ANALYZING CHROMATOGRAPHIC DATA USING MULTILEVEL (HIERARCHICAL) MODELS

making sense of complex data

Paweł Wiczling

Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Poland

ABSTRACT

It is relatively easy to collect chromatographic measurements for a large number of analytes especially if one uses chromatographic methods coupled with mass spectrometry detection. Such data have often a hierarchical or clustered structure. For example, analytes with the same log P and pK$_a$ tend to be more alike in their retention than analytes chosen at random from the population at large. Multilevel models recognize the existence of such data structures by assigning a model for each parameter with its parameters also estimated from the data. In this work we propose such a multilevel (hierarchical) model to describe the retention times obtained for two series of organic modifier content collected at different pH for a large series of acids and bases. It consisted of (i) the same deterministic equation describing the retention for all analytes, (ii) the covariate relationships relating various physicochemical properties of analyte to the chromatographically-specific parameters trough the Quantitative Structure Retention Relationship (QSRR)-based equations, and (iii) the stochastic components of intra-analyte and inter-analyte (residual) variability. Determination of the parameter, inter- and intra- compound variability characterizing the whole "population" of analytes provides a possibility to use Bayesian inference methods of parameter estimation from the limited set of chromatographic experiments to obtain the parameters' estimates and predictions for the specific analyte (and uncertainty around these values).

1) COLLECT THE GRADIENT DATA FOR A LARGE GROUP OF DIVERSE ANALYTES

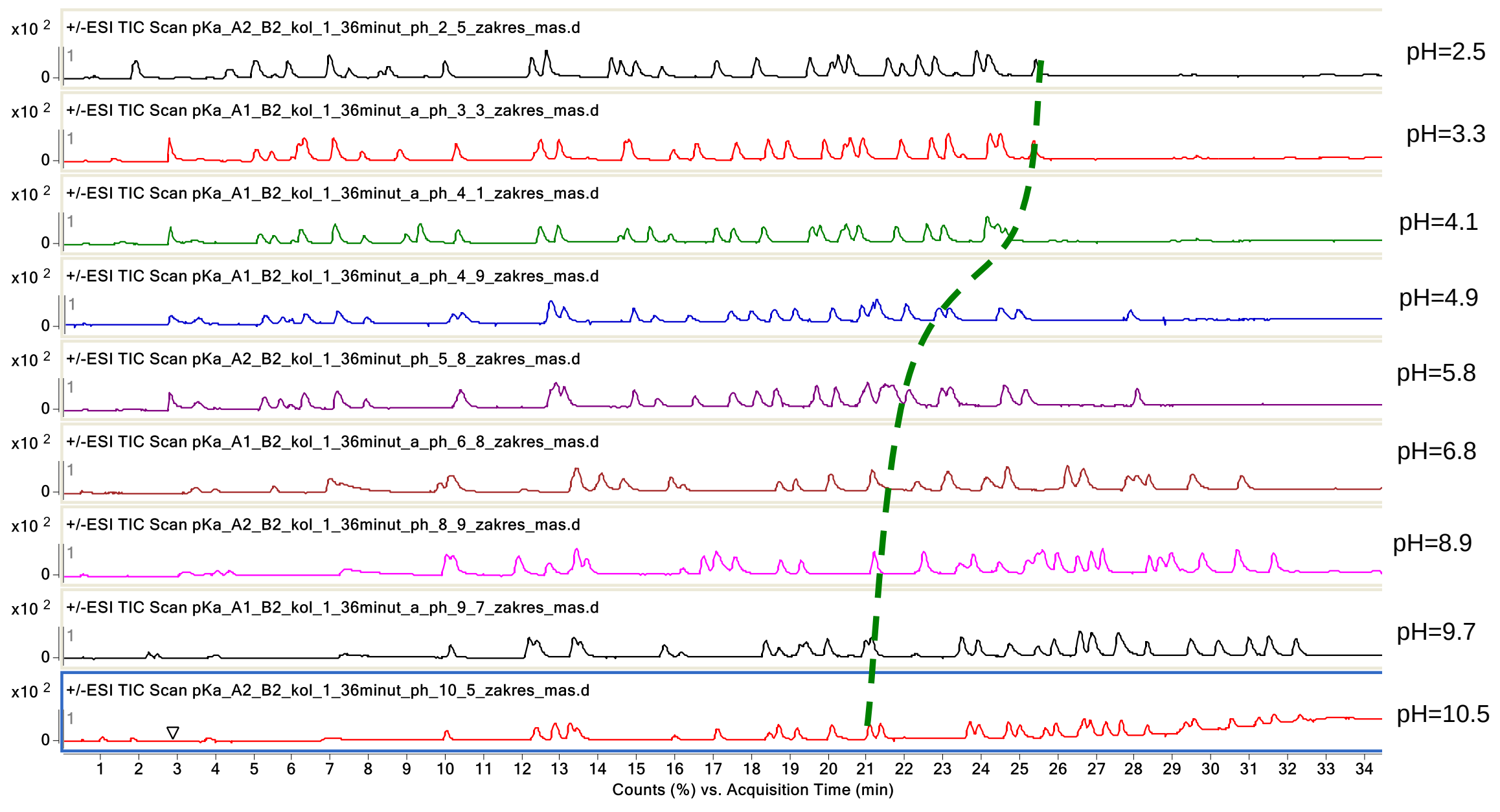
Methods

66 analytes (model building) + 27 analytes (validation)

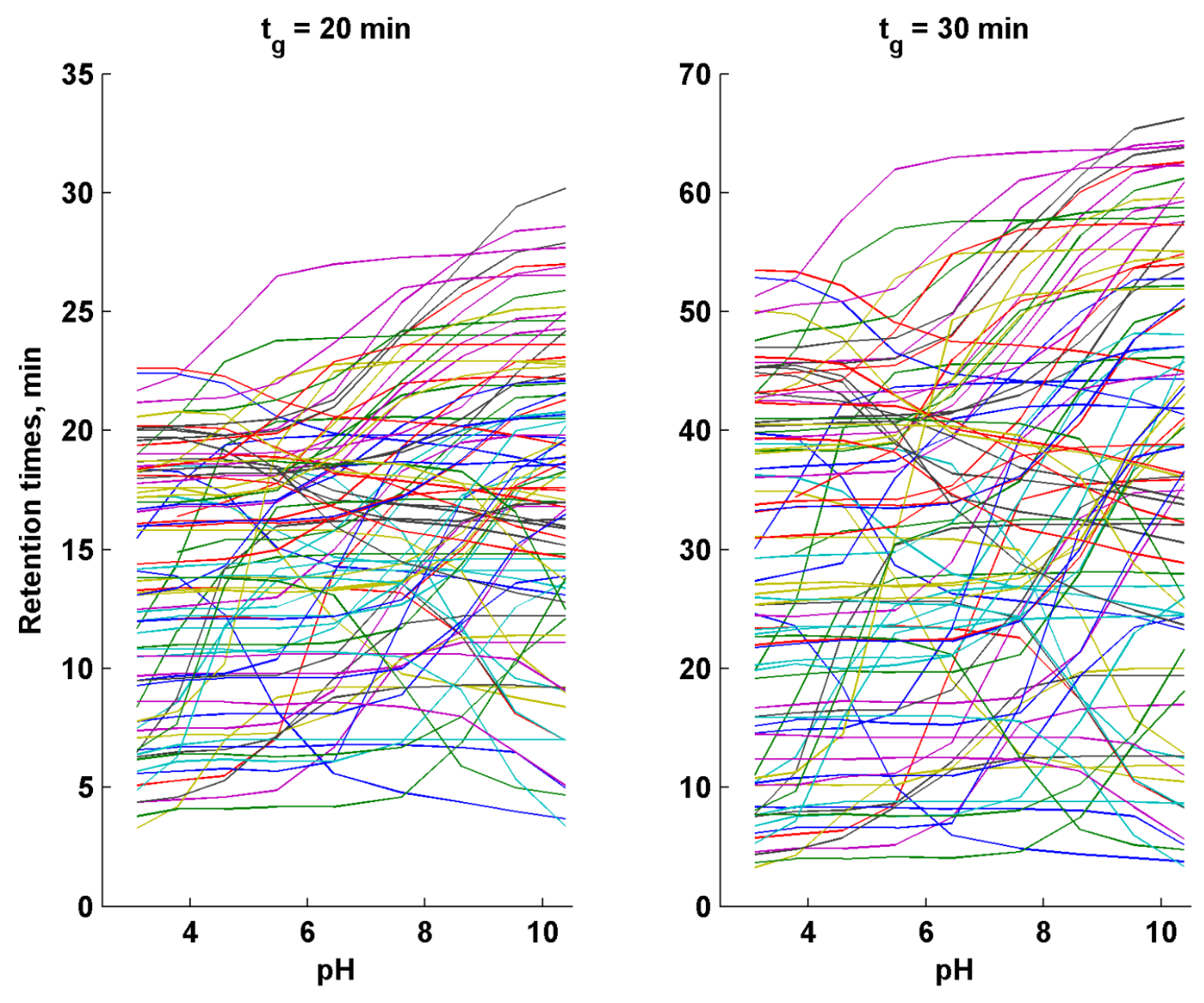
XTerra MS C18 5μm 4.6x150mm **(Waters, USA); F = 1 ml/min; T = 25°C,**

t_G= 20 min (9 values of pH) and 60 min (9 values of pH), φ_0=0.05, φ_f=0.8

Peak tracking using ESI-TOF-MS detection



Raw Data



2) DEVELOP AND VALIDATE HIERARCHICAL MODEL THAT GENERALIZE TO ALL ANALYTES

Model

$$t_{R,ij} = f(D_{ij}, R_i) + \varepsilon_{ij}$$

$$\int_0^{t_{Rij}-t_0} \frac{1}{t_0} \frac{1+10^{{}^s_s pH(t)_{ij} - pK_{a,i}(\varphi(t)_{ij})}}{10^{\log k_{w,N,i} - \frac{S_{1,N,i}\varphi(t)_{ij}}{1+S_{2,i}\varphi(t)_{ij}}} + 10^{\log k_{w,I,i} - \frac{S_{1,I,i}\varphi(t)_{ij}}{1+S_{2,i}\varphi(t)_{ij}}} 10^{{}^s_s pH(t)_{ij} - pK_{a,i}(\varphi(t)_{ij})}} dt = 1$$

t_{Rij} - retention times
D_{ij} - experimental design parameters
R_i - individual (analytes-specific) parameters

$$R_i = h(\theta, X_i) + \eta_{R,i}$$

$$\log k_{w,N,i} = \theta_{\log kw} + \theta_{\log kw-\log P} \log P_i + \theta_{\log kw-PSA} PSA_i + \eta_{\log kw,N,i}$$

$$S_{1,N,i} = \theta_{SN} + \theta_{SN-\log P} \log P_i + \theta_{\log SN-PSA} PSA_i + \eta_{SN,i}$$

$$\log k_{w,I,i} = \log k_{w,N} + \theta_{\Delta \log kw} + \eta_{\log kw,I,i}$$

$$S_{1,I,i} = S_{1,N} + \theta_{\Delta S} + \theta_{AB-\alpha} AB_i + \eta_{SI,i}$$

$$pK_a(\varphi(t))_i = {}^w_w pK_{a,i} + (\theta_\alpha + \theta_{AB-\alpha} AB_i)\varphi(t) + \eta_{pKa,i}$$

$$\eta_{R,i} \sim MVN\,(0, \mathbf{\Omega})$$

Θ - individual typical values
X_i - covariates {log P, pK_a, PSA - Polar Surface Area}
$\eta_{R,i}$ - inter-analyte variability

$$\mathrm{var}(\varepsilon_{ij}) = (\sigma_{add} + \sigma_{prop} f(D_{ij}, R_i))^2$$

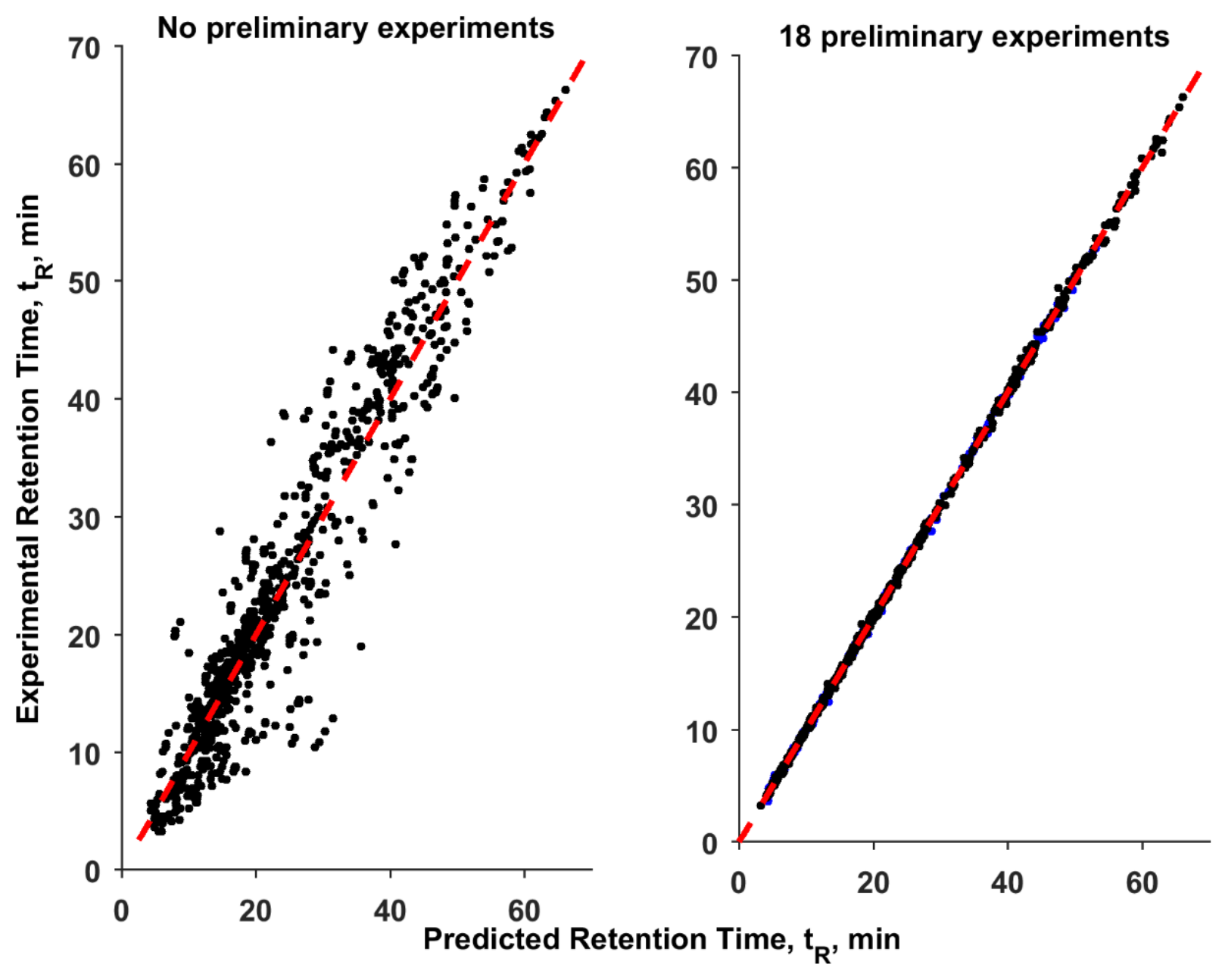
ϵ_{ij} - intra-analyte (residua) variability

Parameter estimates

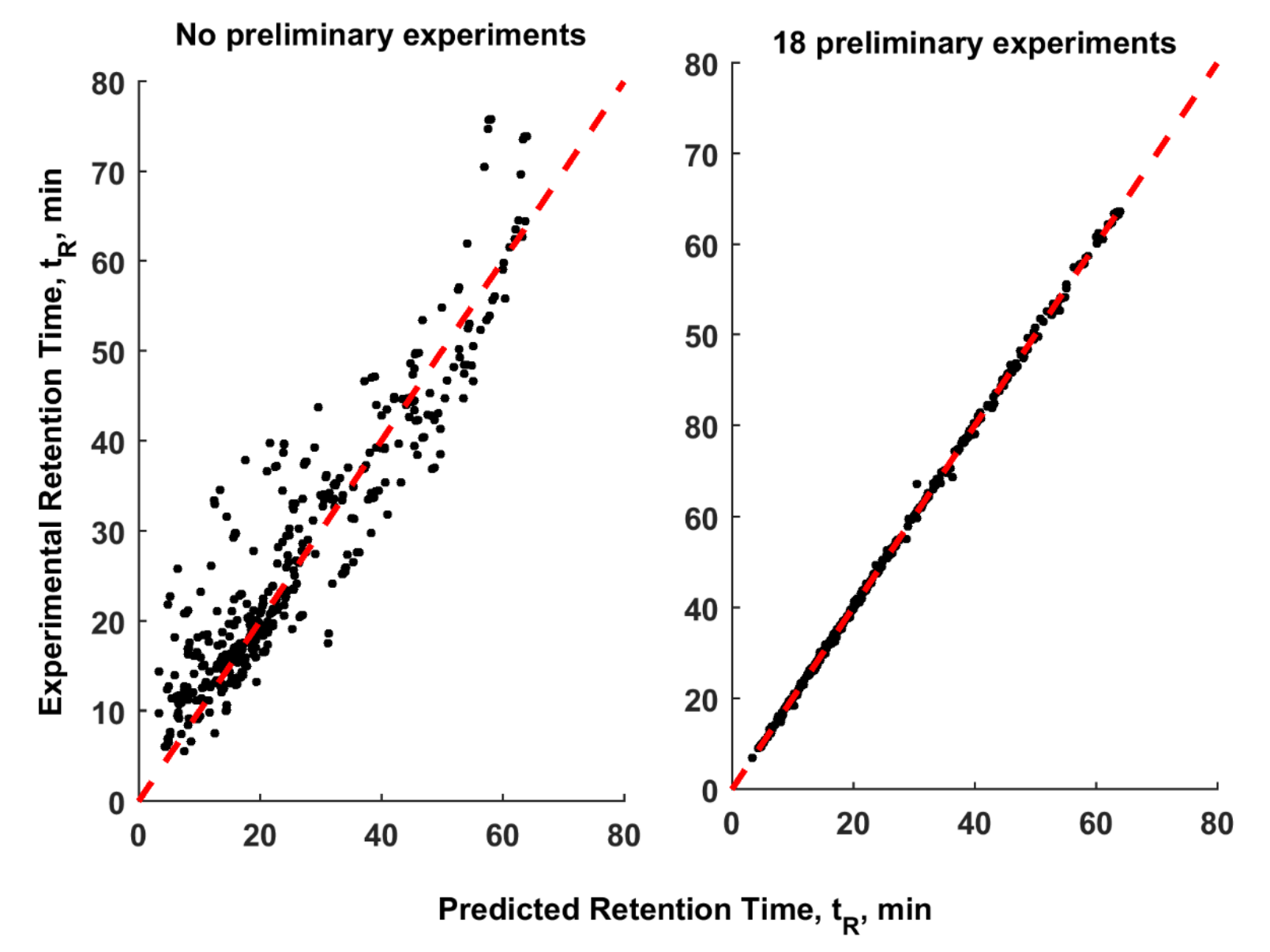
Parameters	Description	Fixed Effects Estimate, θ (%CV)	Random Effects Estimate, Ω (%CV)
log k$_{wN}$ **θ$_{logkw}$** **θ$_{logkw-logP}$** **θ$_{logkw-PSA}$**	Retention factor of non-ionized form of an analyte extrapolated to neat water as an eluent intercept slope for log P slope for PSA	0.433 (44) 0.915 (6) 0.0144 (15)	0.217 (10)
log k$_{wI}$ **θ$_{Δlogkw}$**	Retention factor of ionized form of an analyte extrapolated to neat water as an eluent The difference of log k between the non-ionized and ionized form of an analyte	-1.06 (5)	0.124 (11)
S$_{1,N}$ **θ$_{SN}$** **θ$_{SN-logP}$** **θ$_{SN-PSA}$**	The first slope coefficient for non-ionized form of an analyte intercept slope for log P slope for PSA	2.39 (12) 0.756 (11) 0.0281 (12)	0.437 (11)
S$_{1,I}$ **θ$_{ΔS (Acids)}$** **θ$_{ΔS (Bases)}$**	The first slope coefficient for ionized form of an analyte The difference between S$_1$ of ionized and non-ionized form of acid The difference between S$_1$ of ionized and non-ionized form of acid	-0.831 (29) 1.01 (15)	0.503 (15)
S$_2$	The second slope coefficient	0.183 (17)	
pK$_a$(φ(t)) **Acids: θ$_α$** **Bases: θ$_α$ + θ$_{AB-α}$**	The pK$_a$ value The slope of pK$_a$ vs organic modifier content for acids The slope of pK$_a$ vs. organic modifier content for bases	1.61 (10) -0.365 (19)	0.193 (9)
a	The empirical parameter accounting for the influence of pH on retention of anions due to non-hydrophobic interactions	-0.0172 (5)	
$\mathrm{cov}(\eta_{\log kw,N}, \eta_{S,N})$	Covariance between logk$_{w,N}$ and S$_N$		0.248 (6)
σ$_{add}$ **σ$_{prop}$**	Additive error model component Proportional error model component	0.137 (6) 0.00628 (10)	

Validation

Model predictions (66 analytes used to build the model)

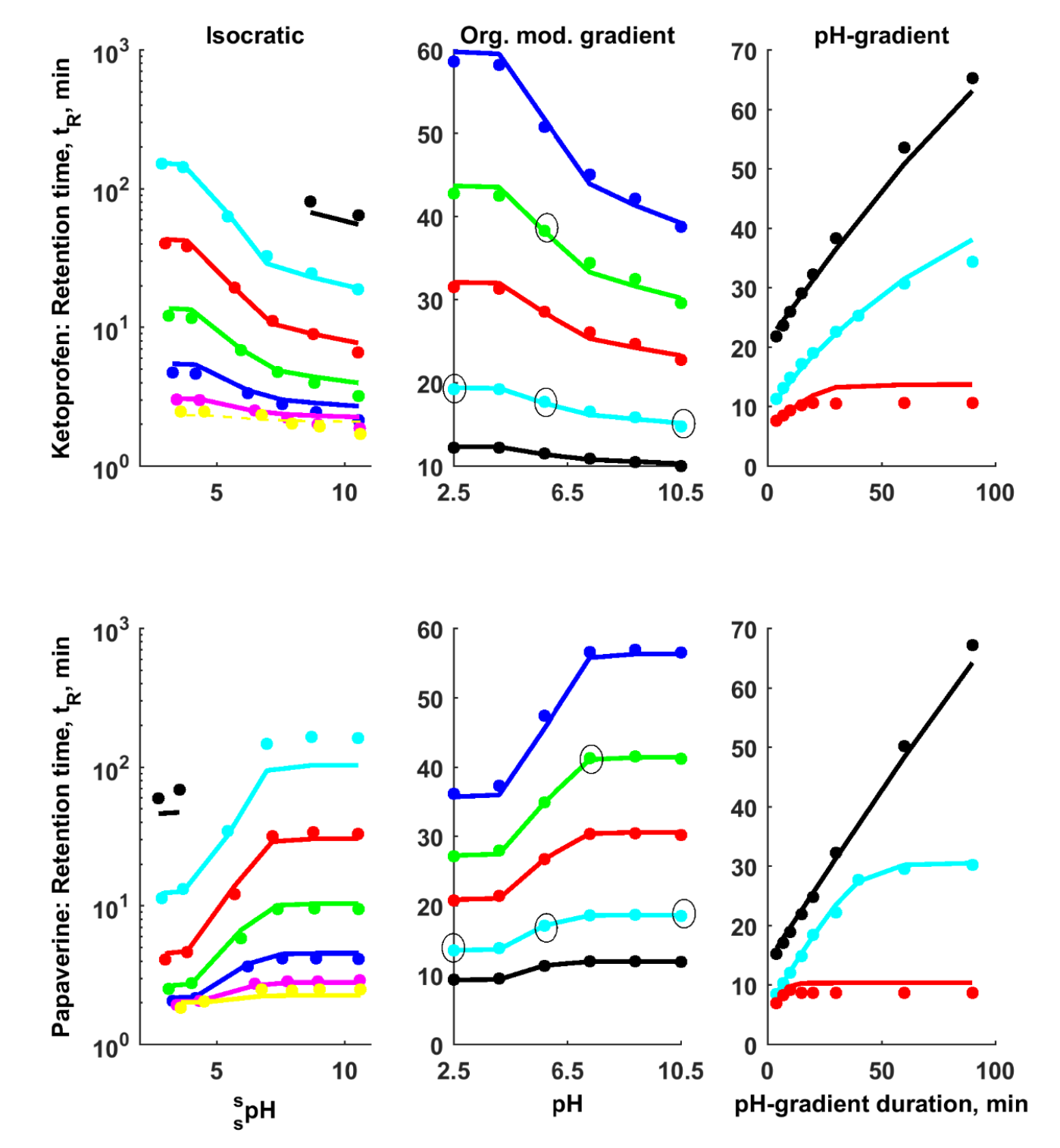


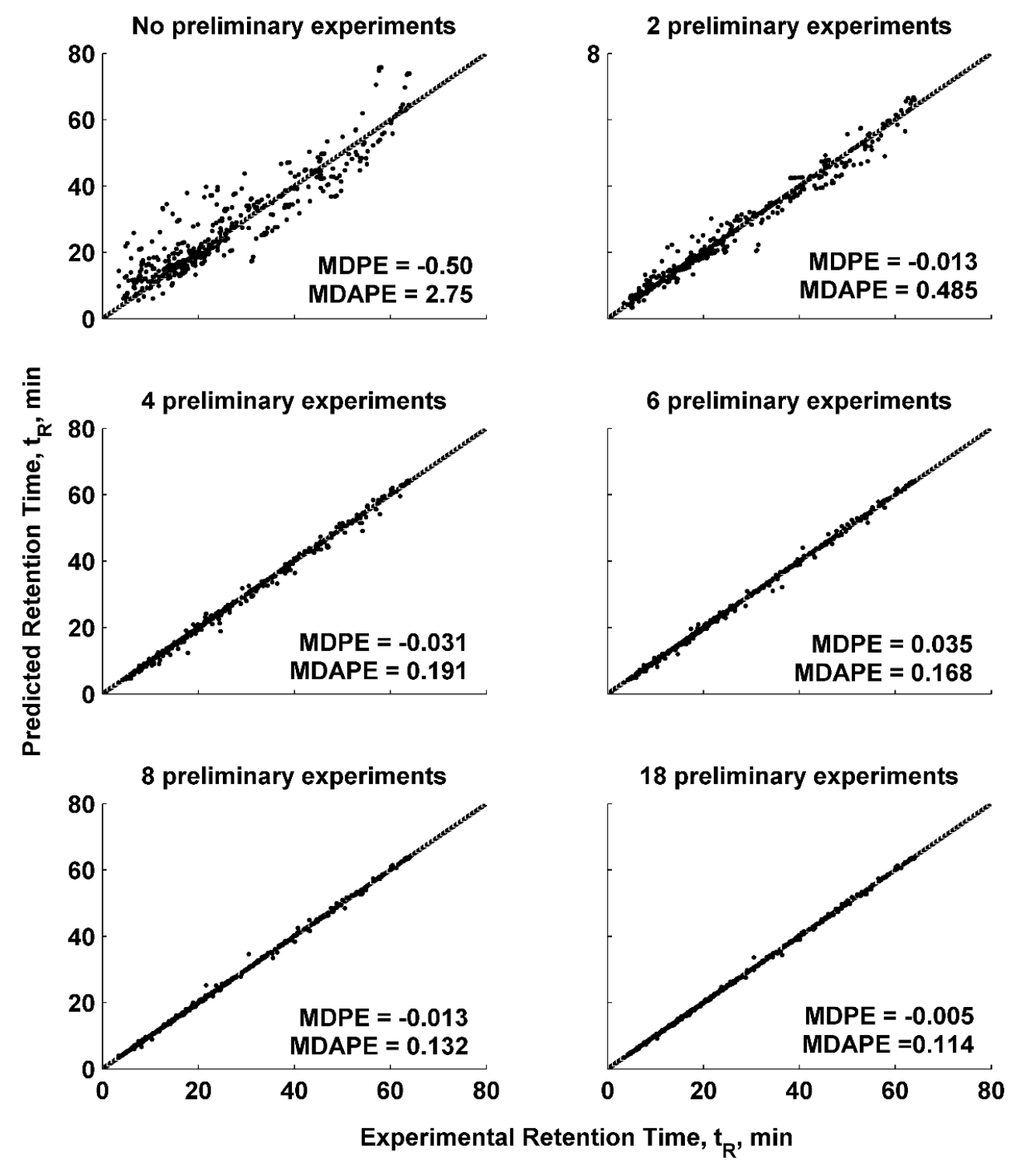
Model predictions (27 analytes used to validate the model)



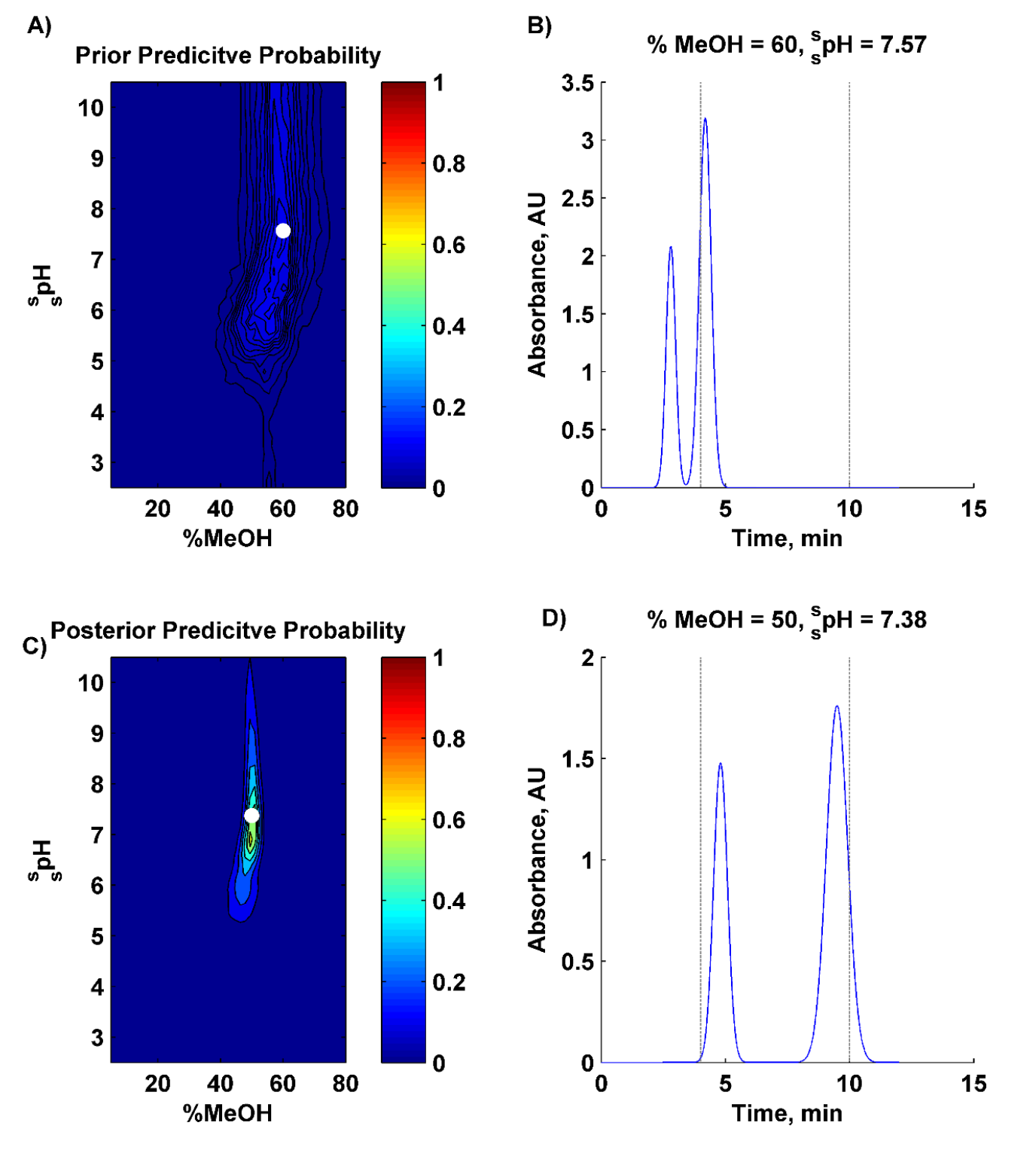
3) USE IT IN YOUR LAB FOR INFERENCE, PREDICTIONS, AND DECISION MAKING

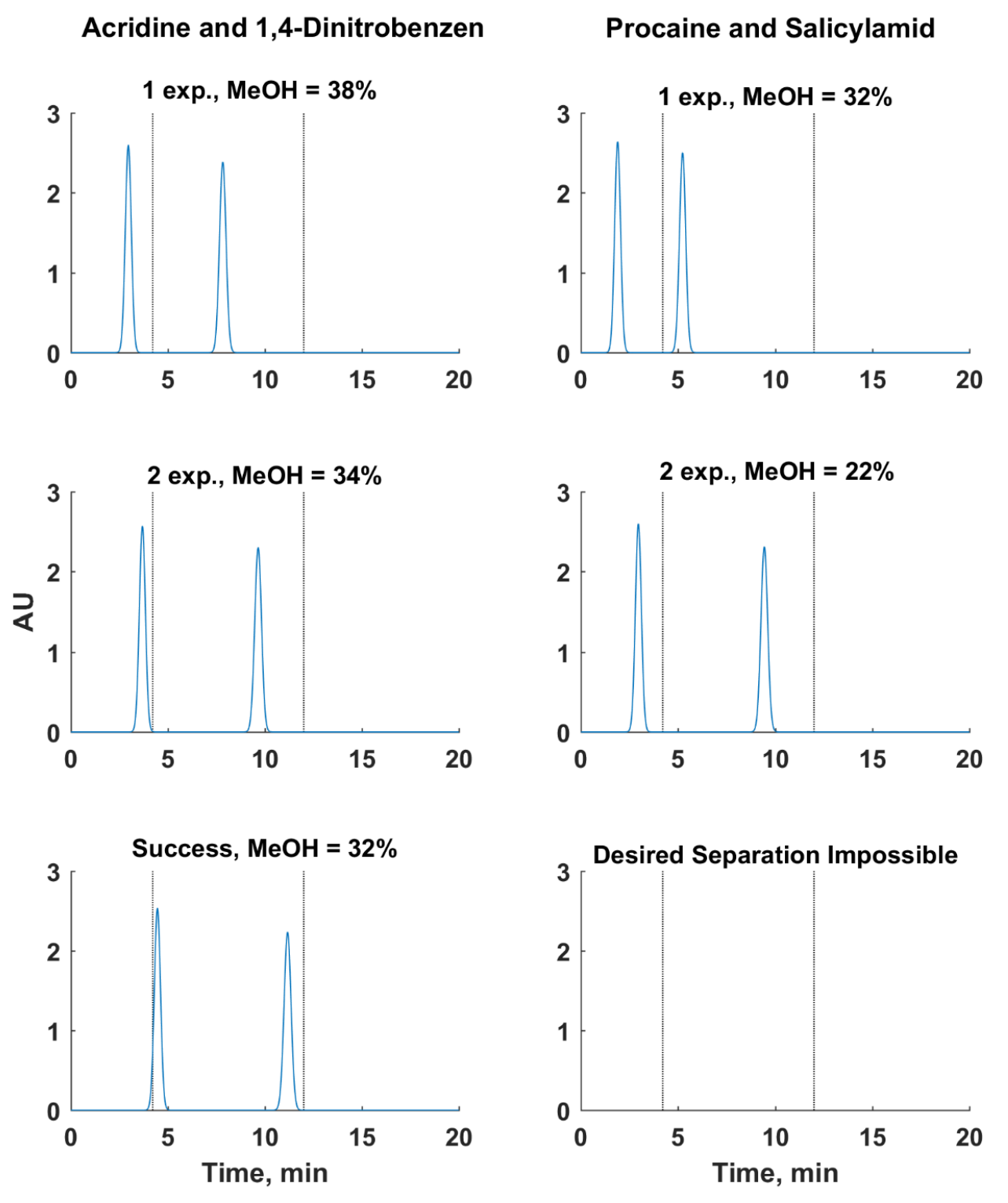
Parameter estimation and predictions from the limited set of chromatographic experiments





Decision making. What are the best chromatographic conditions for the next experiments?





ACKNOWLEDGMENT & REFERENCES

Wiczling P, Kubik Ł, Kaliszan R, Anal Chem, 2015, 87, 7241-9. This project was supported by the National Science Centre, Poland (nr. 2015/18/E/ST4/00449)