molecules

MDPI

Article

Statistical Methods in the Study of Protein Binding and Its Relationship to Drug Bioavailability in Breast Milk

Karolina Wanat * and Elżbieta Brzezińska

Department of Analytical Chemistry, Faculty of Pharmacy, Medical University of Lodz, 90-419 Łódź, Poland; elzbieta.brzezinska@umed.lodz.pl

* Correspondence: karolina.wanat@umed.lodz.pl; Tel.: +48-42-677-92-11

Abstract: Protein binding (PB) is indicated as the factor most severely limiting distribution in the organism, reducing the bioavailability of the drug, but also minimizing the penetration of xenobiotics into the fetus or the body of a breastfed child. Therefore, PB is an important aspect to be analyzed and monitored in the design of new drug substances. In this paper, several statistical analyses have been introduced to find the relationship between protein binding and the amount of drug in breast milk and to select molecular descriptors responsible for both pharmacokinetic phenomena. Along with descriptors related to the physicochemical properties of drugs, chromatographic descriptors from TLC and HPLC experiments were also used. Both methods used modification of the stationary phase, using bovine serum albumin (BSA) in TLC and human serum albumin (HSA) in HPLC. The use of the chromatographic data in the protein binding study was found to be positive —the most effective application of normal-phase TLC and $HPLC_{HSA}$ data was found. Statistical analyses also confirmed the prognostic value of affinity chromatography data and protein binding itself as the most important parameters in predicting drug excretion into breast milk.

Keywords: protein binding; breast milk; M/P ratio; statistical modeling; molecular descriptors; chromatographic descriptors; affinity chromatography

Citation: Wanat, K.; Brzezińska, E. Statistical Methods in the Study of Protein Binding and Its Relationship to Drug Bioavailability in Breast Milk. *Molecules* **2022**, *27*, 3441. https://doi.org/10.3390/molecules27113441

Academic Editors: Giovanni Ribaudo and Laura Orian

Received: 24 April 2022
Accepted: 19 May 2022
Published: 26 May 2022

1. Introduction

Excretion of drugs into breast milk is an important aspect to be considered in the pharmacotherapy of breastfeeding women. Due to ethical considerations, in vivo studies are very rare and it is difficult to obtain the milk-to-plasma (M/P) ratio of many active pharmaceutical compounds (APIs). A mathematical model capable of calculating M/P values using the available data will greatly facilitate the study of the bioavailability of new APIs.

In the previous articles [1,2], we presented a comparison of statistical methods in the study of drug excretion into breast milk with the use of the M/P descriptor. It was shown that the multiple linear regression (MLR) and random forest (RF) analyses were most effective in describing this pharmacokinetic phenomenon, with the use of chromatographic data and physicochemical properties of the tested compounds. These analyses did not deviate from the known principles of bioavailability to breast milk and showed a close relationship between M/P and the level of drug–protein binding (PB) as well as the state of ionization of the API in the bloodstream.

The papers also describe the most effective conditions for thin layer chromatography (TLC) as an analytical model for predicting the penetration of drugs into breast milk. According to these results, it can be assumed that the use of drug–protein binding indices, together with chromatographic data, will make it possible to predict the level of drug distribution into breast milk.

The main aim of this study is to provide supplementary analyses, which include: determination of physicochemical parameters related to drug protein binding; searching for

a mathematical model of PB and/or M/P prediction; and the use of affinity chromatography data as an index of pharmacokinetic properties.

The goal of developing such a model is its further utility in predicting the PB of newly developed active pharmaceutical ingredients. Only easily available API properties are needed to use the model. It can facilitate the process of introducing a new drug to use and reduce expensive in vivo testing.

In this study the following statistical methods were used: cluster analysis (CA), discriminant function analysis (DFA) and principal component analysis (PCA) random forest regression (RF). All molecular descriptors used in this study are listed and described in Table 1.

Table 1. List of molecular and chromatographic descriptors used in statistical analyses.

Descriptor	Description	Reference/Database/Software
a/b/*n* code	acidic, basic or neutral character of the compound; describes the division into groups: a, b and n	CHEMBL database [3]
B1	calculation parameter B2, describes the bioavailability in the CNS and determines penetration through the blood-brain barrier: log bb = 0.139 + 0.152 log P	reference [4]
B2	calculation parameter B2, describes the bioavailability in the CNS and determines penetration through the blood-brain barrier: log bb = 0.547 − 0.016 PSA	reference [5]
B3	calculation parameter related to protein binding: log (bound fraction/unbound fraction) = 0.5 log P–0.665	reference [6]
CNS+/−	ability to penetrate into the central nervous system (+ or −)	DrugBank database [7]
DM	dipole moment	HyperChem, Hypercube, Inc.
eH	energy of the highest occupied molecular orbital (HOMO)	HyperChem, Hypercube, Inc.
eH-eL	ionization capacity	HyperChem, Hypercube, Inc.
eL	energy of the lowest unoccupied molecular orbital (LUMO)	HyperChem, Hypercube, Inc.
HA	number of hydrogen bond acceptors	ACD/Labs
HD	number of hydrogen bond donors	ACD/Labs
log D	distribution coefficient	ACD/Labs
log M/P	logarithm of M/P	
log MW	logarithm of MW	
log P	partition coefficient	HyperChem, Hypercube, Inc.
log U/D	the ratio of neutral to ionized form; determines the degree of ionization	Calculated using: pK_a-pH for acids; pH-pK_a for bases
M/P	milk/plasma drug concentration ratio	references [8–13]
MW	molecular weight	HyperChem, Hypercube, Inc.
PB	percentage of plasma protein binding	DrugBank
PhCharge	the charge of the API under physiological conditions	DrugBank
pK_a	negative logarithm of the acid dissociation constant (K_a)	ACD/Labs
PSA	polar surface area	ACD/Labs
Sa	the surface area of the molecule	HyperChem, Hypercube, Inc.
V	the volume of the molecule	HyperChem, Hypercube, Inc.
NP; RP	R_f (retention factor) obtained from TLC using impregnated with bovine serum albumin (BSA) plates in normal and reversed phase	TLC experiment
NP/C; RP/C	R_f from impregnated NP or RP plate/control R_f	TLC experiment
k_{HSA}	retention factor from HPLC using column with immobilized human serum albumin (HSA)	HPLC experiment
log k_{HSA}	logarithm of the retention coefficient obtained from $HPLC_{HSA}$	HPLC experiment
log k_{IAM}	logarithm of the retention coefficient obtained from $HPLC_{IAM}$ (column with immobilized artificial membrane)	HPLC experiment

2. Results

2.1. Correlation Analyses

The experiment investigated the results of using data from several chromatographic analysis experiments ($HPLC_{HSA}$, NP TLC, RP TLC and, additionally, $HPLC_{IAM}$) in predicting drug binding to protein, and thus bioavailability to breast milk. A group of 165 APIs was analyzed, in which acidic, basic and neutral drugs were observed. The best correlation with PB values was shown in the results of the $HPLC_{HSA}$ and NP TLC experiments, in the form of log k and R_f values, ($HPLC_{HSA}$: n = 165, R = 0.39); (NP TLC: n = 162, R = 0.31). The relationship is directly proportional. This is the result for all kinds of relationships. Much better results were obtained for acidic drugs (R = 0.50), even considering the smaller number of cases (n = 34) (Table A1, Appendix A).

Then the effect of the most frequently mentioned molecular descriptors, related to drug distribution into breast milk and protein binding, was investigated. In all groups of APIs, molecular descriptors related to the hydro-lipophilic nature of drugs play a dominant role. The most important parameters are the partition coefficient and the distribution coefficient (log P and log D). The ability to form hydrogen bonds (HD, HA) is visible here and the correlation with PB is significant. The ratio of neutral to dissociated form (log U/D), dissociation constant (pKa), ionization capacity of compounds (eH-eL) and other electron descriptors: eL and eH, show no significance. The influence of hydrophobic parameters (Sa, V, MW) is visible only in the form of the surface area to volume ratio (Sa/V). As can be seen above, this factor correlates inversely with all types of cases (Table A2, Appendix A).

2.2. Discriminant Function Analysis

All of the descriptors most strongly related to the variability of the PB, which at the same time did not limit the number of cases studied, were introduced into the discriminant function analysis (DFA). All cases were tested using the a/b/n code.

In the stepwise DFA, the discriminant variables included 9 out of 16 entered variables: PhCharge, B2, pKa, M/P, log k_{HSA}, log k_{IAM}, NP, eL and log U/D (Table 2).

Table 2. Classification matrix for the model using discriminant variables: PhCharge, B2, pKa, M/P, log k_{HSA}, log k_{IAM}, NP, eL, log U/D.

API Group	Correctly Classified Cases (%)	a p = 0.17895	n p = 0.52632	b p = 0.29474
a	100,00	17	0	0
n	96,00	0	48	2
b	92,86	0	2	26
all	95,80	17	50	28

The PC1 factor discriminates the groups of APIs the most (PC 1 eigenvalue = 3.61). The variables PhCharge and pKa have the most important share in its value. The PC2 factor (PC2 eigenvalue = 0.81) was shaped by the chromatographic descriptors and the ability to ionize (log U/D). The means of the canonical variables (PC1) for group a = −3.52, for group n = 0.03 and for group b = 2.08, therefore PC1 most strongly discriminates between groups a and b. The means of the canonical variables (PC2) for group a = −0.93, for group n = 0.86 and for group b = −0.97. In this case, the centroids of groups a and b are almost equal, and the group of neutral compounds (n) is the most discriminated against (Figure 1).

2.3. Principal Component Analysis

PCA was performed to determine the effect of the primary descriptors on the characteristics of the drug's ability to pass into breast milk. In order to better visualize the obtained results from the analysis, the M/P values were converted into the scale of the drug penetration into milk—M/P_{code}. The values of this indicator are in the range 1–4. Code 1 corresponds to drugs with an M/P value <0.40—completely safe; 2 corresponds to

the range of 0.40–0.80—at the safety limit; 3 range 0.81–1.20—possibly over the safety limit; and 4 is M/P >1.20—dangerous.

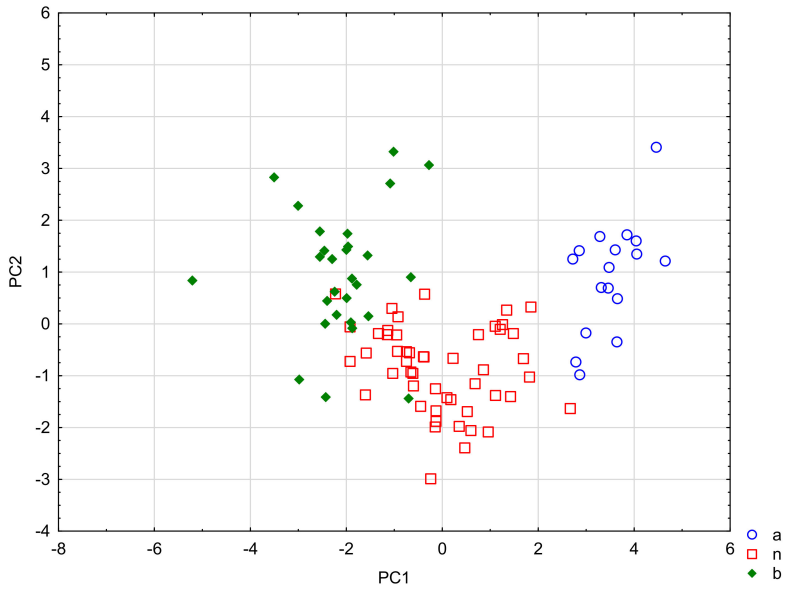

Figure 1. Discrimination against acidic (**a**), basic (**b**) and neutral drugs (*n*). The scatter plot of canonical values for root 1 relative to root 2. Discriminating variables: PhCharge, B2, pKa, M/P, log k_{HSA}, log k_{IAM}, NP, eL, log U/D.

In the course of the analysis, the smallest number of principal components explaining the maximum range of the total variance in the group was initially established. Five factors explain 100% of the variability in the levels of drug excretion into breast milk. The first two factors, PC1 and PC2 (principal components), are described by all used descriptors. As a result, two main components explaining a total of 72% of the variability were obtained. The $HPLC_{HSA}$, $HPLC_{IAM}$, NP TLC and RP TLC chromatographic data is responsible for the first component, PC1 (43.26%), the second component, PC2 (28.66%), is determined by the PB value.

The projection of cases on the PC1 × PC2 plane is presented below (Figure 2):

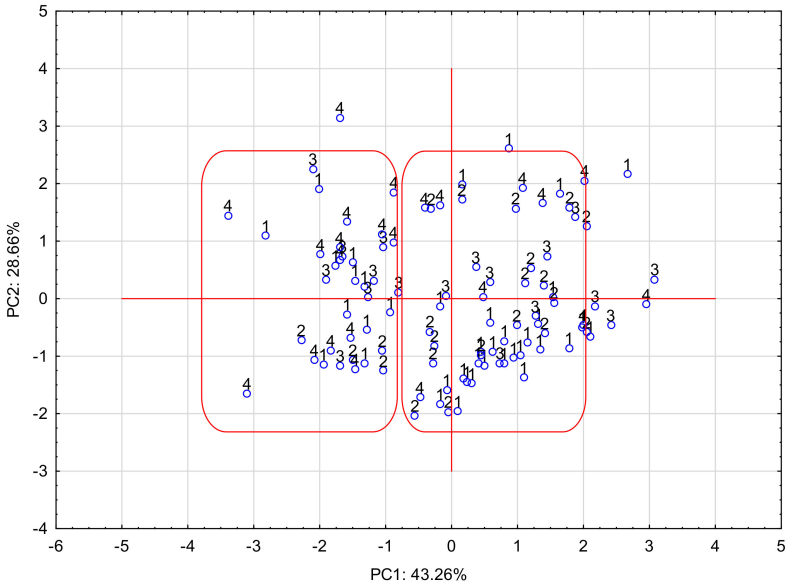

Figure 2. Projection of cases onto the PC1 × PC2 plane.

In the graph of the projection of cases onto the PC plane, where the grouping variable is the scale of drug penetration into breast milk (M/P_{code}), it can be seen that the tested APIs can be divided into two groups (surrounded by a box in the graph). One group included drugs with a lower level of M/P (1–2) penetration—safe, and the other group, M/P 3–4—dangerous. This division is not entirely obvious. It was created on the basis of factors explaining 75% of the variability. Few examples of misclassification are visible. The distinction between these groups is related to PC1. Derivatives with a low M/P are located on the right side of the plot and are clearly related to the positive values of PC1. APIs easily excreted into milk are on the left side of the chart and have negative PC1 values. The share of variables in this component, determined by the PC1-variable correlation (factor loadings), reveals the parameters of the greatest importance for the investigated pharmacokinetic feature of drugs. They are: log k_{HSA}, log k_{IAM}, NP and RP. Thus, affinity chromatography, based on protein binding, can predict the bioavailability of an API into breast milk.

The graph of the projection of variables onto the PC plane shows graphically the relationship between the component and the variable. The graph shows the so-called unit circle, i.e., the maximum correlation of 1 between the variable and the factor. The closer a given variable is to the unit circle line, the greater its correlation with the observed phenomenon (Figure 3).

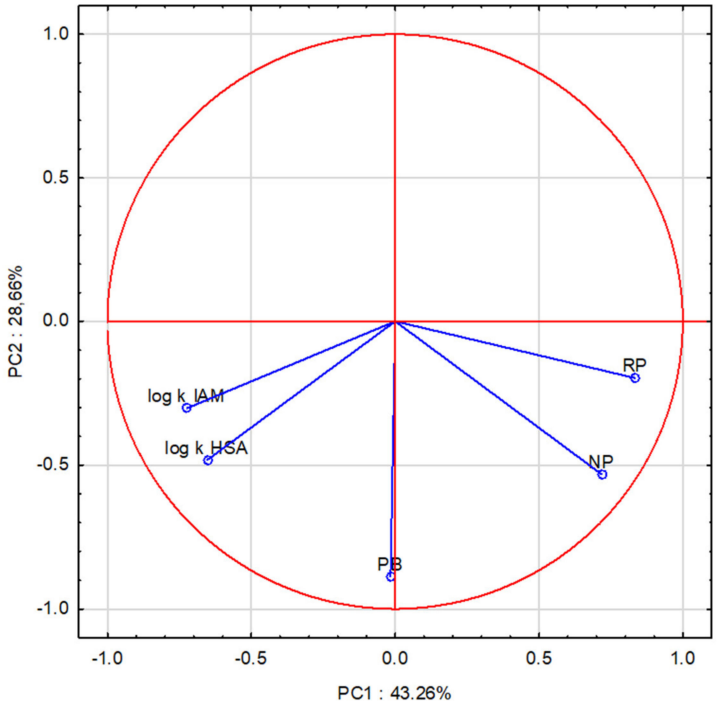

Figure 3. Projection of variables on the plane of factors PC1 × PC2.

2.4. Cluster Analysis

In order to emphasize the diagnostic value of the experiment and to determine the difference in the values of the parameters determining the ability of drugs to penetrate into breast milk, cluster analysis (CA) was also performed. CA was conducted in the proposed M/P_{code} scale, using the k-means method. The means of the most important biological descriptors (CNS +/−, B1, PhCharge, acid/base, NP, RP, log k_{HSA}, log k_{IAM} and PBcode) were compared for groups M/P_{code} 1–4. As shown, all drug biological parameters showed a group variability (see Figure 4). The M/P code values range from 1 to 4 with a clear distinction between relatively safe and unsafe groups. Physicochemical parameters: PB, acid/base, HD, log P, eL, log D also show differentiation, but not in all cases. Unfortunately, M/P_{code} is too clustered here, which indicates a smaller influence

of the tested properties on the observed feature (Figure 5). The descriptors: log D and eL show the highest differentiation.

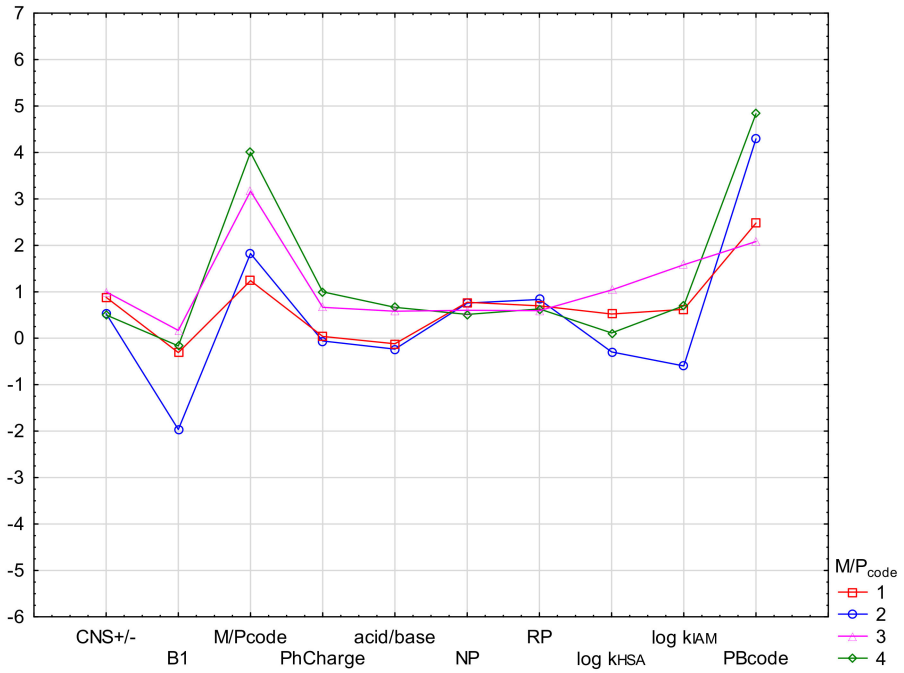

Figure 4. Mean descriptor values in M/P_{code} cluster analysis (k-means method) using biological and chromatographic descriptors.

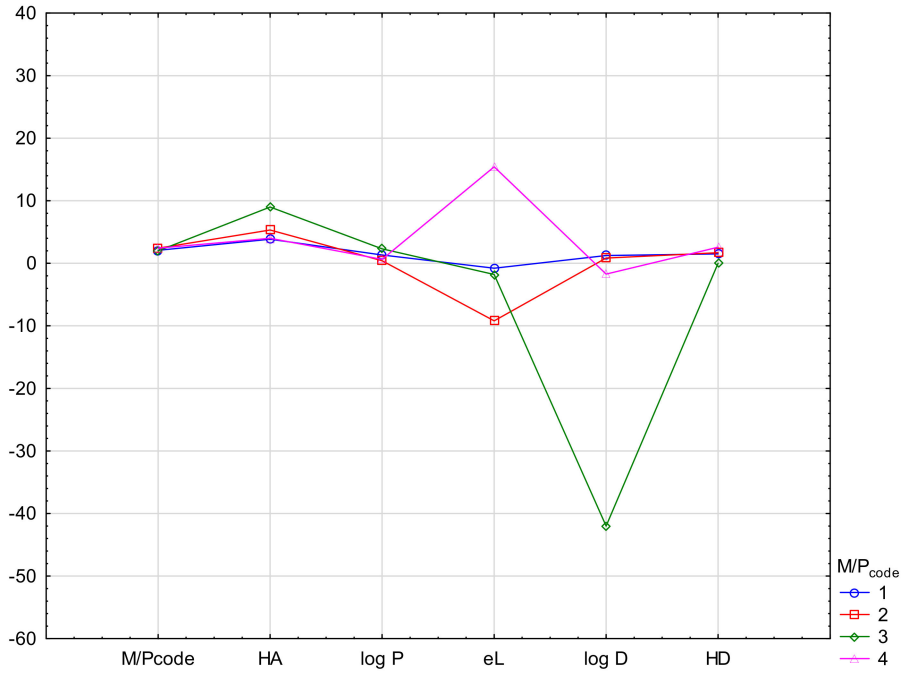

Figure 5. Mean descriptor values in M/P_{code} cluster analysis (k-means method) using physicochemical descriptors.

The above analyses confirmed the values of the parameters HA, log P, log D and eL. The parameters of log D, HA and eL show the greatest differentiation. Unfortunately, the M/P_{code} values are poorly differentiated and their values do not correspond to the variability of other descriptors.

2.5. Regression Methods

MLR failed to create a reliable PB prediction model, therefore an attempt was made to analyze protein binding by other regression methods. A total of 165 test compounds and 22–23 independent variables were used to perform partial least squares (PLS) and random forest regression (RF). The variables used are listed for each model (Tables 3 and 4). During the analyses, 165 compounds were randomly divided into a training set, 70% of the total (TRAIN, n = 115 compounds,) and a test set for external validation, 30% of the total (TEST, n = 50).

Table 3. Twenty-three independent variables with NP TLC data used to create the RF and PLS model for PB.

No.	Independent Variable	No.	Independent Variable	No.	Independent Variable
1.	B3	9.	NP/B2	17.	eH
2.	PhCharge	10.	NP/log P	18.	eL
3.	acid/base	11.	MW	19.	eH-eL
4.	pKa	12.	log MW	20.	logD
5.	log U/D	13.	PSA	21.	Sa
6.	C	14.	HD	22.	V
7.	NP	15.	HA	23.	logP
8.	NP/C	16.	DM		

Table 4. Twenty-two independent variables with $HPLC_{HSA}$ data used to create the RF and PLS model for PB.

No.	Independent Variable	No.	Independent Variable	No.	Independent Variable
1.	B3	9.	log k_{HSA}/log P	17.	eL
2.	PhCharge	10.	MW	18.	eH-eL
3.	acid/base	11.	log MW	19.	log D
4.	pKa	12.	PSA	20.	Sa
5.	log U/D	13.	HD	21.	V
6.	k_{HSA}	14.	HA	22.	log P
7.	log k_{HSA}	15.	DM		
8.	log k_{HSA}/B2	16.	eH		

2.5.1. Partial Least Squares Regression

The PLS model using 23 independent variables, including NP TLC data (Table 3) showed low values of R^2 and Q^2, approximately 0.40, and even lower results of external validation, approximately 0.22–0.24 (Figure A1, Appendix A). Even lower values are achieved with the $HPLC_{HSA}$ chromatographic data. This indicates that, as in the case of breast milk prediction models, the PLS method is again not widely applicable here and is not an appropriate method to analyze this type of data.

2.5.2. Random Forest Regression

RF regression was performed with the use of 150 generated random trees. NP TLC data was used first. The independent variables used for the analysis of all 165 cases (independent variable, PB_{abn}) are listed in Table 3.

The obtained model (Figure 6) showed satisfactory results, especially for the training set (n = 115): R^2_{train} = 0.81; Q^2_{train} = 0.73. The results of external validation using the test kit (n_{abn} = 50) were lower: R^2_{test} = 0.65; Q^2_{test} = 0.56. The Monte Carlo permutation test (MCPT) showed the average value of the Q^2_{test} parameter was equal to 0.56 (Appendix A, Figure A2), which is similar to that in the presented model. The influence of individual independent

variables on the model is presented in the chart below (Appendix A, Figure A3). The order of the descriptors presented there is as shown in Table 3. The log D parameter shows the strongest influence on the model using NP TLC data.

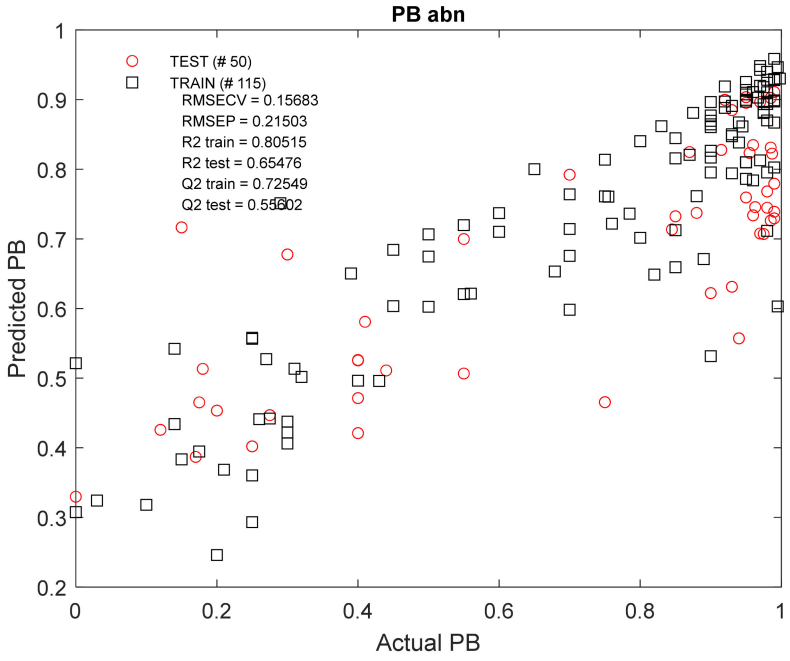

Figure 6. Actual versus predicted PB_{abn} values using RF regression modelling of molecular descriptor set containing 23 variables. $RMSE_{CV}$ = root-mean-square error of cross-validation, $RMSE_P$ = root-mean-square error of prediction, R^2 train/test = coefficient of determination for train/test set models, Q^2 train/test = coefficient of determination for the cross-validated models.

The data from the $HPLC_{HSA}$ experiment were then used for the RF regression (Table 4). The obtained model (Appendix A, Figure A4) again shows good results of the training set (n = 115): R^2_{train} = 0.81; Q^2_{train} = 0.78 but much lower parameters were obtained with external validation (n_{abn} = 50): R^2_{test} = 0.57; Q^2_{test} = 0.53. In the MCPT, the Q^2_{test} value was already at a low level and amounted to 0.35 (Appendix A, Figure A5).

Then, individual groups of compounds were dealt with, either separately, (a), (b) and (n), or combined, (an), (bn) and (ab). The results are shown in Table 5. Only the NP TLC data (Table 3) were used to construct the models, which gave the best results when tested for the complete set of compounds (n_{abn} = 165).

Table 5. Random forest regression results on individual drug combinations.

API Group	Train Set	Test Set
PB_a	n = 24 R^2 = 0.78; Q^2 = 0.62	n = 11 R^2 = 0.29; Q^2 = 0.11
PB_b	n = 35 R^2 = 0.88; Q^2 = 0.80	n = 15 R^2 = 0.33; Q^2 = 0.29
PB_n	n = 57 R^2 = 0.85; Q^2 = 0.81	n = 25 R^2 = 0.62; Q^2 = 0.59
PB_{an}	n = 82 R^2 = 0.82; Q^2 = 0.74	n = 35 R^2 = 0.60; Q^2 = 0.55
PB_{bn}	n = 92 R^2 = 0.85; Q^2 = 0.80	n = 40 R^2 = 0.44; Q^2 = 0.44
PB_{ab}	n = 59 R^2 = 0.80; Q^2 = 0.72	n = 26 R^2 = 0.38; Q^2 = 0.33

RF models for PB_a (n_a = 35) and PB_b (n_b = 50) gave poor results, especially in the external validation, similarly to their combined group (n_{ab} = 85), where the external validation results were in the range of Q^2 = 0.4–0.3.

The best results were obtained for the PB_n (n_n = 82) and PB_{an} (n_{an} = 117) groups. The R^2 and Q^2 values of the test kits ranged between 0.55 and 0.62 (Appendix A: Figures A6 and A7). In both models, the log D values are the most important in their creation (Appendix A: Figures A8 and A9).

3. Discussion

On the basis of the DFA analysis, it was possible to determine the influence of the acidic, basic and neutral properties of APIs on their protein binding capacity and to decide whether the analysis of the pharmacotherapy of nursing mothers (M/P predictions) should be divided into groups: a, *n* and b. The division into acidic, basic and neutral drugs is strongly related to the PB-related descriptors, so the use of groups a, b and *n* seems to bring value for further analysis. The low values of Wilks lambda for both roots, PC1 and PC2, confirm the value of the obtained results (0.11 and 0.54, respectively).

As the DFA analysis revealed a group of physicochemical and chromatographic parameters important for the bioavailability of drugs to milk, the use of CA emphasized the differentiation of their mean values in the M/P 1–4 groups. The above analyses confirmed the values of the parameters HA, log P, log D and eL. The parameters of log D, HA and eL show the greatest differentiation. Unfortunately, the M/P_{code} values are poorly differentiated and their values do not correspond to the variability of other descriptors. Based on the PCA, it can be concluded that the data of the drug–protein binding affinity chromatography, in the form of the proposed analytical models and the protein binding itself as the basis for the experimental design, are the most important parameters in predicting drug excretion into breast milk.

The final step in this study was to construct a model capable of predicting PB value, used as a trait strongly correlated with the bioavailability of breast milk. Unfortunately, it was not possible to obtain an MLR or PLS algorithm for protein binding prediction, that was reproducible for different groups. Models created by regression using the random forest method show a significant relationship, visible in the scatter plots (Figures 6, A4, A6 and A7). The influence of the determination coefficient (log D) and chromatographic parameters from the NP TLC and $HPLC_{HSA}$ experiments in each model are also noticeable. Unfortunately, they do not show the best predictive ability (external validation at the level of Q^2_{test} = 0.56 and 0.35 in MCPT tests).

The best results using random forest regression were obtained for the entire set of compounds, PB_{abn}, and for the PB_n and PB_{an} groups. It is the acidic and neutral compounds that bind primarily to albumin, which constitutes the majority of plasma proteins, so the literature values of protein binding (PB) refer mainly to the binding of drugs to HSA.

4. Materials and Methods

4.1. Molecular Descriptors

All tested drugs are listed in Supplementary Materials, along with molecular descriptors. Active pharmaceutical ingredients were extracted from pharmaceutical formulations, purchased in a generally accessible pharmacy. The main criterion used in composing the drug set was the availability of protein binding values (PB) along with milk-to-plasma ratios for each API, as these were the main pharmacokinetic phenomena studied.

The molecular descriptors selected for statistical analyses, which should have a significant effect on the penetration into breast milk and protein binding, are listed in Table 1. Some were taken from the literature, including M/P ratio obtained in vivo [8–13] or from online databases DrugBank [7] and CHEMBL [3]. Most of the physicochemical data were calculated in the following programs: HyperChem (HyperChem for Windows version 7.02, HyperCube Inc, Gainesville, FL, USA, 2002) and ACD/Labs (ACD/LabsTM Log D Suite 8.0, pKa dB 7.0, Advanced Chemistry Development Inc., Toronto, Canada, 2004).

Chromatographic descriptors were obtained in experiments, thin layer chromatography in normal (NP TLC) and reversed mode (RP TLC). The stationary phase was modified with bovine serum albumin (BSA). TLC was the source of retention factor (R_f) values,

denoted in statistical models as NP and RP. High performance liquid chromatography was performed using immobilized human serum albumin column ($HPLC_{HSA}$) and immobilized artificial membrane ($HPLC_{IAM}$). HPLC was the source of the log k values (logarithm of retention factor), log k_{HSA} and log k_{IAM}. The TLC and HPLC experiments are detailed in Appendix B.

4.2. Statistical Analyses

DFA, PCA and CA were performed in STATISTICA 13.1 (TIBCO Software Inc., Palo Alto, CA, USA). DFA is a classification analysis determining which descriptors best define the assignment of individual cases to each of the predetermined groups. Wilks' lambda is a parameter used to evaluate the discriminant power of the entire model, i.e., all the independent variables used, and takes values from 0 to 1; the closer these values are to zero, the more discriminatory the model becomes.

PCA is used to combine highly correlated variables with one another into one new variable called the principal component (PC). The calculation of new factors consists in diagonalizing the correlation or covariance matrix. The choice of matrix depends on whether the original variables require standardization or centering to mean values. In this way, a reduced number of new variables is generated, but explaining the original variance as much as possible.

The purpose of cluster analysis (CA) is to combine cases into groups so that the association within the same group is as large as possible, and with cases from other groups as small as possible. The method of grouping the data used in the presented studies was the k-means method, in which the means for each cluster and in each dimension are examined, which allows assessment of to what extent the created clusters are different from each other. In the analysis of variance, the size of the F statistic performed in each of them shows how well a given dimension separates individual clusters. In the best situation, very different means are obtained for most of the dimensions analyzed.

PLS and RF regression were performed with MATLAB ver. 2019a (The MathWorks, Natick, MA, USA). The performance of the models was assessed by a double cross-validation. The statistical significance was then evaluated using permutation testing.

In the PLS method, the matrix of independent variables is analyzed for latent variables (LVs) that best describe the covariance between X and Y. Then these transformed independent variables are used in regression to predict the Y response. The RF method uses many decision trees which, based on the entered X variables, repeatedly "make a decision" about the predicted value of Y for each case, from which the mean value is then taken.

In regression analyses, it is good practice to divide the set of cases into two sets: training and testing, in order to perform external validation, which will demonstrate the predictive capacity of the model. The training set accounts for approximately 70% of all collected cases and is used to build a regression equation (training model). The rest, i.e., about 30% of cases, are included in the test set on which the equation is validated. The training and test sets are distributed randomly. In order to check the stability of the model and exclude random effects, it is worth carrying out such a division into two subsets and the construction of the equation several times. The Monte Carlo permutation test (MCPT) is used for this. For the training and test sets, RF regression was performed and RMSECV, Q^2 and R^2 were calculated. Then this procedure was repeated 100 times, each time the training and test sets were drawn anew. Furthermore, the distribution of Q^2 in the original and permuted models was compared and a one-way ANOVA was performed. In the next step, 100 training (70%) and test sets (30%) were prepared by randomly splitting the original data matrix. A similar MCPT (100 perm.) was then performed on the training and test sets that were derived from the permuted data matrix. The results of the original and permuted models were obtained and their Q^2 values were compared.

5. Conclusions

Positive results were obtained on the expediency of using chromatographic data in the study of protein binding and the penetration of drugs into breast milk. The presented statistical analyses showed a close relationship between HPLC and TLC analytical data (under set conditions) with the bioavailability of the drug into breast milk. The correlation of the PB and M/P ratios with these chromatographic data is high, also in the group of all cases (acidic, basic and neutral drugs) together. The most effective application of NP TLC and $HPLC_{HSA}$ data was found. There is also a greater correlation between PB and the chromatographic data in the group of acidic drugs (a), i.e., for specific binding to albumin.

The PCA and DFA analyses identified a group of physicochemical and chromatographic parameters important for the bioavailability of drugs in breast milk. The use of CA emphasized the differentiation of their mean values in groups M/P_{code} 1–4.

NP TLC was proved to be the most useful chromatographic method in statistical analyses. In the case of $HPLC_{HSA}$ data, the relatively large share of the results from the column in the creation of the RF model turned out to be interesting. The second factor that emerges in almost all analyses is the high proportion of the log D parameter, i.e., lipophilicity associated with ionization.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/molecules27113441/s1, Tables S1–S9 contain all data used in statistical analyses.

Author Contributions: Conceptualization, E.B. and K.W.; methodology, E.B.; software, E.B and K.W.; validation, E.B.; formal analysis, K.W.; investigation, E.B.; resources, K.W.; data curation, E.B and K.W.; writing—original draft preparation, K.W.; writing—review and editing, E.B.; visualization, K.W.; supervision, E.B.; project administration, E.B.; funding acquisition, E.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by internal grant from Medical University of Lodz number 503/3-016-03/503-31-001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Chromatographic data from TLC and HPLC experiments and their derivatives used in the analysis of analytical models.

Descriptor	n_{abn}	n_b	n_n	n_a	PB_{abn} *	PB_b *	PB_n *	PB_a *
NP	162	49	79	34	**0.31**	**0.31**	0.15	**0.50**
NP/C	162	49	79	34	0.00	−0.11	−0.02	**0.50**
NP/PSA	162	49	79	34	0.19	**0.28**	0.17	**0.37**
NP/B2	162	49	79	34	−0.10	0.02	0.18	**−0.69**
NP/log P	162	49	79	34	0.12	0.02	**−0.20**	**−0.44**
RP	162	49	79	34	0.01	−0.05	**−0.20**	0.17
RP/C	162	49	79	34	0.12	0.17	0.19	−0.10
RP/PSA	162	49	79	34	0.11	**0.21**	0.11	−0.03
RP/B2	162	49	79	34	−0.08	0.02	0.07	**−0.44**
RP/log P	162	49	79	34	0.12	0.09	0.18	0.16
log k_{HSA}	165	49	80	34	**0.39**	**0.28**	**0.45**	**0.55**
log k_{HSA}/B2	165	49	80	36	0.01	0.05	−0.04	0.08

Table A1. *Cont.*

Descriptor	n_{abn}	n_b	n_n	n_a	PB_{abn} *	PB_b *	PB_n *	PB_a *
log k_{HSA}/log P	165	49	80	36	−0.11	−0.04	−0.16	0.09
log k_{HSA}/PSA	165	49	80	36	0.16	0.11	**0.25**	**0.51**
log k_{IAM}	159	49	74	36	**0.20**	0.17	**0.41**	**0.28**
log k_{IAM}/PSA	159	49	74	36	−0.05	−0.04	0.07	−0.07
log k_{IAM}/log P	159	49	74	36	0.04	0.11	−0.05	−0.04
log k_{IAM}/B2	159	49	74	36	−0.03	−0.06	−0.04	−0.06

* correlation with chromatographic data.

Table A2. Physicochemical parameters of APIs and their correlation with data on PB.

Descriptor	n_{abn}	n_b	n_n	n_a	PB_{abn} *	PB_b *	PB_n *	PB_a *
acid/base	166				−0.15			
B1	129	34	66	29	**0.28**	**0.36**	**0.48**	0.13
B2	166	50	81	35	0.12	0.13	**0.27**	0.05
B3	166	50	81	35	0.13	0.11	**0.21**	0.05
log U/D	160	50	75	35	0.05	0.16	0.02	**0.22**
DM	160	47	79	34	−0.02	0.04	−0.04	−0.16
Sa/V	160	47	79	34	**−0.29**	**−0.34**	**−0.32**	−0.04
eH	160	47	79	34	0.05	0.13	−0.02	0.17
MW	162	48	79	35	−0.17	0.14	0.24	0.00
HD	166	50	81	35	**−0.23**	−0.07	**−0.39**	**−0.23**
HA	166	50	81	35	−0.14	−0.16	**−0.23**	−0.13
eL	160	47	79	35	0.03	0.14	0.00	−0.015
eH-eL	160	50	79	35	0.01	−0.08	−0.01	0.12
log P	160	49	79	35	**0.31**	0.10	**0.34**	**0.41**
log D	160	50	81	35	**0.28**	0.19	**0.38**	**0.30**
MW/V	160	47	79	35	0.03	0.18	0.09	0.03
PhCharge	165	50	80	35	−0.13	−0.05	0.06	**−0.20**
pKa	160	50	75	35	−0.05	−0.15	0.08	**0.22**
M/P	104	30	55	19	**−0.29**	**−0.20**	**−0.35**	0.11
CNS+/−	154	49	72	33	−0.18	−0.05	0.16	**0.33**

* correlation with physicochemical data.

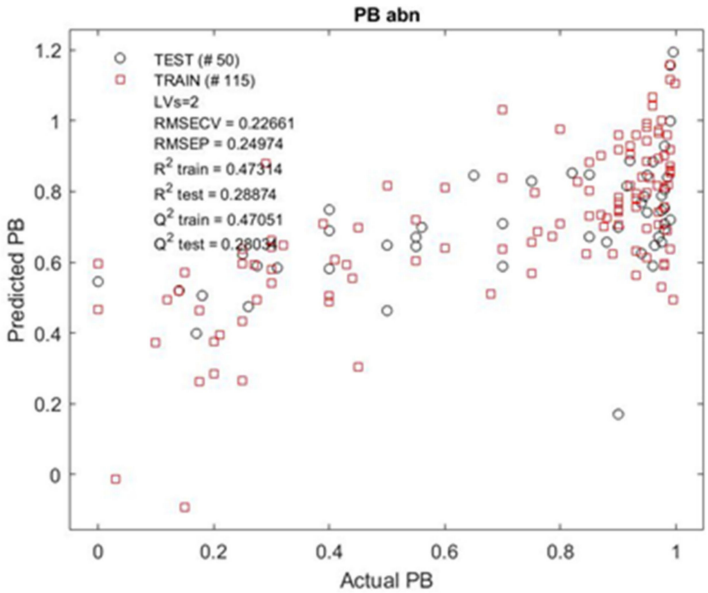

Figure A1. Actual versus predicted PB_{abn} values using PLS modelling and 23 molecular descriptors including NP TLC data. LVs = latent variables, $RMSE_{CV}$ = root-mean-square error of cross-validation, $RMSE_P$ = root-mean-square error of prediction, R^2 train/test = coefficient of determination for train/test set models, Q^2 train/test = coefficient of determination for the cross-validated models.

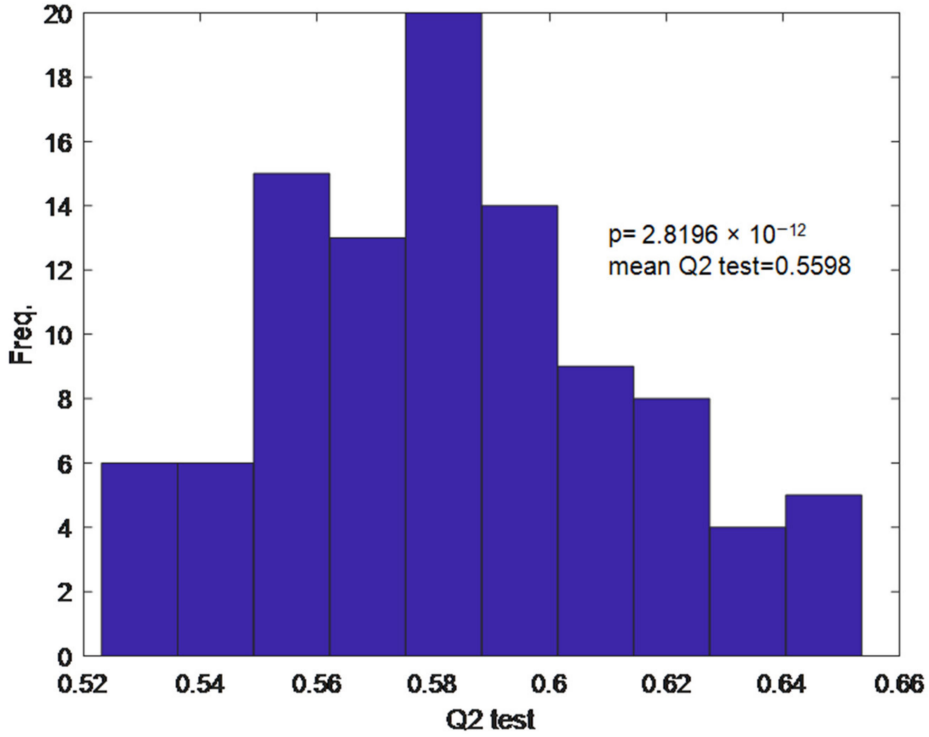

Figure A2. Monte Carlo permutation test (MCPT) showing Q^2 obtained from RF regression models developed on the test set, the number of repetitions was $n = 100$. The mean value of Q^2 was 0.5598 at the significance level $p = 2.8196 \times 10^{-12}$.

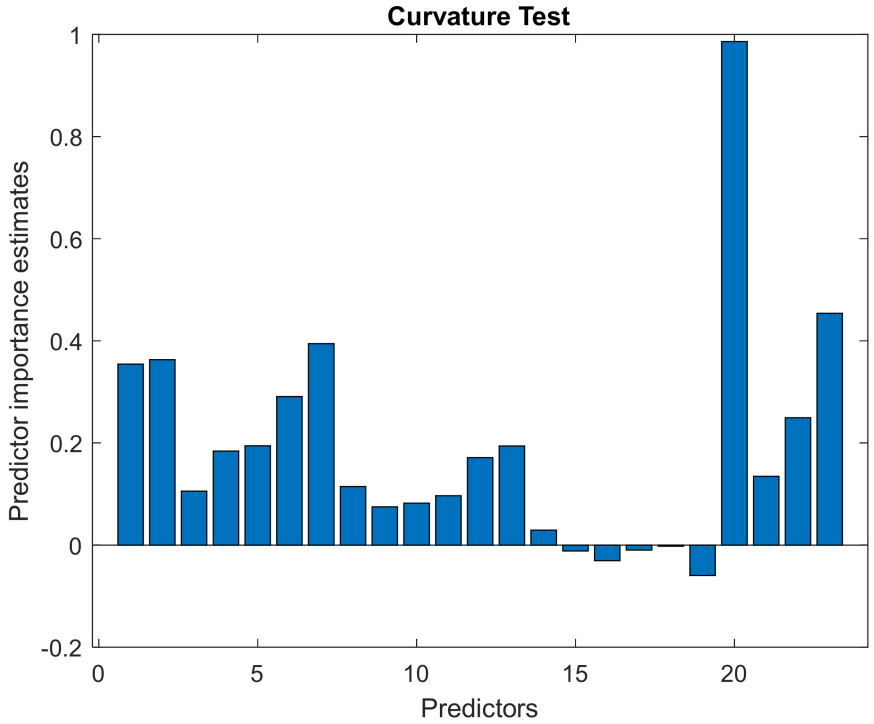

Figure A3. Contribution of individual descriptors to the generation of the RF regression model for PB_{abn}. The greatest influence is shown by the descriptor no. 20, i.e., log D.

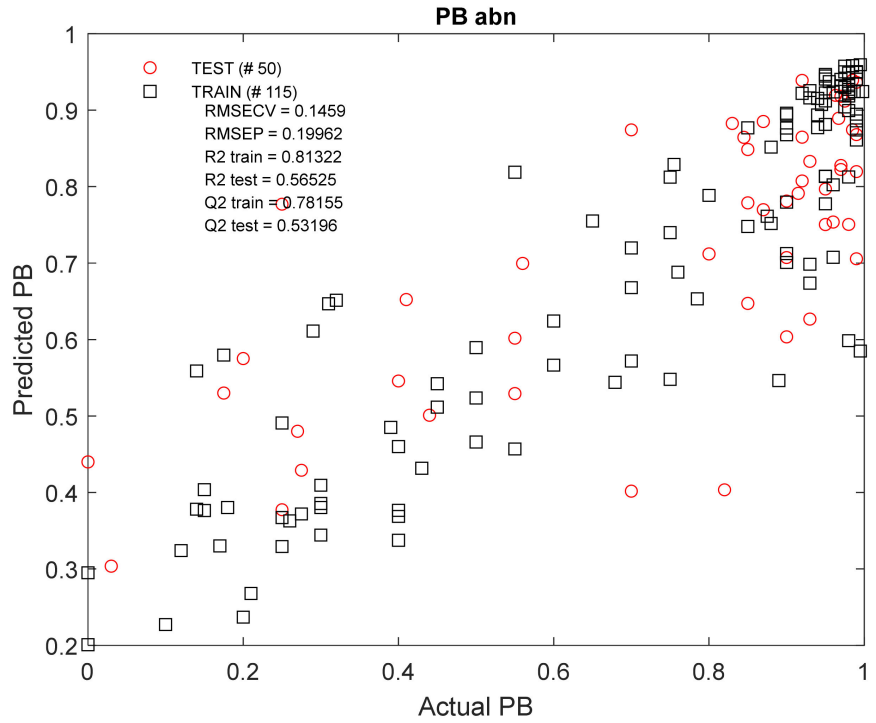

Figure A4. Actual versus predicted PB_{abn} values, using RF regression modelling of molecular descriptor set containing 22 variables along with $HPLC_{HSA}$ data. $RMSE_{CV}$ = root-mean-square error of cross-validation, $RMSE_P$ = root-mean-square error of prediction, R^2 train/test = coefficient of determination for train/test set models, Q^2 train/test = coefficient of determination for the cross-validated models.

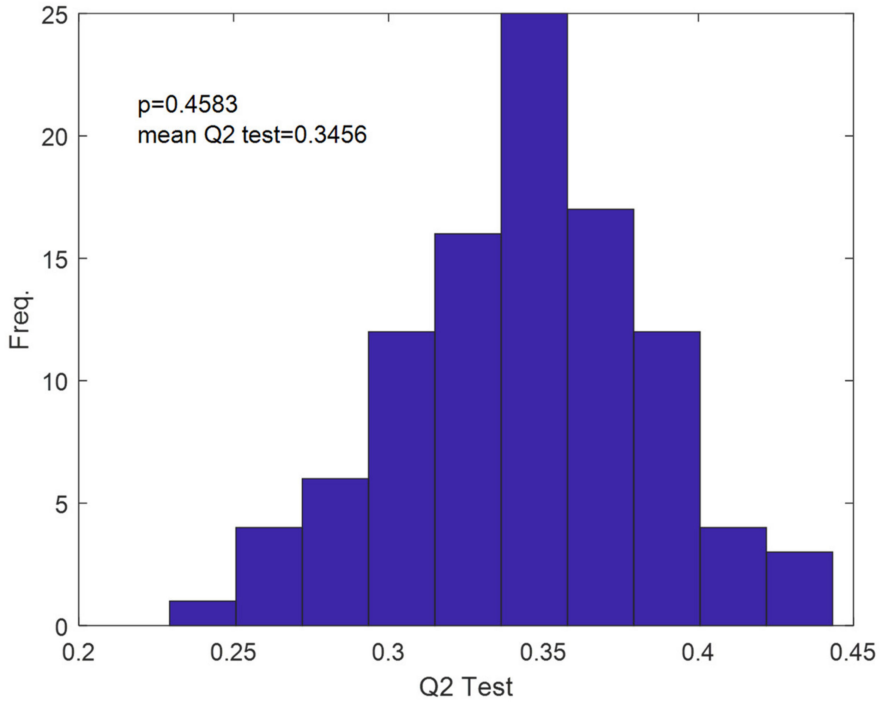

Figure A5. Monte Carlo permutation test (MCPT) showing Q^2 obtained from RF regression models developed on the test set, the number of repetitions was n = 100. The mean value of Q^2 was 0.3456 at the significance level p = 0.4583.

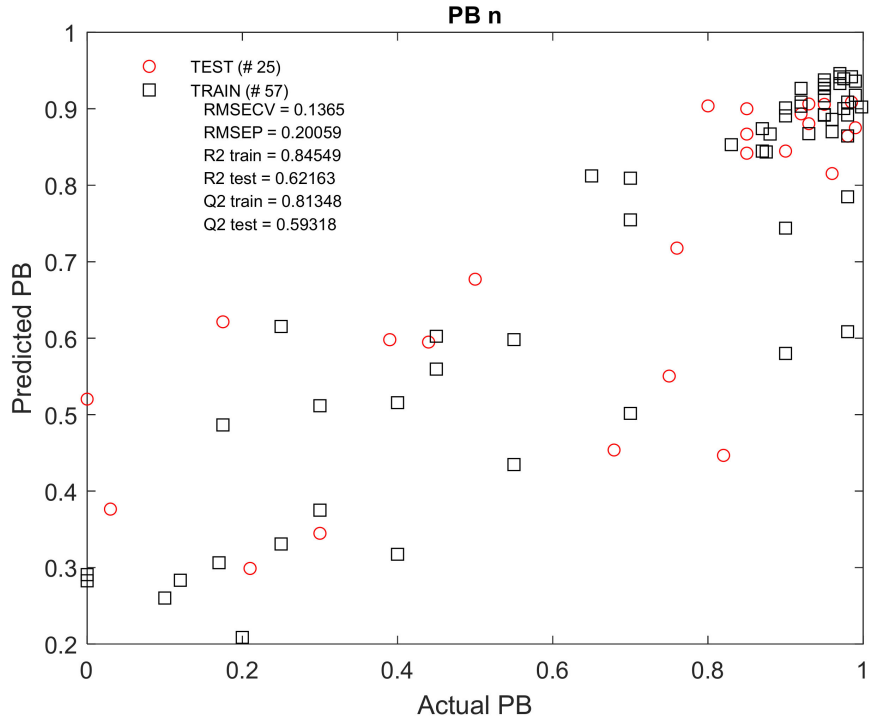

Figure A6. Actual versus predicted PB_n values using RF regression modelling of molecular descriptor set containing 23 variables along with NP TLC data. $RMSE_{CV}$ = root-mean-square error of cross-validation, $RMSE_P$ = root-mean-square error of prediction, R^2 train/test = coefficient of determination for train/test set models, Q^2 train/test = coefficient of determination for the cross-validated models.

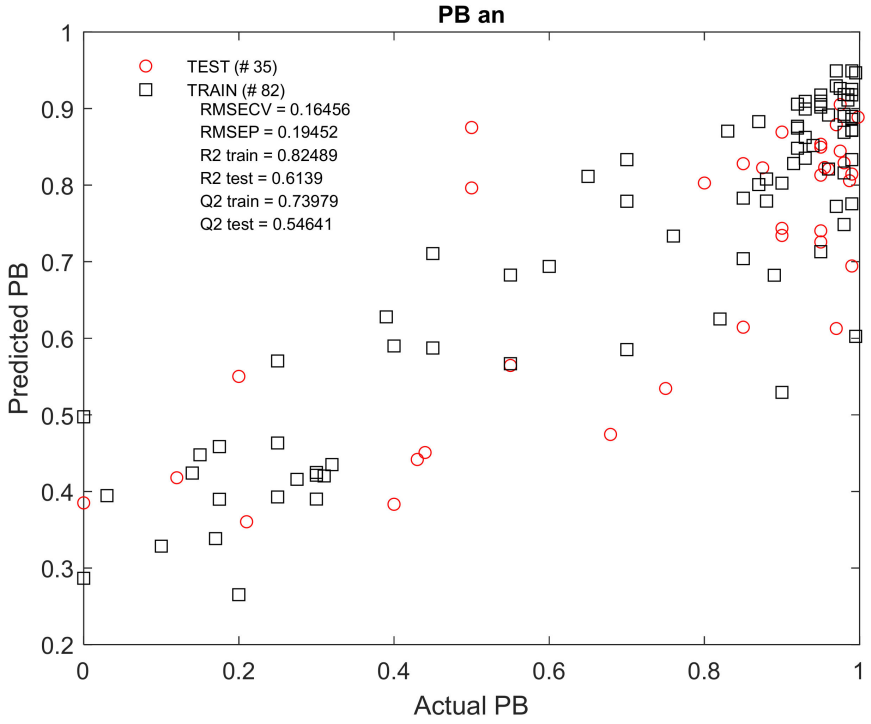

Figure A7. Actual versus predicted PB_{an} values, using RF regression modelling of molecular descriptor set containing 23 variables along with NP TLC data.

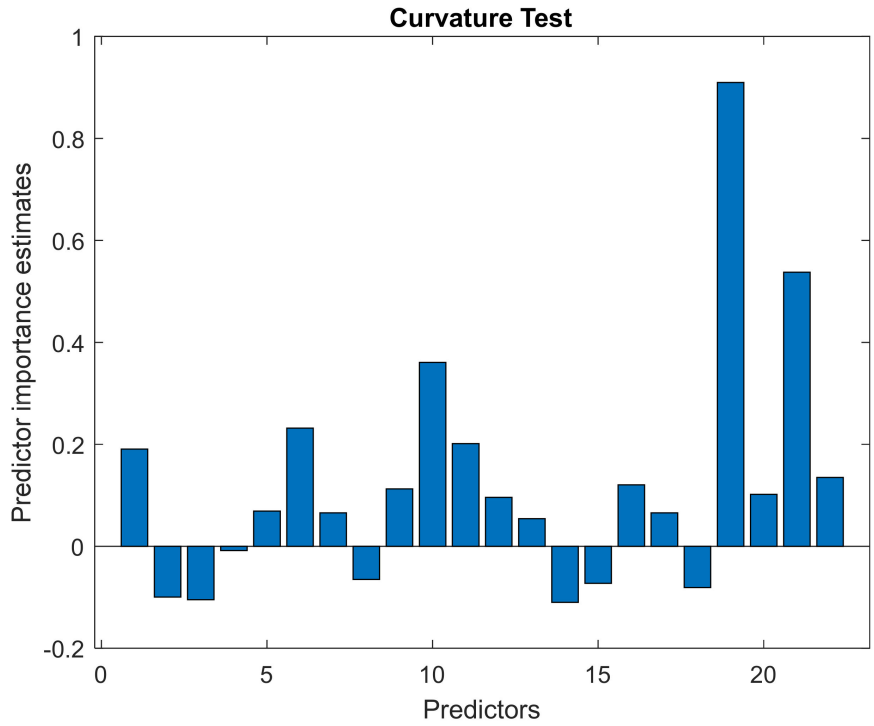

Figure A8. Contribution of individual descriptors to the development of the RF regression model for PB_n. The greatest influence is shown by the descriptor no. 19, i.e., log D, besides this, the molar weight (MW) and molar volume (V) are important.

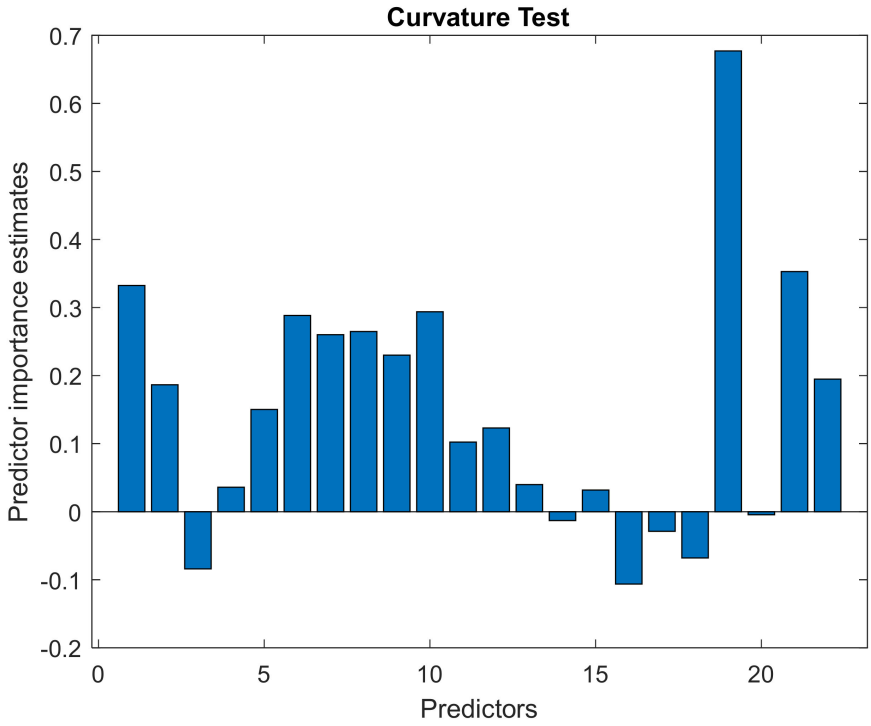

Figure A9. Contribution of individual descriptors to the development of the RF regression model for PB_{an}. The greatest influence is shown by the descriptor no. 19, i.e., log D, in this case a greater share of chromatographic parameters can be seen (descriptors nos. 6–9).

Appendix B Chromatographic Experiments

Appendix B.1 Materials and Reagents

For TLC chromatography, glass plates 20 × 20 cm from Merck, covered with silica gel with the addition of a fluorescent indicator, were used. Normal phase (NP) plates were used with standard Merck TLC Silica gel 60 F254 plates, while in reverse phase (RP) silanized plates RP-2: Merck TLC Silica gel 60 RP-2 F254 were used.

Solvents from J.T. Baker-Water, Methanol and Acetonitrile, with an HPLC gradient grade. Ammonium acetate p.a. was used to prepare an acetate buffer at pH 7.4.

The stationary phase of the plates, both NP and RP, was modified with an aqueous solution of bovine serum albumin purchased from Sigma Aldrich (bovine serum albumin, lyophilized powder).

Human serum albumin immobilized chromatography column was from Daicel: CHIRALPAK®HSA, 5 µm; 4 × 10 mm while column with IAM artificial membrane from Regis Technologies Inc.: IAM.PC.DD.2, 10 µm; 4.6 × 10 mm.

In HPLC chromatography, the organic solvents used (acetonitrile and methanol) and water were also obtained from J.T. Baker (HPLC gradient). LACH-NER ammonium acetate, ammonium acetate p.a. were used to prepare the acetate buffer (HPLCHSA), while to prepare the phosphate buffer (HPLCIAM) a ready-made reagent in the form of tablets (Sigma, Phosphate buffered saline, tablets)was used to be dissolved in a strictly defined amount of water for HPLC.

Appendix B.2 Isolation of Active Pharmaceutical Ingredients (APIs)

A total of 167 active pharmaceutical ingredients (APIs), isolated from pharmaceutical preparations, usually tablets or hard capsules, were used in the chromatographic experiments. Tablets (without coatings) or the contents of capsules crushed in a mortar were placed in 100 mL of 99.8% methanol, mixed with a magnetic stirrer for approximately 30 min and then passed to crystallization tanks through a funnel with a filter. The vessel with the filtrate was allowed to evaporate the solvent and the crystallized active substance was transferred to sealed vials, kept under refrigerated conditions.

The purity of the isolated substances was checked by TLC chromatography and densitometric scanning. All substances isolated gave single densitometric peaks and were used without further purification. The obtained API was dissolved in 99.8% methanol to give 1 mg/mL solutions which were then used in TLC and HPLC.

Appendix B.3 Impregnation of TLC Plates

The surface-modifying protein of the stationary phase of thin-layer chromatography plates was bovine serum albumin (BSA), which is a cheaper substitute for human albumin, with 76% homology and similar drug binding properties [14–18].

The impregnation of the plates was carried out with a 2 mg/mL solution applied to the surface using a Desaga SG 1 hand sprayer; the plates were then air dried. The best concentration was selected earlier—on NP plates impregnated with 1, 2 and 4 mg/mL BSA solutions, active substances were applied at a concentration of 1 mg/mL (solutions in 99.8% methanol), characterized by a different degree of protein binding described in the literature. Retention values differed significantly between plates coated with 1 and 2 mg/mL BSA, but no difference was found between 2 and 4 mg/mL. Therefore, it was decided to use a ratio of 1:2, drug concentration to BSA concentration on the plate.

Appendix B.4 TLC Chromatography

Normal and reversed phase thin layer chromatography (NP TLC and RP TLC respectively) was performed using silica-gel-coated glass plates. Half of them were covered with 2 mg/mL bovine serum albumin solution and half remained pure.

Solutions of the isolated APIs in 99.8% methanol (1 mg/mL) were applied to the plates using a Desaga HPTLC-Applicator AS 30 automatic applicator. The mobility of the

compounds was also determined on the plates with no protein as a modifier. They have been marked as controls (C) and will allow evaluation of the influence of the modifier on API mobility. The plates were then developed in a mobile phase consisting of acetonitrile, acetate buffer pH 7.4 and methanol in the ratio 60:20:20 (*v/v/v*). The acetate buffer (20 mM) was prepared by dissolving 1.54 g of ammonium acetate in 1 L of distilled water. The pH was then adjusted with a concentrated ammonia solution using a pH meter. The plates were developed in standard, vertical chromatographic chambers, each time using 100 mL of the mobile phase, after the chamber was previously saturated with solvent vapors for approximately 1 h.

The unfolded-protein-impregnated plates and the control plates were scanned with a Desaga CD 60 densitometer. The values of the delay factor (R_f) were collected, i.e., the ratio of the distance traveled by the substance to be analyzed to the distance traveled through the front of the mobile phase. The analytical wavelengths were selected individually for each API using the multi-wavelength scanning option (values ranged from 200 to 300 nm). The experiment was repeated (for both BSA-coated and control plates) and the R_f values pooled are the mean of both series of experiments.

Appendix B.5 HPLCHSA Chromatography

High performance liquid chromatography was performed using a chromatography column with immobilized human serum albumin. The assay was performed on a Perkin Elmer Series 200 instrument connected to a UV-VIS spectrometer as detector. The analytical wavelength was the same for all compounds at 210 nm. The experiment was carried out with the 1 mg/mL methanolic solutions of active substances previously described. The mobile phase was a mixture of 10 mM acetate buffer pH 7.4, acetonitrile and methanol in the ratio 85:10:5 (*v/v/v*). The acetate buffer was prepared by dissolving 0.77 g of ammonium acetate in 1 L of distilled water. The pH was then adjusted with a concentrated ammonia solution using a pH meter.

The phase flow through the system was set to 0.9 mL/min as recommended by the column manufacturer. The solutions were delivered to the column using an autosampler syringe, the injection size was 10 µL. Since the column could not be thermostated, the room was kept at a constant temperature of 25 degrees Celsius.

Chromatographic data (retention coefficient, k, and derivative, log k) were obtained with TotalChrom software connected to an HPLC instrument. The k coefficient, which is the ratio between the amount of analyte in the stationary phase and its amount in the mobile phase, was obtained from the equation k = (tR - tM)/tM, where tR is the retention time of the analyzed substance and tM is the dead time (the dead time marker was 99.8% methanol). The experiment was then repeated and the collected retention rates were the mean values of both series.

Appendix B.6 HPLCIAM Chromatography

The second experiment was performed using an immobilized artificial membrane (IAM) column. The assay was also performed on a Perkin Elmer Series 200 instrument connected to a UV-VIS spectrometer as the detector. The analytical wavelength was the same for all compounds, at 210 nm. The experiment was carried out with the 1 mg/mL methanolic solutions of active substances previously described. The mobile phase was a mixture of 10 mM phosphate buffer pH 7.4 and acetonitrile in the ratio 80:20 (*v/v*). The phosphate buffer was obtained by dissolving the finished tablet in the appropriate amount of distilled water (1 tablet per 200 mL). In this case, it was not necessary to adjust the pH of the buffer using a pH meter.

The phase flow through the system was set to 0.5 mL/min as recommended by the column manufacturer. The solutions were delivered to the column using an autosampler syringe, the injection size was 10 µL.

The collected chromatographic data, similar to the HSA column experiment, was the retention coefficient, k, and derivative, log k, which were obtained using TotalChrom

software connected to the HPLC instrument. The experiment was then repeated and the collected retention rates were the mean values of both series.

References

1. Wanat, K.; Khakimov, B.; Brzezińska, E. Comparison of statistical methods for predicting penetration capacity of drugs into human breast milk using physicochemical, pharmacokinetic and chromatographic descriptors. *SAR QSAR Environ. Res.* **2020**, *31*, 457–475. [CrossRef] [PubMed]
2. Wanat, K.; Żydek, G.; Hekner, A.; Brzezińska, E. In silico plasma protein binding studies of selected group of drugs using TLC and HPLC retention data. *Pharmaceuticals* **2021**, *14*, 202. [CrossRef] [PubMed]
3. Chembl Database. Available online: https://www.ebi.ac.uk/chembl/ (accessed on 1 March 2022).
4. Norinder, U.; Haeberlein, M. Computational approaches to the prediction of the blood-brain distribution. *Adv. Drug Deliv. Rev.* **2002**, *54*, 291–313. [CrossRef]
5. Yang, F.; Zhang, Y.; Liang, H. Interactive association of drugs binding to human serum albumin. *Int. J. Mol. Sci.* **2014**, *15*, 3580–3595. [CrossRef] [PubMed]
6. Ozeki, S.; Tejima, K. Drug Interactions. II. Binding of Some Pyrazolone and Pyrazolidine Derivatives to Bovine Serum Albumin. *Chem. Pharm. Bull.* **1974**, *22*, 1297–1301. [CrossRef] [PubMed]
7. Drugbank. Available online: https://www.drugbank.ca/drugs/DB01174 (accessed on 29 March 2022).
8. Agatonovic-Kustrin, S.; Tucker, I.G.; Zecevic, M.; Zivanovic, L.J. Prediction of drug transfer into human milk from theoretically derived descriptors. *Anal. Chim. Acta* **2000**, *418*, 181–195. [CrossRef]
9. Hale, T.W. *Medications and Mother's Milk*, 15th ed; Pharmasoft Medical Publishing: Amarillo, TX, USA, 2012.
10. Katritzky, A.R.; Dobchev, D.A.; Hür, E.; Fara, D.C.; Karelson, M. QSAR treatment of drugs transfer into human breast milk. *Bioorg. Med. Chem.* **2005**, *13*, 1623–1632. [CrossRef] [PubMed]
11. Meskin, M.S.; Lien, E.J. QSAR analysis of drug excretion into human breast milk. *J. Clin. Pharm. Ther.* **1985**, *10*, 269–278. [CrossRef] [PubMed]
12. Wilson, J.T.; Brown, R.D.; Cherek, D.R.; Dailey, J.W.; Hilman, B.; Jobe, P.C.; Manno, B.R.; Manno, J.E.; Redetzki, H.M.; Stewart, J.J. Drug Excretion in Human Breast Milk: Principles, Pharmacokinetics and Projected Consequences. *Clin. Pharmacokinet.* **1980**, *5*, 1–66. [CrossRef] [PubMed]
13. Abraham, M.H.; Gil-Lostes, J.; Fatemi, M. Prediction of milk/plasma concentration ratios of drugs and environmental pollutants. *Eur. J. Med. Chem.* **2009**, *44*, 2452–2458. [CrossRef] [PubMed]
14. Tunç, S.; Çetinkaya, A.; Duman, O. Spectroscopic investigations of the interactions of tramadol hydrochloride and 5-azacytidine drugs with human serum albumin and human hemoglobin proteins. *J. Photochem. Photobiol. B Biol.* **2013**, *120*, 59–65. [CrossRef] [PubMed]
15. Carter, D.C.; Ho, J.X. Structure of serum albumin. *Adv. Protein Chem.* **1994**, *45*, 153–203. [CrossRef] [PubMed]
16. Ayranci, E.; Duman, O. Binding of fluoride, bromide and iodide to bovine serum albumin, studied with ion-selective electrodes. *Food Chem.* **2004**, *84*, 539–543. [CrossRef]
17. Ayranci, E.; Duman, O. Binding of Lead Ion to Bovine Serum Albumin Studied by Ion Selective Electrode. *Protein Pept. Lett.* **2004**, *11*, 331–337. [CrossRef] [PubMed]
18. Raoufinia, R.; Mota, A.; Keyhanvar, N.; Safari, F.; Shamekhi, S.; Abdolalizadeh, J. Overview of albumin and its purification methods. *Adv. Pharm. Bull.* **2016**, *6*, 495–507. [CrossRef] [PubMed]