SoftwareX 25 (2024) 101602

Contents lists available at ScienceDirect

SoftwareX

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

Original Software Publication

SEC2MWD: A MATLAB toolbox for derivation of molecular weight distributions from size exclusion chromatography

Ingrid Måge *, Josipa Matić, Katinka Riiser Dankel

Nofima AS - Norwegian Institute of Food, Fisheries and Aquaculture Research, Tromsø, Norway

ARTICLE INFO

Keywords:
Size exclusion chromatography
SEC
Molecular weight distribution
Average molecular weight

ABSTRACT

Size exclusion chromatography (SEC) is a type of liquid chromatography used for separating molecules based on their size. The pipeline for converting a raw chromatogram to a molecular weight distribution involves multiple steps and require various parameters to be defined for each step. Commercial software lack transparency in terms of methods and algorithms, and it may be cumbersome to explore effects of different parameter settings. We have therefore developed a MATLAB toolbox that reproduces the main functionality of commercial software in a transparent and flexible manner. The toolbox consists of seven main functions, each representing a step in the calculation pipeline. The modular architecture makes it easy to modify or replace individual steps of the pipeline if necessary.

Metadata

Nr.	Code metadata description	*Please fill in this column*
C1	Current code version	*v1*
C2	Permanent link to code/repository used for this code version	https://github.com/ingridmage/SEC2MWD
C3	Permanent link to reproducible capsule	
C4	Legal code license	GNU General Public License v3.0
C5	Code versioning system used	*git*
C6	Software code languages, tools and services used	*MATLAB*
C7	Compilation requirements, operating environments and dependencies	
C8	If available, link to developer documentation/manual	
C9	Support email for questions	Ingrid.mage@nofima.no

1. Motivation and significance

Size exclusion chromatography (SEC) is a type of liquid chromatography used for separating molecules based on their size. SEC separates molecules by passing a sample through a column containing porous beads with a specific pore size. Smaller molecules can diffuse into the pores of the beads and therefore take longer to travel through the column, while larger molecules cannot enter the pores and pass through the column more quickly. The larger molecules are therefore eluted first, while the smaller molecules are eluted later. SEC is commonly used to purify and characterize macromolecules such as proteins and other polymers. The chromatogram can also be used to calculate the molecular weight distribution of a sample, by applying a model that link retention time to molecular weight.

The calculation pipeline for determining the molecular weight distribution from size exclusion chromatograms involves multiple steps that require various parameters to be defined for each step. While commercial software offers user-friendly procedures through graphical interfaces and a wide range of functionalities, they lack transparency in terms of methods and algorithms, and are therefore not well suited for research purposes. The primary motivation for developing this software is to reproduce the main functionality of commercial software in a transparent and flexible manner. There is an openly available MATLAB toolbox for chromatography in general [1], but not for the special case of size exclusion chromatography.

2. Software description

The software is written in MATLAB language. It is released under the GNU General Public License, and its technical documentation and examples are available on GitHub: https://github.com/ingridmage/SEC2MWD.

* Corresponding author.
E-mail address: ingrid.mage@nofima.no (I. Måge).

https://doi.org/10.1016/j.softx.2023.101602
Received 14 September 2023; Received in revised form 8 November 2023; Accepted 28 November 2023
Available online 8 December 2023

2.1. Software architecture

The toolbox consists of seven main functions, corresponding to the calculation pipeline shown in Fig. 1. Data is first imported into a MATLAB table where each row represents a SEC run and columns contain the raw data and additional meta data. For each subsequent step in the pipeline, the table is populated with more columns containing processed data and calculated results.

In addition to the seven main functions, there is one supporting function that converts retention time values to molecular mass values.

An example of the full workflow is given in the script *workflowExample.m*, including plots to check the results in each step. The modular architecture makes it easy to modify or replace individual steps of the pipeline.

2.2. Software functionalities

This section describes in more detail the different steps of the pipeline.

2.2.1. Data import – importCDF.m

The toolbox contains a function for importing data from Common Data Format (CDF) files. The function is a wrapper for the native MATLAB function “cdfread”. The contents of these files may vary, and the function might need modifications depending on the specific file contents.

For other raw data file formats, the importCDF function need to be exchanged with an appropriate import function. The imported data must be collected in a MATLAB table, where each row represents one measurement. The table must have the following columns:

- sampleID (cell array)
- RetentionTimeRaw (matrix)
- SignalRaw (matrix)

The table may also contain any number of additional variables.

2.2.2. Crop and interpolate to desired resolution – cropAndInterpolate.m

This function crops the chromatograms to the relevant retention time interval and interpolates to the desired resolution using linear interpolation. It requires two parameters to be defined:

- Retention time limits: a two-element vector defining the minimum and maximum retention time.
- Resolution: The number of data points representing each cropped chromatogram. The default value is 1800.

2.2.3. Estimate baseline – estimateBaseline.m

The optimal baseline estimation method depends on characteristics of the chromatogram, such as the peak shapes (sharp versus distribution), the signal-to-noise ratio and the nature of the baseline (flat versus diverging). We have chosen to implement the well-established Asymmetric Least Squares method [2], which has proven to work well for many different data types such as chromatograms, Raman- and NMR spectra [3]. Our implementation is a modified version of the baseline correction method in the Chromatography toolbox for MATLAB [1]. The

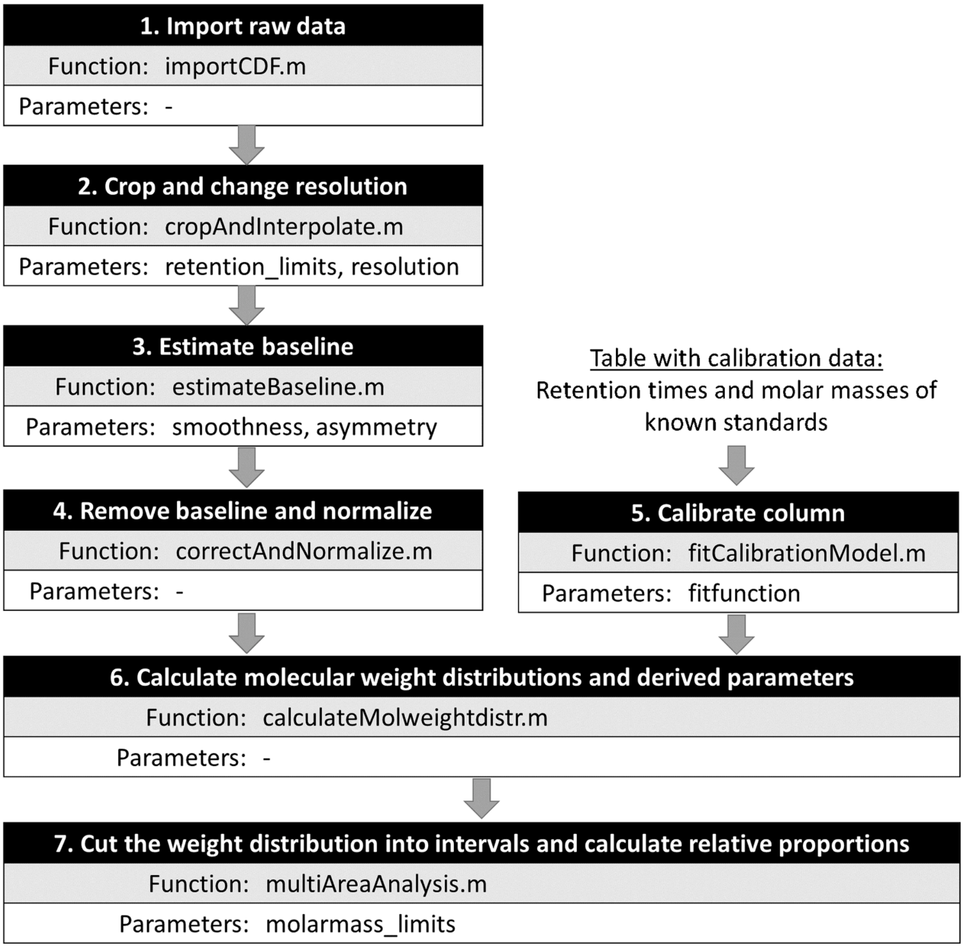

1. Import raw data
Function: importCDF.m
Parameters: -
2. Crop and change resolution
Function: cropAndInterpolate.m
Parameters: retention_limits, resolution
3. Estimate baseline
Function: estimateBaseline.m
Parameters: smoothness, asymmetry
Table with calibration data:
Retention times and molar masses of known standards
4. Remove baseline and normalize
Function: correctAndNormalize.m
Parameters: -
5. Calibrate column
Function: fitCalibrationModel.m
Parameters: fitfunction
6. Calculate molecular weight distributions and derived parameters
Function: calculateMolweightdistr.m
Parameters: -
7. Cut the weight distribution into intervals and calculate relative proportions
Function: multiAreaAnalysis.m
Parameters: molarmass_limits

Fig. 1. The calculation pipeline consists of seven main functions, here represented by boxes and arrows showing the workflow order. For each main function, the changeable parameters are also listed.

function requires two parameters to be set: smoothness and asymmetry. As their names imply, they both contribute to controlling the shape of the estimated baseline and need to be tuned according to the complexity and shape of the chromatogram. The suggested range for smoothness is 10^3–10^9 (default 10^6), and for asymmetry 10^{-6}–10^{-1} (default 10^{-4}), according to [1]. It is important to assess the results and modify the parameters if needed.

2.2.4. *Correct and normalize the chromatograms – correctAndNormalize.m*

This function simply subtracts the estimated baseline from each chromatogram and normalizes to unit area. Samples containing solvents with absorbances lower than the mobile phase will cause negative peaks. This is corrected for by setting all negative values to zero before normalization.

2.2.5. *Calibrate the column – fitCalibrationModel.m*

A calibration model is needed to convert retention times to molar masses. This is usually done by collecting chromatograms from a set of compounds with known molecular weights, identifying the retention times for the peaks, and fitting a function of the form:

$$log(M) = f(RT),$$

Where M is molar mass (g/mol) and RT is the retention time. The function f is usually an odd-numbered polynomial. The function *fitCalibrationModel.m* takes log(M) and RT as inputs, and fits either a first order polynomial (i.e. a straight line), a third order polynomial, or the average between first and third order. The latter is used in the WinGPC software [4] denoted “PSS Poly3”. A systematic comparison of first and third order polynomials showed that the optimal choice cannot be decided a priori; it needs to be established from the data [5]

2.2.6. *Calculate weight distribution and derived parameters – calculateMolweightdist.m*

Molecular weight distributions are calculated by applying the calibration function fitted by *fitCalibrationModel.m* to the normalized chromatograms. The calculations are thoroughly described in [6] and more recently in [7]. The following distributions and derived parameters are calculated:

- FV – cumulative weight fraction as function of retention time
- WM – cumulative weight fraction as function of molecular weight
- xM – differential log weight distribution
- wM – weight distribution
- nM – number distribution
- Mw – weight average molecular weight
- Mn – number average molecular weight
- PDI – Polydispersive index

2.2.7. *Calculate relative proportions of weight fractions – multiAreaAnalysis.m*

The chromatogram can be divided in subsections, or fractions, and the relative proportion of each fraction can be estimated from the cumulative weight distribution WM. The function takes a vector of molar mass limits as input. If the user prefers to set limits based on retention times, the corresponding molar masses may be calculated from the function *retentiontimeToMolarmass.m*, which takes a vector of retention times and the calibration model (obtained by *fitCalibrationModel.m*, Section 2.2.5) as input.

3. Illustrative examples

3.1. *Simple benchmark example*

A verification of the calculation pipeline was done by processing the chromatogram given as supplementary material in reference [7]. This is a polymer sample with two distinctive components. Our data processing pipeline reproduced the same molecular weight distributions (xM, wM and nM) and the same estimates of Mw, Mn and PDI, as provided in the supplementary material of reference [7]. See Fig. 2 for results. The data is provided in the github repository, along with a script for reproducing the results.

3.2. *Analysis of complex protein samples*

The size exclusion chromatograms of four protein hydrolysates are used to demonstrate the functionality of the software. The samples were produced from poultry raw materials as described in [8]. Further fractioning was performed using centrifugal filters with a cutoff of 3 kDa (Amicon Ultra 3K, Merck Millipore, MA, USA) before the permeates were analyzed by SEC. Poultry protein hydrolysis results in highly complex matrices with protein and peptide fragments of a wide range of sizes, originating from both muscle and connective tissue proteins. Free amino acids, small metabolites and enzyme stabilizing agents will also be present in the samples.

SEC is commonly used to assess protein hydrolysates [9–12], despite certain well-known method limitations [13]. Ideally, the analyte elution time should only depend on its hydrodynamic volume, but secondary interactions between the analytes and the stationary phase do occur. In the samples provided in this work, this is evident since a range of metabolites are retained beyond the elution time of the sample solvent ($t_{\text{water}} = 12.2$ min). However, when assessing relative differences between samples, SEC can give highly valuable insight into hydrolysis processes.

In the example provided, the lyophilized hydrolysate samples were rehydrated to 10 mg/mL in ultrapure water and analyzed according to the method described in [12]. Chromatographic separation of chromatographic standards and samples was performed with a Dionex UltiMate 3000 Standard System (Thermo Fisher Scientific, Waltham, MA, USA). An injection volume of 15 µL was used and separation was performed at room temperature using a BioSep-SEC-s2000 column with 300 × 7.8 mm i.d., 5 µm particle size and 145 Å pore size (Phenomenex, Torrance, CA, USA). According to the specification data sheet from the manufacturer, the stationary phase of the SEC-s2000 column consists of a silica-type resin and the column has an exclusion range of 200–75,000 g/mol with a denaturing mobile phase. The mobile phase consisted of a mixture of acetonitrile and ultrapure water in a proportion of 30:70 (V/V), containing 0.05 % TFA. Isocratic elution was carried out using a flow rate of 0.9 mL/min for 20.0 min. The data is provided in the github repository, along with a script for running all the calculations.

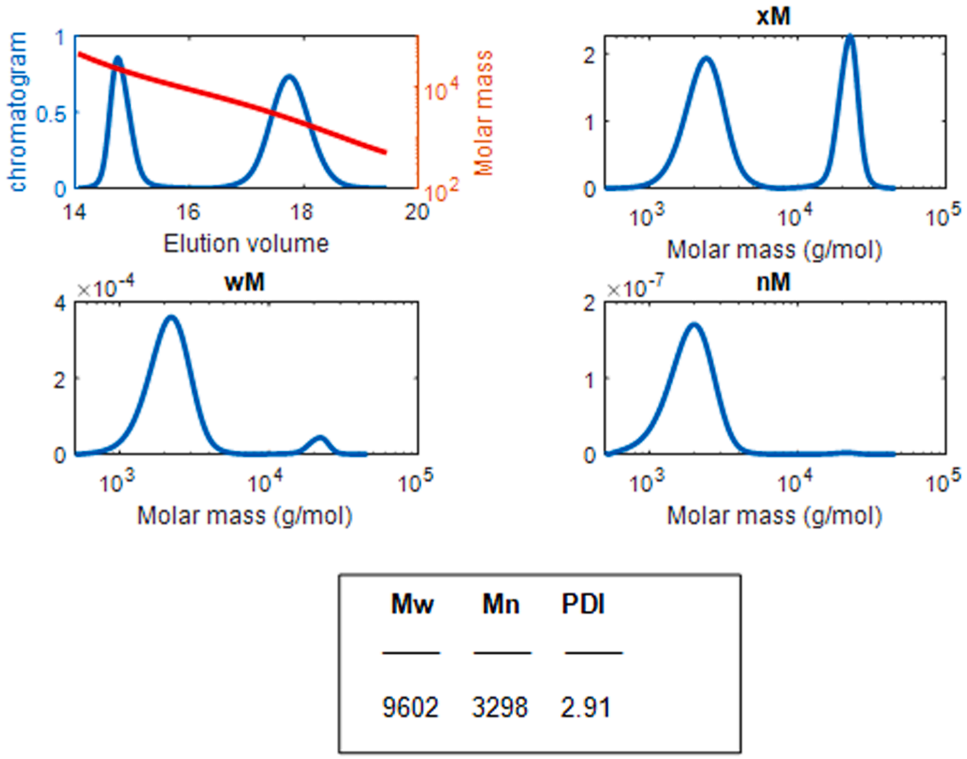
The first step of the pipeline is to import the data. The *importCDF* function creates a MATLAB table with 4 rows (samples) and columns representing sampleID, retention times, raw detector signal and various other meta data (see Fig. 3)

3.2.1. *Steps 2–4: Cropping, baseline correction and normalisation*

The raw chromatograms were first cropped to the retention time 5–20 min and resampled to resolution 1800 using the function *cropAndInterpolate*. Afterwards, baselines were estimated using the default parameter settings (smoothness $= 10^6$ and asymmetry $= 10^{-4}$). The resulting baselines are presented in Fig. 4, alongside with the cropped raw chromatograms. Note that all the estimated baselines are relatively high in the area 6–14 min. This is because the chromatograms have many overlapping peaks in this region. Adjusting the smoothing parameter could mitigate this phenomenon. Fig. 5 shows the outcomes when the smoothing parameter was increased from 10^6 to 10^8. This led to lower baselines with less curvature. We proceed our calculations with these revised baselines.

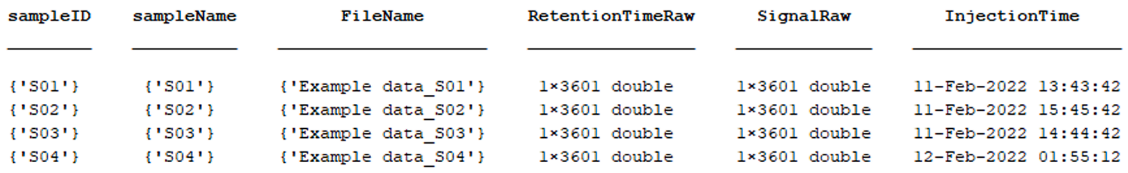
3.2.2. *Step 5: Column calibration*

Data for column calibration is given in the Excel file ‘calibration.xlsx’ on github. It contains the retention times and molecular weights of



Mw	Mn	PDI
9602	3298	2.91

Fig. 2. Results from the simple benchmark example. The normalized chromatogram and the calibration curve are given in the top left subplot, followed by the resulting weight distributions (xM, wM and nM) and the derived parameters Mw, Nm and PDI.



```
sampleID    sampleName          FileName            RetentionTimeRaw     SignalRaw         InjectionTime
________    __________    ______________________    ________________    _____________    ____________________

{'S01'}      {'S01'}      {'Example data_S01'}       1×3601 double      1×3601 double    11-Feb-2022 13:43:42
{'S02'}      {'S02'}      {'Example data_S02'}       1×3601 double      1×3601 double    11-Feb-2022 15:45:42
{'S03'}      {'S03'}      {'Example data_S03'}       1×3601 double      1×3601 double    11-Feb-2022 14:44:42
{'S04'}      {'S04'}      {'Example data_S04'}       1×3601 double      1×3601 double    12-Feb-2022 01:55:12
```

Fig. 3. MATLAB table, output from the function importCDF. This table is the input to all other functions in the pipeline, which each add new results (columns) to the table.

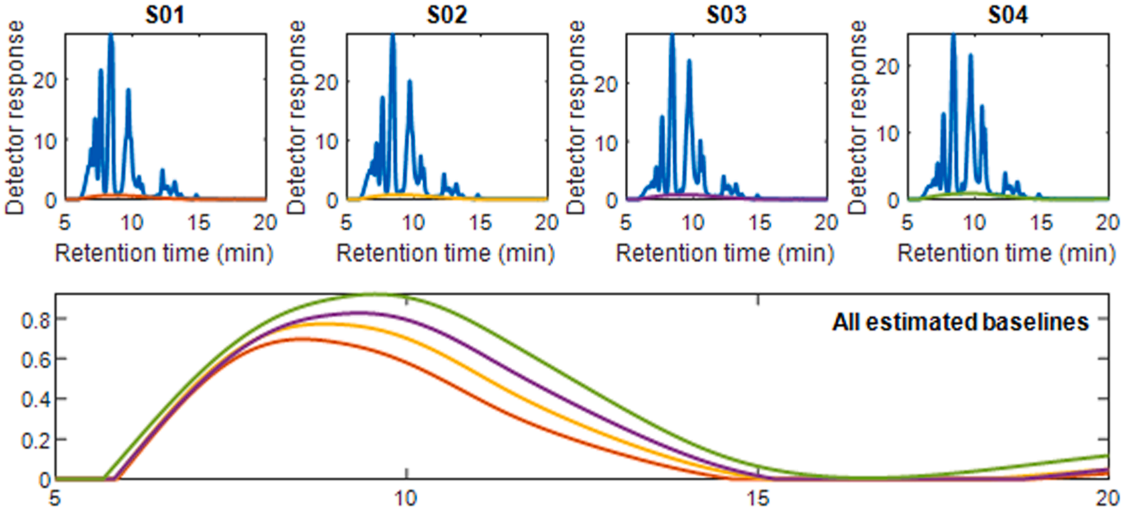


Fig. 4. Cropped raw chromatograms and estimated baselines using default baseline parameter settings, i.e. smoothness = 10^6 and asymmetry is 10^{-4}.

eleven pure proteins, peptides and free amino acids. The function *fitCalibrationModel* may fit either of three functions, as described in Section 2.2.5. The calibration data and three alternative fitted curves are shown in Fig. 6, along with the derivative of the curves. The derivative is important as it is used to calculate weight distributions in step 6 of the pipeline. Large absolute values of the derivative will affect the shape of

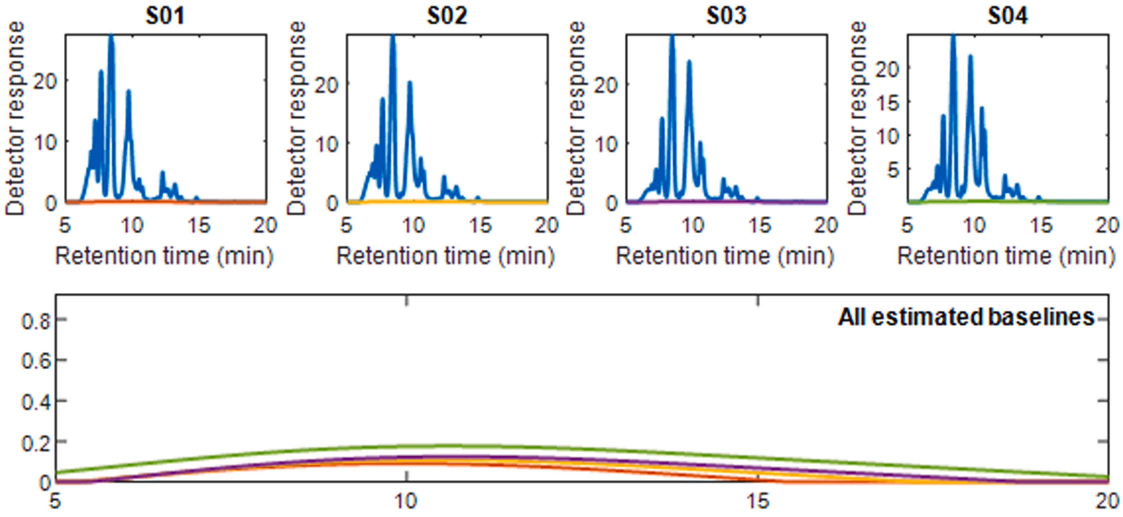


Fig. 5. Cropped raw chromatograms and estimated baselines using smoothness = 10^8 and asymmetry is 10^{-4}.

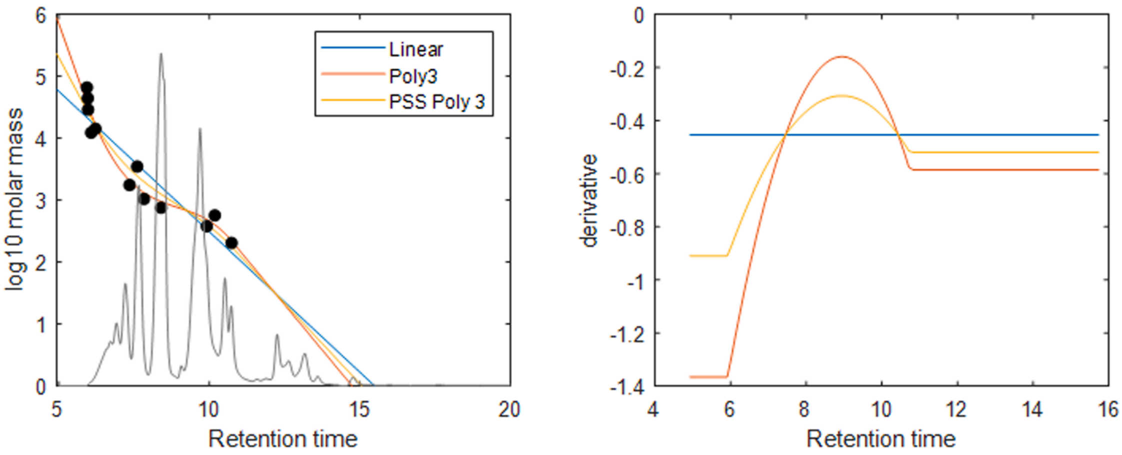


Fig. 6. Left: Calibration data (black dots) with three alternative fits. The average normalized chromatogram is overlayed the calibration curves. Right: the derivative of fitted curves.

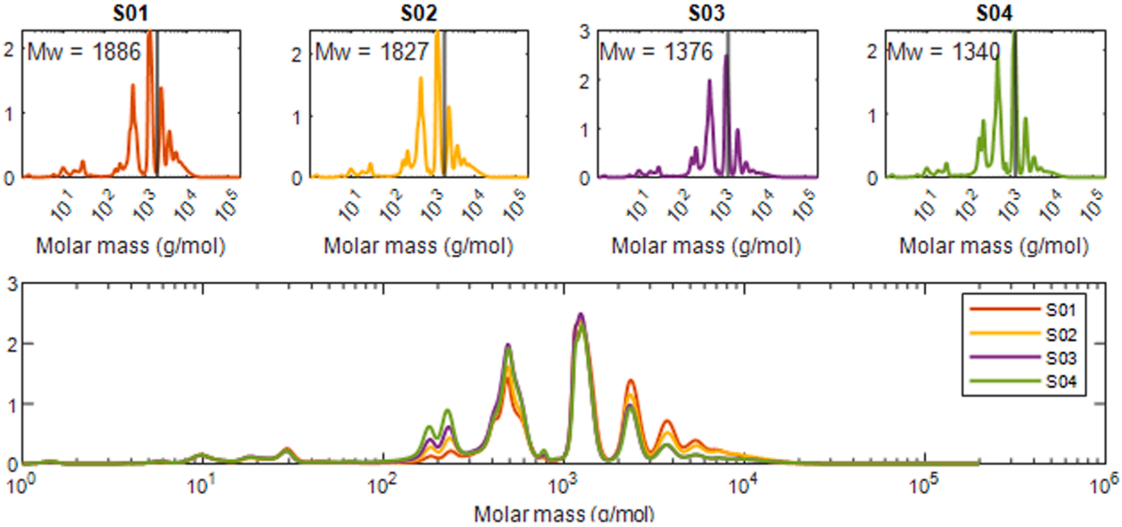


Fig. 7. Top row: Differential log molecular weight distributions (xM) for each sample, with vertical lines indicating the weight average molecular weight (Mw). Bottom: All four chromatograms overlayed.

the distribution substantially. It is therefore desirable to choose a curve that fits the calibration samples well, at the same time as the variation of the derivative is low. Because of this, PSS Poly3 is often a robust choice of fit function, and we continue our calculations using PSS Poly3.

The average normalized chromatogram is shown together with the calibration curves in the left subplot of Fig. 6.

The exclusion range for the column used is stated to be 200–75 000 g/mol for denatured proteins and peptides, and the calibration samples span the range 204–66 000 g/mol. Proteins that are larger than the exclusion range will travel through the column without any retention and elute together at the shortest retention time. Although the separation of the samples should be based solely on their hydrodynamic volume, electrostatic interactions with the stationary phase do occur. Secondary interactions particularly affect the smaller molecules, i.e. those with long retention times. Hence, the estimated molecular weights for both the early and late elution fraction are unreliable. Still, the calibration curve is linearly extrapolated outside the range of the calibration samples to provide a rough estimate of the molecular weights in these regions.

3.2.3. *Weigh distributions and average molecular weights*

Size distributions were calculated with the function *calculateMolweightdistr.m*. Fig. 7 shows the differential log weight distribution (xM), which is the most useful distribution for samples with a large span in molecular weights. The wM is more useful for relatively pure samples (dispersive index ~ 1) and is not shown here.

In Fig. 7, the weight average molecular weight (Mw) values are indicated by vertical lines. We observe an approximate correspondence between the Mw values and the highest peak at 1250 g/mol. However, this agreement doesn't always hold true for complex chromatograms. The Mw value is a *weight* average and is therefore determined mainly by the larger molecules. Even if the peak at 1250 g/mol is equally large for S01, S02 and S04, the Mw value for S04 is notably lower. This is because S04 has a smaller tail in the right (high weight) region of the chromatogram.

It is often useful to assess the relative ratios of specific weight fractions. This can be done with the function *multiArea.m*, which requires molecular weight boundaries as input. For our analysis, we set limits at 900, 1800 and 3000 g/mol, corresponding to valleys in the chromatograms. The results are shown in Fig. 8. It is now clear that S01 and S02 contain more of the large molecules (the two fractions >1800), while the range 900–1800 g/mol is approximately equal for all samples.

4. Impact

First and foremost, this software enables open science by providing a transparent and reproducible pipeline for calculating molecular weight distributions and other derived parameters, such as average molecular weight and PDI, from raw size exclusion chromatographic data.

The software has changed the daily practice of our research group working with SEC, by enabling us to easily inspect all the steps in the processing pipeline and carefully evaluate the respective results. This is facilitated by graphical representations of results from different steps in the pipeline. It can be especially important to scrutinize the chromatographic data after the baseline correction and normalization is performed. Additionally, the software enables researchers to efficiently investigate the effects of different parameter choices on the results. Size-exclusion chromatography is a robust technique, but with low resolution and lacking specificity. Thus, size-exclusion chromatograms usually comprise broad peaks, often with no baseline separation. It is also well-known that the chromatograms might vary between different chromatographic columns. This is especially problematic when analysing complex mixtures with broad, multi-peak molecular weight distributions. It is therefore advantageous to compare the relative amounts of individual fractions of the chromatogrqm, defined either by ranges of molecular weights or by retention times. In this way, the software can be used to quantify and assess the variability caused by different parameter settings in the calculation pipeline.

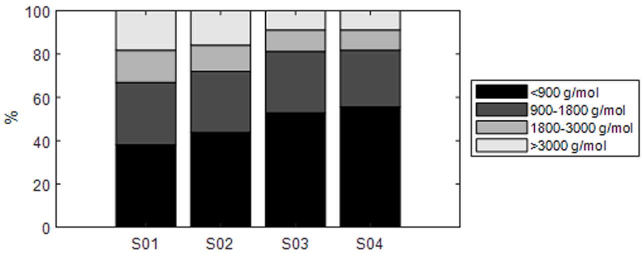


Fig. 8. Relative proportions of four weight fractions in each sample.

The entire analysis is quick and suitable for processing a large number of chromatograms. The layout and format of results makes it convenient for calculations of parameters such as sample recovery or dn/dc values, for further statistical analysis, and for creating desired graphical outputs.

Finally, the toolbox may easily be extended beyond the scope of our work, by exchanging the implemented methods or adding new steps in the pipeline. For instance, it could be relevant to implement alternative baseline estimation methods such as automated iterative moving averaging (AIMA) [3] or FastChrom [14]. Also, support for triple detection (a combination of concentration, viscosity, and light scattering detectors) would be a relevant extension for some application areas. The modular design of the software facilitates such developments.

5. Conclusions

We have developed a flexible and transparent MATLAB toolbox for converting raw SEC chromatograms to molecular weight distributions and calculating derived parameters such as average molecular weights. The toolbox reproduces the main functionality of commercial software. It enables researchers to easily investigate effects of different parameter settings, and to modify or further develop the pipeline.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the data on GitHub.

Acknowledgements

This work was funded by the Foundation for Research Levy on Agricultural Products through the project Precision Food Production (project number 314111), and by The Research Council of Norway through the project SFI Digital Food Quality (project number 309259).

During the preparation of this work the authors used ChatGPT to improve the readability of some sections. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] J. Dillon, “Chromatography toolbox,” https://github.com/chemplexity/chromatography.

[2] Eilers PHC. A perfect smoother. Anal Chem 2003;75(14):3631–6. https://doi.org/10.1021/AC034173T/SUPPL_FILE/AC034173TSI20030422_052903.ZIP. Jul.

[3] Prakash BD, Wei YC. A fully automated iterative moving averaging (AIMA) technique for baseline correction. Analyst 2011;136(15):3130–5. https://doi.org/10.1039/C0AN00778A. Jul.

[4] Agilent, "WinGPC software," https://www.agilent.com/en/product/liquid-chromatography/hplc-ce-software/application-specific-software/wingpc-software.

[5] Vander Heyden Y, Popovici ST, Schoenmakers PJ. Evaluation of size-exclusion chromatography and size-exclusion electrochromatography calibration curves. J Chromatogr A 2002;957(2):127–37. https://doi.org/10.1016/S0021-9673(02)00311-4. May.

[6] Shortt DW. Differential molecular weight distributions in high performance size exclusion chromatography. J Liq Chromatogr 1993;16(16):3371–91. https://doi.org/10.1080/10826079308019695.

[7] Gavrilov M, Monteiro MJ. Derivation of the molecular weight distributions from size exclusion chromatography. Eur Polym J 2015;65:191–6. https://doi.org/10.1016/J.EURPOLYMJ.2014.11.018. Apr.

[8] Kristoffersen KA, et al. Post-enzymatic hydrolysis heat treatment as an essential unit operation for collagen solubilization from poultry by-products. Food Chem 2022;382. https://doi.org/10.1016/j.foodchem.2022.132201. Jul.

[9] Matić J, Bøgwald I, Tengstrand E, Rønning SB, Afseth NK, Wubshet SG. Calanus finmarchicus as a novel source of health-promoting bioactive peptides: enzymatic protein hydrolysis, characterization, and in vitro bioactivity. Biocatal Agric Biotechnol 2023;52:102820. https://doi.org/10.1016/J.BCAB.2023.102820. Sep.

[10] Barbana C, Boye JI. Angiotensin I-converting enzyme inhibitory properties of lentil protein hydrolysates: determination of the kinetics of inhibition. Food Chem 2011; 127(1):94–101. https://doi.org/10.1016/J.FOODCHEM.2010.12.093. Jul.

[11] Newman J, Egan T, Harbourne N, O'Riordan D, Jacquier JC, O'Sullivan M. Correlation of sensory bitterness in dairy protein hydrolysates: comparison of prediction models built using sensory, chromatographic and electronic tongue data. Talanta 2014;126:46–53. https://doi.org/10.1016/J.TALANTA.2014.03.036. Aug.

[12] Wubshet SG, et al. FTIR as a rapid tool for monitoring molecular weight distribution during enzymatic protein hydrolysis of food processing by-products. Anal. Methods 2017;9(29):4247–54. https://doi.org/10.1039/C7AY00865A. Jul.

[13] Štulík K, Pacáková V, Tichá M. Some potentialities and drawbacks of contemporary size-exclusion chromatography. J. Biochem. Biophys. Methods 2003;56(1–3):1–13. https://doi.org/10.1016/S0165-022X(03)00053-8. ElsevierJun. 30.

[14] Johnsen LG, Skov T, Houlberg U, Bro R. An automated method for baseline correction, peak finding and peak grouping in chromatographic data. Analyst 2013;138(12):3502–11. https://doi.org/10.1039/C3AN36276K. May.