Prediction of the performance of pre-packed purification columns through machine learning

Citation for published version:
Jiang, Q, Seth, S, Scharl, T, Schroeder, T, Jungbauer, A & Dimartino, S 2022, 'Prediction of the performance of pre-packed purification columns through machine learning', *Journal of Separation Science*, vol. 45, no. 8, pp. 1445-1457. https://doi.org/10.1002/jssc.202100864

Digital Object Identifier (DOI):
10.1002/jssc.202100864

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Separation Science

Prediction of the performance of pre-packed purification columns through machine learning

Journal:	*Journal of Separation Science*
Manuscript ID	jssc.202100864.R1
Wiley - Manuscript type:	Original Paper
Date Submitted by the Author:	31-Jan-2022
Complete List of Authors:	Jiang, Qihao; The University of Edinburgh, School of Engineering Sesh, Sohan; The University of Edinburgh, The School of Informatics Scharl-Hirsch, Theresa; University of Natural Resources and Life Sciences Vienna, Institute of Statistics; Austrian Centre of Industrial Biotechnology Schroeder, Tim Jungbauer, Alois; University of Natural Resources and Life Sciences Vienna, Institute of Bioprocess Science and Engineering Dimartino, Simone; The University of Edinburgh, The School of Engineering
Keywords:	Asymmetry, HETP, Machine Learning, Pre packed columns, Porous media

Prediction of the performance of pre-packed purification columns through machine learning

Qihao Jiang[1], Sohan Seth[2], Theresa Scharl[3,4], Tim Schroeder[5], Alois Jungbauer[3,6], Simone Dimartino[1,*]

[1]Institute of Bioengineering, The School of Engineering, The University of Edinburgh, Edinburgh, EH9 3DW, UK

[2]The School of Informatics, The University of Edinburgh, Edinburgh, EH9 3DW, UK

[3]Austrian Centre of Industrial Biotechnology, Vienna, Austria

[4]Institute of Statistics, University of Natural Resources and Life Sciences Vienna, Vienna, Austria

[5]Atoll, Weingarten, Germany

[6]Department of Biotechnology, University of Natural Resources and Life Sciences, Vienna, Austria

*Corresponding author: Dr Simone Dimartino, Faraday Building, Colin McLaurin Road, The King's Buildings, Edinburgh EH9 3DW, UK, Phone: +44 131 6507305, Email: simone.dimartino@ed.ac.uk

Keywords:

Asymmetry, Plate Height, Machine learning, Pre packed columns, Porous media

Abstract

Pre-packed columns have been increasingly used in process development and biomanufacturing thanks to their ease of use and consistency. Traditionally, packing quality is predicted through rate models, which require extensive calibration efforts through independent experiments to determine relevant mass transfer and kinetic rate constants. Here we propose machine learning as a complementary predictive tool for column performance. A machine learning algorithm, extreme gradient boosting, was applied to a large data set of packing quality (plate height and asymmetry) for pre-packed columns as a function of quantitative parameters (column length, column diameter, particle size) and qualitative attributes (backbone and functional mode). The machine learning model offered excellent predictive capabilities for the plate height and the asymmetry (90% and 93%, respectively), with packing quality strongly influenced by backbone (~70% relative importance) and functional mode (~15% relative importance), well above all other quantitative column parameters. The results highlight the ability of machine learning to provide reliable predictions of column performance from simple, generic parameters, including strategic qualitative parameters such as backbone and functionality, usually excluded from quantitative considerations. Our results will guide further efforts in column optimization, e.g. by focusing on improvements of backbone and functional mode to obtain optimised packings.

1. Introduction

Pre-packed chromatography columns are widely employed in process development and biomanufacturing. Their biggest advantage is to take away the burden of costly and time consuming packing procedures and associated validation protocols, ultimately ensuring a consistent product [1–4] . The production of pre-packed columns should be simple, cost-effective, and robust over the long term (decades) to ensure consistent quality of columns.

The performance of pre-packed columns is assured by the manufacturer before the sale, with packing quality measured in terms of the height equivalent to a theoretical plate (HETP) and asymmetry. Both parameters are calculated from the response of the column following a pulse injection of a non-binding tracer, i.e. residence time distribution (RTD) experiments. The HETP corresponds to the column length over the number of theoretical plates (N), with efficient columns characterized by relatively large N and small HETP values. According to the general rate model, the RTD response of a "well-packed" column is a symmetrical Gaussian peak. To better assess packing quality, RTD experiments are usually run under conditions for which hydrodynamic dispersion is the dominant contribution to mass transfer (negligible intraparticle mass transfer, no adsorption). Under these conditions (reduced velocity of about 1 to 10), the HETP The minimum HETP value theoretically depends only on the properties of the tracer, the velocity of the mobile phase, and the size of the chromatographic particles [5]. However, the general rate model is unable to capture how the HETP is influenced by key factors of practical relevance such as column size (column diameter and length) or ease of packing across different chromatographic resins [6]. For example, Scharl et al. [7] qualitatively discussed the importance of material backbone on packing quality of a range of pre-packed columns. Deviations from symmetrical peaks are often observed in practice, with peak fronting or tailing associated with a number of non-idealities such as wall effects, inhomogeneous packing, inhomogeneous distribution of the solute over the bed at the column inlet/distributor and/or at the outlet/collector, and dispersion in the extra column volumes [8–12]. Such deviations are measured through the asymmetry, an empirical parameter used to quantify the degree of peak skewness and employed to assess packing quality in tandem to the HETP [13].

Mathematical models to predict column performance and chromatographic processes, including the general rate model, are generally based on first principles. In particular, they include details of mass transfer phenomena and binding kinetics to describe peak profiles and breakthrough curves [14,15]. While the predictive power of these models is often excellent, they require extensive calibration efforts through independent experiments, e.g. to determine key model parameters such as mass transfer and kinetic coefficients [16,17]. Flow non-idealities such as wall effects and distribution/collection of the fluid at the column inlet/outlet also require independent experiments for them to be accounted for in the models. These additional experiments are specific to the chromatographic system (external column volumes) and column (diameter, length) employed, therefore cannot be extrapolated to different systems or different columns. Finally, such models based on first principles do not take into account qualitative variables such as resin backbone and functional chemistry by design.

Machine learning (ML) could represent an alternative modelling approach to analyze and predict column performance. The main advantage of ML is the ability to extract information from large data sets using no or only minimum assumptions, eventually determining generalizable predictive patterns between multiple inputs (including quantitative, qualitative and categorical parameters), and the output variables [18,19]. A number of algorithms, e.g. support vector machine, decision tree, gradient boosting, and deep neural networks have been developed over the years, and have proved their ability in dealing with complex data problems in a practical manner [20,21]. ML has been applied to chromatography systems, with many successful applications e.g. in peak observation [22–24], retention modelling [25–28], process optimization[29–31], and real-time process monitoring [32,33]. The main challenge associated to the application of ML is the availability of very large experimental data sets for the ML algorithm to draw meaningful correlations.

In this work, we consider a large data set of around 25,000 quality assurance experiments of pre-packed columns manufactured and tested under standardized conditions for a period of over 10 years [7]. We first examine the time series of the data set using correlation and autocorrelation analysis to ensure the data are self-consistent and time-independent. We then employ ML methods to find a correlation between column performance (measured in terms of HETP and asymmetry) and qualitative as well as qualitative column variables, namely resin backbone, functionalization chemistry, column size (length and diameter) and particle size. The results are finally commented in relation to the main key variables affecting column performance.

2. Materials and Methods

2.1 Experimental Data Set

The data set employed in this work is a subset of that previously employed by Scharl et al consisting of 24,951 quality control runs of pre-packed small-scale columns over a period of about 10 years [7]. The data contain relevant column parameters (i.e. column length and diameter, particle size, backbone material, functional mode, and date of testing) together with reduced HETP (h) and asymmetry (A_s). Column diameter and length ranged between 5 – 11.3 mm and 10 – 100 mm, respectively, while particle diameter varied between 15 to 400 μm. 2232 experimental runs (approximately 10%) were removed from the original data set as they lacked one or more column parameter inputs, reducing the data set to a total of 22,359 tests. Columns with same attributes were manufactured and tested more than once over the ten year monitored, with some popular types examined hundreds of times (e.g. see table 1 in SI). All experiments having same set of input features were treated as a single entry, with h and A_S averaged over the available runs for that column type. This step was necessary to prevent data leakage in the ML model, i.e. the use of same column type in both the training and testing data sets (see 2.3), as well as to prevent overfitting of the most popular column types over the ones infrequently produced. The standard error for h and A_S was always lower than 10%, indicating that the average h and A_S are representative output indicators of column performance for any given column type. After the averaging process, the data set contained a total of 546 independent runs.

All columns used to generate the data set were packed by slurry packing under vibration following a standardized procedure developed by the packing company (Atoll, now Repligen). The packing quality of the columns was evaluated using a standardised experimental set up and experimental protocol as reported in Scharl et al [7]. Briefly, the response of the column following an acetone or sodium nitrate injection was measured, and the resulting chromatographic peak analysed to extract h and A_S. This simple experiment allowed to isolate the contribution to band broadening associated with hydrodynamic dispersion (which in turn depends on packing quality and extra column dispersion) as the tracers employed are both non-retained (i.e. zero retention factor), with practically same diffusion coefficients (1.2×10^{-5} and 1.3×10^{-5} cm^2/s for acetone [34] and sodium nitrate [35] , respectively), and tested under reduced velocities comprised between 1 and 20 for which the minimum HETP is obtained [14].

2.2 Extreme Gradient Boosting

Extreme gradient boosting (XGBoost) is a scalable ML system for tree boosting [36]. XGBoost is a decision-tree-based ensemble learning method [37] that provides a systematic solution to a given problem by combining the predictive power of several different or same ML algorithms. The algorithm used in XGBoost is the Classification and Regression Tree (CART) [38] which employs a binary tree that can be constantly segmented by data features, thus enabling dynamic growth of the tree. The characteristics of the input data will eventually fall into the leaf nodes n the tree, where each leaf node corresponds to a specific score, and the sum of the scores in all the leaf nodes computes the final prediction value of a certain feature, e.g., h or A_s.

Essential details of the mathematical formulation of the XGBoost model are presented in the following, with additional details in the SI. For a given data set with n examples and m features $\mathcal{D} = \{(x_i,y_i)\}(|\mathcal{D}| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$, the tree ensemble model uses K additive functions to predict the output.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}\#(1)$$

where $\mathcal{F} = \{f(x) = w_{q(x)}\}(q:\mathbb{R}^m \to T, w \in \mathbb{R}^T)$ is the space of the regression trees. The q represents the structure of each tree that maps an example to the corresponding leaf index. T is the number of leaves in the tree. Each f_k corresponds to an independent tree structure q and tree weights w. More mathematical details can be found from the original XGBoost paper [36].

The regularized objective function defined for XGBoost, $\mathcal{L}$, can be written as:

$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)\#(2)$$

$$\Omega(f) = \gamma \cdot T + \frac{1}{2}\lambda\|w\|^2\#(3)$$

Here l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i. The second Ω term prevents unnecessary large trees by penalizing the complexity of the model, in turn avoiding overfitting. The additional regularization term $\frac{1}{2}\lambda\|w\|^2$ helps smooth the final learnt weights. The shrinkage parameter γ is an additional design to prevent over-fitting. The γ is utilized to multiply the score of each leaf node by a reduction weight during the iteration, which ensures that the influence of each tree is not too large, leaving more space for the tress generated later to optimize.

XGBoost is also used to determine the relative importance of the input features. The definition of the relative importance is followed by the study of H. Friedman [39]. For a tree model whose number of terminal nodes is J, the relative importance of a given input feature, I, is calculated by the sum of the corresponding empirical improvements, i^2, with t referring to a non-terminal node and v_t acting as splitting variable for that node. The i^2 term is determined from the two sub-region R_l and R_r, where $\bar{y}_l$ and $\bar{y}_r$ are the response means, respectively, and w_l and w_r are the corresponding sums of the weights. In Python, the contribution of each input features can be automatically transferred into the percentage version.

$$I_j^2(T) = \sum_{t=1}^{J-1} i_t^2(v_t = j)\#(4)$$

$$i^2(R_l,R_r) = \frac{w_l w_r}{w_l + w_r}(\bar{y}_l - \bar{y}_r)^2\#(5)$$

2.3 Data Pre-Processing and Model Implementation

Functional modes and backbone are two categorical features which cannot be operated by many ML algorithms directly. One-hot encoding was applied to transfer them into numerical values [40], with each feature normalized between 0 and 1. All other numerical parameters were also normalized between 0 and 1 before input into the ML model as most ML algorithms perform better or converge faster with features on relatively similar scale [41].

An XGBoost regression model was created in Python 3.6 combining i) GridSearchCV (ten-folds) to select and determine the model's hyper-parameters (e.g. learning rate, maximum tree depth, and minimum child weight) [42] and ii) XGBRegressor as the main package to process our data set [43]. The whole data set was then separated randomly into a training set (66.7%) and testing set (33.3%), with the training set utilized for training the ML model and the testing set used for inspecting the final model accuracy. Mean absolute error (MAE) [44] was used as the evaluation metric during model training. The final prediction precision of the model is reported by the mean absolute percentage error (MAPE) between the prediction results and the testing data set. The overall model prediction capability remained the same when changing initial seeding to randomly generate different training and testing data sets.

3. Results and Discussion

The main goal of this study was the identification of a general relationship between column parameters (column length, column diameter, particle diameter, functional mode, backbone material) and chromatographic performance (reduced HETP, $h$and peak asymmetry, A_s) using ML algorithms as an alternative to classical rate models for chromatography. Classical rate models are derived from first principles and thus tend to be the preferred choice when it comes to the modelling of chromatographic separations. However, some of the parameters entering rate models often are either determined through empirical expressions (e.g. the Wilson Geankopolis correlation for the estimation of the mass transfer coefficient [45]) or simply adjusted to best-fit experimental results (e.g. diffusion or dispersion coefficients [46]).

The introduction of a certain degree of empiricism in physical models is necessary to capture important elements of the model hard to describe in mathematical terms. For example, the three dimensional configuration of chromatographic beds deviates from the theoretical close random packing limit [47], with the resulting bed arrangement strongly influenced by attributes linked to the material and column properties (e.g. Young modulus, friction factor, wall roughness) as well as the packing procedure itself [47,48] . For example, Knox demonstrated that hydrodynamic dispersion in columns packed with smooth non-porous glass beads is smaller than those measured in columns packed with porous glass [49]. Knox explained this result in terms of bed homogeneity, and speculated that smooth glass particles are able to form relatively regular packings, while porous glass particles are affected by greater interparticle friction forces, in turn resulting in particle bridging and the formation of pockets where local mixing occurs. These insights were demonstrated experimentally by Patel and coauthors [50], who confirmed that the A term in the van Deemter equation is primarily associated with radial heterogeneities in the bed. On the opposite front, Malkin et al. showed that submicrometer silica particles tend to pack close to the limit of a face centered cubic arrangement [51], resulting in reduced plate heights below 1. Khirevich et al. also reported that the local microscopic disorder in packings was highly correlated with eddy dispersion, directly affecting column performance [52]. Along the same line, Gritti et al [53] reported the outstanding performance of columns packed with core-shell particles, partly attributing these results to the propensity that these particles have to create homogeneous beds. More recent studies on 3D printed ordered beds further confirm the advantages of perfectly ordered packing, with simulated reduced plate heights below 0.1 for specific arrangements (e.g. octahedral particles in simple cubic configuration) of non-porous stationary phases under non-retained conditions [54].

The concept of “goodness of packing” as proposed by Knox is strongly correlated to the A term of the van Deemter equation [55], with lower A values associated to lower reduced plate heights and hence higher chromatographic efficiency. According to the general rate model for chromatography, the A term can be expressed as [16]:

$$A = 2\chi d_p\#(6)$$

or in dimensionless terms:

$$a = 2\chi\#(7)$$

where d_p is the average particle diameter and χ is the dispersivity of the stationary phase. The dispersivity is a characteristic determined by the hydrodynamics in the column, in turn defined by type of particles and their packing. For a given column, the dispersivity can be determined through estimation of the plate height under conditions suppressing both axial diffusion (i.e. large velocity, negligible B term) and mass transfer and kinetic resistances (i.e injection of a small, fast diffusing non-adsorbing tracer, negligible C-term) for which the van Deemter equation reduces to:

$$h = a = 2\chi\#(8)$$

While this equation represents a relatively rapid method to assess the hydrodynamic properties of a given column, lack of correlations for the estimation of the dispersivity coefficient represents a limitation to predict band broadening due to axial dispersion. In particular, there exist no quantitative method to assess how the dispersivity depends on different column properties such as:

- backbone material and functional mode, closely related to the propensity of the particles to generate regular packing;
- column and particle diameters, i.e. the column to particle ratio, in turn determining the importance of non-homogeneities close to the column wall with respect to the rest of the column volume;
- column length and column diameter, which are associated to both bed compressibility [56], as well as defining the relative influence of extra-

column dispersion effects, e.g. due to non-uniformity of the velocity profile resulting from non-idealities in the extra-column volumes.

Fronting or tailing deviations from the ideal symmetrical peak are often observed in chromatographic practice, negatively impacting the separation performance. Such deviation is often quantified through the asymmetry factor, A_s, defined as the ratio between the width of the tailing end and of the peak front at 10% peak height [57,58]. Large asymmetry factors are associated with the heterogeneity of the column packing [59,60], making A_s another excellent descriptor for "goodness of packing". However, search for a quantitative relationship between asymmetry and column parameters has been elusive so far. In this context, ML is an excellent tool to extract poorly understood links between variables such as the column input parameters and the asymmetry factor.

The data set of pre-packed column performance offers an opportunity to quantitatively analyse the dependence of the dispersivity on a range of qualitative and quantitative column attributes. The two performance parameters, h and A_s, are measured from the experimental response of an injection of a small non-retained tracer (acetone or sodium nitrate). Same experimental and data analysis methods were used to generate the entire data set [7]. Only resins intended to separate proteins or other larger biomolecules were tested, ensuring much larger pores than that of the tracers. Such conditions ensure only the hydrodynamic dispersion is captured in the experiments and that the Van Deemter equation can be simplified into Eq. 8.

In short, we propose here to employ ML as a powerful alternative to traditional chromatographic models to investigate a correlation between the different column input parameters and the output performance parameters. ML is especially valuable in this context given the complexity of the problem described and the qualitative nature of some of the relevant variables such as column backbone and functional mode. ML is also able to suggest the relative importance of the different inputs with respect to the outputs, thus helping the identification of the key descriptors for the performance parameters.

3.1 Time Series of Reduced Plate Height and Asymmetry

Column performance can change over time due to variations in the manufacturing line, e.g. improvement in the packing procedures, change of suppliers of raw materials, and ageing of the production line. Scharl et al. qualitatively observed that the plate height of the prepacked columns tested was stable over ten years [7]. However, any interdependence between h and A_s with time needs to be either identified or excluded in quantitative terms to avoid any input bias to the ML model. In other words, it is first necessary to determine if time represents an input variable to the ML model, as well as if sampling and testing of the columns changed significantly over time. Autocorrelation and partial autocorrelation analysis was employed onto the data set to address these two aims, respectively. In particular, the autocorrelation function (acf) aims to detect cross-similarities of a signal with itself at a different time (time lag) [61]. In this context, acf helps detect changes in the manufacturing line and in the quality assurance protocols employed over time. The partial autocorrelation function (pacf) instead aims to identify the possibility of confounding variables which are correlated to both variables [62]. In this instance, pacf aims to identify a correlation between time and performance parameters, in turn suggesting if a specific pattern of column types was manufactured over time. Additional details on acf and pacf are also provided in the SI.

The h and A_s time series were first resampled by averaging the data set in day intervals, irrespective of the other column parameters. Other than reducing noise, resampling is customary when autocorrelation analysis is executed over large time periods [58,61].

Figure 1 shows the time series of the two performance parameters, h and A_s. Over the 10 year time considered, the h values varied between about 7.8 and 2.2, with an average of around 4.5. Variability reduced significantly from 2011 onwards, with a slight decrease of plate height in 2012–2013. The asymmetry ranged between about 2 and 0.8, with average of 1.1. Similar to plate height, the scatter in the asymmetry over the first five years is larger than after 2011. According to Scharl et al. [7], industrial quality assurance tests require a column to have h comprised should be smaller than 5 in industry, while the acceptable range for A_s is between 0.8 and 1.6 [7]. The observed variability is a natural consequence of industrial manufacturing, yet the columns produced were within specifications in terms of both h and A_S.

Figure 2 shows the results from the autocorrelation and partial autocorrelation analysis on h and A_s using lag time of days up to one year. Other lag times were also examined (i.e. weekly, monthly as well as over 2, 3 months) with no significant difference. For both h and A_s, almost all of the acf and pacf coefficients lie within the 95% confidence interval. The low acf demonstrates that the dataset does not have a specific pattern with time, quantitatively confirming that the manufacturing line was stable over the ten year period here investigated [63]. In addition, low pacf rules out the existence of confounding variables such as certain patterns in terms of column sampling and testing over time. In other words, pacf analysis confirms that column manufacture was unbiased, excluding the possibility that a certain column type (e.g. having specific size and packed with a specific particle) was manufactured predominantly over other columns over time. Overall, acf and pacf demonstrate that all performance tests were time independent, making the data set solely dependent on the five input parameters of particle size, column diameter, column length, column backbone, and functional mode.

3.2 Influence of column parameters on packing quality

XGBoost was utilized to assess the influence of the column parameters (i.e. the inputs to ML algorithm: particle size, column length, column diameter, functional mode, and resin backbone) on packing quality (i.e. ML outputs of h and A_s). Other ML algorithms such as artificial neural networks and decision-tree were also employed in a preliminary model assessment (refer to SI for additional information on ML models). XGBoost consistently provided the highest predictive precision, mainly due to its regularization and shrinkage terms (Eqs. 2 and 3) being capable of curbing over-fitting, the main cause of poor prediction.

Figure 3 summarizes the results obtained with the XGBoost model to predict the experimental data. In particular, Figure 3a and 3b compare the predicted h and A_s, respectively, against the observed data of the testing data set. The predictions are in good agreement with the experimental results, where the mean absolute percentage error (MAPE) of predicted results to the observed values are 10% for h and 7% for A_s, with a few outliers in the 40% $\sim$ 50% range. These acceptable errors confirm that the XGBoost model can be applied to this problem with good prediction accuracy. Figure 3c and 3d report the contribution importance, J (Eqs. 4 and 5), of the various input parameters to predict the model outputs. Interestingly, column backbone resulted as the most important descriptor of packing quality, accounting for 68.4% and 77.0% for the prediction of h and A_s, respectively. Functional mode was the second most significant descriptor for the estimation of packing quality, accounting for about 15% contribution importance, followed in various order by the other parameters (particle size, column diameter and column length). Violin plots were employed to further analyze the correlation between input features and column performance (Figure 4). A violin plot is an extension of a box and whisker plot, clearly recognizable inside the “violins”, decorated with a curve whose width is related to the probability density.

3.2.1 Resin backbone

Resin backbone was the most influential parameter for the prediction of packing quality. The material making up the resin backbone can be either inorganic , synthetic polymer, or natural polymer. The nature of the material employed determines a number of properties such as surface roughness of the particles [64], particle size distribution (linked to the manufacturing method)[65], occurrence of microstructural defects, and other mechanical properties such as Young modulus and density [63,64]. All these factors impact column packing, either directly or indirectly, in turn influencing the homogeneity of the resulting chromatographic bed, i.e. packing quality. Johnson et al. examined a range of resin materials (agarose, cellulose, ceramic) through X-ray computed tomography (CT) and focused ion beam (FIB) [66]. They highlighted clear variations in the chemical, physical and mechanical properties of the different materials. Our analysis with the XGBoost model also confirms that resin characteristics strongly influence chromatographic performance.

Figure 4a and 4b present violin plots of *h* and A_s, respectively, over the eight different backbones tested. It is possible to observe that certain backbones have worse performance than others as measured by both of the two packing quality parameters *h* and A_s. For example, polystyrene-divinilbenzene (PS/DVB), inorganic support (IS) and dextran (DEX) have data widely distributed, with average *h* above 5 and average A_s above 1.2. On the other hand, agarose, cellulose and PVE hydrophilic (PVE) demonstrated consistent results (little data scatter) with average *h* and A_s well below the arbitrary thresholds of 5 and 1.2, respectively. This analysis clearly demonstrates the importance of backbone selection, e.g. during process or method development.

It is worth noting that inorganic support (IS) was relatively popular in the first three years of our data set, while polyvinyl-ether hydrophilic (PVE) matrices were little used at first, becoming more mainstream after 2011. This change in backbone population over time can partly explain the slight decrease of the absolute value of h, as well as the reduced scatter of h and A_s observed from 2011 onwards (Figure 1).

3.2.2 Functional mode

Functional mode was the second most important parameter to predict packing quality. Figure 4c and 4d show the relation between h and A_s over the different functional modes. The influence of the functionalization chemistry on column packing is less intuitive than for chromatographic backbone. Stickel and Fotopoulos [67] reported the difference of the pressure-flow profiles between sepharose and phenyl sepharose, which was associated to the differing hydrophobic and electrostatic character of the resin beads. Electrostatic and hydrophobic interactions might promote local or temporary bonding of two or more particles into clusters, decreasing the degrees of freedom of the slurry, and thus influencing column packing [68]. Also, functionalization procedures can change the mechanical and surface properties of the beads, e.g. as a consequence of the different solvents, chemicals and temperatures employed for ligand immobilization. This in turn influences the packing process [69], ultimately determining packing quality.

The possibility of a correlation between column functionality and backbone was tested both qualitatively (mosaic plot in Figure 5) and statistically by employing the chi-squared test. The size of the mosaic tiles in Figure 5 is proportional to the number of chromatographic columns in the data set having a certain combination of backbone and functional mode. Some of the tiles are predominant over the others, e.g. agarose and methacrylate based materials are employed across affinity, ion exchange and hydrophobic interaction chromatography (AF, AIEC, CIEC, HIC, IMAC, MMC in Figure 5). Such columns are indeed ubiquitous in downstream processing of biopharmaceuticals. Other backbones find use in specific application domains, e.g. dextran is predominantly employed for SEC, and HCIC is purely carried out with cellulosic adsorbents. In addition, a number of combinations of functional mode and backbone are not represented in the data set, indicating some resin materials do not find use for certain chromatographic modalities. A chi-squared test of independence with 63 degrees of freedom, i.e. (8 backbones – 1) x (10 functional modes – 1), and with a sample size of 546 tests indeed showed a significant relationship between the two input variables, $\chi^2(63, N = 546) = 693, p < 0.01$. While a correlation between resin material and functionalisation is apparent, its influence in the ML model was eliminated by averaging all experimental results measured under the same input conditions (see section 2.1), especially important step to prevent same samples being present in both the training and testing set thus overestimating the accuracy.

3.2.3 Column length

The influence of column length on h is presented in Figure 4e. It is possible to observe that the median for h, as well as its propensity to data scatter and relatively large values (h above 10) increase with column length. This observation can be explained by a combination of packing consolidation and wall effects. The former is relevant during column manufacture, i.e. when compression forces transfer through the packing via inter-particle friction as well as friction between particles and the column wall [70]. The uneven stress distribution created between particles in the bulk and at the periphery of the column negatively affect bed consolidation and packing homogeneity. The presence of the wall constrains the resin particles to pack in configurations with higher local porosity in the immediate vicinity of the column wall. The columns investigated in this work were small scale purification columns (column volume about 1 and 10 ml) with relatively large particle diameters (15 to 400 μm) and small column diameters (5 – 11.3 mm). The resulting column diameter to particle diameter ratio was in general around 80, down to 20 for some columns. In this context, Maier et al. [71] reported wall effect on axial dispersion can be observed even for columns with column dimeter to particle diameter ratio greater than 100. Reising et al. [72] and Fabrice Gritti [73] studied the dependence of fluid velocity with radial position, and concluded that the velocity close to the column wall can be up to 2.2 times the bulk velocity, significantly contributing to band broadening and early breakthrough. Flow non-idealities arising from both uneven packing difficulties and wall effects scale with column length, with packing quality and column performance inversely related to it.

The contribution of column length on A_s is reported in Figure 4f. No significant difference can be observed across the data, other than a minor decrease in the median asymmetry with column length. Asymmetry is heavily determined by extra column band broadening, i.e. related to all flow non-idealities present in the extra column volumes such as tubing, fitting, column distributor and collectors, pumps, valves etc. This effect becomes more prominent for smaller columns, as described by Kaltenbrunner et al. [74] who reported extra column volumes accounting for more than 90% band broadening in small columns.

3.2.4 Column diameter

According to ML results the contribution of column diameter to the prediction of h is 5.2%, while it is only 0.9% for A_s (Figure 3c), and no clear relationship can be observed between column diameter and the two performance output parameters (Figure 4g and 4h). All the three column diameters considered in this work fall in the same order of magnitude (5, 8 and 11.3 mm), thus hiding any potential correlation between column diameter and packing quality. Schweiger et al [3] analyzed the band broadening arising from the extra-column and in-column contributions of pre-packed columns with different column diameters, and concluded that an increase in column diameter can lead to an increase in peak width as caused by flow non-idealities in the flow distributor and collector. Experimental data for wider columns is required to identify and eventually quantify any possible relationship between column diameter and column performance.

3.2.5 Particle diameter

The correlation between particle diameter and h is reported in Figure 4i. Accordingly to the reduced form of the van Deemter equation (Eq 8), the magnitude of h is not dependent on particle diameter. ML results indicate that the importance contribution of particle diameter to h is 10.7% (Figure 3c). In Figure 4i the median h slightly drops with particle size, possibly resulting from packing difficulties with smaller particles, as also reported by Scharl et al. [7]. No trend between A_s and particle size could be observed (Figure 4j).

4. Conclusions

Traditional statistical analysis (e.g. autocorrelation analysis, chi square analysis) and machine learning were applied to a large data set (546 different combinations of column features) of packing quality (reduced plate height, h, and asymmetry, A_s) for pre-packed columns manufactured with different column sizes (column length and column diameter) and packed with different resins (backbone, functional mode, and particle diameter) over a ten year period.

Autocorrelation and partial autocorrelation provided a quantitative framework to analyze column quality over time. The results indicate that packing quality was indeed not correlated with time, indicating that column manufacture, sampling and testing was consistent over the ten year period.

The XGBoost represented an excellent ML model to predict column performance, with mean absolute percentage error (MAPE) of 10% and 7% on h and A_s, respectively. According to the ML tool employed, column backbone contributed the most to its predictive capability. In other words, the resin material employed had the most significant impact on column performance. A trend between column length and performance was also observed, with h raising slightly as the length increased, consistent with a larger contribution to band broadening due to wall effects and axial dispersion.

Overall, this work demonstrates the capability of ML to evaluate and predict column performance solely from the knowledge of some basic column characteristics (column length and diameter, particle size, backbone material, functional mode). These results could be employed to extrapolate the expected performance characteristics on new and existing columns types, help set QA protocols for new and existing manufacturing lines for pre-packed chromatography columns, or as a reference benchmark for columns packed traditionally in lab settings, especially for hard to pack columns such as PS-DVB and inorganic supports. The results presented here can guide further efforts in column optimization, e.g. informing potential inefficiencies in the packing process, and suggesting improvements of backbone and functional modes to obtain easy to pack resins prone to form ordered packing arrangements with high chromatographic performance.

More in general, ML provides a quantitative tool to describe complex problems with multiple input features, including categorical features such as resin backbone and functional mode. ML methods can also be employed in other chromatographic areas, e.g. for generating accurate retention models, resolving complex chromatography peaks and for searching column structures with improved performance.

References

[1] Brenac Brochier, V., Chabre, H., Lautrette, A., Ravault, V., Couret, M. N., Didierlaurent, A., Moingeon, P., High throughput screening of mixed-mode sorbents and optimisation using pre-packed lab-scale columns for the purification of the recombinant allergen rBet v 1a. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* 2009, 877, 2420–2427.

[2] Shukla, A. A., Gottschalk, U., Single-use disposable technologies for biopharmaceutical manufacturing. *Trends Biotechnol.* 2013, 31, 147–154.

[3] Schweiger, S., Jungbauer, A., Scalability of pre-packed preparative chromatography columns with different diameters and lengths taking into account extra column effects. *J. Chromatogr. A* 2018, DOI: 10.1016/j.chroma.2018.01.022.

[4] Schweiger, S., Hinterberger, S., Jungbauer, A., Column-to-column packing variation of disposable pre-packed columns for protein chromatography. *J. Chromatogr. A* 2017, 1527, 70–79.

[5] Glueckauf, B. Y. E., Theory of chromatography. 1954, 34–44.

[6] Kaltenbrunner, O., Watler, P., Yamamoto, S., Column Qualification in Process Ion-Exchange Chromatography. Elsevier Masson SAS 2000.

[7] Scharl, T., Jungreuthmayer, C., Dürauer, A., Schweiger, S., Schröder, T., Jungbauer, A., Trend analysis of performance parameters of pre-packed columns for protein chromatography over a time span of ten years. *J. Chromatogr. A* 2016, 1465, 63–70.

[8] Fornstedt, T., Zhong, G., Guiochon, G., Peak tailing and mass transfer kinetics in linear chromatography. 1996, 741.

[9] Fornstedt, T., Zhong, G., Guiochon, G., Peak tailing and slow mass transfer kinetics in nonlinear chromatography. *J. Chromatogr. A* 1996, 742, 55–68.

[10] Wakamatsu, A., Morimoto, K., Shimizu, M., Kudoh, S., A severe peak tailing of phosphate compounds caused by interaction with stainless steel used

for liquid chromatography and electrospray mass spectrometry. *J. Sep. Sci.* 2005, 28, 1823–1830.

[11] Kirkland, J. J., Yau, W. W., Stoklosa, H. J., Dilks, C. H., Sampling and extra-column effects in high-performance liquid chromatography; influence of peak skew on plate count calculations. *J. Chromatogr. Sci.* 1977, DOI: 10.1093/chromsci/15.8.303.

[12] Chapel, S., Heinisch, S., Strategies to circumvent the solvent strength mismatch problem in online comprehensive two-dimensional liquid chromatography. *J. Sep. Sci.* 2022, 45, 7–26.

[13] Mitchell, N. S., Hagel, L., Fernandez, E. J., In situ analysis of protein chromotography and column efficiency using magnetic resonance imaging. *J. Chromatogr. A* 1997, DOI: 10.1016/S0021-9673(97)00457-3.

[14] Carta, G., Jungbauer, A., Protein Chromatography. 2017.

[15] Guiochon, G., Felinger, A., Shirazi, D. G., Fundamentals of Preparative and Nonlinear Chromatography. Elsevier 2006.

[16] Dimartino, S., Boi, C., Sarti, G. C., A validated model for the simulation of protein purification through affinity membrane chromatography. *J. Chromatogr. A* 2011, 1218, 1677–1690.

[17] Dimartino, S., Boi, C., Sarti, G. C., Scale-up of affinity membrane modules: Comparison between lumped and physical models. *J. Mol. Recognit.* 2015, 28, 180–190.

[18] Neal, R. M., Pattern Recognition and Machine Learning. *Technometrics*

2007, DOI: 10.1198/tech.2007.s518.

[19] O'Rourke, J., Toussaint, G. T., Handbook of Discrete and Computational Geometry, Third Edition. 2017.

[20] Rasouli, K., Hsieh, W. W., Cannon, A. J., Daily streamflow forecasting by machine learning methods with weather and climate inputs. *J. Hydrol.* 2012, 414–415, 284–293.

[21] Xu, X., Zhang, Y., Zou, L., Wang, M., Li, A., A gene signature for breast cancer prognosis using support vector machine. *2012 5th Int. Conf. Biomed. Eng. Informatics, BMEI 2012* 2012, 928–931.

[22] Bos, T. S., Knol, W. C., Molenaar, S. R. A., Niezen, L. E., Schoenmakers, P. J., Somsen, G. W., Pirok, B. W. J., Recent applications of chemometrics in one- and two-dimensional chromatography. *J. Sep. Sci.* 2020, 43, 1678–1727.

[23] Risum, A. B., Bro, R., Using deep learning to evaluate peaks in chromatographic data. *Talanta* 2019, 204, 255–260.

[24] Kantz, E. D., Tiwari, S., Watrous, J. D., Cheng, S., Jain, M., Deep neural networks for classification of LC-MS spectral peaks. *Anal. Chem.* 2019, 91, 12407–12413.

[25] Marengo, E., Gianotti, V., Angioi, S., Gennaro, M. C., Optimization by experimental design and artificial neural networks of the ion-interaction reversed-phase liquid chromatographic separation of twenty cosmetic preservatives. 2004, 1029, 57–65.

[26] Hervás, C., Martínez, A. C., Silva, M., Serrano, J. M., Improving the

quantification of highly overlapping chromatographic peaks by using product unit neural networks modeled by an evolutionary algorithm. *J. Chem. Inf. Model.* 2005, 45, 894–903.

[27] Vasiljević, T., Onjia, A., Čokeša, D., Laušević, M., Optimization of artificial neural network for retention modeling in high-performance liquid chromatography. *Talanta* 2004, 64, 785–790.

[28] Kensert, A., Collaerts, G., Efthymiadis, K., Desmet, G., Cabooter, D., Deep Q-learning for the selection of optimal isocratic scouting runs in liquid chromatography. *J. Chromatogr. A* 2021, 1638, 461900.

[29] Ben Hameda, A., Elosta, S., Havel, J., Journal of Chromatography A. 2005.

[30] Wang, G., Briskot, T., Hahn, T., Baumann, P., Hubbuch, J., Estimation of adsorption isotherm and mass transfer parameters in protein chromatography using artificial neural networks. *J. Chromatogr. A* 2017, 1487, 211–217.

[31] Narayanan, H., Seidler, T., Luna, M. F., Sokolov, M., Morbidelli, M., Butté, A., Hybrid Models for the simulation and prediction of chromatographic processes for protein capture. *J. Chromatogr. A* 2021, 1650, 462248.

[32] Narayanan, H., Luna, M. F., Stosch, M., Cruz Bournazou, M. N., Polotti, G., Morbidelli, M., Butté, A., Sokolov, M., Bioprocessing in the Digital Age: The Role of Process Models. *Biotechnol. J.* 2020, 15, 1900172.

[33] Narayanan, H., Behle, L., Luna, M. F., Sokolov, M., Guillén-Gosálbez, G., Morbidelli, M., Butté, A., Hybrid-EKF: Hybrid model coupled with extended

Kalman filter for real-time monitoring and control of mammalian cell culture. *Biotechnol. Bioeng.* 2020, 117, 2703–2714.

[34] Cussler, E. L., Cussler, E. L., Diffusion: Mass Transfer in Fluid Systems. Cambridge university press 2009.

[35] Yeh, H. S., Wills, G. B., Diffusion coefficient of sodium nitrate in aqueous solution at 25.deg. as a function of concentration from 0.1 to 1.0M. *J. Chem. Eng. Data* 1970, 15, 187–189.

[36] Chen, T., Guestrin, C., XGBoost: A scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 2016, 13-17-Augu, 785–794.

[37] Freund, Yoav, and L. M., The alternating decision tree learning algorithm. *Int. Conf. Mach. Learn.* 1999, DOI: 10.1093/jxb/ern164.

[38] Lewis, R. J., Ph, D., Street, W. C., An Introduction to Classification and Regression Tree (CART) Analysis. *2000 Annu. Meet. Soc. Acad. Emerg. Med.* 2000, DOI: 10.1.1.95.4103.

[39] Friedman, J. H., Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* 2001, 29, 1189–1232.

[40] Knapp, S. K., Accelerate FPGA macros with one-hot approach. *Electron. Des.* 1990.

[41] Singh, D., Singh, B., Investigating the impact of data normalization on classification performance. *Appl. Soft Comput. J.* 2019, 105524.

[42] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos,

A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 2011.

[43] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Mu, L., Xie, J., Lin, M., Geng, Y., Li, Y., Package "Xgboost." 2019.

[44] Chai, T., Draxler, R. R., Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* 2014, 7, 1247–1250.

[45] Wilson, E. J., Geankoplis, C. J., Liquid mass transfer at very low reynolds numbers in packed beds. *Ind. Eng. Chem. Fundam.* 1966, 5, 9–14.

[46] Sarwar, M. S., Simon, U., Dimartino, S., Experimental investigation and mass transfer modelling of 3D printed monolithic cation exchangers. *J. Chromatogr. A* 2021, 1646, 462125.

[47] Guiochon, G., Gritti, F., Shell particles, trials, tribulations and triumphs. *J. Chromatogr. A* 2011, 1218, 1915–1938.

[48] Gritti, F., Leonardis, I., Abia, J., Guiochon, G., Physical properties and structure of fine core–shell particles used as packing materials for chromatography. *J. Chromatogr. A* 2010, 1217, 3819–3843.

[49] Knox, J. H., Band dispersion in chromatography - A universal expression for the contribution from the mobile zone. *J. Chromatogr. A* 2002, 960, 7–18.

[50] Patel, K. D., Jerkovich, A. D., Link, J. C., Jorgenson, J. W., In-depth

characterization of slurry packed capillary columns with 1.0-μm nonporous particles using reversed-phase isocratic ultrahigh-pressure liquid chromatography. *Anal. Chem.* 2004, 76, 5777–5786.

[51] Malkin, D. S., Wei, B., Fogiel, A. J., Staats, S. L., Wirth, M. J., Submicrometer Plate Heights for Capillaries Packed with Silica Colloidal Crystals. *Anal. Chem.* 2010, 82, 2175–2177.

[52] Khirevich, S., Daneyko, A., Höltzel, A., Seidel-Morgenstern, A., Tallarek, U., Statistical analysis of packed beds, the origin of short-range disorder, and its impact on eddy dispersion. *J. Chromatogr. A* 2010, 1217, 4713–4722.

[53] Gritti, F., Leonardis, I., Shock, D., Stevenson, P., Shalliker, A., Guiochon, G., Performance of columns packed with the new shell particles, Kinetex-C18. *J. Chromatogr. A* 2010, DOI: 10.1016/j.chroma.2009.12.079.

[54] Dolamore, F., Dimartino, S., Fee, C. J., Numerical Elucidation of Flow and Dispersion in Ordered Packed Beds: Nonspherical Polygons and the Effect of Particle Overlap on Chromatographic Performance. *Anal. Chem.* 2019, 91, 15009–15016.

[55] Knox, J. H., Band dispersion in chromatography - A new view of A-term dispersion. *J. Chromatogr. A* 1999, 831, 3–15.

[56] Lan, T., Gerontas, S., Smith, G. R., Langdon, J., Ward, J. M., Titchener-Hooker, N. J., Investigating the use of column inserts to achieve better chromatographic bed support. *Biotechnol. Prog.* 2012, 28, 1285–1291.

[57] Jaulmes, A., Ignatiadis, I., Cardot, P., Vidal-Madjar, C., Characterization of

peak asymmetry with overloaded capillary columns. *J. Chromatogr. A* 1987, 395, 291–306.

[58] Pápai, Z., Pap, T. L., Analysis of peak asymmetry in chromatography. *J. Chromatogr. A* 2002, 953, 31–38.

[59] Miyabe, K., Guiochon, G., Estimation of the column radial heterogeneity from an analysis of the characteristics of tailing peaks in linear chromatography. *J. Chromatogr. A* 1999, 830, 29–39.

[60] Miyabe, K., Guiochon, G., Peak tailing and column radial heterogeneity in linear chromatography. *J. Chromatogr. A* 1999, 830, 263–274.

[61] Madsen, H., Time Series Analysis. 2007.

[62] Palma, W., Long-Memory Time Series. John Wiley & Sons, Inc., Hoboken, NJ, USA 2007.

[63] Mills, T. C., ARMA Models for Stationary Time Series. *Appl. Time Ser. Anal.* 2019, 31–56.

[64] Bendada, K., Hamdi, B., Boudriche, L., Balard, H., Calvet, R., Surface characterization of reservoir rocks by inverse gas chromatography: Effect of a surfactant. *Colloids Surfaces A Physicochem. Eng. Asp.* 2016, 504, 75–85.

[65] Bacskay, I., Sepsey, A., Felinger, A., Determination of the pore size distribution of high-performance liquid chromatography stationary phases via inverse size exclusion chromatography. *J. Chromatogr. A* 2014, 1339, 110–117.

[66] Johnson, T. F., Bailey, J. J., Iacoviello, F., Welsh, J. H., Levison, P. R., Shearing, P. R., Bracewell, D. G., Three dimensional characterisation of chromatography bead internal structure using X-ray computed tomography and focused ion beam microscopy. *J. Chromatogr. A* 2018, 1566, 79–88.

[67] Stickel, J. J., Fotopoulos, A., Pressure-flow relationships for packed beds of compressible chromatography media at laboratory and production scale. *Biotechnol. Prog.* 2001, 17, 744–751.

[68] Kawachi, Y., Ikegami, T., Takubo, H., Ikegami, Y., Miyamoto, M., Tanaka, N., Chromatographic characterization of hydrophilic interaction liquid chromatography stationary phases: Hydrophilicity, charge effects, structural selectivity, and separation efficiency. *J. Chromatogr. A* 2011, 1218, 5903–5919.

[69] McCue, J. T., Cecchini, D., Hawkins, K., Dolinski, E., Use of an alternative scale-down approach to predict and extend hydroxyapatite column lifetimes. *J. Chromatogr. A* 2007, DOI: 10.1016/j.chroma.2007.07.053.

[70] Dorn, M., Eschbach, F., Hekmat, D., Weuster-Botz, D., Influence of different packing methods on the hydrodynamic stability of chromatography columns. *J. Chromatogr. A* 2017, 1516, 89–101.

[71] Maier, R. S., Kroll, D. M., Davis, H. T., Diameter-dependent dispersion in packed cylinders. *AIChE J.* 2007, 53, 527–530.

[72] Reising, A. E., Schlabach, S., Baranau, V., Stoeckel, D., Tallarek, U., Analysis

of packing microstructure and wall effects in a narrow-bore ultrahigh pressure liquid chromatography column using focused ion-beam scanning electron microscopy. *J. Chromatogr. A* 2017, 1513, 172–182.

[73] Gritti, F., On the relationship between radial structure heterogeneities and efficiency of chromatographic columns. *J. Chromatogr. A* 2018, 1533, 112–126.

[74] Kaltenbrunner, O., Jungbauer, A., Yamamoto, S., Prediction of the preparative chromatography performance with a very small column. *J. Chromatogr. A* 1997, 760, 41–53.

Figure Captions:

Figure 1 Time series of a) reduced plate height, h, and b) asymmetry, A_s, for pre-packed purification columns manufactured over the 10 year period monitored.

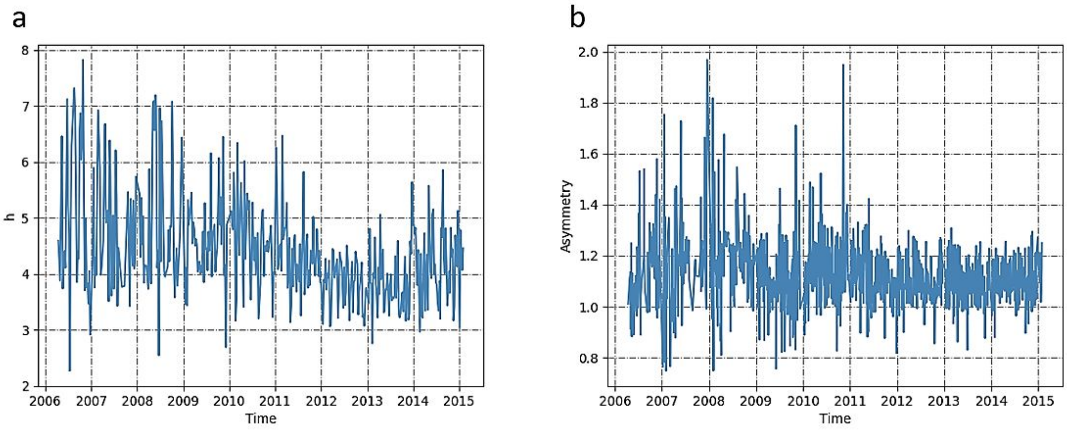


Figure 2 Autocorrelation (acf) and partial autocorrelation (pacf) analysis of reduced plate height, h, and asymmetry, A_s. a) Acf of h; b) pacf of h; c) acf of A_s; d) pacf of A_s. The blue shaded areas correspond to 95% confidence interval.

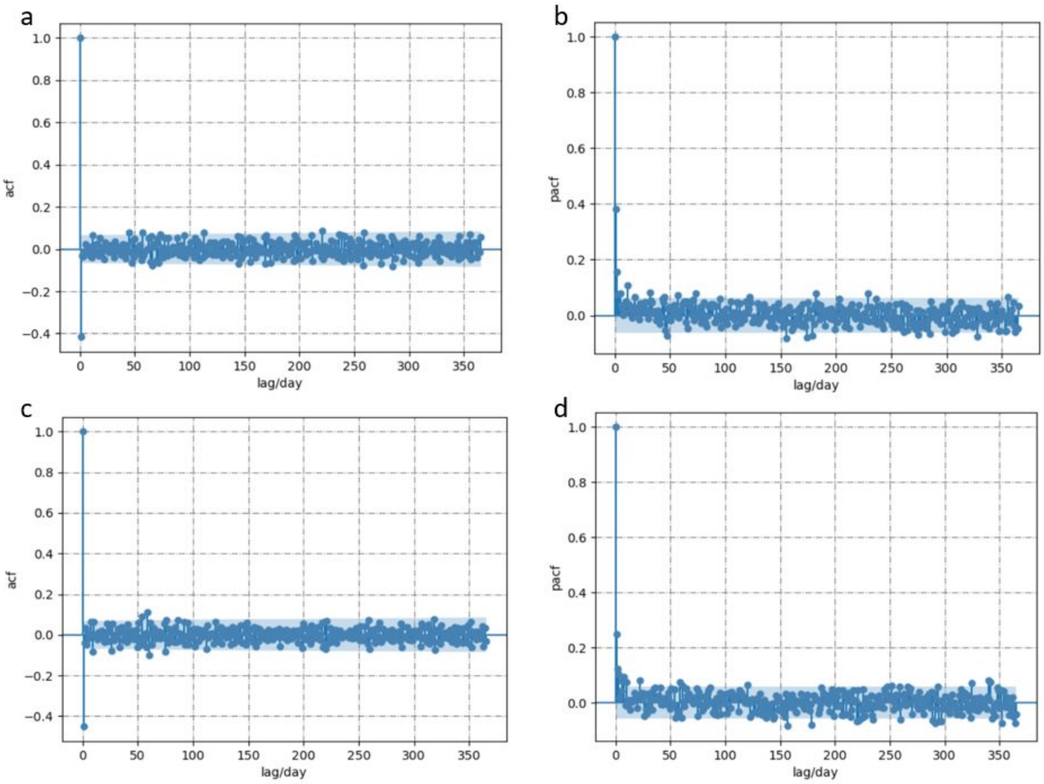


Figure 3. XGBoost prediction results for a) h and b) A_s over the testing data set. Variable importance contributions of c) h and d) A_s are reported. The importance is calculated based on the improvement of the performance measured by each attribute split points, weighted by the number of the observations the node is responsible for. The importance contributions, named by Gain in XGBoost (refer to Eq 4&5), were transferred into percentage.

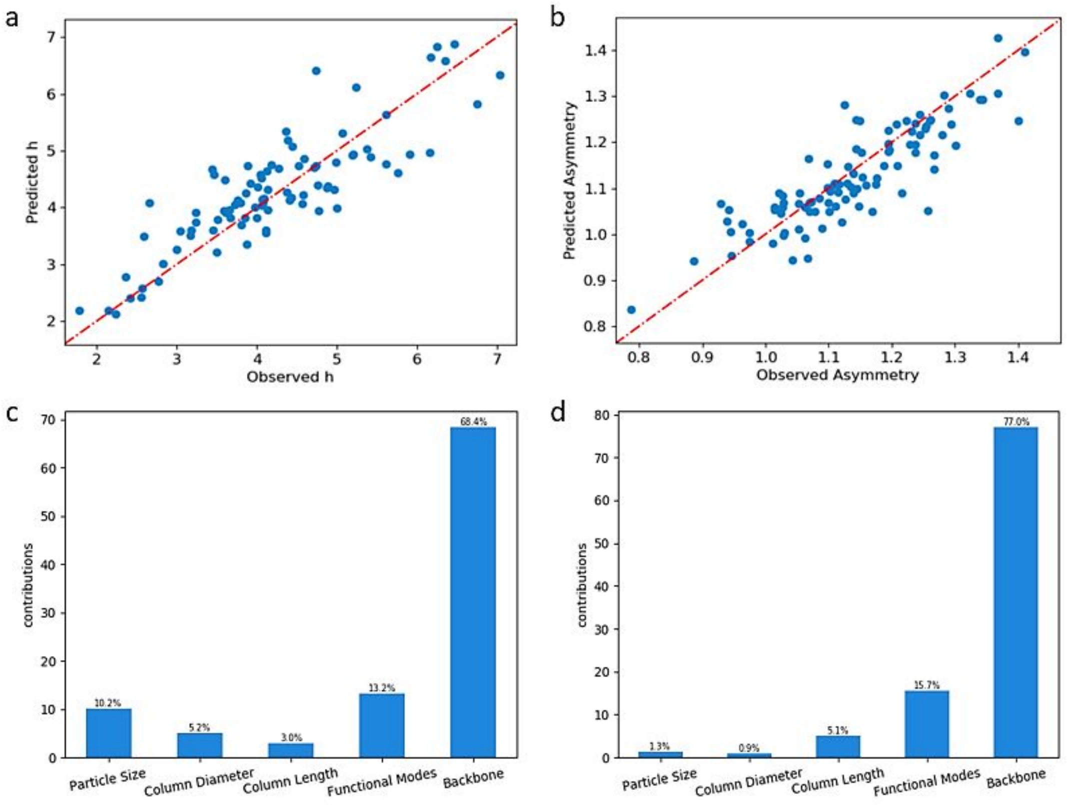


Figure 4. Violin plots of h and A_s against input parameters (backbone, functional mode, column length, column diameter, particle diameter). a) h vs backbone (PS/DVB: polystyrene divynilbenzene; IS: Inorganic support; MET: Methacrylate; AGR: Agarose; POL: Polymer grafted; DEX: Dextran; CEL: Cellulose; PVE: polyvinyl-ether hydrophilic). b) A_s vs backbone. c) h vs functional mode (CIEC: cation exchange chromatograph; AF: affinity chromatography; HA: hydroxyl-apatite chromatography; AIEC: anion exchange chromatography; HIC: hydrophobic interaction chromatography; SEC: size-exclusion chromatography; IMAC: immobilized metal affinity chromatography; MMC: mixed-mode chromatography; FA: fluorophore adsorption chromatography; HCIC: hydrophobic charge induction chromatography). d) A_s vs functional mode. e) h vs column length. f) A_s vs column length. g) h vs column diameter. h) A_s vs column diameter. i) h vs particle diameter. j) A_s vs particle diameter.

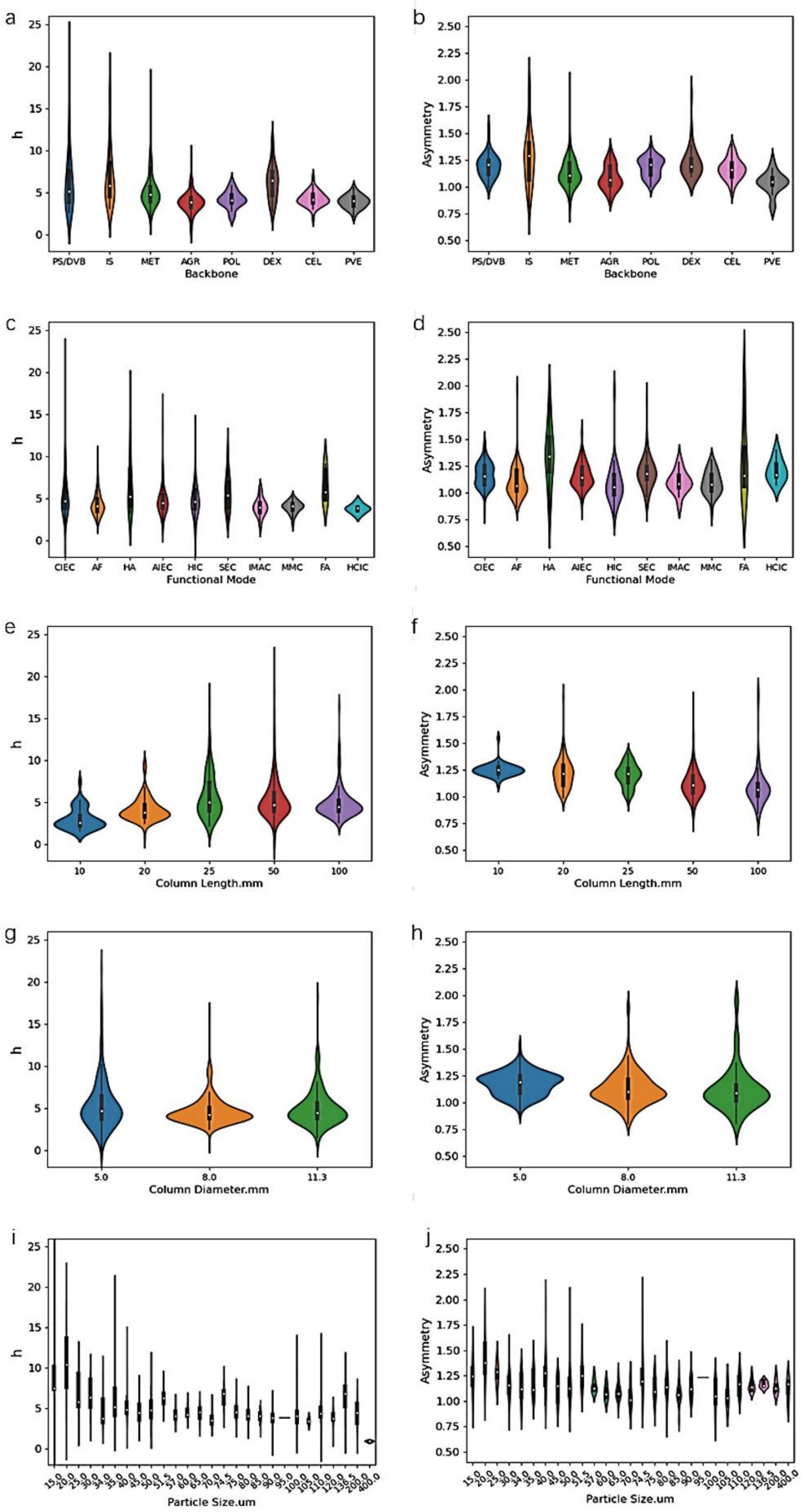
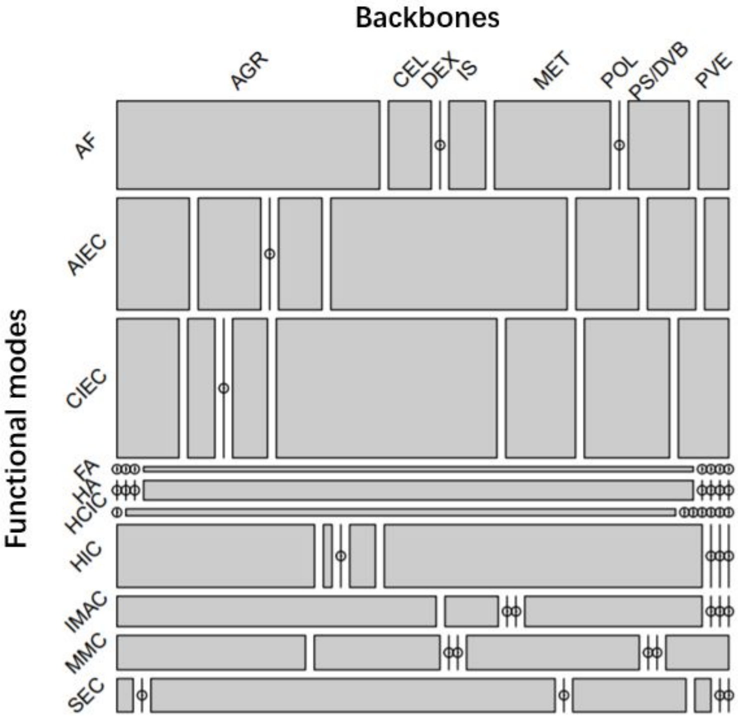

a
h
0
5
10
15
20
25
PS/DVB
IS
MET
AGR
POL
DEX
CEL
PVE
Backbone
b
Asymmetry
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
PS/DVB
IS
MET
AGR
POL
DEX
CEL
PVE
Backbone
c
h
CIEC
AF
HA
AIEC
HIC
SEC
IMAC
MMC
FA
HCIC
Functional Mode
d
Asymmetry
CIEC
AF
HA
AIEC
HIC
SEC
IMAC
MMC
FA
HCIC
Functional Mode
e
h
10
20
25
50
100
Column Length.mm
f
Asymmetry
10
20
25
50
100
Column Length.mm
g
h
5.0
8.0
11.3
Column Diameter.mm
h
Asymmetry
5.0
8.0
11.3
Column Diameter.mm
i
h
15.0
20.0
25.0
30.0
34.0
35.0
40.0
45.0
50.0
51.5
57.0
60.0
65.0
70.0
74.5
75.0
80.0
85.0
90.0
95.0
100.0
105.0
110.0
120.0
136.5
200.0
400.0
Particle Size.um
j
Asymmetry
15.0
20.0
25.0
30.0
34.0
35.0
40.0
45.0
50.0
51.5
57.0
60.0
65.0
70.0
74.5
75.0
80.0
85.0
90.0
95.0
100.0
105.0
110.0
120.0
136.5
200.0
400.0
Particle Size.um

Figure 5. Mosaic plot of the combinations of functional mode and backbone material tested. The size of the tiles represents the relative frequency of each combination. PS/DVB: polystyrene divynilbenzene; IS: Inorganic support; MET: Methacrylate; AGR: Agarose; POL: Polymer grafted; DEX: Dextran; CEL: Cellulose; PVE: polyvinyl-ether hydrophilic; CIEC: cation exchange chromatograph; AF: affinity chromatography; HA: hydroxyl-apatite chromatography; AIEC: anion exchange chromatography; HIC: hydrophobic interaction chromatography; SEC: size-exclusion chromatography; IMAC: immobilized metal affinity chromatography; MMC: mixed-mode chromatography; FA: fluorophore adsorption chromatography; HCIC: hydrophobic charge induction chromatography.



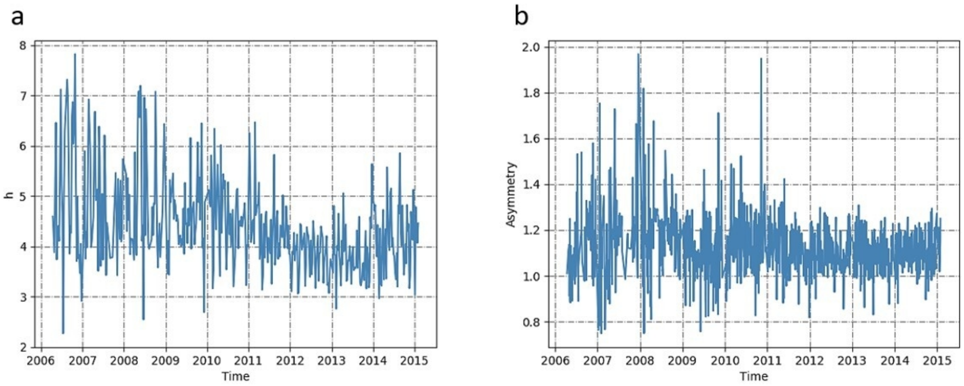


Figure 1 Time series of a) reduced plate height, h, and b) asymmetry, A_s.

99x40mm (300 x 300 DPI)

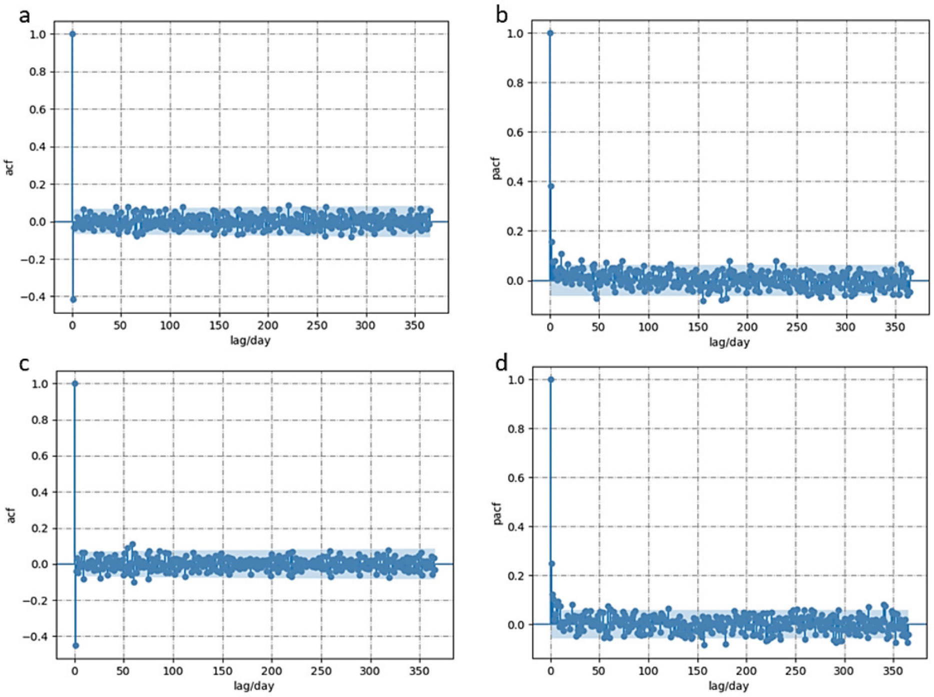


Figure 2 Autocorrelation (acf) and partial autocorrelation (pacf) analysis of reduced plate height, h, and asymmetry, A_s. a) Acf of h; b) pacf of h; c) acf of A_s; d) pacf of A_s. The blue shaded areas correspond to 95% confidence interval.

101x74mm (300 x 300 DPI)

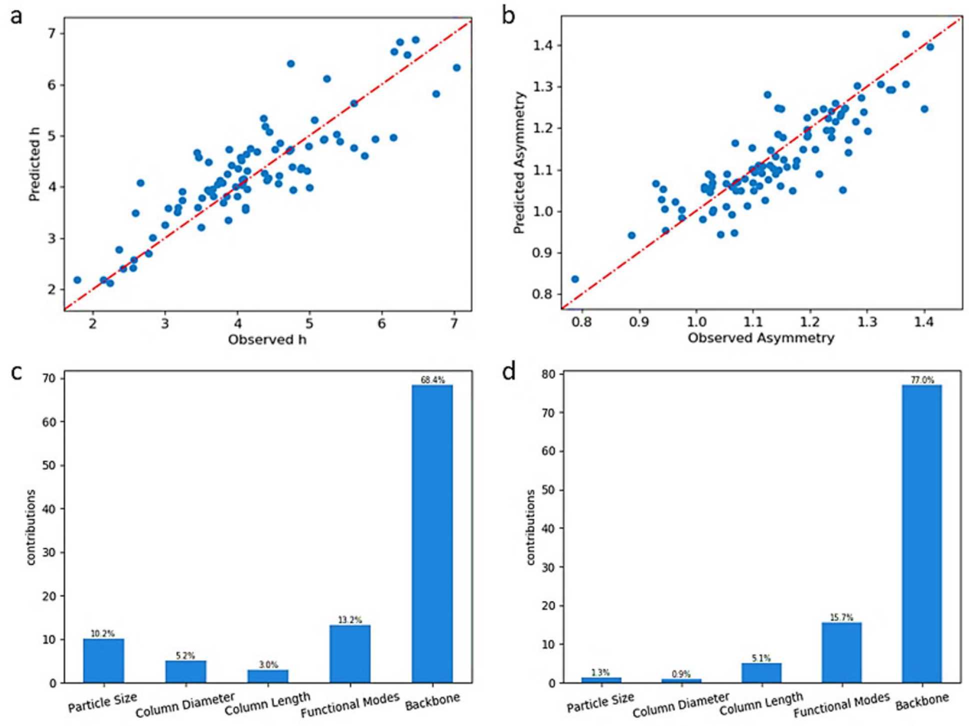


Figure 2 Autocorrelation (acf) and partial autocorrelation (pacf) analysis of reduced plate height, h, and asymmetry, A_s. a) Acf of h; b) pacf of h; c) acf of A_s; d) pacf of A_s. The blue shaded areas correspond to 95% confidence interval.

57x42mm (500 x 500 DPI)

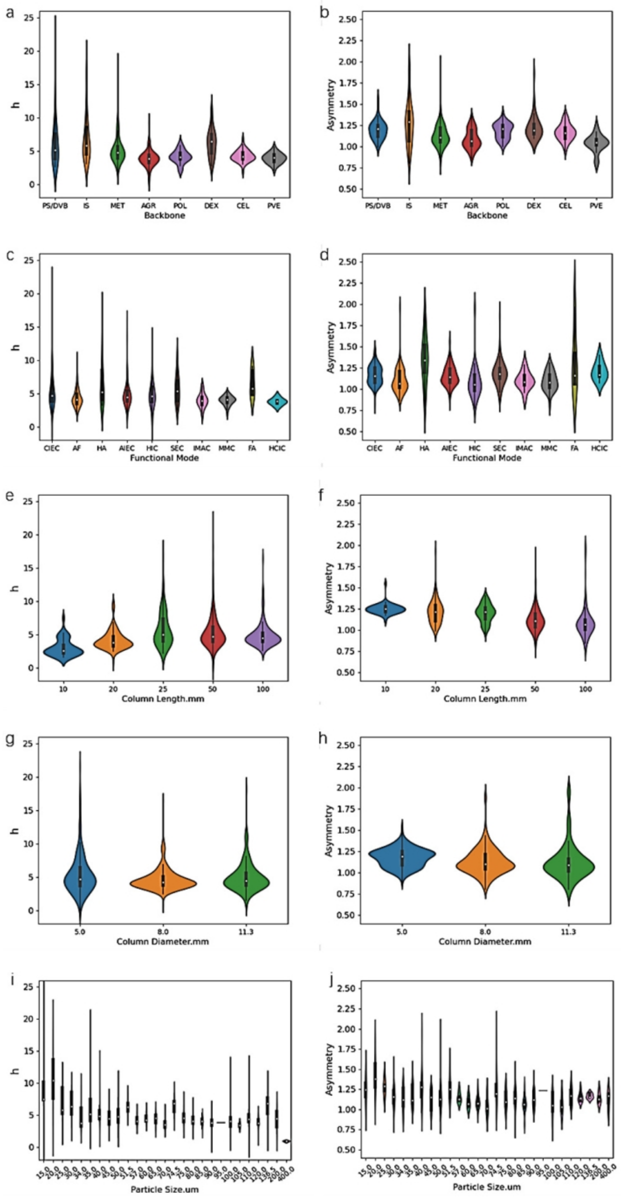


Figure 4. Violin plots of h and As against input parameters (backbone, functional mode, column length, column diameter, particle diameter). a) h vs backbone (PS/DVB: polystyrene divynilbenzene; IS: Inorganic support; MET: Methacrylate; AGR: Agarose; POL: Polymer grafted; DEX: Dextran; CEL: Cellulose; PVE: polyvinyl-ether hydrophilic). b) A_s vs backbone. c) h vs functional mode (CIEC: cation exchange chromatograph; AF: affinity chromatography; HA: hydroxyl-apatite chromatography; AIEC: anion exchange chromatography; HIC: hydrophobic interaction chromatography; SEC: size-exclusion chromatography; IMAC: immobilized metal affinity chromatography; MMC: mixed-mode chromatography; FA: fluorophore adsorption chromatography; HCIC: hydrophobic charge induction chromatography). d) A_s vs functional mode. e) h vs column length. f) A_s vs column length. g) h vs column diameter. h) A_s vs column diameter. i) h vs particle diameter. j) A_s vs particle diameter.

26x48mm (1000 x 1000 DPI)

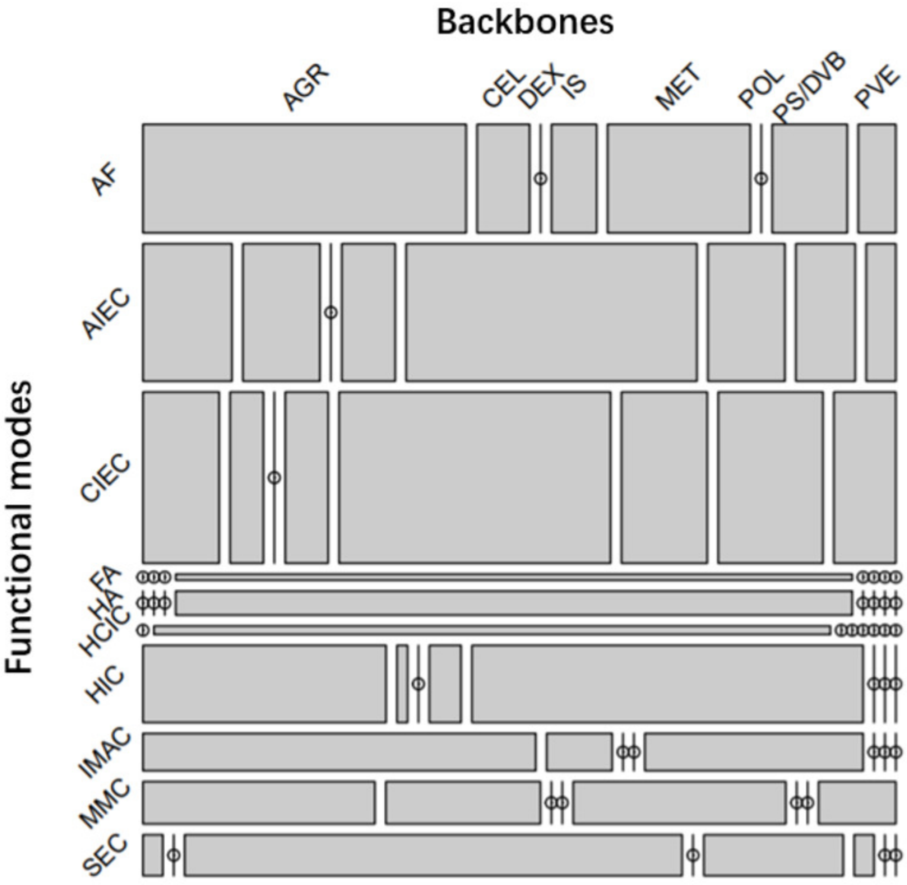


Figure 5. Mosaic plot of the combinations of functional mode and backbone material tested. The size of the tiles represents the relative frequency of each combination. PS/DVB: polystyrene divynilbenzene; IS: Inorganic support; MET: Methacrylate; AGR: Agarose; POL: Polymer grafted; DEX: Dextran; CEL: Cellulose; PVE: polyvinyl-ether hydrophilic; CIEC: cation exchange chromatograph; AF: affinity chromatography; HA: hydroxyl-apatite chromatography; AIEC: anion exchange chromatography; HIC: hydrophobic interaction chromatography; SEC: size-exclusion chromatography; IMAC: immobilized metal affinity chromatography; MMC: mixed-mode chromatography; FA: fluorophore adsorption chromatography; HCIC: hydrophobic charge induction chromatography.

21x20mm (1000 x 1000 DPI)