x_k)^T$ denotes the number of times a particular isotope is chosen ($\sum_{l=1}^{k}x_l=\rho_i$) and $t=(t_1,\ldots,t_k)'$ is the corresponding variable of the MGF. By rearrangement of the MGFs of all elements, it is possible to separate integer and real-valued contributions, yielding the common averagine model $\hat{\boldsymbol{\rho}}=(\lfloor\rho_1\rfloor,\lfloor\rho_2\rfloor,\ldots,\lfloor\rho_5\rfloor)^T$ for the integers and the fractional averagine correction $\tilde{\boldsymbol{\rho}}=(\rho_1-\lfloor\rho_1\rfloor,\rho_2-\lfloor\rho_2\rfloor,\ldots,\rho_5-\lfloor\rho_5\rfloor)^T$ for the remaining fractional masses. The theoretical isotope distribution for $\tilde{\rho}_i$ is given by the linear combination of a peak of intensity one at mass zero and the theoretical isotope distribution of the ith averagine element, weighted by $1-\tilde{\rho}_i$ and $\tilde{\rho}_i$, respectively. Thus, efficient calculation of the theoretical isotope distribution of the stoichiometry $\hat{\boldsymbol{\rho}}$ is carried out based on the Mercury7 algorithm [14], and the theoretical isotope distribution for the fractional stoichiometry $\boldsymbol{\rho}$ is subsequently obtained with five additional convolution steps.

Basis function selection

Given the set of basis functions $\boldsymbol{\Phi}=[\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2\cdots\boldsymbol{\phi}_K]$, basis function selection and subsequent determination of the contribution coefficients c_i provides a solution to eq. (1). Thus, as the modeling parameters and, in particular, the monoisotopic masses for all basis function are known, one can determine which isotope distributions are present and in what abundance (assuming $\sum_k\phi_{ki}=1$).

In practice, basis functions are calculated for each possible monoisotopic mass and each expected charge state, yielding model matrices $\boldsymbol{\Phi}$ with $K>N$ (in the case of one basis function per mass/charge bin and charge, we have $K=n_ZN$, where n_Z corresponds to the number of charge states observable in the experiment; hence, for $n_Z>1$, there exists an infinite number of solutions for eq. (1)). This is a problem intrinsic to the proposed mixture modeling approach and has been observed previously [23, 28, 30].

The *least absolute shrinkage and selection operator* (LASSO) [32] enjoys favorable properties of regularization and subset selection. Because the LASSO is capable of shrinking coefficients to exactly zero, it offers a non-greedy way to gain sparse models. The LASSO solution $\hat{\boldsymbol{c}}$ for equation (1) is given by

$$\begin{aligned}\hat{\boldsymbol{c}}=&\arg\min_{\boldsymbol{c}}\{\|\boldsymbol{s}-\boldsymbol{\Phi}\boldsymbol{c}\|^2\}\\&\text{s.t.}\quad\sum_{i=1}^{K}|c_i|\leq t,\end{aligned}\tag{3}$$

where $t \geq 0$ is a user-defined tuning parameter [31,32]. Mass spectra intensities s_i, basis function values ϕ_{ji}, and basis function contributions c_k are strictly non-negative, thus adding a non-negativity constraint to the solution space of $\hat{\boldsymbol{c}}$, yielding

$$\begin{aligned} \hat{\boldsymbol{c}} &= \arg\min_{\boldsymbol{c}} \{\|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{c}\|^2\} \\ \text{s.t.} \quad & \sum_{i=1}^{K} |c_i| \leq t, c_i \geq 0. \end{aligned} \tag{4}$$

For fixed t, this is a quadratic programming problem with linear inequality constraints which can be solved by an active set algorithm, sequentially introducing the inequality constraints and seeking a feasible solution satisfying the Kuhn-Tucker conditions [32,35,36]. Equation (4) corresponds to $\hat{\boldsymbol{c}}(\lambda) = \arg\min_{\boldsymbol{c}}\{\|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{c}\|^2 + \lambda \sum_{i=1}^{K} |c_i|\}$ with $c_i \geq 0$ where the parameter t is related to the Lagrangian multiplier λ which determines the number of free parameters $df(\lambda)$ in the linear model [32,36–38]. Common procedures for the optimal selection of λ or $df(\lambda)$ are based on the minimization of the prediction error. This involves estimation of training optimism via C_p-statistics, the Akaike Information Criterion (AIC), or the Bayesian Information Criterion (BIC) [37]. Alternatively, direct estimation of prediction error can be carried out via cross-validation or generalized cross-validation (GCV) [37]. All these methods require the LASSO trace $\hat{\boldsymbol{c}}(\lambda_l)$, where $\lambda_l \in \mathcal{L}$ and $\mathcal{L} = \{\lambda_1, \ldots, \lambda_{|\mathcal{L}|}\}$ defines the set of LASSO regularization parameters for which the prediction error is calculated. In general, the calculation of the LASSO trace is computationally intensive and it is not clear how the elements of $\mathcal{L}$ should be selected [36]. *Least angle regression* (LARS) [39] is an algorithmically different approach to variable selection which can be modified such that the LARS algorithm implements the non-negative LASSO from equation (4). The LASSO-modified LARS is a constructive active set procedure which constructs the LASSO regularization path in a stepwise manner. Denote by $\mathcal{A}(\lambda)$ the set of indices $i \in \{1, \ldots, K\}$ of those $\boldsymbol{\phi}_i$ which are in the active set for a particular choice of λ. Starting from $\lambda = \infty$ and letting $\lambda \rightarrow 0$, the algorithm computes non-negative LASSO solutions for all λ for which the active set changes, thus implicitly defining $\mathcal{L}$. The LASSO-modified LARS guarantees $\mathcal{A}(\lambda_j) \neq \mathcal{A}(\lambda_{j+1})$, but it allows for the deletion of previously selected basis functions, and hence $|\mathcal{A}(\lambda_j)|$ need not increase monotonically for increasing j. Basis functions can be required to enter the active set in their predefined directions [39] which allows the implementation of a non-negativity constraint. Necessary matrix inversions are constrained to $|\mathcal{A}(\lambda)| \times |\mathcal{A}(\lambda)|$-sized scatter matrices $\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}$ and can be implemented as iterative updates, thus the procedure is computationally efficient.

Complexity estimation

It is desirable to terminate active set updates as soon as the basis functions in the active set are able to explain the observed data sufficiently well, i.e. until the increase in explanatory power does not justify the increase in model complexity anymore. We now describe a modification to the non-negative LASSO-modified LARS, which enables us to sequentially build a BIC trace along the LASSO regularization path and to identify minima along this trace. Upon termination, the proposed procedure returns the estimate $\hat{\boldsymbol{c}}_{\mathcal{A}}$ and the set $\mathcal{A} = \{i | \hat{c}_{\mathcal{A}_i} > 0\}$ of active basis functions.

BIC *measure*

The LARS C_p-type risk reestimation formula [39] for optimal selection of λ does not hold under the non-negative LASSO modification. Instead, we recalculate a BIC measure

$$\mathrm{BIC}(\lambda) = \frac{1}{\sigma^2} \underbrace{\|\boldsymbol{s} - \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)}\|^2}_{N \cdot \mathrm{MSE}(\lambda)} + df(\lambda) \log N, \tag{5}$$

in each LARS iteration [40]. For the calculation of the unbiased training error MSE(λ) in eq. (5) we require an additional non-negative least squares fit

$$\begin{aligned} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} &= \arg \min_{\boldsymbol{c}_{\mathcal{A}(\lambda)}} \|\boldsymbol{s} - \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \boldsymbol{c}_{\mathcal{A}(\lambda)}\|^2 \\ \text{s. t.} \quad & \left(\hat{c}^q_{\mathcal{A}(\lambda)}\right)_i \geq 0. \end{aligned} \tag{6}$$

The noise variance σ^2 in eq. (5) is estimated as the mean residual sum of squares of a low-bias non-negative least squares estimate [37].

Estimation of $df(\lambda)$

The calculation of BIC(λ) in eq. (5) requires an estimate for the degrees of freedom $df(\lambda)$, which can be obtained via the generalized degrees of freedom (GDF) [38]. The GDF of an NN-LASSO-modified LARS model based on an active set $\mathcal{A}(\lambda)$ are given by

$$\mathrm{GDF}(\lambda) = \frac{1}{\sigma^2} \boldsymbol{s}^T \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)}. \tag{7}$$

Because the coefficients $(\hat{c}^q_{\mathcal{A}(\lambda)})_i > 0$ are nonnegative, the estimate $\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)}$ solves

$$\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} = \arg \min_{\boldsymbol{c}^q_{\mathcal{A}(\lambda)}} \left\{ \|\boldsymbol{s} - \boldsymbol{\Phi} \boldsymbol{c}^q_{\mathcal{A}(\lambda)}\|^2 + \lambda \sum_{i=1}^{K} \left(c^q_{\mathcal{A}(\lambda)}\right)_i \right\} \tag{8}$$

which is differentiable with respect to $\left(c^q_{\mathcal{A}(\lambda)}\right)_i$. Setting the derivative to zero, we obtain

$$\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} = (\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\boldsymbol{\Phi}_{\mathcal{A}(\lambda)})^{-1}(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\boldsymbol{s} - \frac{1}{2}\lambda\mathbf{1}_{\mathcal{A}(\lambda)}). \tag{9}$$

Hence, given an active set $\mathcal{A}(\lambda)$, the generalized degrees of freedom from eq. (7) can be written as

$$\begin{aligned} &\mathrm{GDF}(\lambda) \\ &= \boldsymbol{s}^T\frac{1}{\sigma^2}(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)}) \\ &= \boldsymbol{s}^T\frac{1}{\sigma^2}\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\boldsymbol{\Phi}_{\mathcal{A}(\lambda)})^{-1}(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\boldsymbol{s} - \frac{1}{2}\lambda\mathbf{1}_{\mathcal{A}(\lambda)}) \end{aligned} \tag{10}$$

which is monotonously increasing for decreasing λ (see appendix B for a proof).

Optimal termination

The minimal possible training error of the model is attained when all variables are in the active set, in which case the respective coefficients $\hat{\boldsymbol{c}^q}$ are given by $\hat{\boldsymbol{c}}^q = \arg\min_{\boldsymbol{c}} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{c}\|^2$ subject to $c_i \geq 0$, and the corresponding error is $\mathrm{MSE} = \frac{1}{N}\|\boldsymbol{s} - \boldsymbol{\Phi}\hat{\boldsymbol{c}}^q\|^2$. Thus, a lower bound for $\mathrm{BIC}(\lambda)$ is given by

$$\mathrm{BIC}_{min}(\lambda) = \frac{N}{\sigma^2}\,\mathrm{MSE} + \mathrm{GDF}(\lambda)\log N \tag{11}$$

(see appendix C for a proof). In general, $\mathrm{BIC}(\lambda)$ will have several minima for increasing values of $\mathrm{GDF}(\lambda)$, hence we track the minimum $\mathrm{BIC}(\lambda_{min})$ through the NN-LASSO-modified LARS cycles and accept λ_{min} as a minimizer as soon as the lower bound $\mathrm{BIC}_{min}(\lambda)$ of a subsequent LARS step exceeds the current best estimate $\mathrm{BIC}(\lambda_{min})$, i.e. $\mathrm{BIC}(\lambda_{min}) < \mathrm{BIC}_{min}(\lambda)$ (see figure 2).

Regression on selected models

The sum constraint in equation (4) is ultimately responsible for the sparseness property of the LASSO. Its regularizing effect is similar to the one of the regularization term found in ridge regression, especially with respect to the fact that all LASSO estimates $\hat{c}_i$, $i = 1, \ldots, K$ are subject to shrinkage [32, 37] and represent biased versions of the least squares estimates. Given an active set $\mathcal{A}$, the shrinkage bias on the $\hat{c}_i$ can effectively be removed by introducing a subsequent non-negative least squares regression step after the basis functions have been selected by the LASSO procedure [32]. This also holds true for the NN-LASSO-modified LARS procedure, and the corresponding unbiased quantification estimate $\hat{\boldsymbol{c}}^q_{\mathcal{A}}$ is given by equation (6) with $\mathcal{A}(\lambda) = \mathcal{A}$.

Postprocessing

The estimate $\hat{\boldsymbol{c}}$ is subject to modeling errors and these shortcomings lead to suboptimal NN-LASSO-modified LARS estimates and active sets. In particular, the estimation depends on the match between the observed and theoretical peak shape function. Especially in high mass resolution experiments, one can frequently observe spurious peak detections in bins directly adjacent to monoisotopic mass bins of true peaks [30]. A possible remedy is a local maximum detection implemented as a postprocessing filter $\varphi(\cdot)$ applied to the active basis function index set $\mathcal{A}$:

$$\begin{aligned}\mathcal{A}' &= \varphi(\mathcal{A}|G) \\ &= \{j \in \mathcal{A}|(\hat{c}^q_{\mathcal{A}})_j = \max\{(\hat{c}^q_{\mathcal{A}})_l|l \in \nu_G(j)\}\}\end{aligned} \tag{12}$$

where $\nu_G(j) = \{k \in \mathcal{A}||b_k - b_j| \leq \frac{G-1}{2}\}$ defines an m/z -neighborhood of size G around each peak and b_j is the mass/charge bin index of the monoisotopic mass m_0 of the jth theoretical isotope distribution $\boldsymbol{\phi}_j$. If $\mathcal{A} \neq \mathcal{A}'$, $\hat{\boldsymbol{c}}^q_{\mathcal{A}}$ is reestimated using eq. (6) with $\mathcal{A}(\lambda) = \mathcal{A}'$.

Results

Stoichiometry models

The fractional averagine stoichiometry model was compared against the classical averagine model based on the analysis of their respective approximation errors using simulated theoretical peptide isotope distributions.

Data Set

All UniProt (version 51.4.) [41] human proteins were subjected to *in silico* tryptic digestion. For each of the R digestion product peptides $\mathcal{P}_r$, $r \in \{1, \ldots, R\}$, exact element stoichiometries $\boldsymbol{\rho}^x_r$ and exact theoretical isotope distributions $\boldsymbol{d}^x_r$ were calculated. Peptides with monoisotopic masses above m/z 5000 were discarded.

Comparison of deviations

Classical and fractional averagine were used to estimate approximate element stoichiometries $\hat{\boldsymbol{\rho}}_r$ and $\boldsymbol{\rho}_r$, respectively, for all peptides $\mathcal{P}_r$ in the data set. Based on $\hat{\boldsymbol{\rho}}_r$ and $\boldsymbol{\rho}_r$, the corresponding theoretical isotope distribution intensity vectors $\hat{\boldsymbol{d}}_r$ and $\boldsymbol{d}_r$ were calculated. Figure 3 shows the cumulative distribution of the squared differences between the classical averagine and the true theoretical isotope distribution intensity vectors ($||\hat{\boldsymbol{d}}_r - \boldsymbol{d}^x_r||^2_2$, dashed black), and fractional averagine and the true theoretical isotope distribution intensity vectors ($||\boldsymbol{d}_r - \boldsymbol{d}^x_r||^2_2$, solid red).

Peak picking

For peak picking/feature extraction performance evaluation, we determine representative peak picking statistics: we calculate accuracy, sensitivity, specificity, and positive and negative predictive values on simulation data. Further, and in contrast to previous contributions, we explicitly perform manual *validation* on a real-world data set.

Data sets

Simulation data set. For the simulation, all UniProt (version 51.4.) [41] human protein sequences were subjected to *in silico* tryptic digestion. Simulation sets were generated by random drawing of digestion product peptides and intensities. To ensure a fair comparison with the pepex procedure (which was selected for benchmarking as the only publicly available procedure implementing non-greedy feature extraction) which is limited to singly charged data sets, all simulated peptide were endowed with a single charge. *Mercury7* [14] was used for the calculation of the respective theoretical isotopic distributions. After convolution with an m/z -dependent Gaussian aperture function [42], intensity-weighted linear combinations of peptide spectra were calculated and a Poisson noise model (see appendix D) was applied to obtain spectra of different signal to noise (SNR) ratios. Simulations were performed in the densely populated m/z $500 - 700$ range (see Additional file 1 for the data sets).

Real-world data set. Experiments on real-world data were performed using Bovine Serum Albumin (BSA) LC/(ESI-)MS calibration data. The data set was acquired on a QSTAR XL mass spectrometer (Applied Biosystems/MDS Sciex) equipped with microscale capillary HPLC system (Famos Autosampler, LC packings, Agilent 1100 HPLC pump). A mixture spectrum with many overlapping peaks was obtained by integration of the LC/MS data set over the retention time domain (see Additional files 2 and 3). Peak identification was carried out in the m/z $500 - 700$ range and peak shape functions were modeled according to mass-dependent Gaussian distributions with standard deviations $\sigma(m/z) = 0.005m/z$ [42].

Performance estimation

We characterize peak picking performance based on a set of measures from statistical test theory, all of which depend on the availability of the numbers of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

Ground truth is based on knowledge of the complete set of peptide signals present in a mass spectrum. For simulated data sets, this information is available. In real-world experiments, the definition of ground truth is complicated by sample complexity, stochastic sample modification, non-peptidic components and limited dynamic range. As a consequence, TNs, FNs and the overall number of true peaks are not available for real-world data, limiting the available statistical measures to positive predictive values and the ratio of true positives (sensitivity ratios).

Nevertheless, we can determine the number of TPs and FPs in both cases: we check whether a detected peak really exists and if it has been assigned its correct monoisotopic mass m_0 and charge z. If so, it is counted as true positive (TP) or, otherwise, as false positive (FP).

Simulation data. As the complete set of simulated peaks is known, the remaining set of undetected peaks can be determined and its members are counted as false negatives (FN). With the true number of positives and negatives available the calculation of the number of true negatives (TN) is straightforward, thus enabling the use of related statistical test error measures for performance characterization:

- accuracy (ACC) measures the rate of correct peak vs. no peak decisions, i.e. $\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}+\mathrm{TP}+\mathrm{FN}}$
- the negative predictive value (NPV) gives the rate at which there is no peak at positions where the procedure was unable to find a peak, $\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$.
- the positive predictive value (PPV) measures the rate of correct peak detections among all peaks detected by the procedure, $\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
- sensitivity (SE) measures the method's ability to detect a peak if it exists, $\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
- specificity (SP) measures the method's ability to correctly identify the absence of peaks in the spectrum, $\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$

All measures have been computed with and without the application of postprocessing.

Real-world data. Resorting to LC/MS data and creating a semi-artificial data set by integration over the retention time domain was motivated by the fact that this approach yields a data set accessible to human manual validation. With LC resolution power available to the human expert (and resorting to comparatively simple mixtures), all peaks detected in the integrated mixture can still be manually verified. Exemplary peak picking results are illustrated below.

Comparative results

Pepex. We chose to compare NITPICK to a conceptually similar, model-based approach called pepex [30]. In contrast to model-free approaches and in accordance with NITPICK, pepex models observed spectra based on a linear mixture model, which is augmented by a complexity constraint. It uses the averagine model to describe unknown features and is capable of terminating its feature selection routine after a sufficient number of basis functions has been selected. However, as the publicly available implementation of the pepex approach is limited to charge state z=1 data sets, NITPICK comparison against pepex was limited to the simulated data set.

For the analysis, the pepex algorithm was tailored to the problem at hand: its parameters were heavily optimized to maximize peak picking performance on the simulation data set. As a consequence, the reported results underestimate the pepex generalization error and overestimate its performance (see Additional file 4). For NITPICK, no specific parameter optimization was carried out, postprocessing was kept to a minimum ($G = 3$), and the reported results are representative (see Additional file 5).

MarkerView. We also compared NITPICK's ability to extract peak information from a retention time integrated mixture spectrum against the proprietary MarkerView application (Applied Biosystems/MDS Sciex, Concord, Canada) version 1.2, which includes an LC/MS peak picking algorithm. In contrast to NITPICK, MarkerView was provided with the original LC/MS data set and thus had retention time information available. Peak picking was carried out in the *m/z* 400-1400 range and detected peaks were manually validated (see Additional files 6 and 7).

Discussion

Stoichiometry models

In comparison (see figure 3, classical averagine in dashed black, fractional averagine in solid red), Senko's classical averagine [25] features a larger number of very small deviances from the truth than fractional averagine. This is caused by the rounding to integers property of the classical approach, yielding exact models more often. At the same time, the deviance distribution of the fractional averagine model has a significantly lighter tail, i.e. the model generates significantly less stoichiometries whose theoretical isotope distributions have large deviations. The cumulative distribution based on the fractional averagine model approaches 1 more quickly, and its use yields an overall decrease in theoretical isotope distribution deviations. This finding is supported by the corresponding one-sided non-parametric Mann-Whitney test ($p < 2.6 \times 10^{-11}$). Because the overall impact on the peak picking performance depends on the squared mean error magnitude ($7.6 \cdot 10^{-4}$ for classical averagine, $6.3 \cdot 10^{-4}$ for fractional averagine, corresponding to a 17% decrease for fractional averagine), fractional averagine clearly is the preferable model.

Peak picking

Simulation data set

Figure 4 shows the results for the peak detection performance analysis. As expected, ACC, NPV and PPV improve with increasing SNR. Postprocessing causes a decrease in NPV and an increase in PPV for all SNR levels as the removal of spurious peaks decreases FP but also, erroneously, increases FN. The ACC plot (top left) illustrates the fact that NITPICK is successful at simultaneously maximizing PPV and NPV. Postprocessing can then be used to trade specificity for sensitivity as supported by the sensitivity-specificity trace in figure 4 (bottom right). Here, each dot marks sensitivity and specificity of a given NITPICK postprocessing parameterization. Lines connect points of different SNRs. As expected, the introduction of a postprocessing step increases specificity and decreases sensitivity. Further analysis of FNs in the simulated data reveals that false negatives are predominantly due to low-intensity components in complex mixtures (data not shown).

Comparative results

In comparison with pepex, NITPICK exhibits better results with respect to all statistical measures in figure 4. It is especially obvious that pepex suffers from a severe increase in false positives (FPs) for very low SNR situations, yielding significant decreases in accuracy (ACC) and specificity (SP). For PPV, although the pepex approach outperforms NITPICK when no postprocessing is applied, it is inferior to the full NITPICK algorithm with simple spurious peak removal corresponding to eq. (12). With respect to sensitivity (SE) and specificity (SP), figure 4 reveals constant high (above 0.99) and superior specificity values for NITPICK at greatly increased sensitivity. Thus one can conclude that the NITPICK algorithm is more sensitive than pepex and, at the same time, provides picked peaks with higher confidence.

Real-world data set

We give peak picking illustrations for the mass ranges m/z 507–525 (with a zoom on m/z 518–525), m/z 636–646, m/z 695–725 and m/z 775–782, detailing positive and negative peak picking performance aspects. In the m/z 507–525 mass range (figure 5), all picked peaks could be verified, including the monoisotopic masses of the mixture distribution with components located at m/z 523.23 (z=3) and m/z 523.82 (z=5). Upon re-examination of the raw data, we detected a missed low-intensity peak at m/z 515.76.

Figure 6 zooms onto two cases of overlapping isotope distributions in the m/z 518–525 mass range. At m/z 518.22 and m/z 519.11 NITPICK resolves two distinct monoisotopic masses, in spite of their unfavorable mass distance. Although the second isotope peak of the doubly charged ion with monoisotopic mass m/z 518.22 exhibits a heavy overlap with the monoisotopic peak of the ion at m/z 519.11, NITPICK is still able to correctly detect the monoisotopic peaks of the two isotope distributions. NITPICK also separates two isotope distributions located at m/z 523.23 (z=3) and m/z 523.82 (z=4). The detection of the monoisotopic mass at m/z 523.82 is particularly non-trivial because of its heavy overlap with an isotope peak of the isotope distribution located at m/z 523.23 and also because the detected monoisotopic mass peak at m/z 523.82 is not the most abundant peak within its isotope distribution.

In the m/z 636–646 mass range (figure 7) we observe an example of incomplete unmixing: the isotope distribution (z=3) with monoisotopic mass located at m/z 636.29 heavily overlaps the distribution (z=3) located at m/z 636.64 (left triangle marker). The overlap proves inseparable and the monoisotopic mass of the second distribution is wrongly detected at m/z 636.96. Further, due to conservative noise level/complexity estimation, the isotope distribution located at m/z 642.33 (right triangle marker) is not detected. Note that in both of the correctly detected distributions located at m/z 636.29 and m/z 639.65, the monoisotopic mass peak does not correspond to the most prominent peak.

In the m/z 695–725 mass range (figure 8), with one exception, all detected peaks could be verified. The wrongly detected peak at m/z 714.29 corresponds to the first isotope peak of the isotope distribution located at m/z 713.78 (z=2). Especially in the m/z 718 to m/z 724 region the algorithm proves capable of resolving nontrivial low-intensity mixtures.

In the m/z 775–782 range (figure 9), the separation of two heavily overlapping isotope distribution clearly illustrates the benefits of NITPICK's intensity model-based approach to the peak picking/feature extraction problem: the second isotope peak of the isotope distribution located at m/z 779.32 (z=2) and the monoisotopic peak of the distribution located at m/z 780.35 (z=2) overlap completely and can only be distinguished by taking intensity information into account.

Overall, the results obtained on real-world data are in agreement with simulation results: after manual validation of 192 peaks detected in the real-world dataset, we observe 127 true positives, yielding a positive predictive value of PPV = 66.15%.

Comparison with MarkerView

On the BSA data set, MarkerView detected 388 peaks, for 96 (24.7%) of which charge state information was available. Peaks without charge state assignment were counted as true peaks if their detected mass/charge ratio was correct. This resulted in 205 true positives for 82 (40.0%) of which charge state information was available. In comparison to NITPICK, this yields a sensitivity ratio of $\mathrm{SER} = \frac{\mathrm{SE}_{MarkerView}}{\mathrm{SE}_{NITPICK}} = \frac{205}{127} = 1.61$ and a positive predictive value of $\mathrm{PPV} = 0.53$.
As expected, with retention time information available, MarkerView manages to detect a significantly larger number of peaks. Surprisingly, though, retention time information did not contribute to an increased PPV. The partial lack of charge state information also caused the performance interpretation to favor MarkerView: for peaks with correct mass/charge ratio, we assumed completely error-free charge state assignments, which is unlikely to hold true in reality. In contrast, in absence of retention time information, NITPICK delivered charge state information for each and every peak and peaks were counted as true positives if and only if their assigned charge state was correct. MarkerView's PPV and SER are subject to overestimation, whereas NITPICK's PPV is not. Even under this pro-MarkerView bias, if joint maximization of PPV and sensitivity is desired, NITPICK arguably proved competitive with MarkerView: despite the 1.6-fold increase in sensitivity, only slightly more than half of the peaks reported by MarkerView are true positives.

Analysis CPU time on the real-world spectrum was 114s on a 2GHz AMD Opteron machine. Measurements are based on native, interpreted R code. Preliminary tests with an in-house C++ implementation (to be published elsewhere) yielded a speed increase by a factor of ≈ 20.

Conclusions and perspectives

Conclusions

We present NITPICK, an iterative, non-greedy, globally optimal mixture modeling approach for feature extraction from multicomponent mass spectra. The calculation of the set of explanatory theoretical isotope distributions is based on *fractional averagine*, a mass error-free extension to the well-known *averagine* [8] model. Subsequent feature selection is driven by a modified *least angle regression* [39] algorithm for which we derived a suitable, statistically motivated early stopping criterion. Experiments show that NITPICK is able to unmix and deconvolve complex mixture mass spectra. The algorithm was thoroughly evaluated on simulated and real-world data sets and was found to perform better than a conceptually similar algorithm. NITPICK was even found to deliver competitive results when compared against a vendor-supplied algorithm which, in contrast to NITPICK, had retention time resolution available.

We would like to note that although the analysis at hand was confined to a proteomics data set, the application of the proposed methodology is in no way limited to this type of data and can easily be adapted to similar problems outside the field of proteomics.

NITPICK is available as software package for the R programming language and can be downloaded from http://hci.iwr.uni-heidelberg.de/mip/proteomics/.

Perspectives

The constrained least squares regression model in equation (3) implicitly assumes Gaussian noise on the observed spectra. Especially with low-intensity time-of-flight spectra the Gaussian approximation is crude, yielding suboptimal estimates. The incorporation of a data type- and intensity-dependent procedure pursuing a suitable Poisson regression approach [36] in appropriate cases could improve on this shortcoming.

The non-negative least squares step in equation (6) assumes error-free basis functions ϕ_i. Although fractional averagine improves over the classical averagine model, this assumption is still violated. Possible remedies include direct intensity estimation techniques [43, 44] and enhanced sparse feature selection methodology which allows for errors in explanatory variables. Alternatively, extended stoichiometry models could provide problem-tailored basis functions if model bias is not an issue.

For charge states $z < 3$ and mass ranges $m/z \lesssim 1400$, there exist so-called *forbidden regions* [45] within the mass spectrum, i.e. mass ranges which are inaccessible to peptides (including modifications). Such information has been reported to be suitable as a preprocessing filter [31].

Further computational efficiency could be achieved by a complexity-driven hierarchical estimation approach, resorting to subtractive feature extraction for simple signals and to the full mixture modeling for complex samples only.

Appendix

A Computation of fractional averagine

For the computation of the isotopic distribution of fractional averagine, we build on the fact that the distribution of the isotopes of an element follows a multinomial [33]. The multinomial is discrete, hence for fractional counts of events we can interpolate between the two adjacent integer multinomials for each element such that

$$\begin{aligned} &P_{n=c}(X_1 = x_1, \ldots, X_{k-1} = x_{k-1}) \\ &= (\lceil c \rceil - c) P_{n=\lfloor c \rfloor}(X_1 = x_1, \ldots, X_{k-1} = x_{k-1}) \\ &+ (c - \lfloor c \rfloor) P_{n=\lceil c \rceil}(X_1 = x_1, \ldots, X_{k-1} = x_{k-1}) \end{aligned} \tag{13}$$

with $\lceil c \rceil = \min_{j \in \mathbb{N}}(j \geq c)$, $\lfloor c \rfloor = \max_{j \in \mathbb{N}}(j \leq c)$ and X_i representing the number of times the ith isotope of an element occurs. Under the (reasonable) assumption of independence of the atomic distributions of the elements, the resulting joint distribution for a molecule follows from the multiplication of the distributions of its elements.

By changing the order of multiplication and separating the highest possible integer number from the remaining fractional numbers, the calculation of fractional averagine can be related to the Mercury7 algorithm [14], yielding a highly efficient calculation scheme (see eq. (2)). For the convolution of the Mercury integer results and the fractionals we follow [46]: Let $g_p(i)$ represent the ith element of the probability vector of the first and $f_p(j)$ the jth element of the second distribution, then

$$h_p(k) = \sum_i g_p(i) f_p(k-i) \tag{14}$$

can be used to compute $h_p(k)$, the kth element of the new vector of probabilities for the joint distribution. Similarly, the corresponding mass vector h_m can be computed using the probability vectors g_p and f_p and the corresponding mass vectors g_m and f_m using

$$\begin{aligned} h_m(k) = &\left(\sum_i g_p(i) f_p(k-i) \right)^{-1} \\ &\sum_i g_p(i) f_p(k-i) \left(g_m(i) + f_m(k-i) \right). \end{aligned} \tag{15}$$

B Proof of the monotony of the GDF for the non-negative lasso

As long as a given set $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}$ is valid, it can be easily shown that the GDF are monotonous in λ. Starting with the $GDF(\lambda)$ of equation (10),

$$
\begin{aligned}
&\mathrm{GDF}(\lambda)\\
&= \boldsymbol{s}^T \frac{1}{\sigma^2} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} (\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)})^{-1} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} - \frac{1}{2} \lambda \mathbf{1}_{\mathcal{A}(\lambda)} \right)\\
&= \frac{1}{\sigma^2} \sum_{i=1}^{N} \boldsymbol{s}_i\\
&\quad \left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \right)^{-1} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{s} - \frac{1}{2} \lambda 1_{\mathcal{A}(\lambda)} \right) \right)_i\\
&= \frac{1}{\sigma^2} \sum_{i=1}^{N} \boldsymbol{s}_i \left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \right)^{-1} \boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{s} \right)_i\\
&\quad - \lambda \frac{1}{2\sigma^2} \sum_{i=1}^{N} \boldsymbol{s}_i \left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \right)^{-1} 1_{\mathcal{A}(\lambda)} \right)_i
\end{aligned}
\tag{16}
$$

To show that the GDF are monotonously increasing for decreasing values of λ, it suffices to analyze the following part of the formula,

$$
\begin{aligned}
&\sum_{i=1}^{N} \left(\boldsymbol{s}_i \left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \right)^{-1} 1_{\mathcal{A}(\lambda)} \right)_i \right)\\
&= \sum_{i=1}^{N} \left(e_i^T \left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \right)^{-1} 1_{\mathcal{A}(\lambda)} \right) \right)^T \left(e_i^T \boldsymbol{s} \right)\\
&= \sum_{i=1}^{N} \left(1^T_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \right)^{-1} \boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \right) e_i e_i^T \boldsymbol{s}\\
&= \left(1^T_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \right)^{-1} \boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \right) I_N \boldsymbol{s}\\
&= \left(1^T_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \right)^{-1} \boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \right) \boldsymbol{s}\\
&= \sum_{j=1}^{K} \hat{\boldsymbol{c}}_j^{LS_{\mathcal{A}(\lambda)}}
\end{aligned}
\tag{17}
$$

where e_i denotes the ith canonical unit vector of length N and $I_N = \sum_{i=1}^{N} e_i e_i^T$ is the identity matrix of size N.

$\hat{\boldsymbol{c}}_j^{LS_{\mathcal{A}(\lambda)}}$ is the least squares regression coefficient for the corresponding least squares problem of the active set. It is known that all non-negative lasso coefficients $\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_j}$ are greater or equal zero, so

$$\begin{aligned}
&\sum_{j=1}^{K} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_j} \geq 0\\
&\overset{(eq.\ 9)}{\Longleftrightarrow} \left(1^T_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\right)^{-1} \boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\right)\boldsymbol{s}\\
&\quad - 1^T_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\right)^{-1} \frac{1}{2}\lambda 1_{\mathcal{A}(\lambda)} \geq 0\\
&\Longleftrightarrow \left(1^T_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\right)^{-1} \boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\right)\boldsymbol{s}\\
&\quad \geq \frac{1}{2}\lambda 1^T_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\right)^{-1} 1_{\mathcal{A}(\lambda)} \geq 0\\
&\Longleftrightarrow \sum_{j=1}^{K} \hat{\boldsymbol{c}}_j^{LS_{\mathcal{A}(\lambda)}}\\
&\quad \geq \frac{1}{2}\lambda 1^T_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\right)^{-1} 1_{\mathcal{A}(\lambda)} \geq 0
\end{aligned}\tag{18}$$

as $\left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)}\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\right)^{-1}$ is the inverse of a covariance matrix and, thus, positive-semidefinite, and λ is by definition always greater or equal 0. Thus, the second part of equation (16) is monotone with regard to λ and therefore the GDFs are monotone as long as a given active set is valid.

It remains to be shown that changes of $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}$ do not influence the monotony, so it needs to be shown that neither the addition of $\boldsymbol{\phi}_j$ to the set $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}$ nor the removal of $\boldsymbol{\phi}_k$ from $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}$ lead to a decrease of $\text{cov}(\boldsymbol{s}, \boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)})$ as given in (10). A formal proof is given further below, nevertheless, this can also be argued intuitively.

In the non-negative LARS implementation as described above and in [39], a variable $\boldsymbol{\phi}_j$ will be added to the active set $\boldsymbol{\phi}_{\mathcal{A}(\lambda)}$ only if it is positively correlated with the remaining residuals, i. e. if

$$\text{cov}\left(\boldsymbol{\phi}_j, \boldsymbol{s} - \boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)}\right) > 0 \tag{19}$$

This obviously leads to an increase of $\text{cov}\left(\boldsymbol{s}, \boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)}\right)$ as less unexplained variation remains. A variable $\boldsymbol{\phi}_k$ is removed from the active set $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}$ only if $\text{cov}\left(\boldsymbol{\phi}_k, \boldsymbol{s} - \boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)}\right) < 0$, so if the residuals are negatively correlated with the variable its removal leads to an increase of $\text{cov}\left(\boldsymbol{s}, \boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)}\right)$as well. Thus, as long as changes of the set $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}$ appear one at a time (which is ensured by the active set implementation), they do not influence the monotonous character of the estimate of the degrees of freedom.

More formally, when a variable $\boldsymbol{\phi}_j$ is added to the current set of variables $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}$, the solution for $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} = \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \cup \boldsymbol{\phi}_j$ can be constructed from the solution of $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}$ in the following manner [39]:

$$\boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+}\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+} = \boldsymbol{\Phi}_{\mathcal{A}(\lambda)}\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} + \hat{\gamma} u_{\mathcal{A}(\lambda)_+} \tag{20}$$

where

$$\hat{\gamma} = \min\nolimits^{+}_{j \in \mathcal{A}(\lambda)^C} \left\{ \frac{\hat{D} - \hat{d}_j}{B_{\mathcal{A}(\lambda)} - b_j} \right\} > 0 \tag{21}$$

is strictly positive by definition and gives the magnitude of the change.

$$\hat{d} = \boldsymbol{\Phi}^T (\boldsymbol{s} - \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)}) \tag{22}$$

is the vector of the current correlation and

$$\hat{D} = \max\nolimits_j \{ \hat{d}_j | \hat{d}_j > 0 \}. \tag{23}$$

In addition,

$$B_{\mathcal{A}(\lambda)} = \left(1^T_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \right)^{-1} 1_{\mathcal{A}(\lambda)} \right)^{-\frac{1}{2}} \tag{24}$$

and

$$u_{\mathcal{A}(\lambda)} = \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} B_{\mathcal{A}(\lambda)} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \right)^{-1} 1_{\mathcal{A}(\lambda)} \tag{25}$$

leading to

$$b = \boldsymbol{\Phi}^T u_{\mathcal{A}(\lambda)}. \tag{26}$$

We need to show that

$$\begin{aligned} &\operatorname{cov}\left(\boldsymbol{s}, \boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+} \right) \geq \operatorname{cov}\left(\boldsymbol{s}, \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} \right) \\ \iff &\operatorname{cov}\left(\boldsymbol{s}, \boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+} - \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} \right) \geq 0 \\ \iff &\boldsymbol{s}^T \left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+} - \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} \right) \geq 0. \end{aligned} \tag{27}$$

Using the construction of $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+}$ from above, this leads to

$$\begin{aligned} \iff \boldsymbol{s}^T \left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} + \hat{\gamma} u_{\mathcal{A}(\lambda)_+} - \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} \right) &\geq 0 \\ \iff \boldsymbol{s}^T \left(\hat{\gamma} u_{\mathcal{A}(\lambda)_+} \right) &\geq 0. \end{aligned} \tag{28}$$

It is known from its definition that $\hat{\gamma}$ is strictly positive, thus it can be dropped from the inequality and

$$\begin{aligned} \boldsymbol{s}^T u_+ &\geq 0 \\ \iff \boldsymbol{s}^T \boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} B_{\mathcal{A}(\lambda)_+} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)_+} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} \right)^{-1} 1_{\mathcal{A}(\lambda)_+} &\geq 0. \end{aligned} \tag{29}$$

It is also known from the idea of the non-negative lasso that all variables in X_A are positively correlated with the remaining residuals, so

$$\begin{aligned} &\operatorname{cov}\left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}, \boldsymbol{s} - \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} \right) \geq 0 \\ \iff &\left(\boldsymbol{s} - \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} \right)^T \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \geq 0 \\ \iff &\boldsymbol{s}^T \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \geq \left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} \right)^T \boldsymbol{\Phi}_{\mathcal{A}(\lambda)}. \end{aligned} \tag{30}$$

Using this result,

$$\begin{aligned} &\boldsymbol{s}^T \boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} B_{\mathcal{A}(\lambda)_+} \left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)_+} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+}\right)^{-1} 1_{\mathcal{A}(\lambda)_+} \\ &\geq \left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+}\right)^T \boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} B_{\mathcal{A}(\lambda)_+} \\ &\left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)_+} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+}\right)^{-1} 1_{\mathcal{A}(\lambda)_+} \end{aligned} \tag{31}$$

holds true and it suffices to show that

$$\begin{aligned} &\left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+}\right)^T \boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} B_{\mathcal{A}(\lambda)_+} \\ &\left(\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)_+} \boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+}\right)^{-1} 1_{\mathcal{A}(\lambda)_+} \geq 0 \\ &\Longleftrightarrow \left(\boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+}\right)^T u_{A_+} \geq 0 \\ &\Longleftrightarrow \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+} 1^T \boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)_+} u_{\mathcal{A}(\lambda)_+} \geq 0. \end{aligned} \tag{32}$$

When further recalling the fact from [39] that $\boldsymbol{\Phi}^T_{\mathcal{A}(\lambda)} u_{\mathcal{A}(\lambda)} = B_{\mathcal{A}(\lambda)} 1_{\mathcal{A}(\lambda)}$, this can be reduced to

$$\left(\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+}\right)^T B_{\mathcal{A}(\lambda)_+} 1_{\mathcal{A}(\lambda)_+} \geq 0, \tag{33}$$

but as $B_{\mathcal{A}(\lambda)_+}$ is strictly positive by definition, it follows that

$$\begin{aligned} &\left(\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+}\right)^T 1_{\mathcal{A}(\lambda)_+} \geq 0 \\ &\Longleftrightarrow \sum_i \left(\hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+}\right)_i \geq 0. \end{aligned} \tag{34}$$

This is always fulfilled for the non-negative lasso as it is the constraint on its initial optimization problem. The case of the removal of $\boldsymbol{\phi}_k$ from $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}$ can be argued almost identically with the only difference being that now

$$\boldsymbol{\Phi}_{\mathcal{A}(\lambda)_+} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)_+} = \boldsymbol{\Phi}_{\mathcal{A}(\lambda)} \hat{\boldsymbol{c}}^q_{\mathcal{A}(\lambda)} + \tilde{\gamma} u_{\mathcal{A}(\lambda)_+} \tag{35}$$

where

$$\tilde{\gamma} = \min_{\gamma_j > 0} \{\gamma_j\} \tag{36}$$

which is also always positive and thus can be dropped from the resulting inequality in exactly the same fashion as $\hat{\gamma}$ could be dropped for the case of the addition of a variable. Consequently, changes in $\boldsymbol{\Phi}_{\mathcal{A}(\lambda)}$ do not change the monotony of the GDF estimate.

C Lower bound properties of BIC_{min}

BIC_{min} is a lower bound for BIC, if $\forall k \geq i$

$$\mathrm{BIC}_{min}(i) \leq \mathrm{BIC}(k), \tag{37}$$

which equals

$$\frac{N}{\sigma_\varepsilon^2}\,\mathrm{MSE} + df(\lambda_i)\log N \leq \frac{N}{\sigma_\varepsilon^2}\,\mathrm{MSE}(\lambda_i) + df(\lambda_k)\log N \tag{38}$$

which is always fulfilled because $\mathrm{MSE} \leq \mathrm{MSE}(\lambda_i)$ and $df(\lambda_i) \leq df(\lambda_k)$ for $i \leq k$ and $N \geq 1$, $\sigma_\varepsilon^2 > 0$.

D SNR definition for simulated spectra

Given the undistorted simulated signal $\mathbf{s}$, the effect of Poisson noise is simulated with $s_i \leftarrow v_i$, where v_i is drawn from a Poisson distribution with mean $ks_i + 1$. The signal-to-noise ratio (SNR) thus depends on the parameter k. In order to determine k for a selected set of SNR values, we consider the definition

$$\mathrm{SNR} \doteq \frac{\sigma_s^2}{\sigma_n^2}. \tag{39}$$

The empirical variance of the original signal $\mathbf{s}$ multiplied by a scalar k is defined as

$$\sigma_s^2(k) \doteq k^2 \sum_{i=1}^{N} (s_i - \bar{s})^2, \tag{40}$$

where $\bar{s}$ denotes the mean over all s_i. For Poisson noise, location and dispersion parameters coincide, i.e. with $X \sim \mathcal{P}(\lambda)$ we have $\mathrm{Var}(X) = \mathbb{E}(X) = \lambda$, and we approximate the variance of a set of Poisson variables $n_i \sim \mathcal{P}(ks_i), i = 1, \ldots, N$ by their average

$$\sigma_n^2(k) \doteq \frac{1}{N}\sum_{i=1}^{N} ks_i. \tag{41}$$

For a given SNR, this allows the estimation of k because

$$\mathrm{SNR} = \frac{\sigma_s^2(k)}{\sigma_n^2(k)} = k\frac{\sigma_s^2}{\sigma_n^2} \tag{42}$$

and thus

$$k = \frac{\sigma_n^2}{\sigma_s^2}\,\mathrm{SNR}\,. \tag{43}$$

Authors' contributions

BYR and MK have developed the methodology, implemented the software, carried out the data analysis and drafted the manuscript. HS and JAJS have contributed to the basic methodology and the manuscript, carried out critical review and provided application feedback and evaluation for the proposed methods. FAH has suggested the fractional averagine approach, and has contributed to the manuscript and the overall project design. All authors have read and approved the final manuscript.

Acknowledgements

The authors would like to thank Yin Yin Lin (Dept. of Pathology, Children's Hospital, Boston, MA, USA) for LC/MS data acquisition, Lyle Burton (AB/MDS Sciex, Concord, Canada) for MarkerView 1.2 evaluation versions, Ullrich Köthe, Linus Görlitz, Björn Menze, Michael Kelm (Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Germany), and Flavio Monigatti (Dept. of Pathology, Children's Hospital, Boston, MA, USA) for comments, suggestions, and fruitful discussions. We gratefully acknowledge financial support by the Hans L. Merkle foundation (M.K.), the Karl Steinbuch Scholarship (B.Y.R.), dm/Filiadata GmbH (B.Y.R.), Robert Bosch GmbH (F.A.H.), the Children's Hospital Trust (J.A.J.S. and H.S.), and the DFG under grant no. HA4364/2-1 (B.Y.R., F.A.H).

References

1. Jensen ON: **Interpreting the protein language using proteomics.** *Nature Reviews Molecular Cell Biology* 2006, **7**(6):391–403, [http://dx.doi.org/10.1038/nrm1939].
2. Beretta L: **Proteomics from the Clinical Perspective: Many Hopes and Much Debate**. *Nature Methods* 2007, **4**(10):785–786.
3. Schwartz SA, Weil RJ, Johnson MD, Toms SA, Caprioli RM: **Protein Profiling in Brain Tumors Using Mass Spectrometry: Feasibility of a New Technique for the Analysis of Protein Expression**. *Clinical Cancer Research* 2004, **10**:981–987.
4. Claydon MA, Davey SN, Edwards-Jones V, Gordon DB: **The Rapid Identification of Intact Microorganisms Using Mass Spectrometry**. *Nature Biotechnology* 1996, **14**:1584–1586.
5. Pineda FJ, Antoine MD, Demirev PA, Feldman AB, Jackman J, Longenecker M, Lin JS: **Microorganism Identification by Matrix-Assisted Laser/Desorption Ionization Mass Spectrometry and Model-Derived Ribosomal Protein Biomarkers**. *Analytical Chemistry* 2003, **75**(15):3817–3822.
6. Zhang Z, Marshall AG: **A Universal Algorithm for Fast and Automated Charge State Deconvolution of Electrospray Mass-to-Charge Ratio Spectra**. *Journal of the American Society for Mass Spectrometry* 1998, **9**(3):225–33.
7. Yu W, Wu B, Lin N, Stone K, Williams K, Zhao H: **Detecting and Aligning Peaks in Mass Spectrometry Data with Applications to MALDI**. *Computational Biology and Chemistry* 2006, **30**:27–38.
8. Senko M, Beu S, McLafferty F: **Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions**. *Journal of the American Society for Mass Spectrometry* 1995, **6**:229–233.
9. Horn DM, Zubarev RA, McLafferty FW: **Automated Reduction and Interpretation of High Resolution Electrospray Mass Spectra of Large Molecules**. *Journal of the American Society for Mass Spectrometry* 2000, **11**(4):320–332.
10. Wehofsky M, Hoffmann R, Hubert M, Spengler B: **Isotopic Deconvolution of Matrix-Assisted Laser Desorption/Ionization Mass Spectra for Substance-Class Specific Analysis of Complex Samples**. *European Journal of Mass Spectrometry* 2001, **7**:39–46.
11. Gras R, Müller M, Gasteiger E, Gay S, Binz PA, Bienvenut W, Hoogland C, Sanches JC, Bairoch A, Hochstrasser DF, Appel RD: **Improving Protein Identification from Peptide Mass Fingerprinting through a Parameterized Multi-Level Scoring Algorithm and an Optimized Peak Detection**. *Electrophoresis* 1999, **20**:3535–3550.
12. Rockwood A, Van Orden S, Smith R: **Rapid Calculation of Isotope Distributions**. *Analytical Chemistry* 1995, **67**:2699–2704.

13. Rockwood A, Van Orden SL, Smith RD: **Ultrahigh-Speed Calculation of Isotope Distributions**. *Analytical Chemistry* 1996, **68**:2027–2030.

14. Rockwood A, Haimi P: **Efficient Calculation of Accurate Masses of Isotopic Peaks**. *Journal of the American Society for Mass Spectrometry* 2006, **17**:415–419.

15. Yergey JA: **A General Approach to Calculating Isotopic Distributions for Mass Spectrometry**. *International Journal of Mass Spectrometry and Ion Physics* 1983, **52**:337–349.

16. Senko M: **Isopro 3.0** 1997, [http://members.aol.com/msmssoft/].

17. Breen EJ, Hopwood FG, Williams KL, Wilkins MR: **Automatic Poisson Peak Harvesting for High Throughput Protein Identification**. *Electrophoresis* 2000, **21**:2243–2251.

18. Chen L, Sze SK, Yang H: **Automated Intensity Descent Algorithm for Interpretation of Complex High-Resolution Mass Spectra**. *Analytical Chemistry* 2006, **78**:5006–5018.

19. Kaur P, O'Connor PB: **Algorithms for automatic interpretation of high resolution mass spectra**. *Journal of the American Society for Mass Spectrometry* 2006, **17**(3):459–468.

20. Szymura JA, Lamkiewicz J: **Band Composition Analysis: a new Procedure for Deconvolution of the Mass Spectra of Organometallic Compounds**. *Journal of Mass Spectrometry* 2003, **38**:817–822.

21. Wehofsky M, Hoffmann R: **Automated Deconvolution and Deisotoping of Electrospray Mass Spectra**. *Journal of Mass Spectrometry* 2002, **37**:223–229.

22. Zhang X, Hines W, Adamec J, Asara JM, Naylor S, Regnier FE: **An Automated Method for the Analysis of Stable Isotope Labeling Data in Proteomics**. *Journal of the American Society for Mass Spectrometry* 2005, **16**:1181–1191.

23. Mason CJ, Therneau TM, Eckel-Passow JE, Johnson KL, Oberg AL, Olson JE, Nair KS, Muddiman DC, Bergen HRI: **A Method for Automatically Interpreting Mass Spectra of ^{18}O Labeled Isotopic Clusters**. *Molecular & Cellular Proteomics* 2006, **6**:305–318.

24. Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH: **Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards**. *Analytical Chemistry* 2003, **75**:4818–4826.

25. Senko MW, Beu SC, McLafferty FW: **Automated Assignment of Charge States from Resolved Isotopic Peaks for Multiply Charged Ions**. *Journal of the American Society for Mass Spectrometry* 1995, **6**:52–56.

26. Tabb DL, Shah MB, Strader MB, Conelly HM, Hettich RL, Hurst GB: **Determination of Peptide and Protein ion Charge States by Fourier Transformation of Isotope-Resolved Mass Spectra**. *Journal of the American Society for Mass Spectrometry* 2006, **17**:903–915.

27. Listgarten J, Emili A: **Statistical and Computational Methods for Comparative Proteomic Profiling Using Liquid Chromatography-Tandem Mass Spectrometry**. *Molecular and Cellular Proteomics* 2005, **4**(4):419–434.

28. Fernández-de-Cossio J, Gonzalez LJ, Satomi Y, Betancourt L, Ramos Y, Huerta V, Besada V, Padron G, Minamino N, Takao T: **Automated Interpretation of Mass Spectra of Complex Mixtures by Matching of Isotope Peak Distributions**. *Rapid Communications in Mass Spectrometry* 2004, **18**:2465–2472.

29. Roussis SG, Proulx R: **Reduction of Chemical Formulas from the Isotopic Peak Distributions of High-Resolution Mass Spectra**. *Analytical Chemistry* 2003, **75**:1470–1482.

30. Samuelsson J, Dalevi D, Levander F, Rögnvaldsson T: **Modular, Scriptable and Automated Analysis Tools for High-Throughput Peptide Mass Fingerprinting**. *Bioinformatics* 2004, **20**:3628–3635.

31. Du P, Angeletti RH: **Automatic Deconvolution of Isotope-Resolved Mass Spectra Using Variable Selection and Quantized Peptide Mass Distribution**. *Analytical Chemistry* 2006, **78**:3385–3392.

32. Tibshirani R: **Regression Shrinkage and Selection via the LASSO**. *Journal of the Royal Statistical Society* 1996, **Series B 58**:267–288.

33. Kaur P, O'Connor PB: **Use of Statistical Methods for Estimation of Total Number of Charges in a Mass Spectrometry Experiment**. *Analytical Chemistry* 2004, **76**:2756–2762.

34. Casella G, Berger RL: *Statistical Inference*. Duxbury Press 2001.

35. Lawson CL, Hanson RJ: *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, N J 1974.

36. Park MY, Hastie T: **An L_1 Regularization-path Algorithm for Generalized Linear Models**. *Journal of the Royal Statistical Society, Series B* 2007, **69**:659–677.

37. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*. Springer Verlag New York 2001.

38. Ye J: **On Measuring and Correcting the Effects of Data Mining and Model Selection**. *Journal of the American Statistical Association* 1998, **93**:120–131.

39. Efron B, Hastie T, Johnstone I, Tibshirani R: **Least Angle Regression**. *Annals of Statistics* 2004, **32**(2):407–499.

40. Zou H, Hastie T, Tibshirani R: **On the "Degrees of Freedom" of the Lasso**. *Annals of Statistics* 2007, **35**(5):2173–2192.

41. Bairoch A, Apweiler R: **The SWISS-PROT Protein Sequence Database and its Supplement TrEMBL in 2000**. *Nucleic Acids Research* 2000, **28**:45–48.

42. Tibshirani R, Hastie T, Narasimhan B, Soltys S, Shi G, Koong A, Le QT: **Sample Classification from Protein Mass Spectrometry, by Peak Probability Contrasts**. *Bioinformatics* 2004, **20**(17):3034–3044.

43. Wallace WE, Kearsley AJ, Guttman CM: **An Operator-Independent Approach to Mass Spectral Peak Identification and Integration**. *Analytical Chemistry* 2004, **76**:2446–2452.

44. Kearsley AJ, Wallace WE, Bernal J, Guttman CM: **A Numerical Method for Mass Spectral Data Analysis**. *Applied Mathematics Letters* 2005, **18**:1412–1417.

45. Mann M: **Useful Tables of Possible and Probable Peptide Masses**. In *43rd Conference on Mass Spectrometry and Allied Topics* 1995.

46. Rockwood AL, Kushnir MM, Nelson GJ: **Dissociation of individual isotopic peaks: predicting isotopic distributions of product ions in MS^n.** *Journal of the American Society for Mass Spectrometry* 2003, **14**(4):311–22.

Additional files

Additional file 1 — simulation_data.zip

A zip folder containing all simulation files (R data files)

- uniprot_sprot-HUMAN-tryptic-equispaced11-500-700-charge-1-N-500-k-20-spectra-SNR-5.rda

  containing 500 simulated spectra with a signal to noise ratio of 5

- uniprot_sprot-HUMAN-tryptic-equispaced11-500-700-charge-1-N-500-k-20-spectra-SNR-10.rda

  containing 500 simulated spectra with a signal to noise ratio of 10

- uniprot_sprot-HUMAN-tryptic-equispaced11-500-700-charge-1-N-500-k-20-spectra-SNR-25.rda

  containing 500 simulated spectra with a signal to noise ratio of 25

- uniprot_sprot-HUMAN-tryptic-equispaced11-500-700-charge-1-N-500-k-20-spectra-SNR-50.rda

  containing 500 simulated spectra with a signal to noise ratio of 50

- uniprot_sprot-HUMAN-tryptic-equispaced11-500-700-charge-1-N-500-k-20-spectra-SNR-100.rda containing 500 simulated spectra with a signal to noise ratio of 100
- uniprot_sprot-HUMAN-tryptic-equispaced11-500-700-charge-1-N-500-k-20.rda contains the ground truth for each simulated spectrum, so the exact m/z-position as well as the amino acid sequence of the peptide
- relevantRegions-uniprot_sprot-HUMAN-tryptic-equispaced11-500-700-charge-1-N-500-k-20.rda list of the relevant regions in which NITPICK and pepex were evaluated
- mz.bins.500-700.equispaced11.rda gives the underlying mass binning for all simulated spectra in the range of m/z 500-700

Additional file 2 — BSA-sample.zip

The zipped original LC/MS .wiff-file on which MarkerView was run (as acquired by the AB/Sciex QStar instrument)

Additional file 3 — TOF-MS-yylBSAstd-sample4-23.817-29.278-rebinned.txt

The original spectrum of BSA-sample.wiff integrated over retention time (23.817-29.278 minutes) on which NITPICK was run

Additional file 4 — pepex_simulation_results.zip

A zip-folder containing all pepex results on the simulated data

- pepex_parameter_optimization describing the gradient descent parameter optimization applied for each SNR starting from the preset SNR-threshold-parameter of 2
- thresh'j' for j in 0.5,1,2,3,4,5,7 as a folder containing a folder for each signal to noise ratio
  - a folder for each SNR containing preSNR'SNR'_'i'_pepex.xml as the output of pepex for the respective preSNR'SNR'_'i'.txt file and SNR'SNR'_'i'.txt, the parsed peak list of pepex
  - TP_SNR'SNR'_'N'.rda containing the number of peaks correctly identified by pepex for a given spectrum, N is the number of spectra included (as optimization was restricted to 50 first spectra)
  - FP_SNR'SNR'_'N'.rda containing the number of peaks incorrectly identified by pepex for a given spectrum, N is the number of spectra included (as optimization was restricted to 50 first spectra)

Additional file 5 — NITPICK_simulation_results.zip

A zip-folder containing all NITPICK results on the simulated data sets and for each SNR (SNR in 5, 10, 25, 50, 100) a R data file called

- resultList_0_’SNR’_0.1.RDA gives the peaks found by NITPICK for all spectra of a certain SNR
- length_ResultList_0_’SNR’_0.1.RDA gives the number of peaks found by NITPICK for each spectrum
- correct_0_’SNR’_0.1.RDA gives the number of correctly identified peaks found by NITPICK for each spectrum
- tooMany_0_’SNR’_0.1.RDA gives the number of incorrectly identified peaks found by NITPICK for each spectrum
- pp_resultList_0_3_0_’SNR’_0.1.RDA gives the peaks found by NITPICK for all spectra of a certain SNR after postprocessing with g=3
- length_ResultList_0_3_0_’SNR’_0.1.RDA gives the number of peaks found by NITPICK for each spectrum after postprocessing with g=3
- correct_0_3_0_’SNR’_0.1.RDA gives the number of correctly identified peaks found by NITPICK for each spectrum after postprocessing with g=3
- tooMany_0_3_0_’SNR’_0.1.RDA gives the number of incorrectly identified peaks found by NITPICK for each spectrum after postprocessing with g=3

Additional file 6 — BSA-sample_NITPICK.xls

Excel sheet containing the peaks detected by NITPICK (mz-position, charge, intensity) as well as their manual validation

Additional file 7 —BSA-sample_MarkerView.xls

Excel sheet containing the peaks detected by MarkerView (mz-position, charge, if available) as well as their manual validation

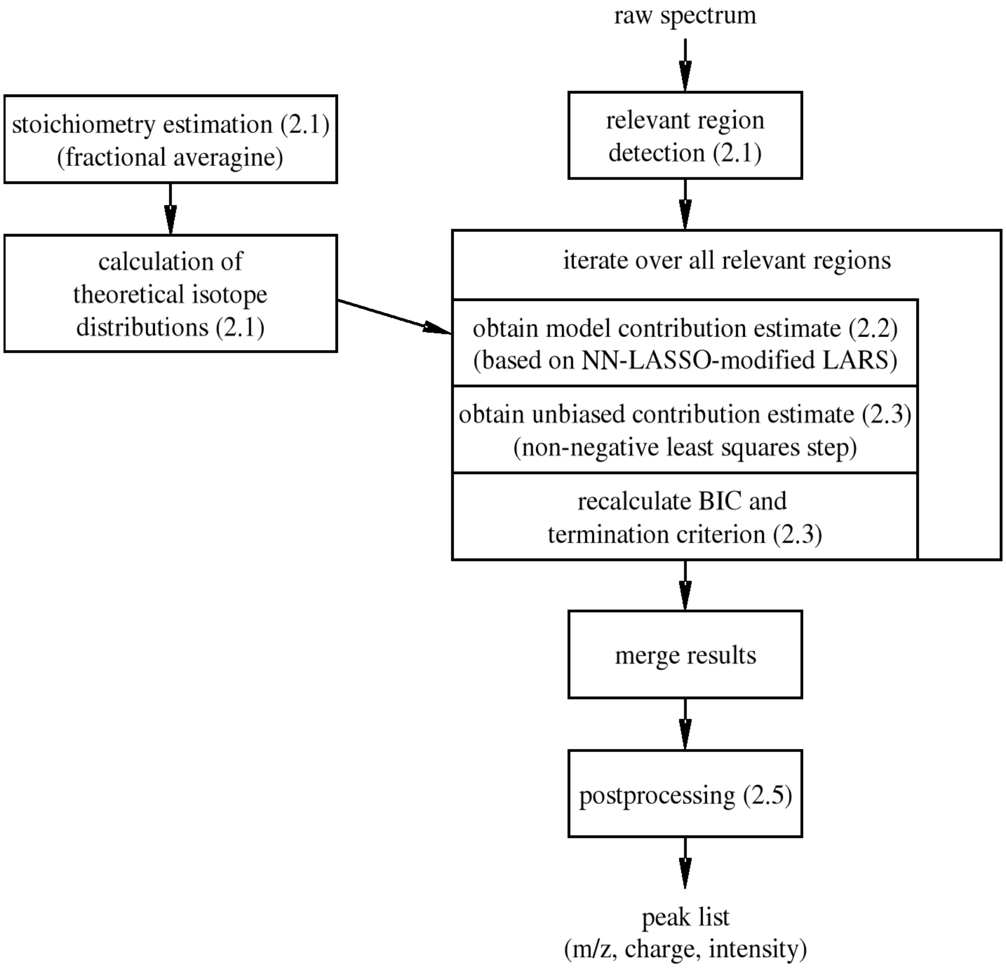
raw spectrum
stoichiometry estimation (2.1)
(fractional averagine)
relevant region
detection (2.1)
calculation of
theoretical isotope
distributions (2.1)
iterate over all relevant regions
obtain model contribution estimate (2.2)
(based on NN-LASSO-modified LARS)
obtain unbiased contribution estimate (2.3)
(non-negative least squares step)
recalculate BIC and
termination criterion (2.3)
merge results
postprocessing (2.5)
peak list
(m/z, charge, intensity)

Figure 1: NITPICK workflow overview: raw spectrum preprocessing, relevant region detection, region-wise peak picking, merging of detected peaks and peak list postprocessing. At the heart of the method lies an iterative feature selection procedure controlled by a statistical termination criterion, as illustrated by the large box in the center. As a tightly interconnected prerequisite to the main workflow, the column on the left depicts the steps required for the calculation of the regression model matrix. Numbers in parentheses give the manuscript sections in which the specific steps are detailed.

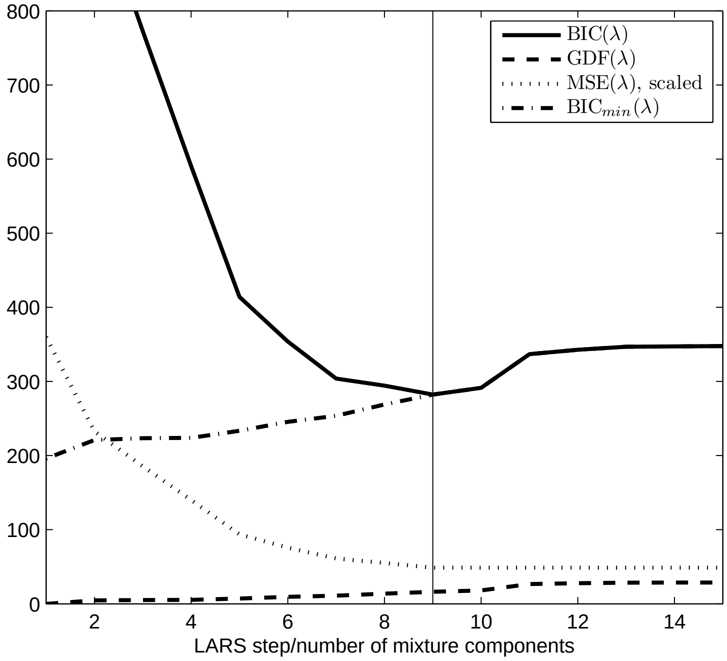

Figure 2: Efficient automated determination of the number of components in an area with overlapping peaks using the $\mathrm{BIC}_{min}(\lambda)$ termination criterion: the mean squared error MSE (scaled, dotted) decreases monotonically over the LARS steps and the generalized degrees of freedom $\mathrm{GDF}(\lambda)$ (dashed) increase monotonically. The resulting $\mathrm{BIC}(\lambda)$ measure (solid) exhibits a minimum $\mathrm{BIC}(\lambda_9)$ in the 9th LARS step and λ_9 is accepted as a minimizer because the lower bound $\mathrm{BIC}_{min}(\lambda_{10})$ exceeds $\mathrm{BIC}(\lambda_9)$ in the 10th LARS step.

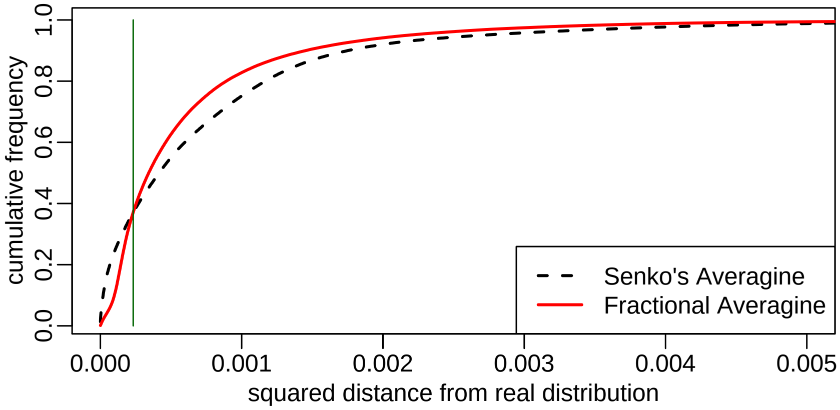

Figure 3: Comparison of the impact of *averagine* and *fractional averagine* stoichiometry estimation errors on the estimation of theoretical isotope distributions: the cumulative histograms of least squares deviations from the true theoretical isotope distribution illustrate the superior overall performance of fractional averagine (solid line) compared to Senko's classical averagine (dashed line): fractional averagine causes a 17% decrease in mean squared error magnitude.

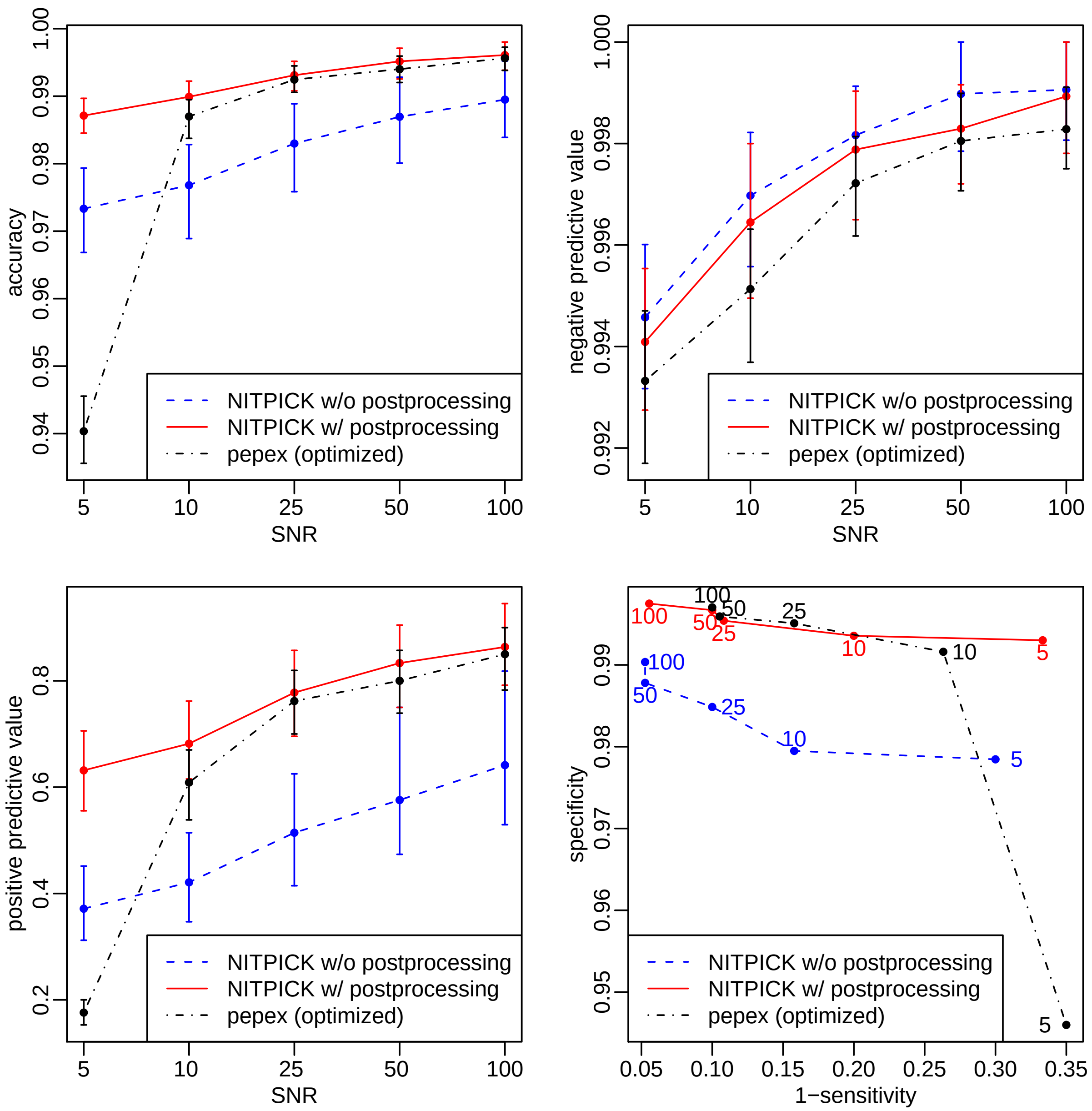

Figure 4: Evaluation and comparison with the pepex algorithm on simulated data: accuracy (top left), negative predictive values (top right), positive predictive values (bottom left) and sensitivity-specificity traces (bottom right). Plots show NITPICK results in solid red, NITPICK results without postprocessing in dashed blue and pepex results (optimized, see text) in dashed-dotted black. NITPICK is clearly superior in terms of accuracy, specificity and sensitivity.

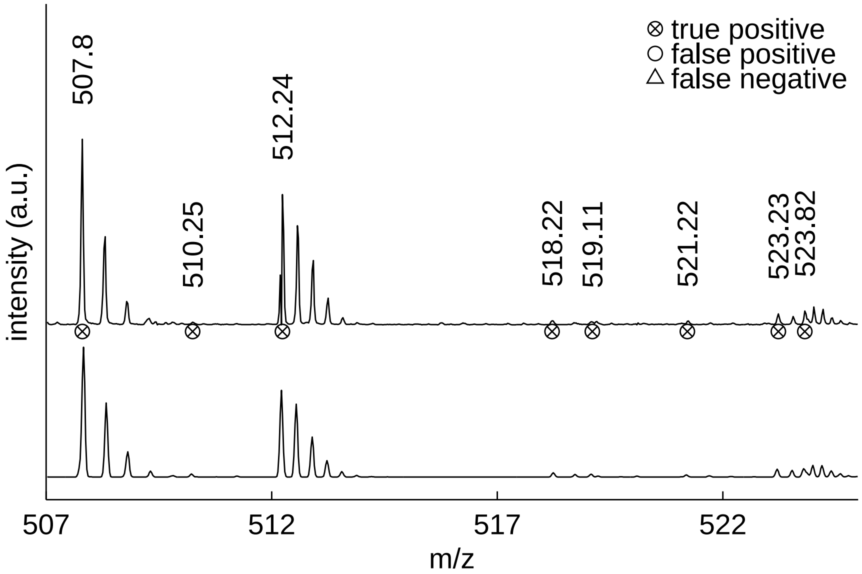

Figure 5: Peak picking in the *m/z* 507-525 mass range: Illustration of observed (top) and reconstructed (bottom) spectra. All detected peaks could be confirmed, including the monoisotopic masses of the mixture distribution with components located at *m/z* 523.23 (z=3) and *m/z* 523.82 (z=5).

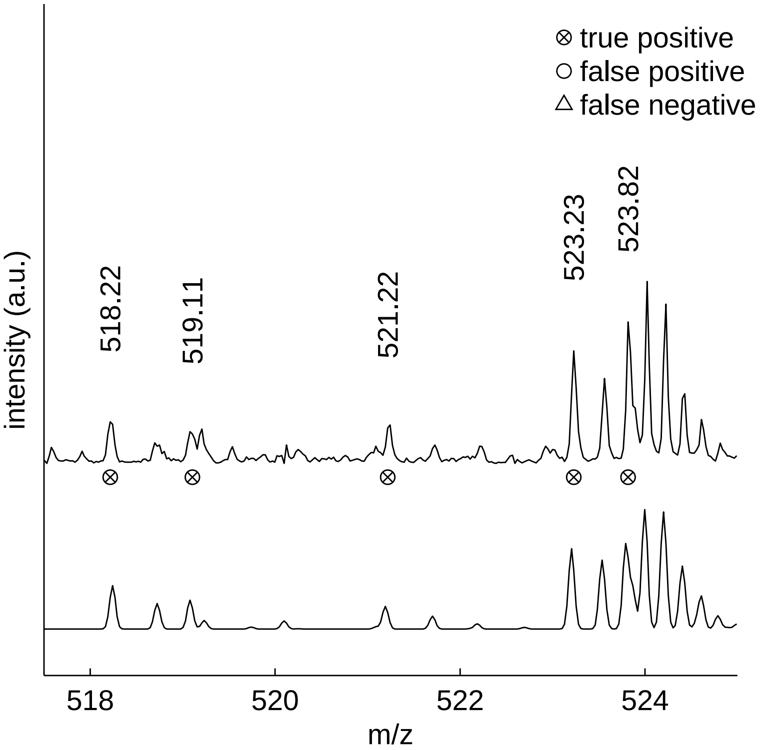

Figure 6: Zoom on the *m/z* 518–525 mass range: NITPICK proves capable of resolving overlapping isotope distributions and assigning correct monoisotopic masses for the distributions located at *m/z* 518.22 and 519.11 and at *m/z* 523.23 and 523.82. See text for details.

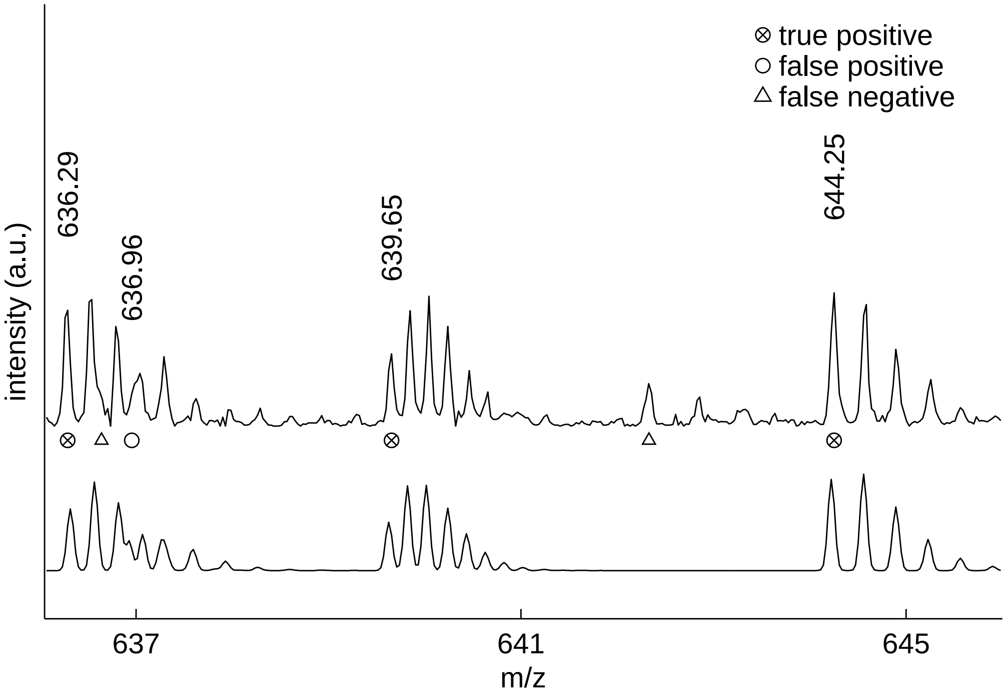

Figure 7: Peak picking results in the *m/z* 636–646 mass range: Illustration of observed (top) and reconstructed (bottom) spectra. At *m/z* 636.64 and *m/z* 636.96 we observe incomplete unmixing: The isotope distribution (z=3) with monoisotopic mass m_0 located at *m/z* 636.29 heavily overlaps the distribution (z=3) with $m_0 = 636.64 m/z$ (left triangle marker). The overlap proves inseparable and the monoisotopic mass of the second distribution is wrongly detected at *m/z* 636.96. Further, due to conservative noise level/complexity estimation, the isotope distribution located at *m/z* 642.33 (right triangle marker) is not detected. Note that in both of the distributions located at *m/z* 636.29 and *m/z* 639.65, the monoisotopic mass peak does not correspond to the most intensive peak.

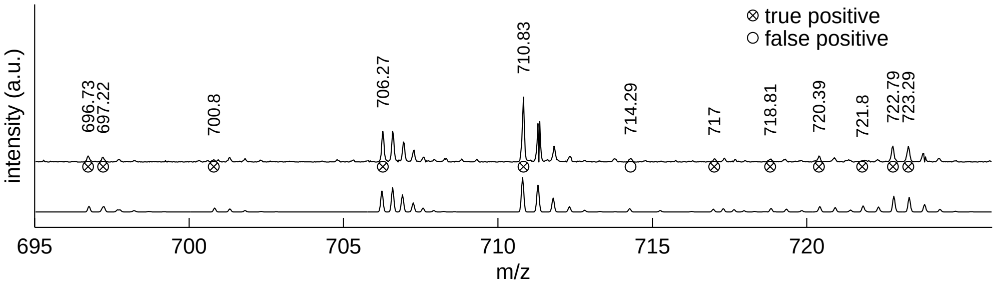

Figure 8: Peak picking in the *m/z* 695–725 mass range: Illustration of observed (top) and reconstructed (bottom) spectra. With a single exception, all detected peaks could be manually confirmed. The peak detected at *m/z* 714.29 corresponds to the first isotope peak of the isotope distribution located at *m/z* 713.78 (z=2). In the *m/z* 718–724 region the algorithm proves capable of resolving nontrivial low-intensity mixtures.

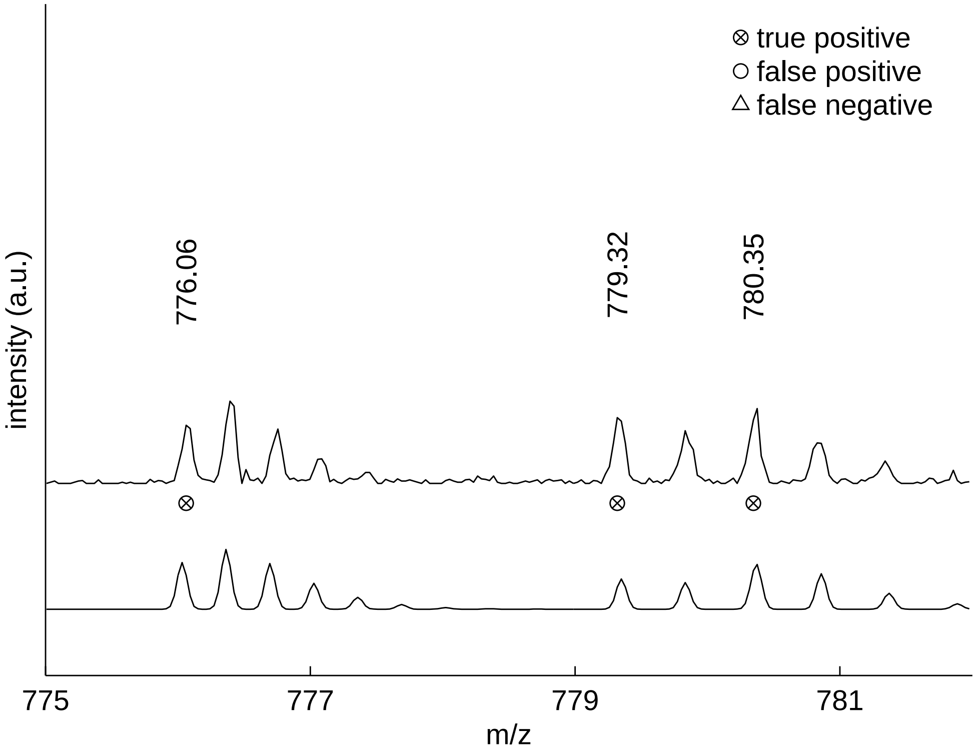

Figure 9: Observed (top) and reconstructed (bottom) mass spectrum in the *m/z* 775–782 range: the separation of two heavily overlapping isotope distribution clearly illustrates the benefits of NITPICK's intensity model-based approach to the peak picking/feature extraction problem: the second isotope peak of the isotope distribution located at *m/z* 779.32 (charge 2) and the monoisotopic peak of the distribution located at *m/z* 780.35 (charge 2) are exactly superimposed and can only be distinguished by taking intensity information into account.