TOWARDS A DIGITAL TWIN FOR ANALYTICAL HPLC

Monica Tirapelle [a], Maximilian O. Besenhard [a], Luca Mazzei[a], Jinsheng Zhou[b], Scott A. Hartzell[b] and Eva Sorensen [a,1]

[a] Department of Chemical Engineering, University College London, Torrington Place, London WC1E 7JE, UK

[b] Eli Lilly and company, 893 Delaware St, Indianapolis 46225, USA

Abstract

Digital twins for industrial process development are quickly gaining popularity in the pharmaceutical industry as an effective alternative to expensive and time-consuming physical experiments. This work describes the digital model element of a digital twin of High-Performance Liquid Chromatography (HPLC). The model is based on a mechanistic model implemented in gPROMS ModelBuilder and integrated into the MATLAB environment. Unlike other models reported in the literature, our model comprises a more accurate prediction of the injection profile and can predict the elution behaviour for a wide range of HPLC conditions given a reduced number of experiments. The model is compared against experimental data performed to separate a mixture of eight small drug molecules on a C18 column, in gradient elution mode, and under nine different operative conditions (i.e. 3 temperatures × 3 solvent gradient). We will show that by considering only two isotherm parameters for each molecule, the digital model can accurately predict the retention behaviour of the eight analytes. Furthermore, it facilitates HPLC in-silico method development, showcased here via method time minimization through a dynamic solvent strength gradient. The proposed model is intended to be integrated into a digital twin architecture for offline decision support and real-time optimization.

Keywords

Digital Twin, Analytical HPLC, Chromatography

Introduction

High-Performance Liquid Chromatography (HPLC) is one of the most employed techniques in pharmaceutical and biopharmaceutical industries for purification of a variety of particles ranging from small molecules to large molecular weight compounds such as peptides and proteins. Since the very early stages of computing, the development of HPLC methods has usually been achieved in-silico (i.e. by means of computer simulations), thus reducing the time and costs required for physical experiments (Chen et al., 2020). However, even though computational HPLC has been widely explored, in-silico HPLC methods have not yet reached their full potential (Besenhard et al., 2021).

The most relevant methods used for predicting retention times and developing in-silico HPLC methods are the linear solvation energy relationships (LSER) (Wang et al., 1999) and the quantitative structure retention relationships (QSRR) (Héberger, 2007). These methods are based on semi-empirical models, statistical analysis and/or machine learning algorithms, and thus are limited to describing the restricted number of experimental settings and operative conditions considered for their development. Attractive alternatives are mechanistic and hybrid models that allow assessing a wide range of operative conditions, and thus can be used for developing digital twins, i.e., virtual representations of real processes integrated either periodically (offline digital twin) or in real-time (online digital twin) with the physical system (Rosen et al., 2015; Kritzinger et al., 2018).

Herein, we present the first step towards a digital model for an analytical reversed-phase HPLC that mimics an experimental HPLC facility. The digital model consists of a transport model, an empirically-based injection profile that accounts for the residence time distribution of the analytes in the tubing system upstream of the chromatographic column, and a linear (in concentration) adsorption isotherm that accounts for the variation of temperature and gradient time. The model is fitted against experimental retention times obtained by separating a mixture of 8 active pharmaceutical in-

[1] Corresponding author. Email: e.sorensen@ucl.ac.uk

gredients in gradient elution mode and under 9 different operative conditions (i.e. 3 temperatures and 3 gradient times). With the estimated isotherm parameters, the in-silico HPLC model is able to predict reasonably well the residence times of the molecules under investigation and thus, it can be used to gain insight into the separation mechanism and to support design decisions at production scale. As an example, this work will show that performing a simple solvent gradient optimization allows the analysis times to be significantly shortened.

This work is organized as follows. In the *methodology* section, we report the mechanistic model implemented in gPROMS ModelBuilder and describe the injection profile designed in MATLAB. We further discuss parameter estimation and model optimization. The *results and discussion* section deals with the description of the benchmark experimental study and reports the outcomes of parameter estimation, sensitivity analysis, and gradient method optimization. Finally, we draw some conclusions and discuss possible future extensions.

Methodology

To replicate the real system, we equipped the model with: i) a transport model to account for the movement of each analyte through the column, ii) an adsorption isotherm that describes the equilibrium between the mobile and stationary phase for each analyte, taking into account the impact of both temperature and solvent strength, and iii) a new and more accurate prediction for the injection profile. The model is then used for in-silico HPLC model development and optimization.

Chromatography model

Several transport models have been developed for modeling chromatographic separations such as the ideal model, the equilibrium dispersive model (EDM), the lumped kinetic model, and the general rate model. Furthermore, each model can be solved for different equilibrium adsorption isotherms (Guiochon et al., 2006). Since this work targets purification of small APIs at low concentrations under conditions where the adsorption process and the mass transfer process between mobile and stationary phases is fast, the EDM with linear isotherm is employed.

In the EDM, the mobile and stationary phases are assumed to be constantly at equilibrium and all the non-equilibrium contributions (such as mass transfer resistances) are lumped into an apparent axial dispersion coefficient, $D_{a,i}$. If no reactions take place in the chromatographic column, the chromatographic process is isothermal and isobaric, and the physicochemical properties of the packing do not depend on the radial position within the column, the mass balance equation of the mobile phase reads:

$$\frac{\partial C_i}{\partial t} + F\frac{\partial q_i}{\partial t} + u\frac{\partial C_i}{\partial z} = D_{a,i}\frac{\partial^2 C_i}{\partial z^2} \tag{1}$$

where C_i and q_i are the local concentrations of the analyte in the mobile and stationary phases, respectively, u is the cross-sectional average mobile phase velocity, z is the axial coordinate and F is the solid phase ratio defined as $F \equiv (1-\varepsilon)/\varepsilon$ being ε the total porosity. At the beginning of a chromatographic run, the chromatographic column contains mobile and stationary phases in equilibrium but it is empty of feed components (Guiochon et al., 2006). Thus, the initial conditions correspond to:

$$C_i(z,0) = 0 \tag{2}$$

for $0 < z < L$. Regarding the inlet and outlet boundary conditions, we assigned the Danckwerts boundary conditions that are written respectively as (Danckwerts, 1953):

$$C_i(0,t) - \frac{D_{a,i}}{u}\left(\frac{\partial C_i(0,t)}{\partial z}\right) = C_{0,i}(t) \tag{3}$$

$$\frac{\partial C_i(L,t)}{\partial z} = 0 \tag{4}$$

where $C_{0,i}$ is the analyte concentration in the feed.

Adsorption isotherm

When the concentration of the analytes is very small, as it is in analytical HPLC, the transport model can be coupled with a linear adsorption isotherm, according to which the concentrations of each analyte in the solid phase is proportional to that in the mobile phase (Guiochon et al., 2006):

$$q_i = a_i C_i \tag{5}$$

The slope of the linear isotherm, a_i, is the Henry's constant of adsorption and, multiplied by the volumetric phase ratio, gives the retention factor of component i under linear conditions:

$$k_i \equiv a_i F \tag{6}$$

The latter depends, among other things, on solvent strength/mobile phase composition, temperature, and pH value.

To account for mobile phase composition, we adopt the linear solvent strength (LSS) theory introduced by Snyder et al. (1979). According to the LSS theory, the variation of the retention factor with the volume fraction of the organic modifier, φ, is described with the following linear relationship:

$$\ln k_i(\varphi) = \ln k_{0,i} - S_{S,i}\varphi \tag{7}$$

where $k_{0,i}$ is the (extrapolated) retention factor at infinite dilution, and $S_{S,i}$ is a constant coefficient that is characteristic of a given analyte-mobile phase system and that accounts for the organic solvent elution strength.

To take into account the effect of temperature on the retention, we resort to the van't Hoff theory of Gibbs free energy ΔG_i^0 (Melander et al., 1978):

$$\ln k_i(T) = -\frac{\Delta H_i^0}{RT} + \frac{\Delta S_i^0}{R} + \ln F \tag{8}$$

where ΔH_i^0 is the standard enthalpy of adsorption, ΔS_i^0 is the standard entropy, R is the universal gas constant, and T is the temperature in Kelvin.

If we assume that the effects of temperature and mobile phase composition on the retention behaviour of an analyte are independent of each other, the retention factor can be approximated as (Jandera et al., 2010):

$$k_i(\varphi, T) \propto \exp\left(\frac{C'_{T,i}}{T}\right) \exp\left(-S'_{S,i}\varphi\right) \tag{9}$$

The adsorption isotherm parameters $C'_{T,i}$ and $S'_{S,i}$ appearing in Eq. 9 have to be determined using additional models or experiments. Our digital model is compared with a real HPLC facility and the unknown isotherm parameters are determined via parameter estimation.

Injection profile

The dispersion of the sample related to the part of the system upstream of the chromatographic column can be accounted for with an accurate prediction of the injection profile. However, developing an accurate injection profile represents one of the biggest challenges in modelling a chromatographic process. For this reason, in the literature, the injection profile is often incorrectly implemented as a rectangular pulse (Samuelsson et al., 2010). In developing our model, instead we consider an empirically-based injection profile that accounts for the residence time distribution of the sample through injection-loop tubing volumes and heat exchangers. The obtained injection profile has a sharp front followed by a tailing decay. Giving a detailed description of the injection profile is out of the aims of this work.

In-silico HPLC model

This work represents the first step towards an accurate in-silico HPLC model. The model has been developed in gPROMS ModelBuilder and is called as a single function inside the MATLAB workspace. This allows accessing a wide range of parameter estimation tools and optimization strategies.

To replicate the real system, we use the model for parameter estimation by fitting the experimental elution times. The model formulates parameter estimation in Matlab as an optimization problem. As minimization method, it uses the routine *fmincon()*, which is a Newton-like gradient-based method. The objective function, which aims to minimize the mean of the percent relative errors done in predicting the elution times of each component under n different operative conditions, reads:

$$Obj = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|t_{R,i} - \hat{t}_{R,i}|}{t_{R,i}} \cdot 100\right) \tag{10}$$

where $t_{R,i}$ and $\hat{t}_{R,i}$ are experimental and simulated retention times of the ith operative condition, respectively. Note that, to perform parameter estimation, we have compared only the retention time, and we have not considered the shape of the eluted peak.

The HPLC model is also used to perform computer-assisted optimization. To optimize the HPLC method, several system parameters and operational conditions may be varied. Furthermore, different goals may be achieved with optimization procedures. Here, we perform a multi-factorial optimization of gradient time t_G, and initial gradient concentration φ_0. We optimize the separation process with the goal to separate faster (i.e. shorter time of analysis) the eight APIs in a single experimental run, without incurring any loss of product quality and by guaranteeing a sufficient resolution, R_s:

$$R_s = 2\frac{\hat{t}_{R,i+1} - \hat{t}_{R,i}}{\hat{w}_i + \hat{w}_{i+1}} \tag{11}$$

where $\hat{w}$ is the simulated peak width at the base. Resolution is indeed a prime concern in optimization and to achieve a adequate resolution there must be baseline separation between two adjacent peaks (i.e. $R_s \geq 1.5$).

To execute gradient method optimization, our optimization program executes a simple minimization problem and implements the *fminsearchcon()* routine (D'Errico, 2022). Although this is a simple method, this optimization procedure represents a good alternative to the trial-and-error method.

Results and discussion

Experimental setup and experimental conditions

The retention behaviour of eight small APIs (i.e. atenolol, indoprofen, naproxen, pindolol, propanolol, retinoic acid, terfenadine, warfarin) has been investigated experimentally under Reversed-Phase High-Performance Liquid Chromatography (RP-HPLC). The experimental data have been obtained for a range of experimental conditions and in gradient elution mode. The concentration of organic modifier changed linearly from 5% to 95% for different gradient times, and the flow rate was 1.2 mL/min. The composition of the mobile phase consisted of water with +0.1% trifluoroacetic acid and acetonitrile +0.1% trifluoroacetic acid, respectively. The mobile phase pH was equal to 1.35. The variables of the full factorial design of the experiment (DoE) were gradient time and temperature. The factor levels for the gradient time were 10, 20 and 30 min, whereas the factor levels for the temperature were 25, 40 and 55°C. Thus, a total of 9 different operative conditions have been investigated, and 72 retention times have been experimentally acquired. All HPLC experiments have been performed on an Xbridge BEH C18 stationary phase column from Agilent Technologies operated at Eli Lilly. The column has particle diameter of 2.5 μm and dimensions of 3.0×100 mm. The peak profiles were detected at 220 nm via ultraviolet-visible (UV-Vis) spectroscopy integrated with a mass spectrometer (MS). The acquired set of experimental data is used to obtain the unknown adsorption isotherm parameters by fitting the mechanistic model.

Estimation of adsorption isotherm parameters

To estimate the unknown adsorption isotherm parameters, we compare the results of the chromatographic model with the available experimental data. Table 1 lists all the estimated parameters, together with the values of the objective function corresponding to the evaluated parameter sets. The significant error in estimating the retention behaviour of atenolol (i.e. 26.56%) probably arises because under certain conditions of temperature and gradient steepness atenolol behaves as an unretained component. Except for atenolol, the mean relative error done in predicting the retention behaviour of each molecule is lower than 10%.

Table 1: Estimated adsorption isotherm parameters and objective function.

Molecule	S'_s	C'_t	Obj.(%)
Atenolol	38.35	7.76	26.56
Pindolol	39.04	8.92	9.91
Propanolol	43.46	11.42	8.04
Indoprofen	38.56	11.34	7.90
Naproxen	34.57	11.32	7.73
Warfarin	34.24	11.58	7.63
Terfenadine	27.67	10.90	6.49
Retinoic Acid	23.87	11.99	6.00

Fig. 1 shows the predicted vs. the experimentally acquired retention times for the 8 molecules under the 9 operative conditions, while the boxplot to the right of Fig. 1 represents the distribution of the predictive relative errors (in %). It can be seen that the median relative error is less than 8%. Although we believe that the predictive performance of the model can be improved, this is in line with most of the predictive models reported in the literature. However, unlike other computational strategies, which allow predicting the retention only in a limited number of HPLC settings (for instance Quantitative-Structure-Retention-Relationships), our model is based on mechanistic models and, once the isotherm parameters are known, it may be used to assess a wide range of operative conditions.

Fig. 2 shows a comparison between the experimental chromatograms after baseline correction (top) and the simulated chromatogram (bottom) obtained at $T = 40°\text{C}$ and $t_G = 20$ min. We can see that the model predicts the retention times accurately; however, because of numerical diffusion, peak widths are overestimated. To simulate the peak shape more precisely, one should use a finer mesh, but mesh refinement leads to high computational requirements.

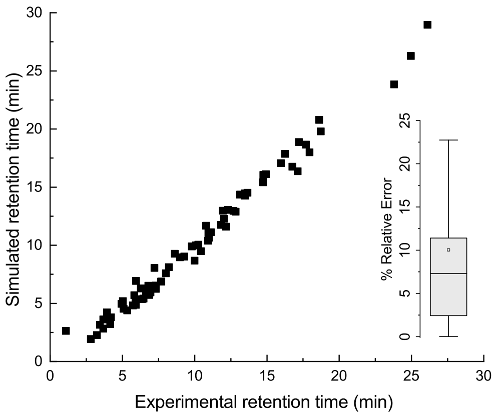

Figure 1: Predicted vs. Experimental retention times and box plot for the distribution of the percentage relative errors. The boxplot represents the median, (25%-75%) interquartile range, and mean (dot) but it excludes outliers.

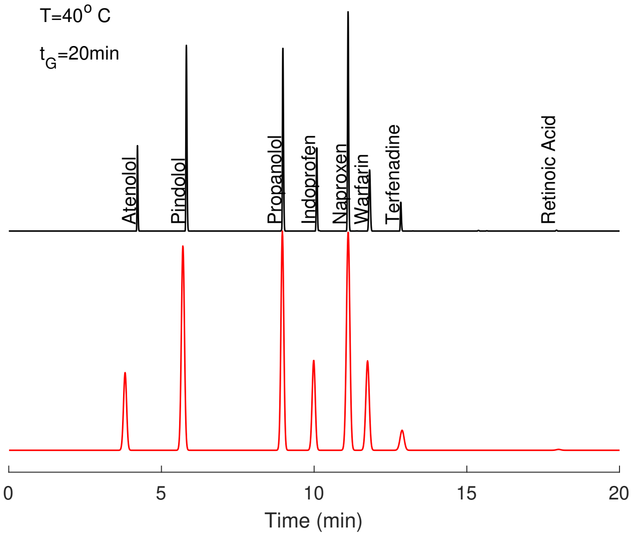

Figure 2: Experimental chromatogram after baseline correction (top) and simulated chromatogram (bottom) obtained at $T = 40°\text{C}$ and $t_G = 20$ min. The initial and final volume fractions of organic modifier are 0.05 and 0.95, respectively, while the flow rate is 1.2mL/min.

Sensitivity analysis

The results obtained from parameter estimation constitute the base for further improvement. Among the other functionalities, the model can be used to perform parameter sensitivity analysis, i.e. to evaluate the impact of packing structure and operative conditions on the retention behaviour of the analytes. As an example, Fig. 3 shows the sensitivity analysis of naproxen obtained by varying gradient time, temperature, flow rate and total porosity by ±10%. Clearly, the factor that most influences the retention behaviour of naproxen is gradient time, followed by temperature. Since similar results have been obtained for all the components under investigation, a good strategy for substantially shortening the time of analysis would be lowering the gradient time and increasing the oven temperature.

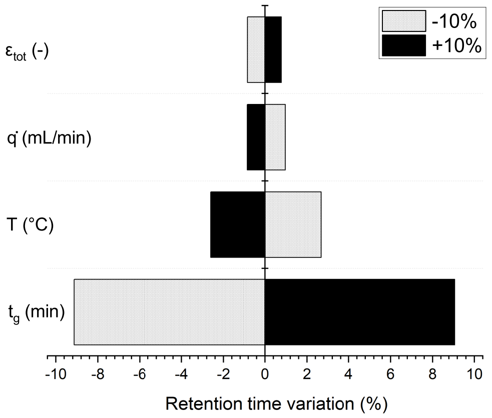

Figure 3: Percentage variation of retention time of naproxen for ±10% relative variations in selected process parameters.

Gradient method optimization

With the previously estimated isotherm parameters for the 8 molecules, the model can also be used to perform computer-based optimization. Since the gradient strongly influences chromatographic separation, in this section we report gradient method optimization. Initial gradient composition and gradient time (i.e. the time required to reach 100% of pure organic modifier) are chosen as the two variables to optimize, while temperature, flow rate and final mobile phase composition are set equal to 40°C, 1.2 mL/min, and 100% respectively. If the optimization objective is to achieve faster separation of the mixture into its components with a sufficient resolution, the results reported in Fig. 4 are obtained. A comparison with Fig. 2 reveals that, by simply changing the solvent gradient, the analysis time reduces by a factor of 6. However, a higher fraction of organic modifier would be required to decrease even more the retention of the last eluting molecule (i.e. retinoic acid). Thus, we also perform optimization of a two step-wise gradient profile. Since each gradient node (step) of a multi-step gradient profile corresponds to two additional variables (Fekete and Molnár, 2018), the optimization becomes a four variables optimization. As Fig. 5 shows, the implementation of the optimal segmented gradient allows the analysis time to be further reduced by 0.4 min. The simple optimization strategy reported here (i.e. it is a simple minimization problem) may represent a first scouting procedure, cheaper and faster than expensive and time-consuming trial-and-error experiments. More sophisticated techniques such as the Monte Carlo methods and Genetic Algorithms, which are required to simultaneously optimize multi-linear gradient separation, temperature, flow rate, column geometry, etc., will be implemented in a future version of this work.

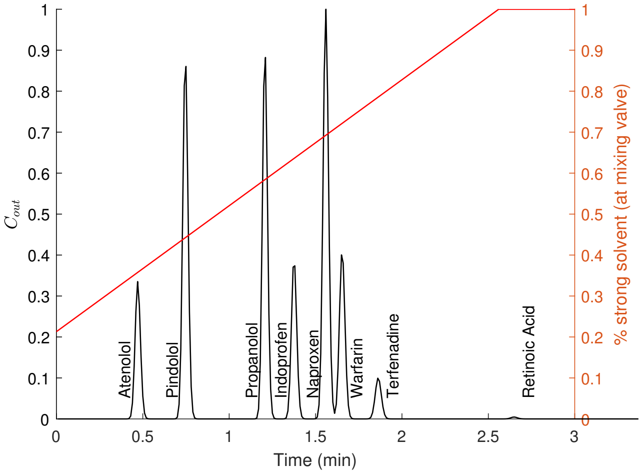

Figure 4: Simulated chromatogram generated from the optimized linear gradient considering 40°C and 1.2 mL/min flow rate.

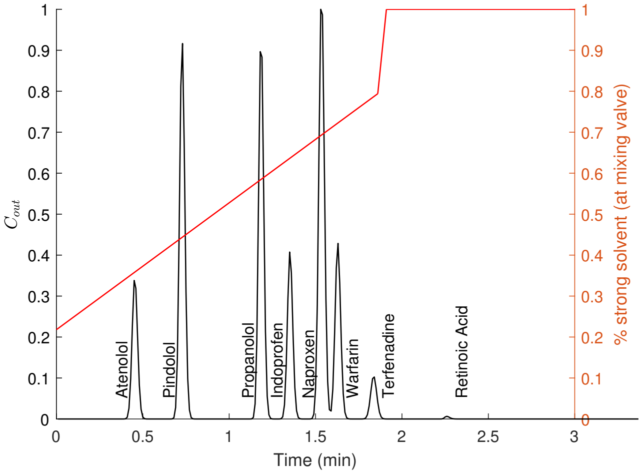

Figure 5: Simulated chromatogram generated from the optimized segmented gradient considering 40°C and 1.2 mL/min flow rate.

Conclusions

In this paper, the first step towards a digital twin of an HPLC facility has been presented. The digital model, which has been developed in gPROMS ModelBuilder and coupled with MATLAB, has been implemented for parameter estimation and gradient method optimization. We have shown that the HPLC model can predict with reasonable accuracy the retention time of different molecules under different conditions of gradient time and temperature by analysing a few experimental results designed with DoE. Furthermore, the model can be used for accelerating HPLC in-silico model development. As an example, we presented here gradient method optimization and found that by optimizing only initial gradient composition and gradient time, we were able to speed up the analysis sixfold.

In the future, we aim to integrate this model into a digital twin architecture to support offline decisions and/or enable real-time optimizations. All molecules loaded within the sample will be detected and recorded continuously upstream of the chromatographic system, and the HPLC will be operated at optimal temperature and optimal gradient, as suggested by the digital twin. As a result, the time for analysis will be reduced, the purity of the sample will be improved, and the process sustainability will be increased, in accordance with the quality-by-design (QbD) paradigm of pharmaceutical development.

Note that the proposed HPLC digital twin must be re-calibrated for different stationary phases and different molecules (i.e., for unknown isotherm parameters). In the future, we aim to predict the unknown isotherm parameters of novel molecules in the same HPLC system through machine learning techniques by using, as input, their molecular structure (i.e. molecular descriptors and fingerprints), and then across new HPLC systems through transfer learning methodologies.

Acknowledgments

The authors wish to acknowledge the financial support given to this research project by the Engineering and Physical Sciences Research Council (EPSRC), grant code EP/T005556/1.

References

Besenhard, M. O., A. Tsatse, L. Mazzei, and E. Sorensen (2021). Recent advances in modelling and control of liquid chromatography. *Current Opinion in Chemical Engineering 32*, 100685.

Chen, Y., O. Yang, C. Sampat, P. Bhalode, R. Ramachandran, and M. Ierapetritou (2020, sep). Digital twins in pharmaceutical and biopharmaceutical manufacturing: A literature review. *Processes 8*(9).

Danckwerts, P. V. (1953). Continuous flow systems. Distribution of residence times. *Chemical Engineering Science 2*, 1–13.

D'Errico, J. (2022). fminsearchbnd, fminsearchcon. url: https://www.mathworks.com/matlabcentral/fileexchange/8277-fminsearchbnd-fminsearchcon.

Fekete, S. and I. Molnár (2018). *Software-assisted method development in high performance liquid chromatography*. World Scientific.

Guiochon, G., A. Felinger, D. G. Shirazi, and A. M. Katti (2006). *Fundamentals of Preparative Chomatography*.

Héberger, K. (2007). Quantitative structure–(chromatographic) retention relationships. *Journal of chromatography A 1158*(1-2), 273–305.

Jandera, P., K. Krupczyńska, K. Vyňuchalová, and B. Buszewski (2010). Combined effects of mobile phase composition and temperature on the retention of homologous and polar test compounds on polydentate C8 column. *Journal of Chromatography A 1217*(39), 6052–6060.

Kritzinger, W., M. Karner, G. Traar, J. Henjes, and W. Sihn (2018). Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine 51*(11), 1016–1022.

Melander, W., D. E. Campbell, and C. Horváth (1978). Enthalpy—entropy compensation in reversed-phase chromatography. *Journal of Chromatography A 158*(C), 215–225.

Rosen, R., G. Von Wichert, G. Lo, and K. D. Bettenhausen (2015). About the importance of autonomy and digital twins for the future of manufacturing. *Ifac-papersonline 48*(3), 567–572.

Samuelsson, J., L. Edström, P. Forssén, and T. Fornstedt (2010). Injection profiles in liquid chromatography. I. A fundamental investigation. *Journal of Chromatography A 1217*(26), 4306–4312.

Snyder, L. R., J. W. Dolan, and J. R. Gant (1979). Gradient elution in high-performance liquid chromatography : I. Theoretical basis for reversed-phase systems. *Journal of Chromatography A 165*(1), 3–30.

Wang, A., L. C. Tan, and P. W. Carr (1999). Global linear solvation energy relationships for retention prediction in reversed-phase liquid chromatography. *Journal of chromatography A 848*(1-2), 21–37.