Supporting information

Physicochemical parameters affecting Electrospray Ionization efficiency of amino acids after acylation

*Jos Hermans[1], Sara Ongay[1], Vadym Markov[2], Rainer Bischoff[1]**

[1]Analytical Biochemistry, Department of Pharmacy, University of Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands.

[2] Department of Chemical Metrology, Kharkov V.N. Karazin National University, Svoboda sq. 4, 61022 Kharkov, Ukraine.

*E-mail: r.p.h.bischoff@rug.nl. Phone +31(0)503633338. Fax 031(0)503637582.

Table of contents:

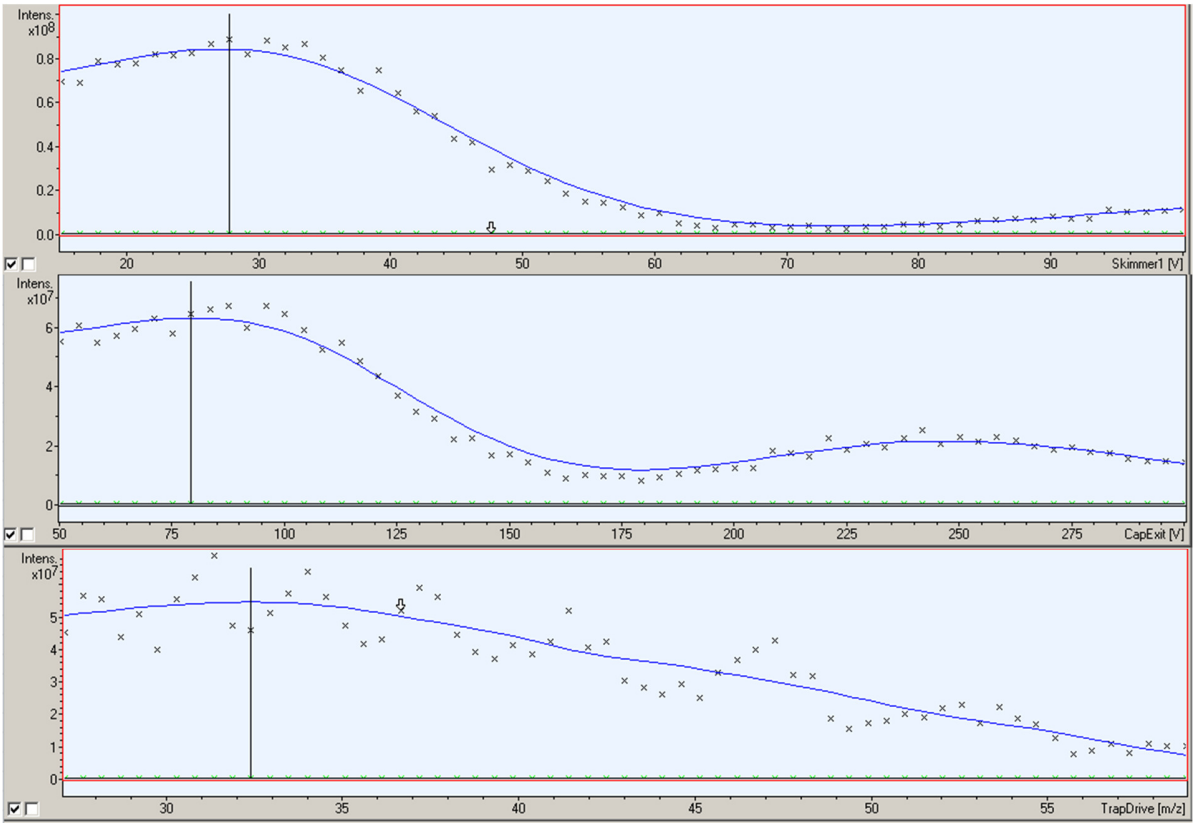

Figure S-1: Effect of instrumental parameters to ion transmission for m/z 118 (Valine). From top to bottom: Skimmer voltage, Capillary Exit voltage and Trap drive.

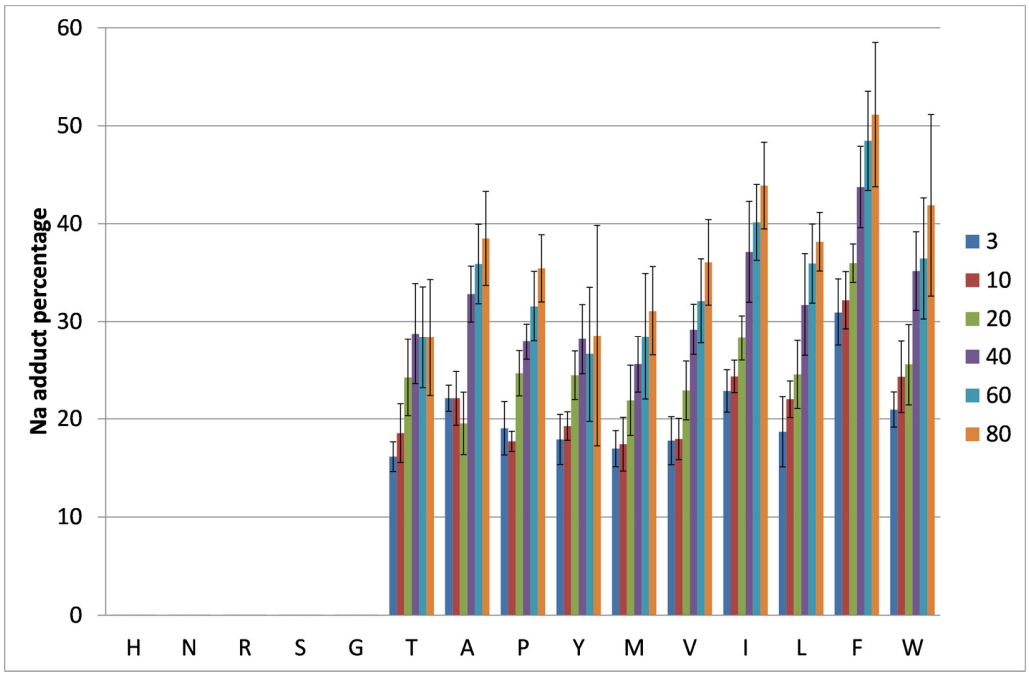

Figure S-2: Increasing formation of sodium adducts for PEG-labeled amino acids at increasing percentages of acetonitrile. The amino acid derivatives are ordered according to RPLC retention time. Analyses were performed by flow injection analysis (n=5) of collected fractions.

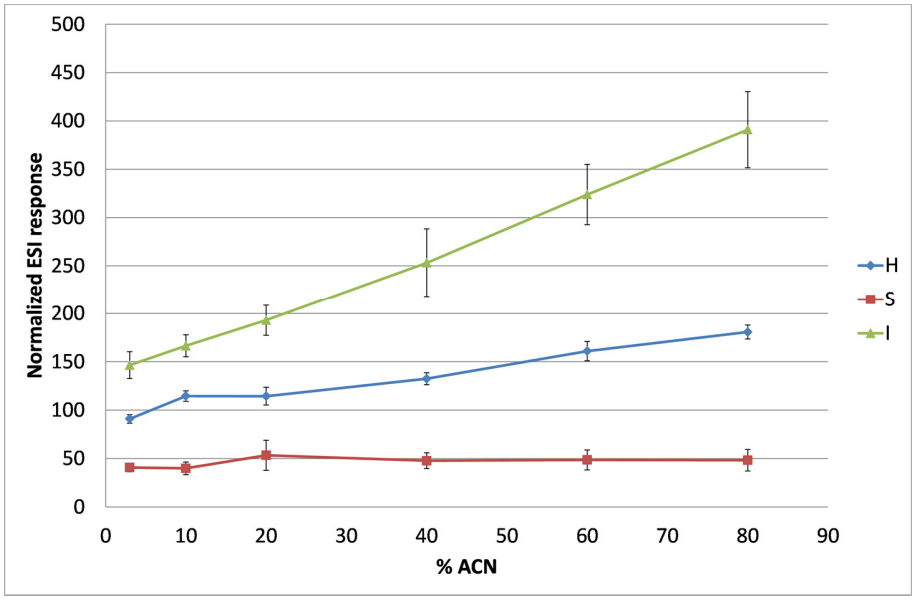

Figure S-3: Effect of acetonitrile concentration on the FIA electrospray response for 5µM PEG-labeled amino acids (n=3, adducts included).

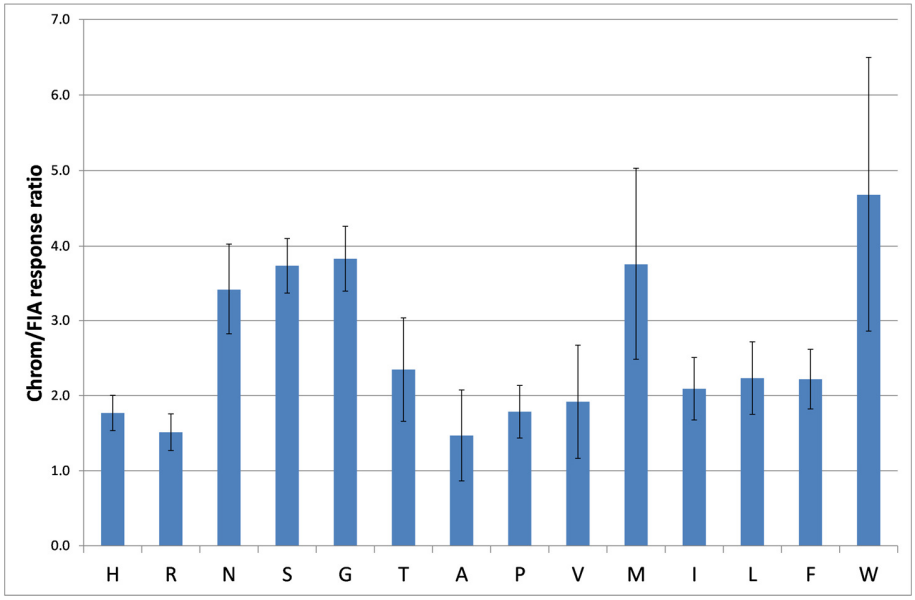

Figure S-4: Ratio of chromatographic and Flow injection ESI responses of PEG-derivatised amino acids after RPLC separation in an acetonitrile gradient to FIA at 20% acetonitrile.

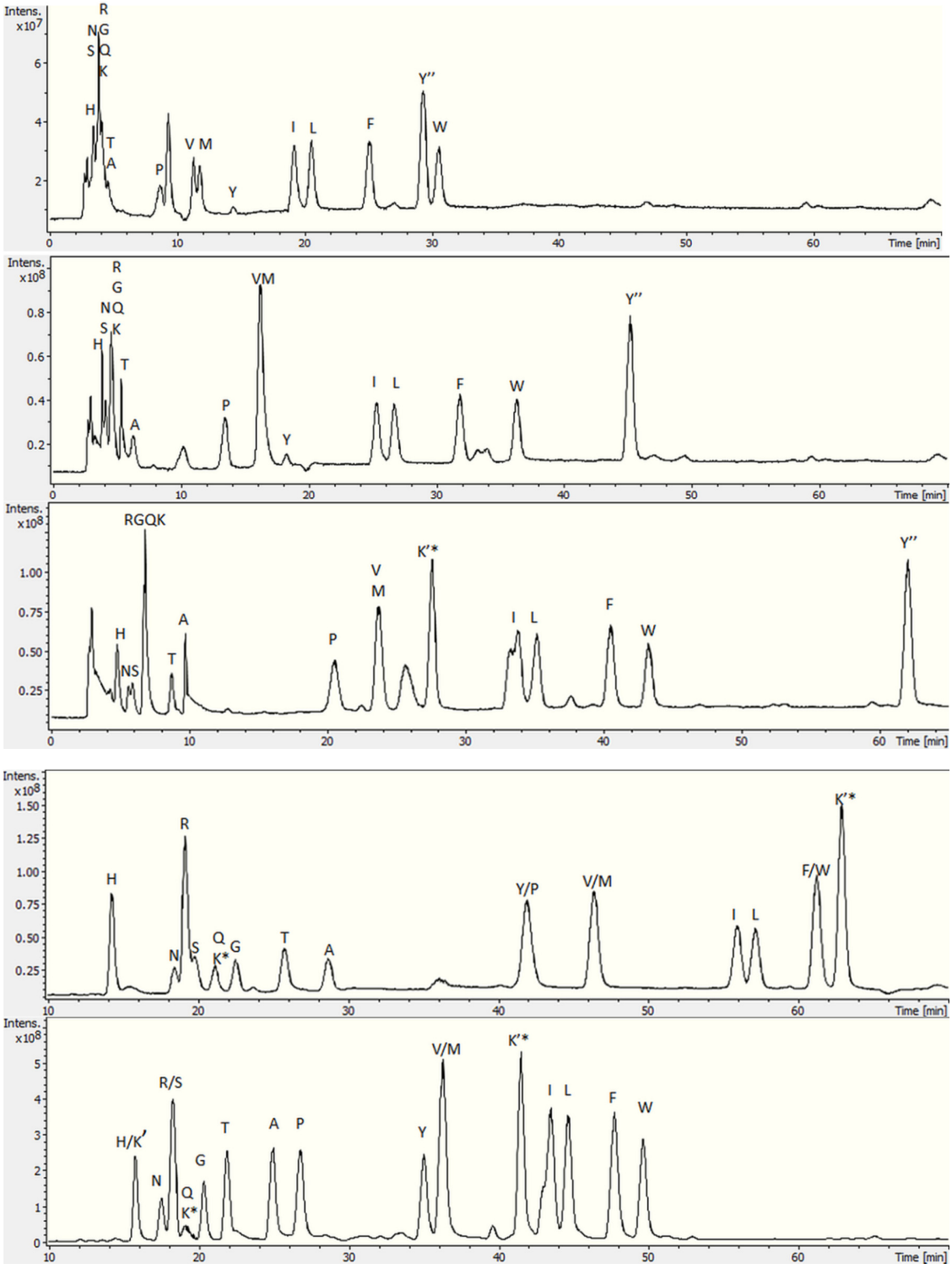

Figure S-5: Total Ion Chromatograms of 50pmol labeled amino acids, from top to bottom C2-, C3-, C4-, C6-acyl-labeled and PEG-labeled.

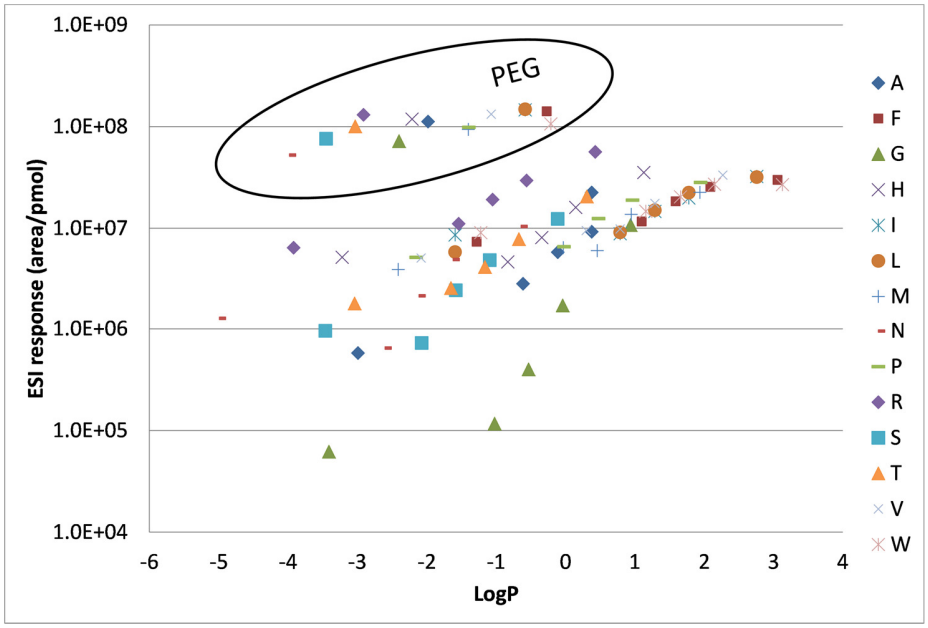

Figure S-6: Electrospray response in relation to the logP of derivatized amino acids with respect to the amino acid moiety.

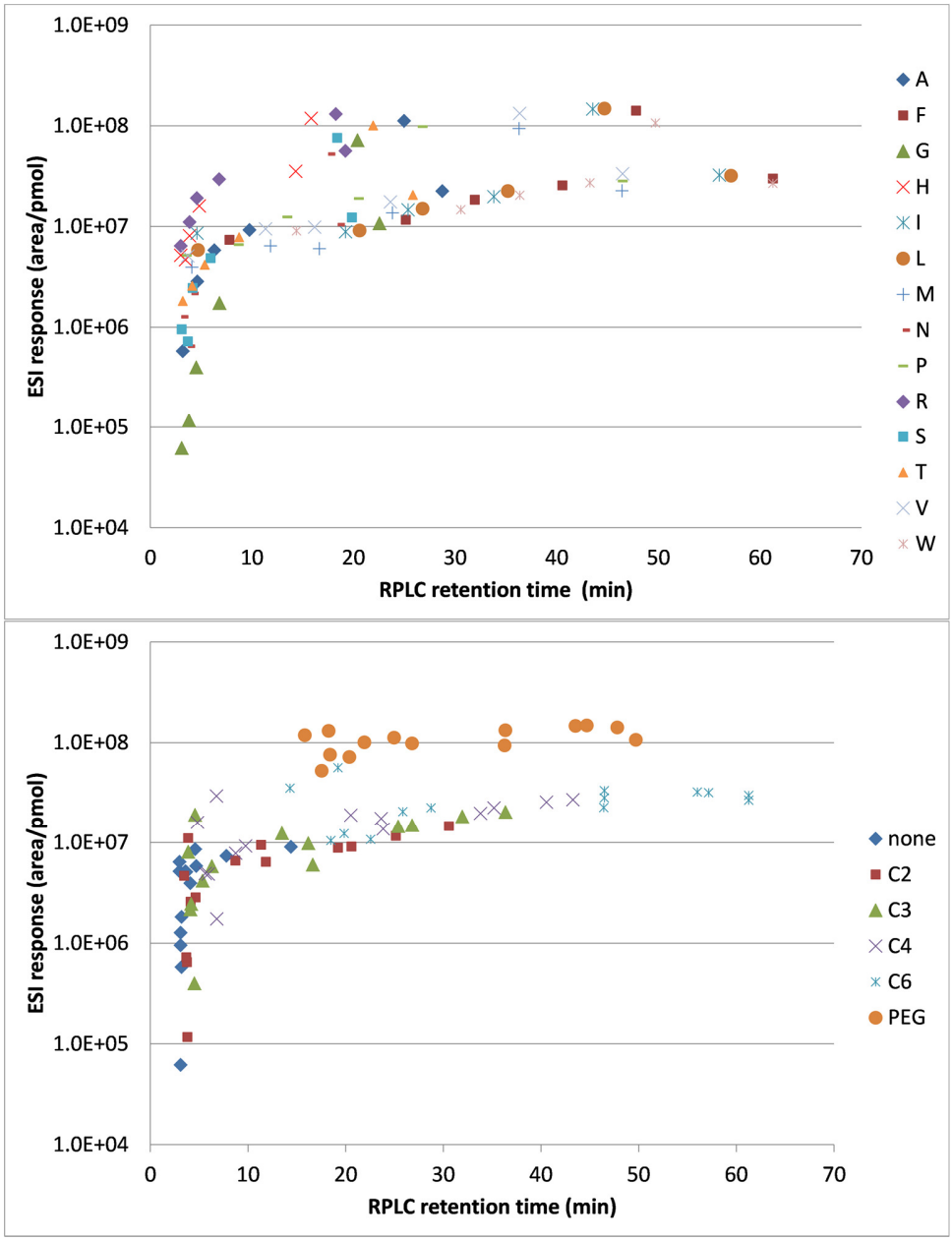

Figure S-7: Electrospray response in relation to the retention time specified per amino acid (top panel) or label (bottom panel).

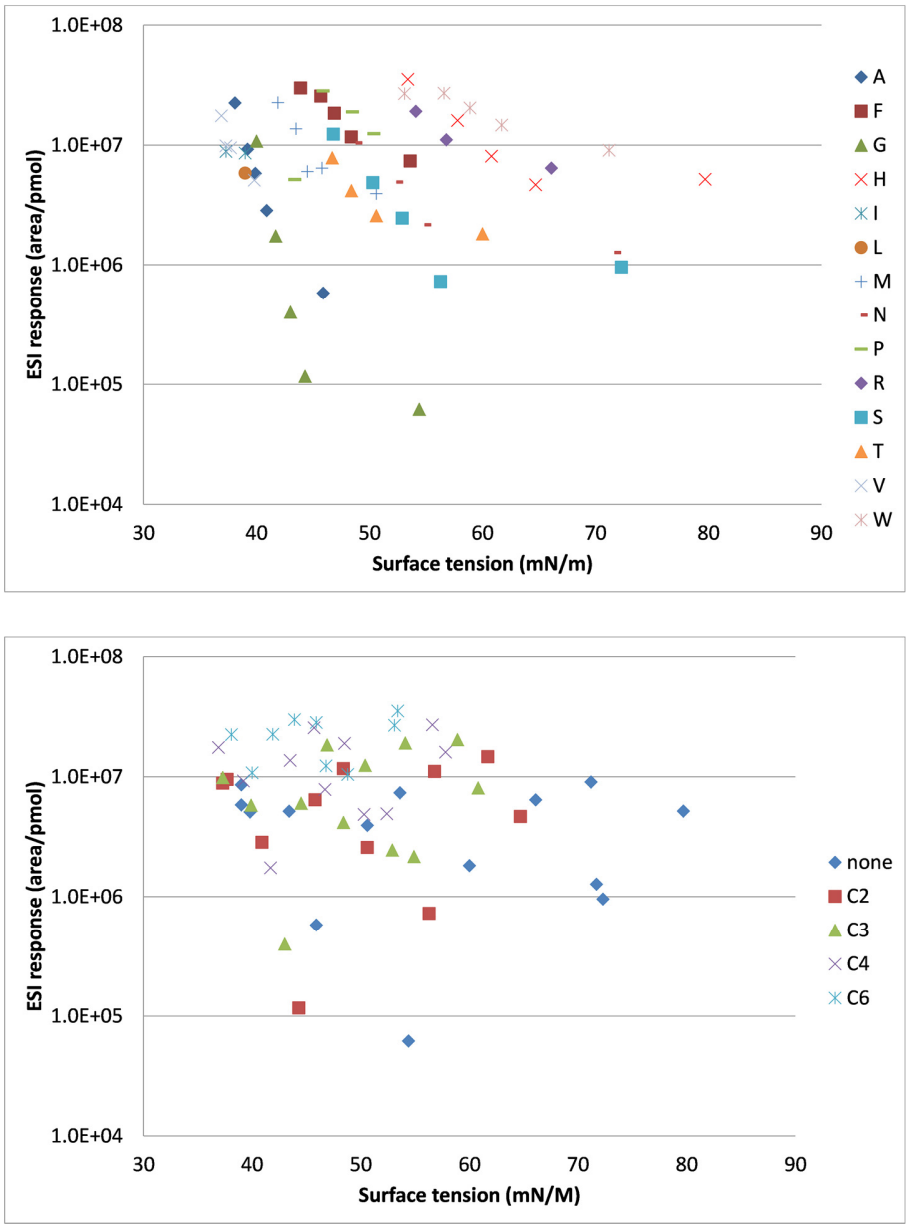

Figure S-8 : Electrospray response relation to the calculated surface tension specified per amino acid (top panel) or label (bottom panel). Please note that some surface tension values are not included, as they were unavailable from the Chemspider database (like for all PEG labeled amino acids)

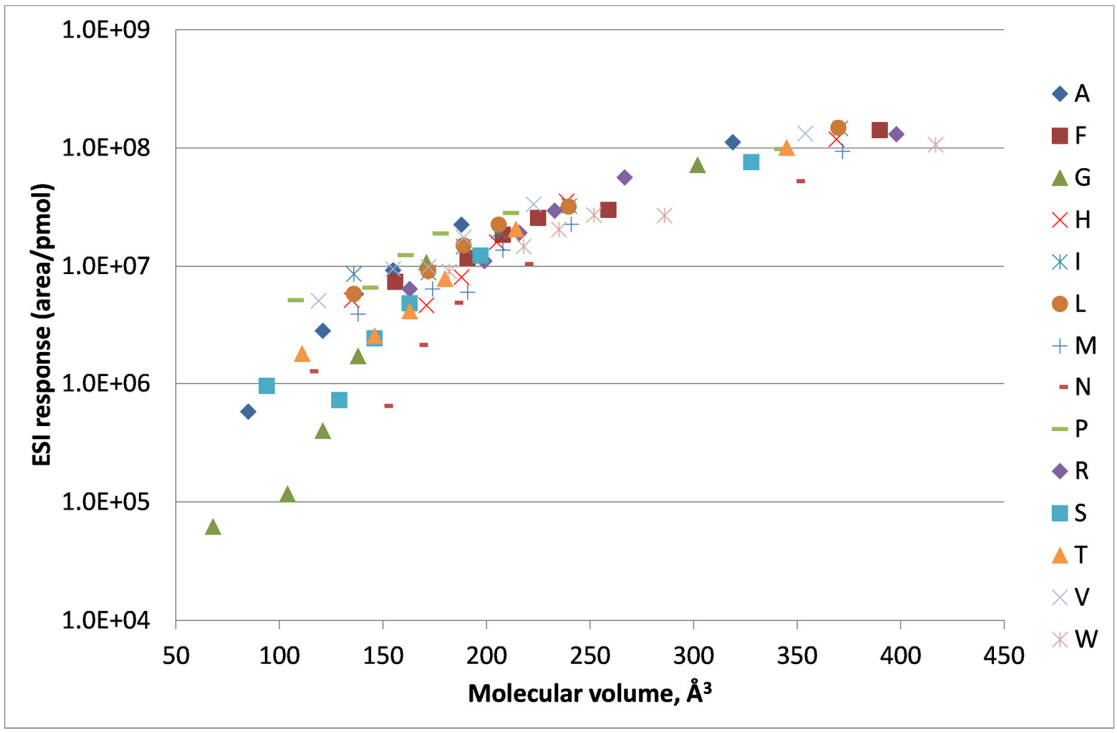

Figure S-9: Electrospray response in relation to the calculated molecular volume of (derivatized) amino acids specified per amino acid.

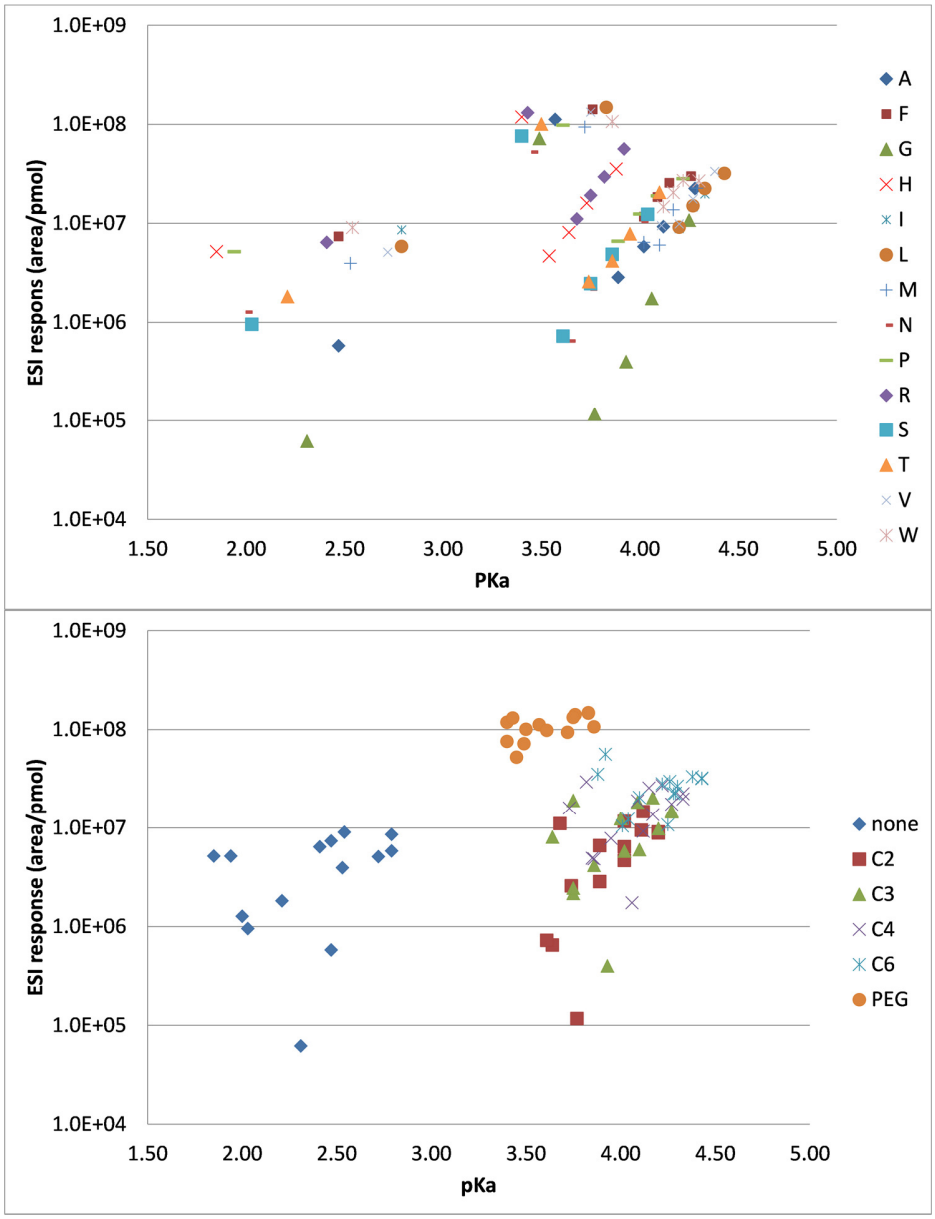

Figure S-10: Electrospray response in relation to the pKa of (derivatized) amino acids specified per amino acid (top panel) and label (bottom panel).

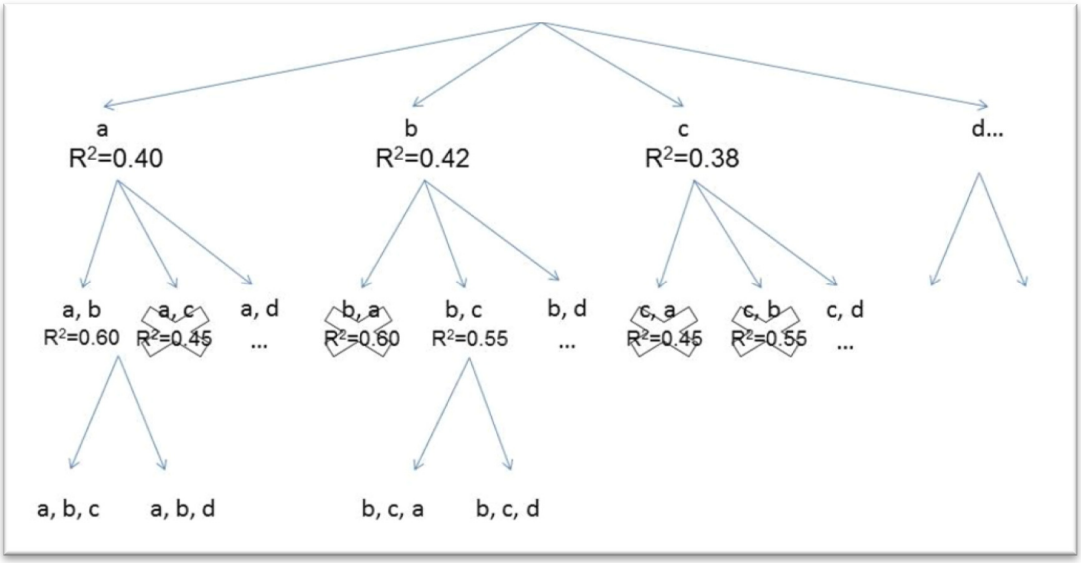

Figure S-11: Fuzzy forward selection procedure schema. Models (a,c) and (c,a) are dropped due to their low R^2 and (b,a) was considered equivalent to (a,b).

substance label	none	C2	C3	C4	C6	PEG
Alanine	+	+	+	+	+	+
Phenylalanine	+	+	+	+	+	+
Glycine	+	+	+	+	+	+
Histidine	+	+	+	+	+	+
Isoleucine	+	+	+	+	+	+
Leucine	+	+	+	+	+	+
Methionine	+	+	+	+	+	+
Aspargine	+	+	+	+	+	+
Proline	+	+	+	+	+	+
Arginine	+	+	+	+	+	+
Serine	+	+	+	+	+	+
Threonine	+	+	+	+	+	+
Valine	+	+	+	+	+	+
Tryptophan	+	+	+	+	+	+
2-amino-5-bromo-Benzoic acid	+	+	+	+	-	+
4-amino-Benzoic acid	+	+	+	+	+	+
Aniline	+	+	+	+	-	+
Cyclohexylamine	+	+	+	+	-	+
p-chloro-Anilne	+	+	+	+	-	+
p-nitro-Aniline	+	+	+	+	+	+
PhenylGlycine	+	+	+	+	+	+
p-Toluidine	+	+	+	+	-	+

Table S-1: Overview of all substances included/excluded (+/-) from the QSPR model for the training- (upper part) and validation- set (lower part).

	RT (min) R^2			Volume (Å³) R^2			Log P R^2			pKa R^2			Surface tension R^2		
	all	**-PEG**	**-PEG, -none**	**all**	**-PEG**	**-PEG, -none**	**all**	**-PEG**	**-PEG, -none**	**all**	**-PEG**	**-PEG, -none**	**all**	**-PEG**	**-PEG, -none**
A	0.71	0.62	0.84	0.89	0.98	0.99	0.08	0.90	0.82	0.30	0.86	1.00		0.95	1.00
F	0.53	0.90	0.79	0.98	0.94	0.86	0.01	0.92	0.86	0.17	0.71	0.89		0.94	0.93
G	0.83	0.78	0.80	0.88	0.94	0.99	0.07	0.78	0.99	0.19	0.57	1.00		0.63	1.00
H	0.88	0.79	0.77	0.91	0.84	0.99	0.01	0.63	0.99	0.14	0.33	1.00		0.60	0.99
I	0.54	0.92	0.91	0.98	0.90	0.97	0.00	0.73	0.97	0.04	0.41	0.99		1.00	
L	0.61	0.93	0.86	0.98	0.97	0.94	0.01	0.90	0.94	0.11	0.65	0.97			
M	0.66	0.91	0.87	0.96	0.91	0.89	0.01	0.79	0.89	0.11	0.51	0.88		0.76	0.86
N	0.76	0.65	0.61	0.87	0.71	0.92	0.01	0.50	0.92	0.09	0.24	0.96			
P	0.49	0.82	0.78	0.95	0.94	0.92	0.02	0.85	0.92	0.19	0.55	0.98			
R	0.81	0.73	0.80	0.89	0.99	0.98	0.09	0.90	0.98	0.25	0.64	1.00			
S	0.77	0.70	0.67	0.91	0.84	0.94	0.00	0.77	0.94	0.12	0.39	0.98		0.56	0.99
T	0.77	0.86	0.89	0.96	0.95	1.00	0.00	0.90	1.00	0.14	0.51	0.99			
V	0.69	0.95	0.95	0.97	0.97	0.95	0.02	0.87	0.95	0.10	0.62	0.91			
W	0.50	0.83	0.63	0.97	0.88	0.73	0.01	0.92	0.73	0.19	0.77	0.78		0.95	0.82
average	**0.68**	**0.81**	**0.80**	**0.94**	**0.91**	**0.93**	**0.02**	**0.81**	**0.92**	**0.15**	**0.56**	**0.95**		**0.80**	**0.94**
s	0.13	0.11	0.10	0.04	0.07	0.07	0.03	0.12	0.08	0.07	0.17	0.06		0.18	0.07
rsd	19	13	13	4	8	8	133	15	8	44	31	7		23	8

Table S-2: Correlation coefficients of log ESI response related to retention time, molecular volume, LogP, pKa and Surface tension for each amino acid taking all labels (first column), excluding the PEG labels (second column) and excluding the PEG as well as the non-labeled compounds. Note that some surface tension results are missing due to lack of surface tension data, especialt for the PEG labeled compounds.

QSPR modelling and fuzzy forward selection procedure

Molecule structure optimization

Molecule structures were imported from SMILES to Chemoffice ChemBio3D Ultra 12.0. At first, every structure was energy minimized by the MM2 algorithm available from this software package using default parameters. This method is fast and generates molecule structures eligible for the slow but accurate PM3 semi-empirical energy minimizen with an R-Closed shell wave function structure optimization applied hereafter. The resulted stuctures were saved as MDL Molfiles and should adequately present at least some molecule conformation.

Descriptor calculation

Descriptors were calculated by the DRAGON 5.5 software using MDL Molfiles as input resulting in a table with rows and columns corresponding to molecules and descriptors, respectively. In total 3224 descriptors were computed excluding unavailable descriptors.

Descriptor preprocessing

1672 descriptors were constant for every molecule in the dataset and were dropped from further computations (e.g. no molecules in the dataset contain iodine, so the nI descriptor is zero for every molecule). Hereafter the cross-correlation matrix of descriptors was generated and Pearson R^2 coefficients were computed for every possible pair of descriptors. It is assumed that if $R^2>0.99$ for a pair of descriptors, we can consider these descriptors as being equivalent and drop any one of them from further computations isolating 788 descriptors. The resulting models thus do not contain descriptors with a cross-correlation $R^2>0.99$. This step just reduced computation time approximately 5-fold.

$$3224\ descriptors \xrightarrow{constant\ drop} 1552\ descriptors \xrightarrow{cross-correlated\ drop} 788\ descriptors$$

Fuzzy forward search algorithm

Since a direct brute force search of the optimal model requires a great amount of computational resources and is too slow for the subject dataset, a fuzzy forward search algorithm was developed roughly following the procedure described below.

Models are defined as a set of descriptors and the coefficients of all models are calculated by multilinear regression using the vector of log(ESI) values as y-variable and a set of descriptor values for every molecule as x-variables. The obtained coefficients were used both for the training as well as the validation datasets. Model quality was defined as the correlation coefficient R^2 between experimental and calculated values for the training dataset.

Additional descriptors were stepwise appended to the best quality models from model **n** to (**n+1**) according the next steps:

1. Generate every possible combination of **n** descriptors
2. Sort obtained models on their quality (R^2)
3. Drop 80% of the worst models
4. Generate every possible combination of the remaining models with a new descriptor not present in this model yet

Most models had a very poor quality and were dropped immediately. 80% of the worst models were rejected before the addition of the next descriptor. Also models which differed only by the order of descriptors are assumed as being equivalent rejecting one of them (see figure S10).