

Supplementary Information

Title: Quantitative structure retention-relationship modelling: Towards an innovative general-purpose strategy

Authors: Priyanka Kumari^{a,b*}, Thomas Van Laethem^{a,b}, Philippe Hubert^a, Marianne Fillet^b, Pierre-Yves Sacré^a, Cédric Hubert^{a*}

a. University of Liège (ULiege), CIRM, Laboratory of Pharmaceutical Analytical Chemistry, Liège, Belgium

b. University of Liège (ULiege), CIRM, Laboratory for the Analysis of Medicines, Liège, Belgium

S1: Illustration of molecular descriptor calculation with an example

There were two tools used for molecular descriptor calculation: [1] RdKit [2] Chemicalize

- Values of LogD was calculated using Chemicalize at each pH

Method of calculation of other descriptors:

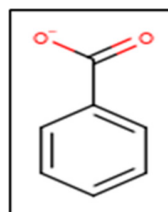
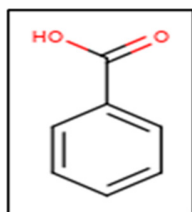
- Step1: Smile strings of compounds into Chemicalize
- Step2: Retrieve Smile strings of all microspecies and their distributions at all pH
- Step3: Calculate molecular descriptors of every microspecies from RdKit
- Step4: Calculate weighted average of molecular descriptors at each pH using formula below

$$FV_{ph} = \frac{\sum_{i=1}^n MS_i * D_i}{\sum_{i=1}^n D_i}$$

Where , FV_{ph} = Weighted average, MS_i = Descriptor value for microspecies and,

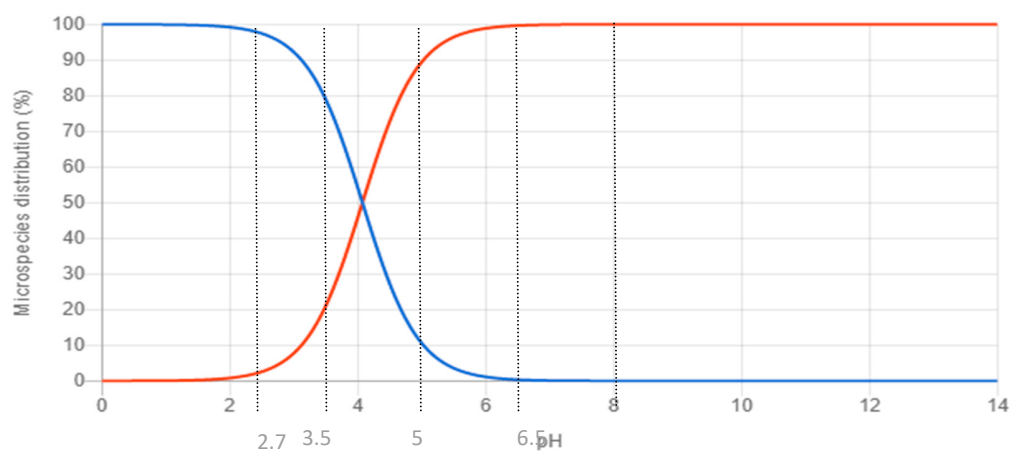
D_i = %Distribution of microspecies, n = no. of microspecies , ph = Specific pH at which final value is being calculated

Step-1



[b]Benzoic Acid (MS1) - C1=CC=C(C=C1)C(=O)O

[b]Benzoate(MS2)- MS1C1=CC=C(C=C1)C(=O)[O-]



Step2: Microspecies distribution calculation using Chemicalize

Condition	MS1	MS2
pH2.50	0.97	0.03
pH3.50	0.79	0.21
pH5.00	0.11	0.89
pH6.50	0.00	1.00
pH8	0.00	1.00

Step3: Calculate molecular descriptors of every microspecies from RdKit

Microspecies	PEOE_VSA7
MS1	12.13
Ms2	5.56

Step4: Calculate weighted average of molecular descriptors at each pH :

Final value of PEOE_VSA7

[1] At pH2.7

$$= (12.13 \cdot 0.97 + 5.56 \cdot 0.03) / 1$$

$$= 11.76 + 0.166$$

$$= 11.92$$

[2] At pH 3.5

$$= (12.13 \cdot 0.79 + 5.56 \cdot 0.21) / 1$$

$$= 9.58 + 1.16$$

$$= 10.74$$

[3] At pH 5

$$= (12.13 \cdot 0.11 + 5.56 \cdot 0.89) / 1$$

$$= 1.33 + 4.94$$

$$= 6.27$$

[4] At pH 6.5

$$= (12.13 \cdot 0.0 + 5.56 \cdot 1.0) / 1$$

$$= 5.56$$

[5] At pH 8

$$= (12.13 \cdot 0.0 + 5.56 \cdot 1.0) / 1$$

$$= 5.56$$

Final Value of molecular descriptors: -

Condition	Final Value PEOE_VSA7
-----------	--------------------------

pH 2.7	11.92
pH 3.5	10.74
pH5	6.27
pH6.5	5.56
pH 8	5.56

Table S2: Name of features used to start QSRR modeling

MolWt	EState_V SA2	fr_ArN	fr_quatN	MinEStateIn dex	PEOE_VSA 10	SlogP_ VSA3	VSA_ EStat e6
logD	EState_V SA3	fr_aryl_met hyl	FractionCSP3	MinPartialCh arge	PEOE_VSA 11	SlogP_ VSA4	VSA_ EStat e7
Asymmetri c.atom.cou nt	EState_V SA4	fr_benzene	FSP3	Molar.refract ivity	PEOE_VSA 12	SlogP_ VSA5	VSA_ EStat e8
Atom.count	EState_V SA5	fr_bicyclic	HallKierAlpha	MolLogP	PEOE_VSA 13	SlogP_ VSA6	VSA_ EStat e9
BalabanJ	EState_V SA6	fr_C_O	Heavy.atom.co unt	MolMR	PEOE_VSA 14	SlogP_ VSA7	
BertzCT	EState_V SA7	fr_C_O_noC OO	HeavyAtomMo lWt	NHOHCount	PEOE_VSA 2	SlogP_ VSA8	
Chi0	EState_V SA8	fr_COO	Hetero.ring.co unt	NOCCount	PEOE_VSA 3	SMR_ VSA1	
Chi0n	EState_V SA9	fr_COO2	Hydrogen.bon d.acceptor.cou nt	NumAliphati cHeterocycle s	PEOE_VSA 6	SMR_ VSA10	
Chi0v	FpDensit yMorgan 1	fr_ether	Hydrogen.bon d.donor.count	NumAliphati cRings	PEOE_VSA 7	SMR_ VSA3	

Chi1	FpDensityMorgan2	fr_halogen	lpc	NumAromaticCarbocycles	PEOE_VSA8	SMR_VSA5	
Chi1n	FpDensityMorgan3	fr_imidazole	Kappa1	NumAromaticHeterocycles	PEOE_VSA9	SMR_VSA6	
Chi1v	fr_Al_CO_O	fr_Ndealkylation1	Kappa2	NumAromaticRings	Polarizability	SMR_VSA7	
Chi2n	fr_Al_OH	fr_NH0	Kappa3	NumHAcceptors	qed	SMR_VSA9	
Chi2v	fr_Al_OH_noTert	fr_NH1	LabuteASA	NumHDonors	Ring.count	TPSA	
Chi3n	fr_amide	fr_NH2	MaxAbsEStateIndex	NumHeteroatoms	Rotatable.bond.count	VSA_EState1	
Chi3v	fr_aniline	fr_Nhpyrrole	MaxAbsPartialCharge	NumRotatableBonds	SlogP_VSA1	VSA_EState10	
Chi4n	fr_Ar_CO_O	fr_para_hydroxylation	MaxEStateIndex	NumSaturatedHeterocycles	SlogP_VSA10	VSA_EState2	
Chi4v	fr_Ar_N	fr_phenol	MaxPartialCharge	NumSaturatedRings	SlogP_VSA11	VSA_EState3	
EState_VSA1	fr_Ar_NH	fr_phenol_nOrthoHbond	MinAbsEStateIndex	NumValenceElectrons	SlogP_VSA12	VSA_EState4	
EState_VSA10	fr_Ar_OH	fr_pyridine	MinAbsPartialCharge	PEOE_VSA1	SlogP_VSA2	VSA_EState5	

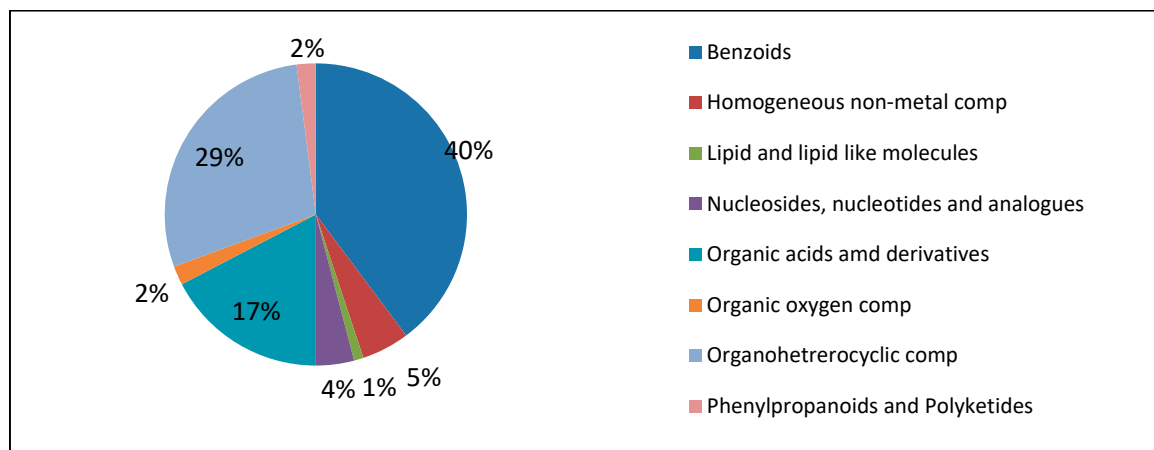
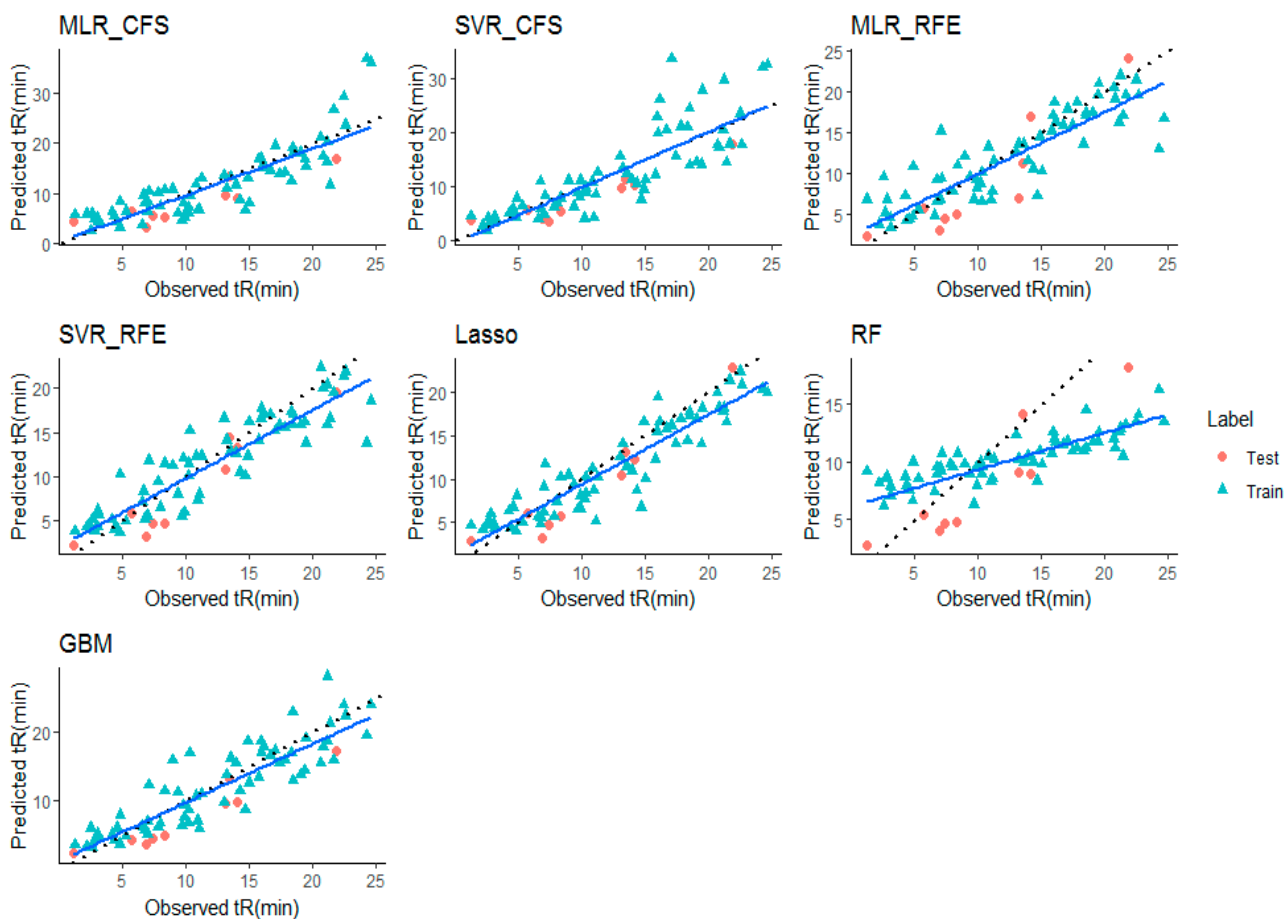


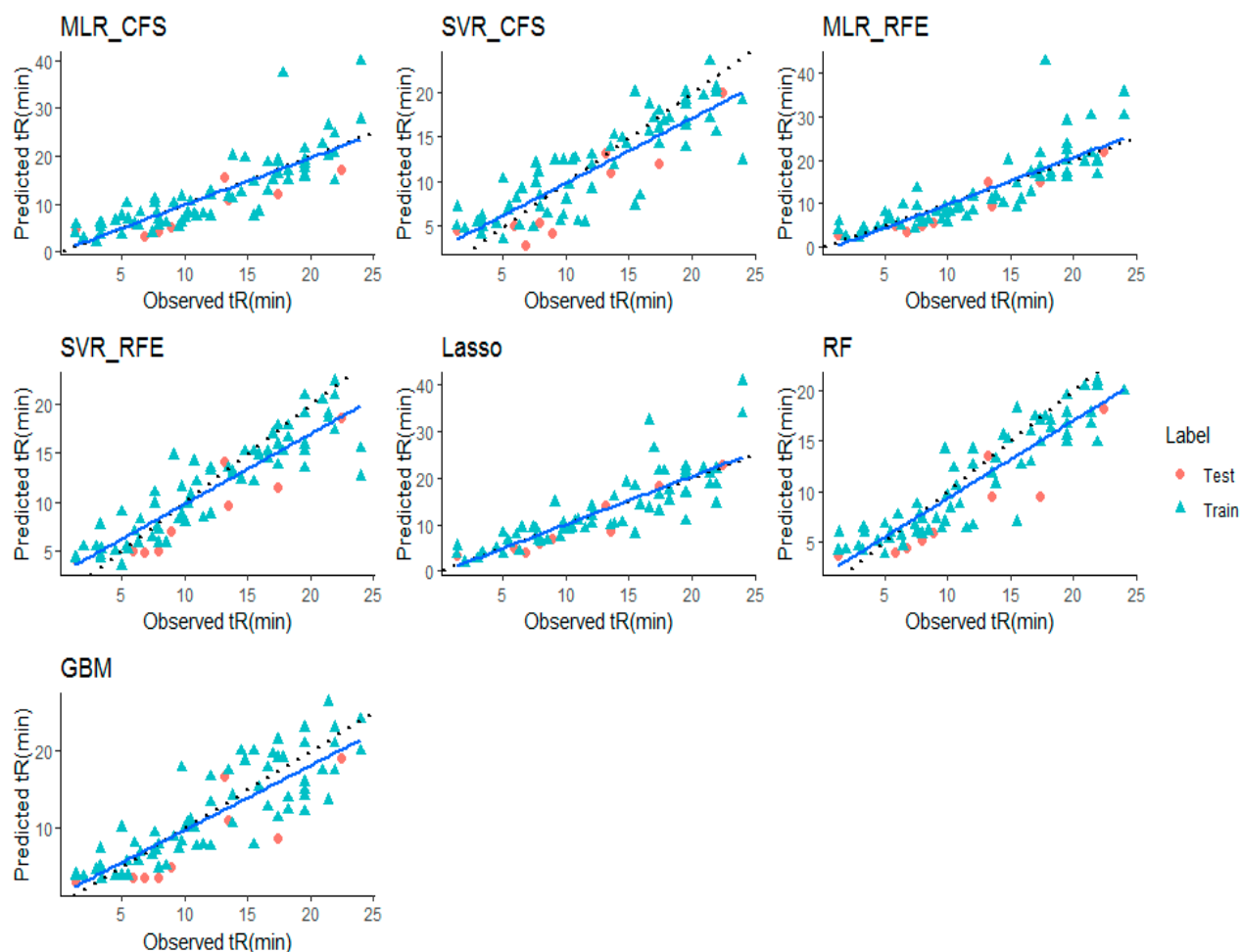
Figure S3:Chemical taxonomy of the molecules in the dataset

Table S4: Important features selected by each algorithm on data at each pH

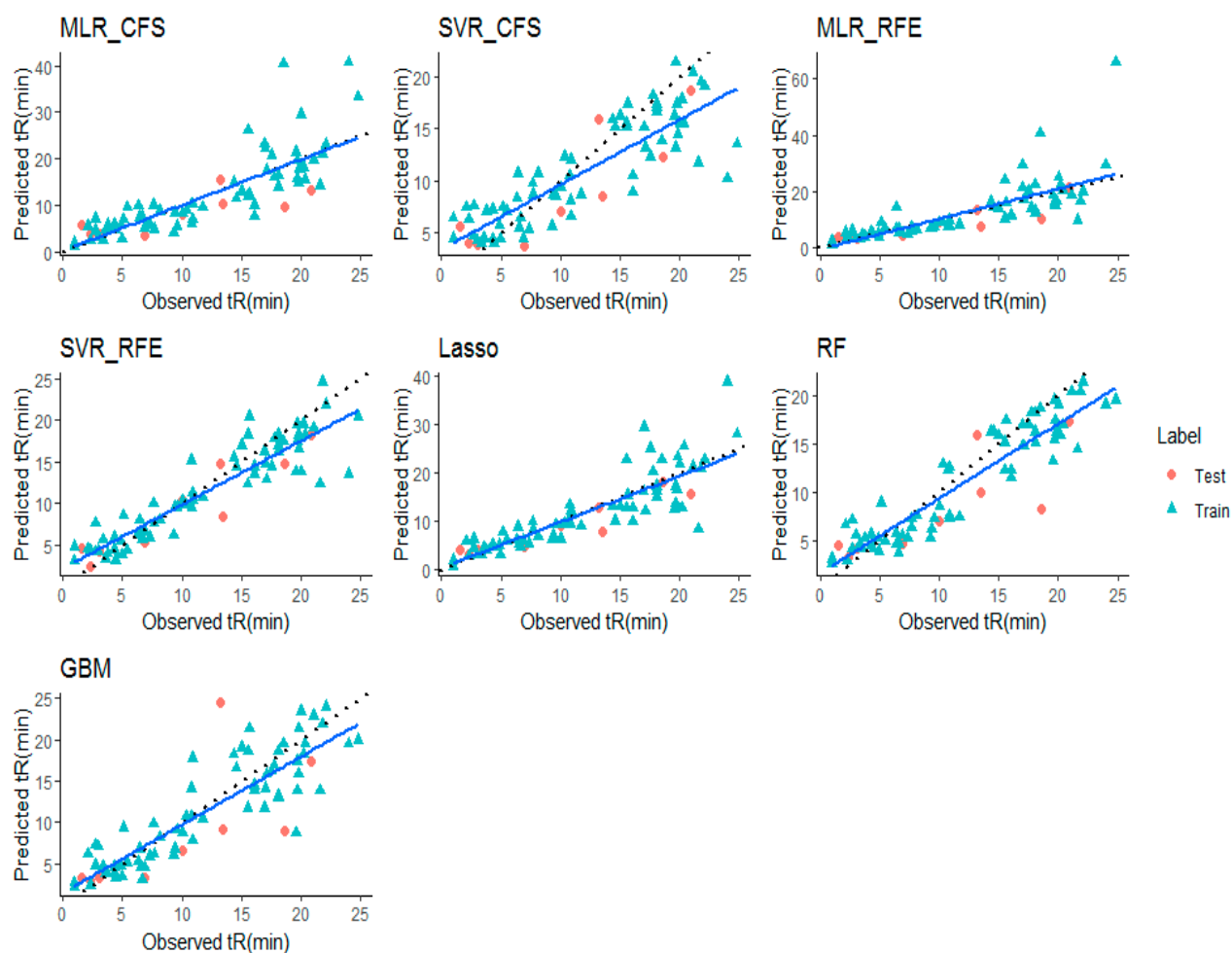
pH	Total	Descriptors
2.7	5	MolLogP, logD , NHOHCount , fr_Ar_NH, PEOE_VSA6
3.5	3	MolLogP, logD, NHOHCount
5	4	Polarizability, MolLogP, logD, PEOE_VSA6
6.5	3	MolLogP,logD,PEOE_VSA6
8	3	MolLogP , logD, PEOE_VSA6



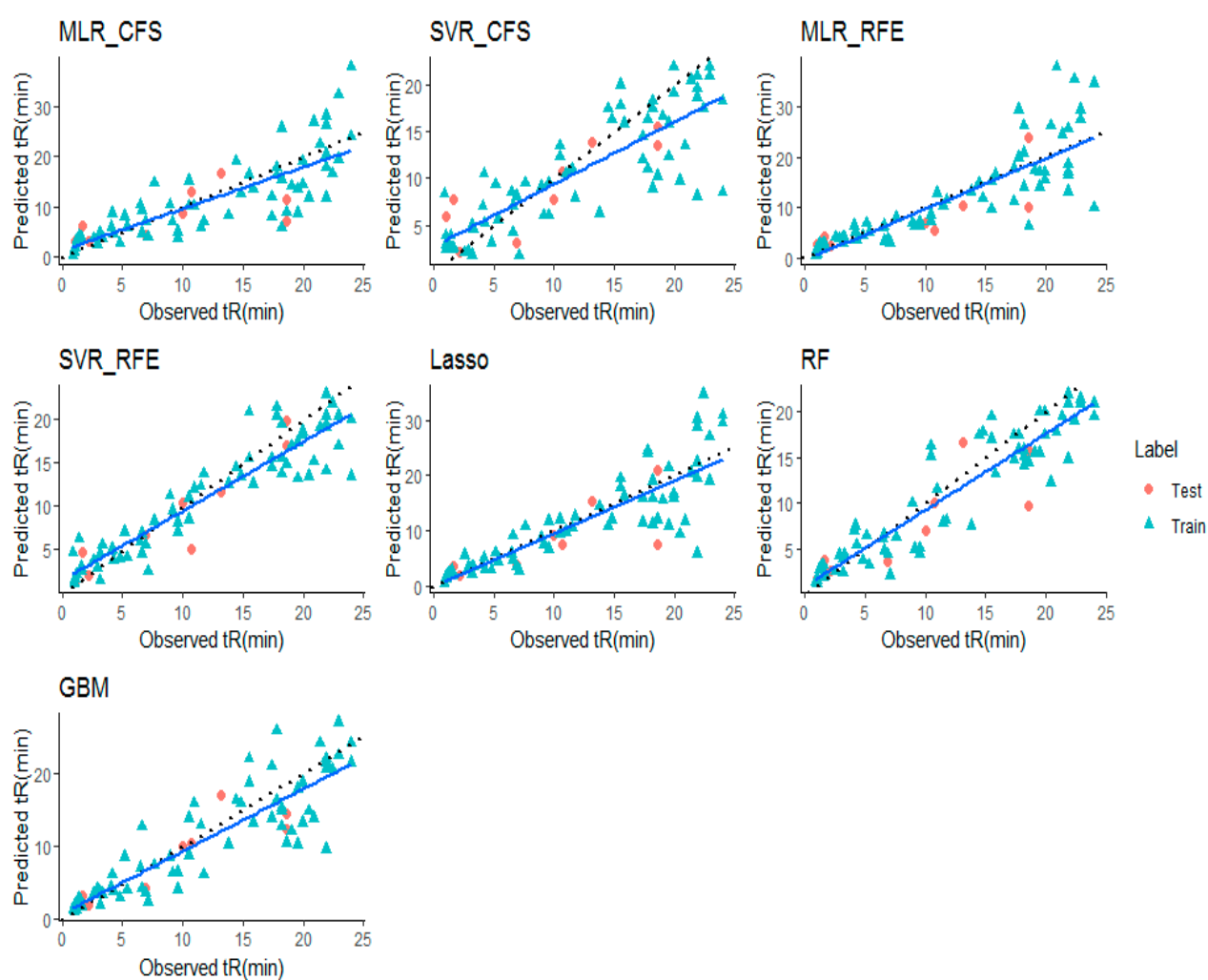
S5: Predicted Vs. Experimental Retention times (in min.) for all models at pH2.7 (Blue line- fit , Black dashed line- identity line)



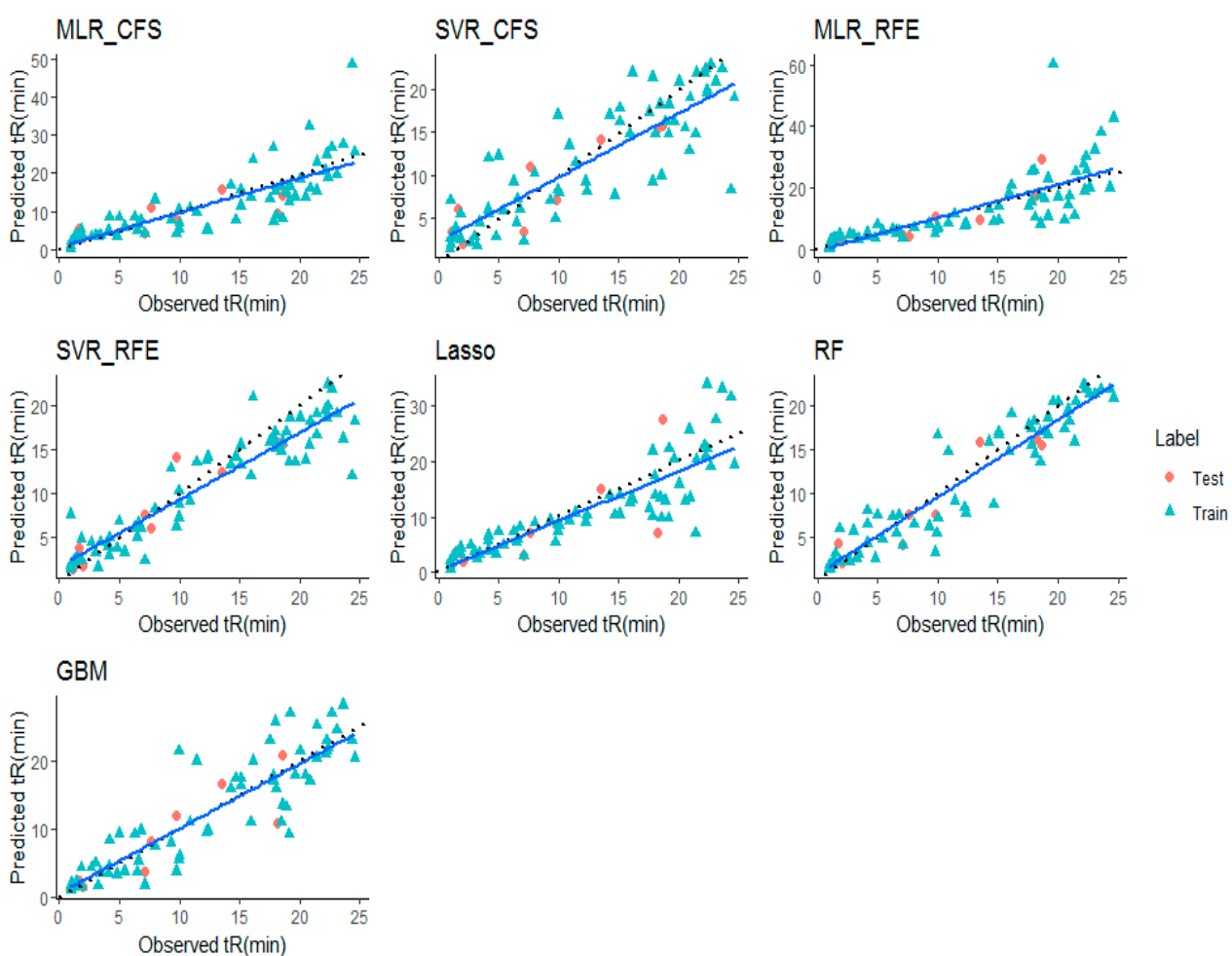
S6: Predicted Vs. Experimental Retention times (in min.) for qsrr models at pH 3.5((Blue line- fit , Black dashed line- identity line)



S7: Predicted Vs. Experimental Retention times (in min.) for qsrr models at pH 5 (Blue line- fit , Black dashed line- identity line)



S8: Predicted Vs. Experimental Retention times for qsrr models at pH 6.5 (Blue line- fit, Black dashed line- identity line)



S9: Predicted Vs. Experimental Retention times for qsrr models at pH 8 (Blue line- fit, Black dashed line- identity line)

S10: Parameters used by prediction models at all pH

Models at pH2.7:

SVR_CFS	Sigma = 0.362	C = 1
SVR_RFE	Sigma = 0.048	C=1
Lasso	Aplha = 1	Lambda = 0.014
RF	Mtry = 73	
GBM	n. trees = 150, Shrinkage = 0.1	Interaction depth =2, n. minobsinnode = 10

Models at pH3.5:

SVR_CFS	Sigma = 0.122	C = 1
SVR_RFE	Sigma = 0.075	C=1
Lasso	Aplha = 0.1	Lambda = 0.048
RF	Mtry = 73	
GBM	n. trees = 150, Shrinkage = 0.1	Interaction depth =2, n.minobsinnode = 10

Models at pH 5:

SVR_CFS	Sigma = 0.353	C = 1
---------	---------------	-------

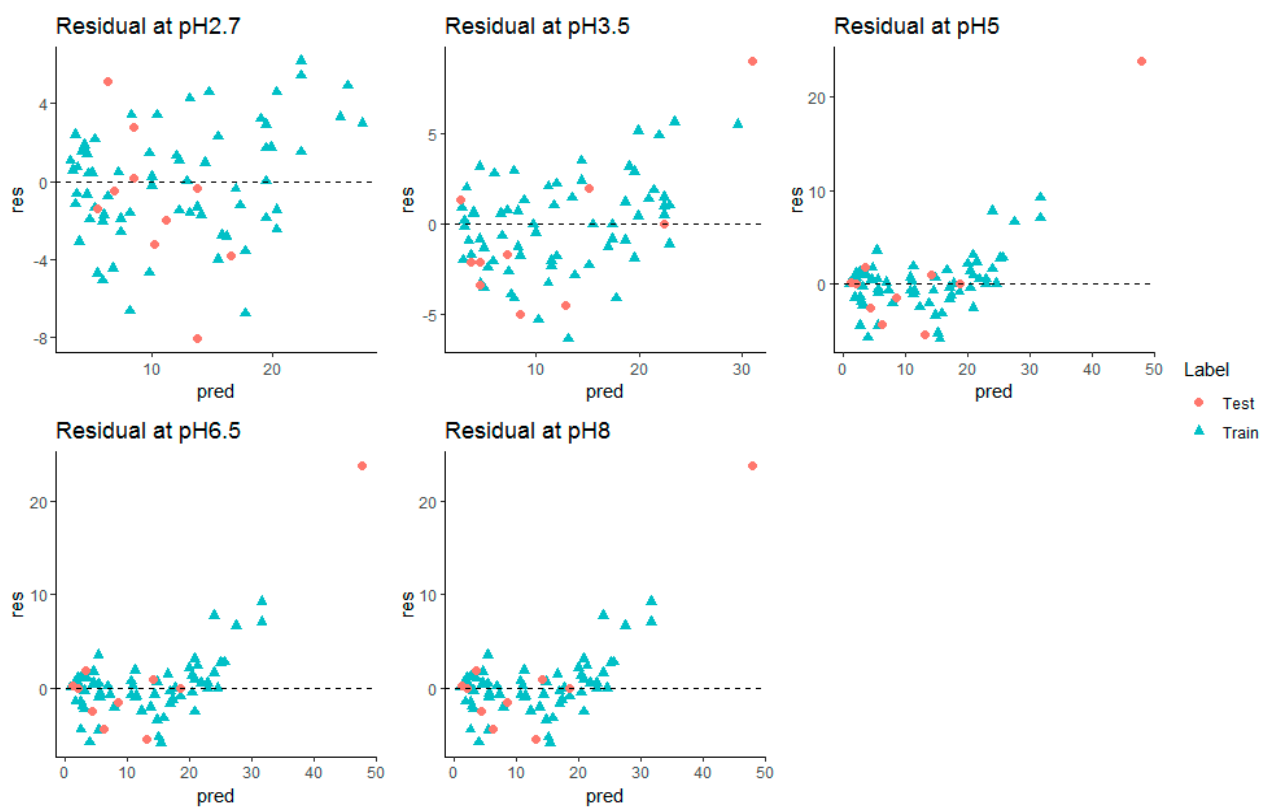
SVR_RFE	Sigma = 0.054	C=1
Lasso	Aplha = 0.1	Lambda = 0.058
RF	Mtry = 73	
GBM	n.trees = 150, Shrinkage = 0.1	Interaction depth =2, n.minobsinnode = 10

Models at pH 6.5:

SVR_CFS	Sigma = 0.313	C = 1
SVR_RFE	Sigma = 0.042	C=1
Lasso	Aplha = 0.1	Lambda = 0.02
RF	Mtry = 147	
GBM	n.trees = 150, Shrinkage = 0.1	Interaction depth =3, n.minobsinnode = 10

Models at pH 8:

SVR_CFS	Sigma = 0.319	C = 1
SVR_RFE	Sigma = 0.110	C=1
Lasso	Aplha = 0.1	Lambda = 0.024
RF	Mtry = 147	
GBM	n.trees = 150, Shrinkage = 0.1	Interaction depth =2, n.minobsinnode = 10



S11: Residual plots (in min) for Stacking model for all dataset(with Miconazole)