Electronic Supplementary Information to

Quantitative profile-profile relationship (QPPR) modelling: a novel machine learning approach to predict and associate chemical characteristics of unspent ammunition from gunshot residue (GSR)

Matteo D. Gallidabino*, Leon P. Barron, Céline Weyermann, Francesco S. Romolo

*Email: matteo.gallidabino@northumbria.ac.uk

Table of contents

Distributions of input and output variables

Table S1 – Comparison of the distribution (n = 36) characteristics for the eight output variables before and after base-10 logarithm transformation of peak areas (PA). The last column shows the differences between %RSD and skewness values (PA – logPA).

Comp.	PA				LogPA				Difference	
	Mean	**SD**	**%RSD**	**Skew.**	**Mean**	**SD**	**%RSD**	**Skew.**	**%RSD**	**Skew.**
O_{4ND}	$1.08x10^2$	$1.10x10^2$	101.9	1.10	1.44	1.02	71.1	-0.60	-30.8	-0.50
O_{AK2}	$5.18x10^3$	$9.88x10^3$	190.9	1.51	1.25	1.86	148.9	0.94	-42.0	-0.57
O_{DOF}	$2.22x10^2$	$6.39x10^2$	288.0	2.62	0.37	1.05	286.9	2.58	-1.1	-0.04
O_{DBP}	$1.04x10^5$	$2.42x10^5$	231.7	2.54	3.68	1.43	38.9	-0.23	-192.9	-2.31
O_{EC}	$1.03x10^5$	$1.24x10^5$	120.1	0.95	3.69	1.85	50.2	-0.73	-69.9	-0.23
O_{DPA}	$6.21x10^4$	$3.60x10^4$	58.0	-0.83	4.17	1.52	36.4	-2.02	-21.6	1.19
O_{NG}	$8.69x10^5$	$6.16x10^5$	70.9	0.81	5.30	1.83	34.5	-2.54	-36.4	1.73
O_{2ND}	$2.38x10^2$	$2.08x10^2$	87.3	0.38	1.84	1.05	57.2	-1.06	-30.1	0.68

SD: standard deviation; %RSD: relative standard deviation in %; Skew.: skewness.

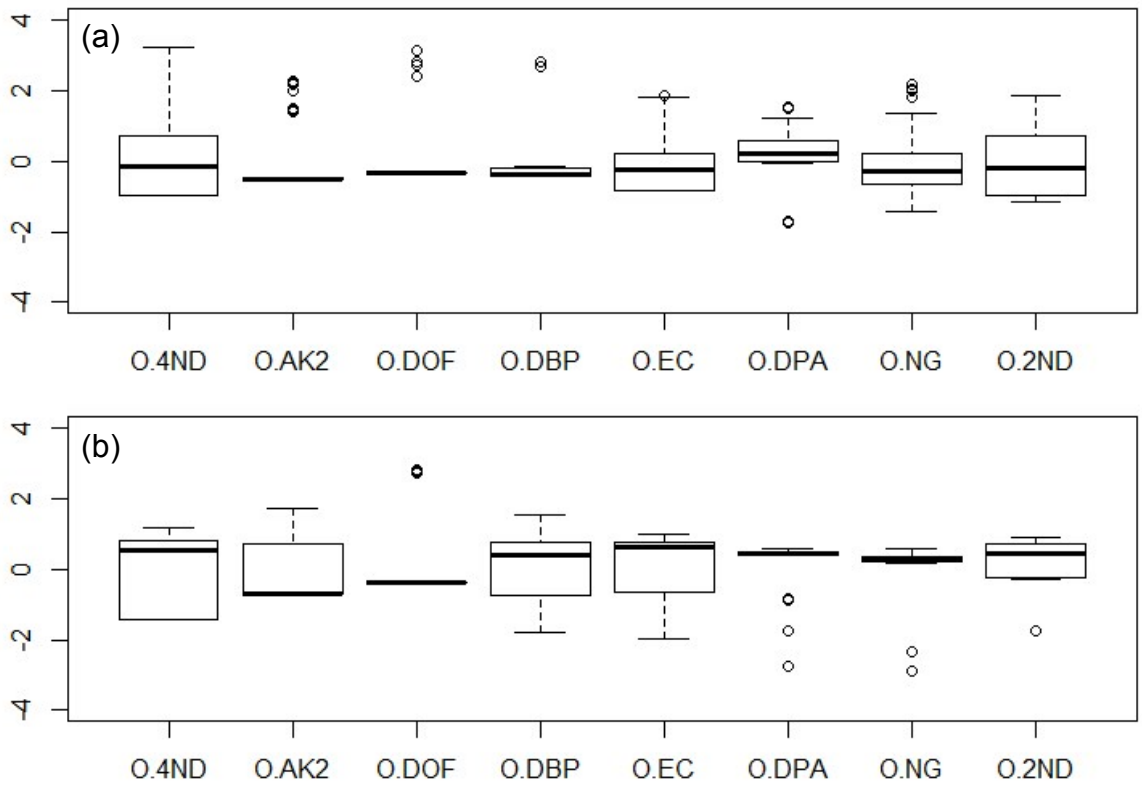


Figure S1 – Boxplots (n = 36) for the eight output variables before (a) and after (b) base-10 logarithm transformation of peak areas. Data were centred and scaled to allow a better comparison of their distribution.

Table S2 – Comparison of the distribution (n = 27) characteristics for the eight input variables before and after base-10 logarithm transformation of peak areas (PA). The last column shows the differences between %RSD and skewness values (PA – logPA).

Comp.	PA				LogPA				Difference	
	Mean	**SD**	**%RSD**	**Skew.**	**Mean**	**SD**	**%RSD**	**Skew.**	**%RSD**	**Skew.**
I_{DND}	9.15×10^3	2.34×10^4	256.4	4.34	1.79	2.07	115.2	0.31	-141.1	-4.03
I_{4ND}	2.49×10^6	5.63×10^6	226.2	2.93	5.06	1.79	35.4	-1.63	-190.9	-1.30
I_{Ak2}	1.80×10^5	3.35×10^5	186.0	1.79	4.42	0.84	19.0	1.06	-167.1	-0.73
I_{2ND}	8.22×10^6	1.31×10^7	159.0	2.26	6.28	0.95	15.2	-0.60	-143.8	-1.67
I_{DBP}	6.03×10^7	7.10×10^7	117.8	1.66	7.39	0.66	9.0	-0.03	-108.8	-1.63
I_{EC}	2.12×10^7	2.66×10^7	125.4	1.69	6.92	0.68	9.9	-0.13	-115.5	-1.56
I_{DPA}	1.55×10^7	1.97×10^7	127.8	0.94	6.56	0.90	13.8	-0.23	-114.0	-0.71
I_{HEX}	4.70×10^6	6.29×10^6	133.8	2.24	6.39	0.51	7.9	0.18	-125.8	-2.06

SD: standard deviation; %RSD: relative standard deviation in %; Skew.: skewness.

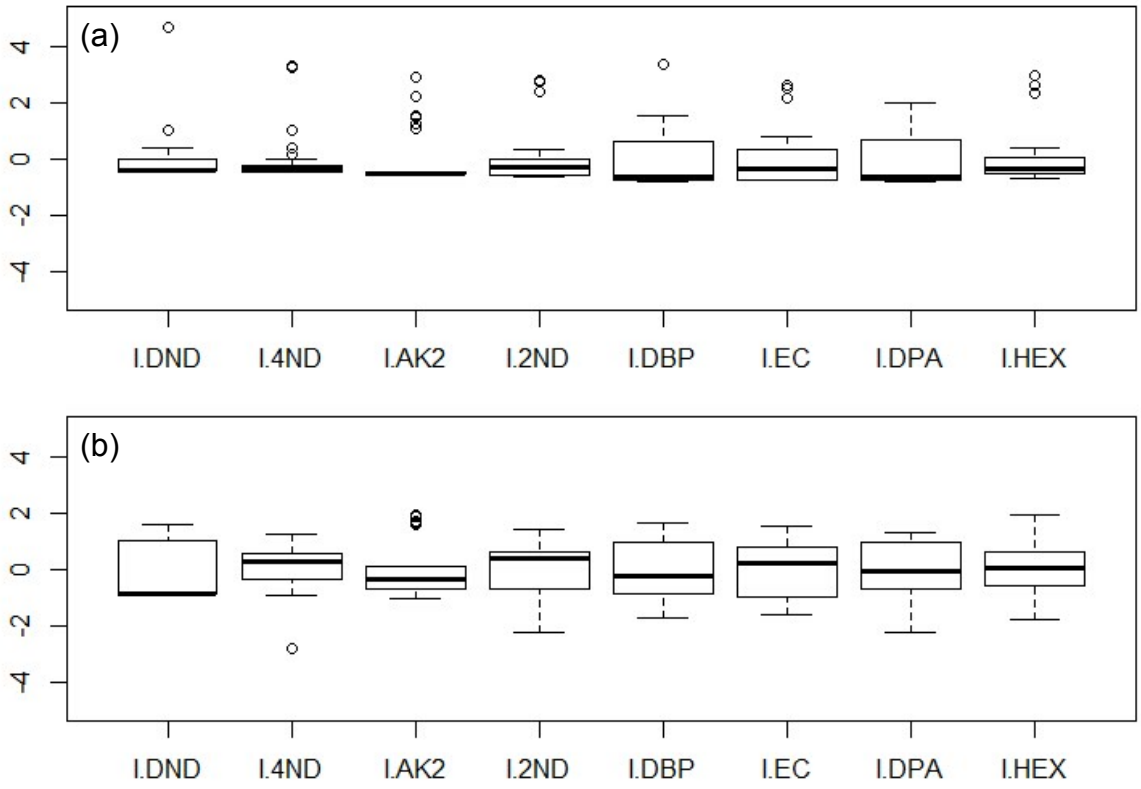


Figure S2 – Boxplots (n = 27) for the eight input variables before (a) and after (b) base-10 logarithm transformation of peak areas. Data were centred and scaled to allow a better comparison of their distribution.

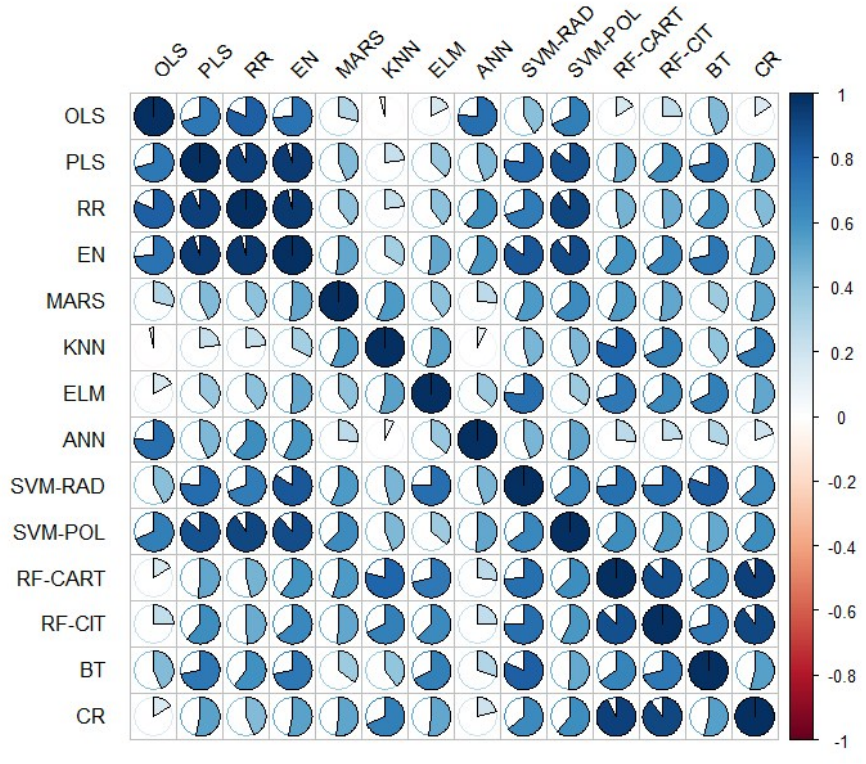
Tested machine learning methods

Table S3 – Tested regression techniques with corresponding method values in the *caret* package, libraries, and tuning parameters.

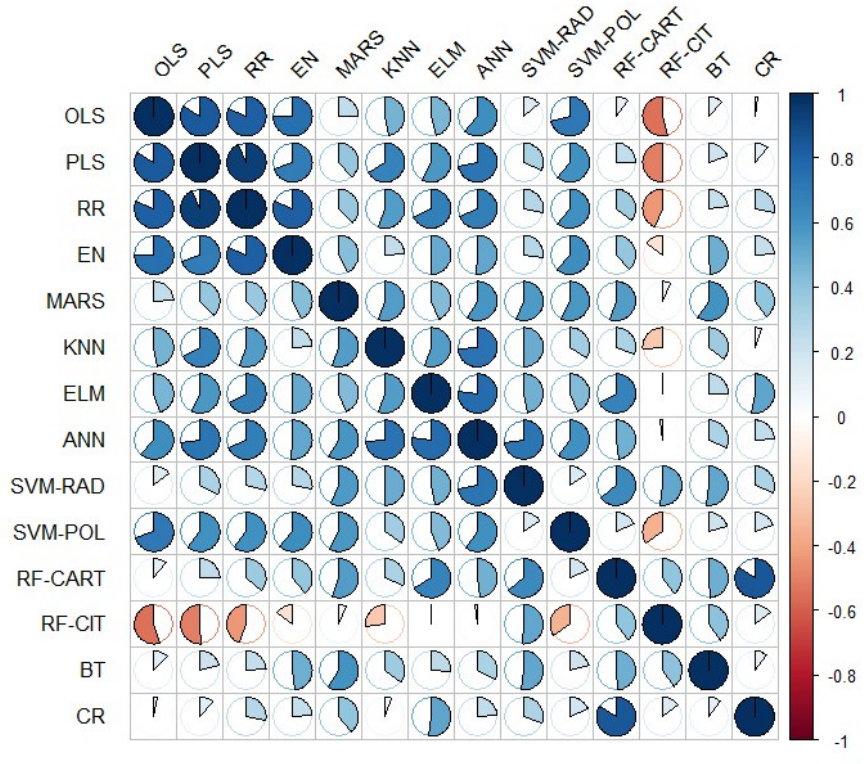
Regression method	***Abbr.***	***CARET method value***	***Library***	***Tuning parameters***
Ordinary least-squares	OLS	lm	stats	-
Partial least-squares	PLS	pls	pls	*ncomp*
Ridge regression	RR	ridge	elasticnet	*lambda*
Elastic net	EN	enet	elasticnet	*lambda*, *fraction*
Multivariate adaptive regression spline (bagged)	MARS	bagEarth	earth	*nprune*, *degree*
k-nearest neighbors	KNN	knn	knn	*k*
Extreme learning machine	ELM	elm	elmNN	*actfun*, *nhid*
Single-layer perceptron neural network (averaged)	ANN	nnet	nnet	*size*, *decay*
Support vector machines (radial basis kernel)	SVM-RAD	svmRadial	kernlab	*C*
Support vector machines (polynomial kernel)	SVM-POL	svmPoly	kernlab	*C*, *degree*, *scale*
Random forest (CART learner)	RF-CART	rf	randomForest	*mtry*
Random forest (CIT learner)	RF-CIT	cforest	party	*mtry*
Boosted trees	BT	gbm	gbm	*interaction.depth*, *n.trees*, *shrinkage*, *n.minobsinnode*
Cubist	CR	cubist	cubist	*committees*, *neighbors*

Correlations between absolute errors

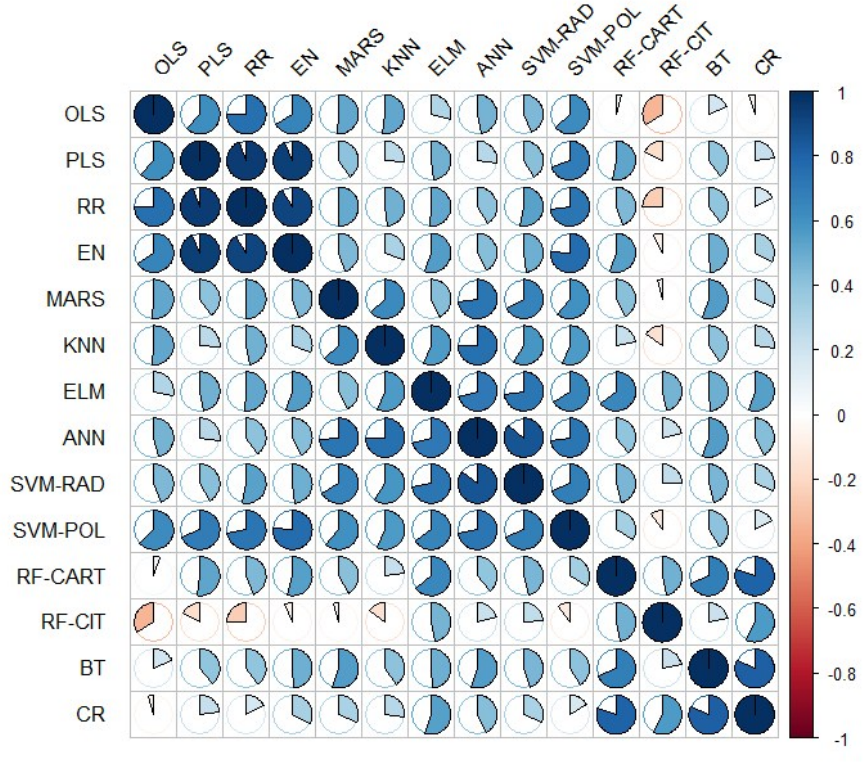
(a) O_{DPA}



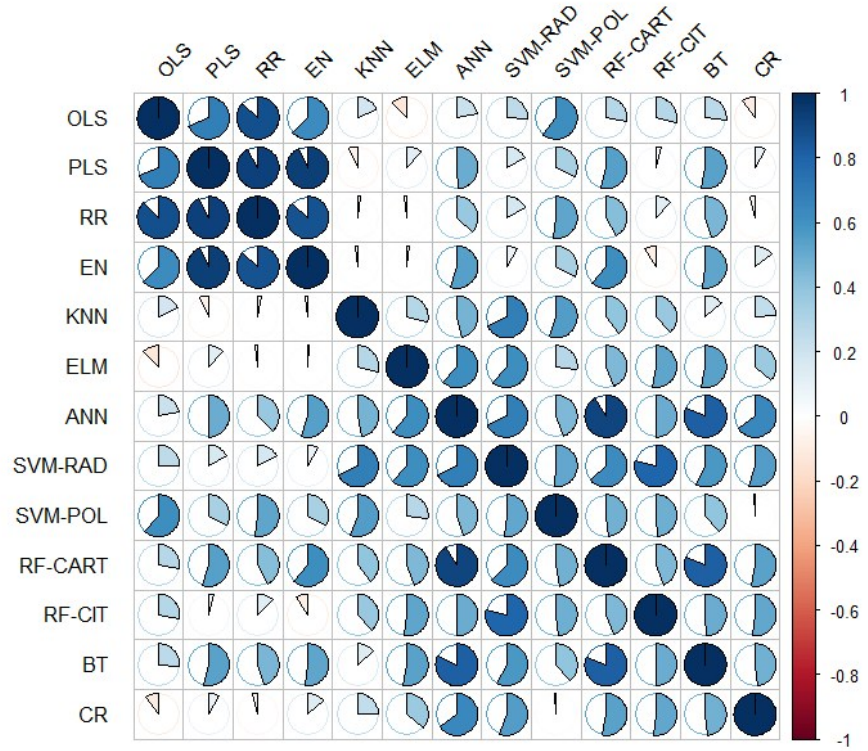
(b) O_{2ND}



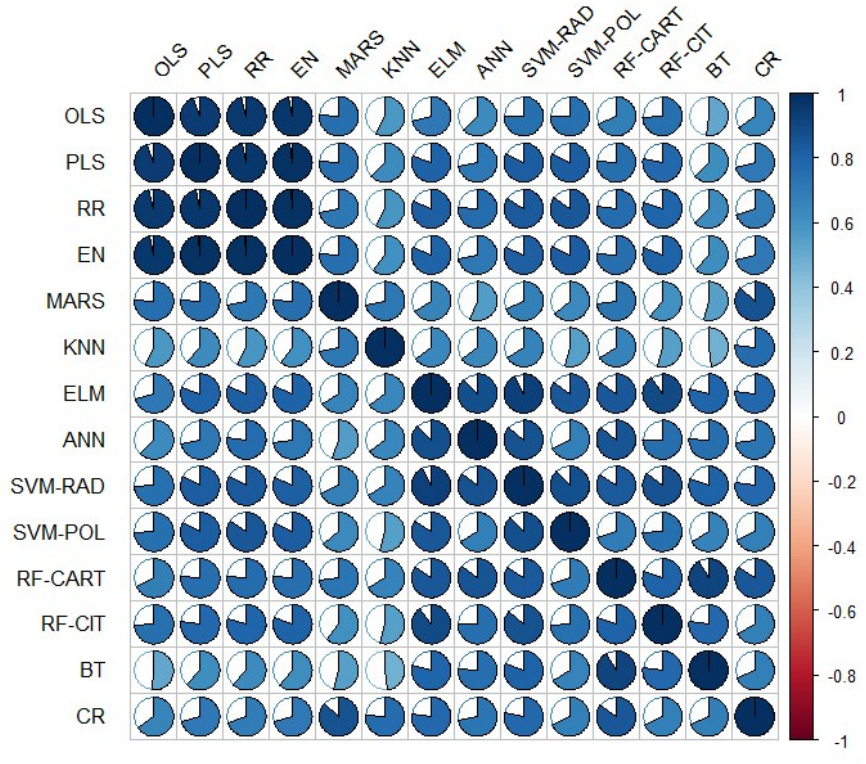
(c) O_{4ND}


OLS
PLS
RR
EN
MARS
KNN
ELM
ANN
SVM-RAD
SVM-POL
RF-CART
RF-CIT
BT
CR
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

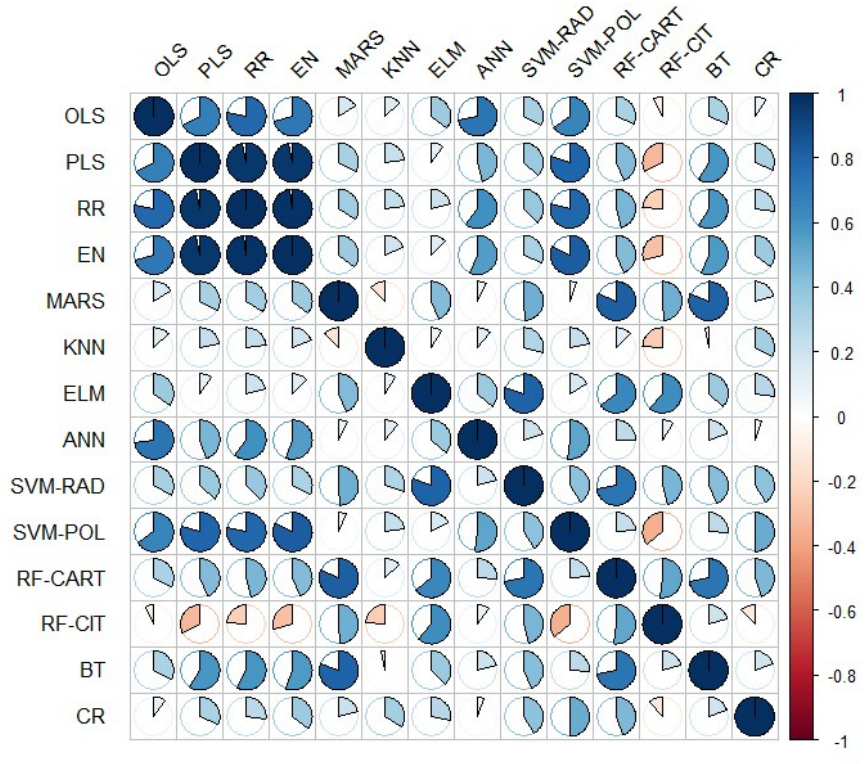
(d) O_{AK2}


OLS
PLS
RR
EN
KNN
ELM
ANN
SVM-RAD
SVM-POL
RF-CART
RF-CIT
BT
CR
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

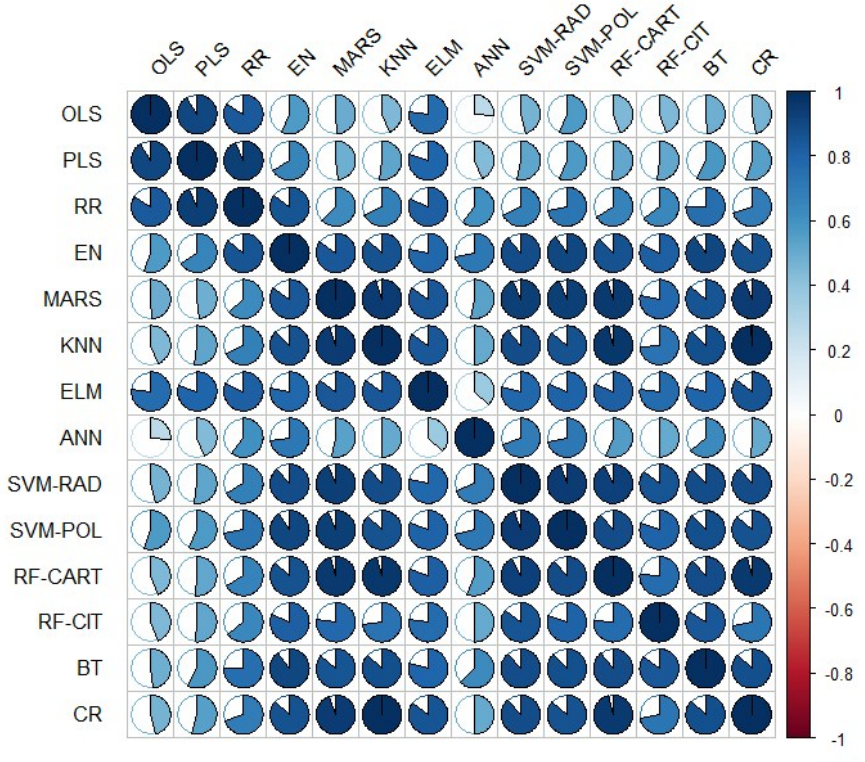
(e) O_{EC}


OLS
PLS
RR
EN
MARS
KNN
ELM
ANN
SVM-RAD
SVM-POL
RF-CART
RF-CIT
BT
CR
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

(f) O_{DBP}


OLS
PLS
RR
EN
MARS
KNN
ELM
ANN
SVM-RAD
SVM-POL
RF-CART
RF-CIT
BT
CR
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

(g) O_{NG}



(h) O_{DOF}

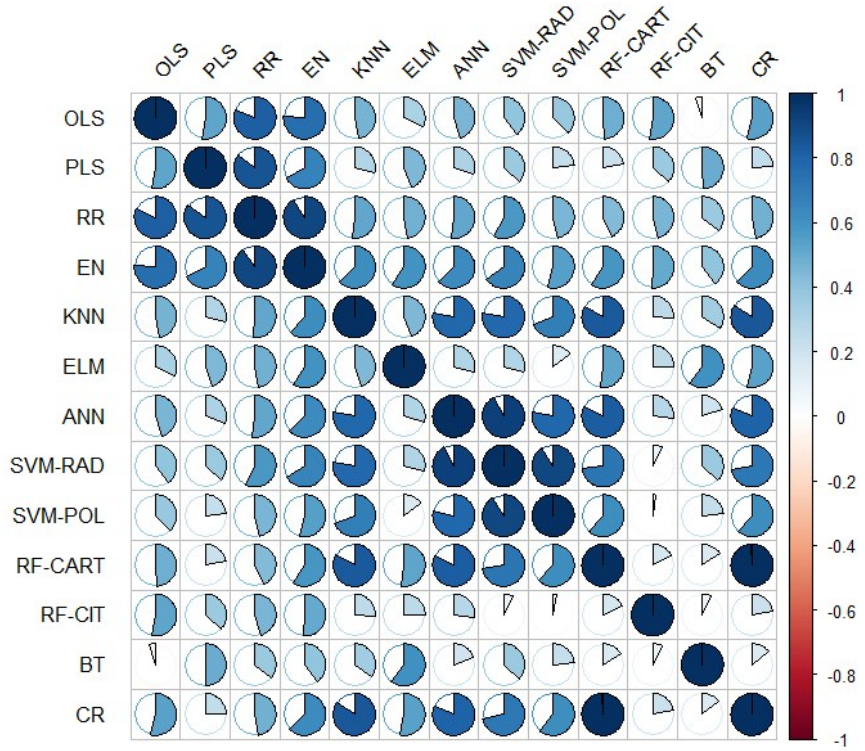


Figure S3 – Plot of Pearson's correlation coefficients (PPCs) between the absolute errors obtained after predictions of the eight outputs with the fourteen tested regression methods (extrapolation mode). For O_{AK2} and O_{DOF}, MARS has been omitted as the algorithm did not converge to solution for some resampled data subsets during model training.

Summary table of RMSEs

Table S4 – Root-mean-square errors (RMSEs) observed after prediction of the logPAs of the eight output compounds using the fourteen tested regression methods (extrapolation mode).

Regression technique	O_{4ND}	O_{AK2}	O_{DOF}	O_{DBP}	O_{EC}	O_{DPA}	O_{NG}	O_{2ND}	***Median***
OLS	1.553	1.245	0.943	1.642	2.015	1.752	3.925	1.299	1.597
PLS	1.235	1.363	1.226	1.754	2.146	1.460	3.606	1.153	1.411
RR	1.334	1.333	1.086	1.604	2.033	1.418	3.161	1.150	1.376
EN	1.232	1.275	0.972	1.547	2.002	1.337	2.600	1.042	1.306
MARS	0.951	NA[a]	NA[a]	1.755	2.252	1.501	2.613	1.014	1.628
KNN	0.881	1.754	1.347	1.541	1.678	1.069	2.803	0.834	1.444
ELM	1.129	1.246	0.538	1.762	1.948	1.353	2.756	1.046	1.299
ANN	0.880	1.380	1.234	1.053	1.904	1.635	3.595	0.873	1.307
SVM-RAD	0.777	1.474	0.882	1.424	1.701	1.125	2.427	0.730	1.274
SVM-POL	1.168	1.343	1.111	1.654	1.934	1.174	2.717	1.064	1.259
RF-CART	0.924	1.781	0.510	1.435	1.385	1.081	2.592	0.633	1.233
RF-CIT	1.100	2.066	0.412	1.582	2.055	1.487	1.969	1.169	1.534
BT	1.039	1.793	0.817	1.830	1.736	1.311	2.693	0.961	1.523
CR	1.030	1.939	1.167	1.556	1.355	1.011	2.761	0.542	1.261

[a] algorithm failed to converge to solution for some ammunition types.

Compound characteristics vs. class RMSEs

Table S5 – Comparison between the characteristics of the eight output compounds and the mean root-mean-square errors (RMSEs) for the two classes of regression technique. Grey-coloured correlation values are specifically discussed in the manuscript.

Output	*Compound characteristics*			*Mean RMSEs*		
	Mean logPA	*Range logPAs*	*Linearity (R^2)[a]*	*Linear models*	*Non-linear models*	*Difference (δ_{RMSE})*
O_{4ND}	1.478	2.463	0.668	1.338	0.988	-0.350
O_{AK2}	1.252	4.427	0.902	1.304	1.642	0.338
O_{DOF}	0.367	3.300	0.656	1.057	0.891	-0.166
O_{DBP}	3.685	4.471	0.717	1.637	1.559	-0.078
O_{EC}	3.692	5.502	0.877	2.049	1.795	-0.254
O_{DPA}	4.232	4.093	0.732	1.492	1.275	-0.217
O_{NG}	5.314	5.927	0.314	3.323	2.693	-0.631
O_{2ND}	1.841	2.785	0.762	1.161	0.886	-0.275
Correlations	*Mean logPA*			*0.809*	*0.748*	*-0.513*
	Range logPAs			*0.809*	*0.917*	*-0.138*
	Linearity			*-0.693*	*-0.493*	*0.768*

[a] degree of linearity between the considered output compound and the full set of input compounds, measured with the coefficients of determination of the reciprocal OLS model.

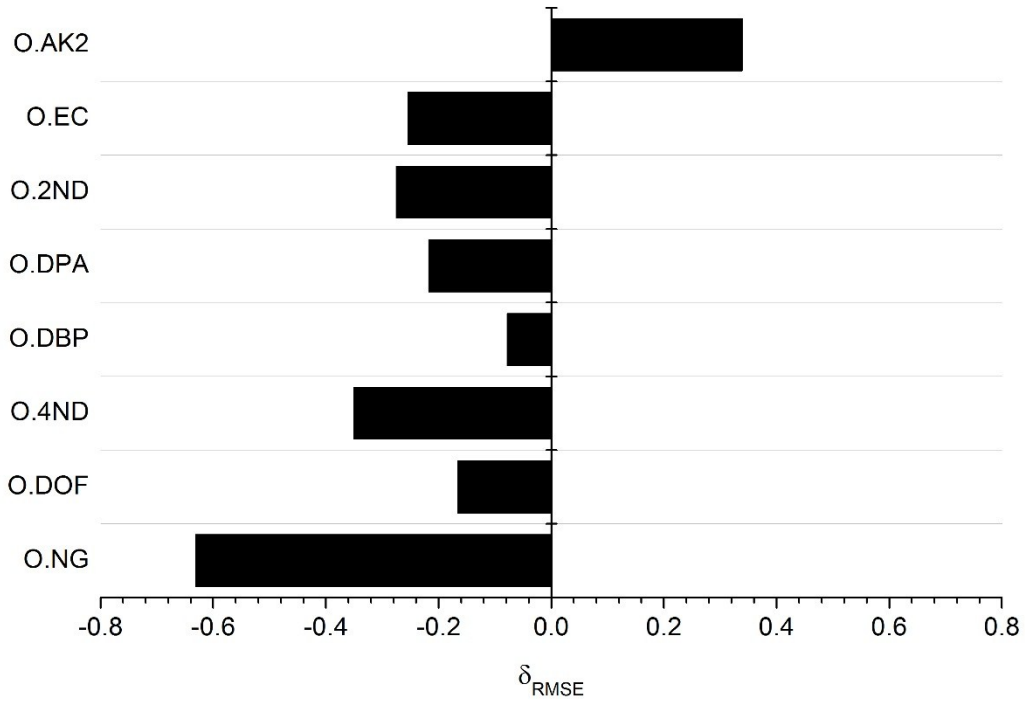


Figure S4 – Visual representation of the observed mean RMSE enhancement after moving from linear to non-linear methods (δ_{RMSE}), as a function of the eight output compounds. The latter are ranked according to their decreasing degree of linearity with the full set of predictors (most linear on top).

Summary tables of PCCs

Table S6 – Observed PCCs after within-ammunition comparisons of profiles predicted using the two tested combinations of models (CoM_1 and CoM_2) and both training schemes (extrapolation and interpolation).

	M-M	***Extrapolation***				***Interpolation***			
		Predicted-Measured		***Predicted-Predicted***		***Predicted-Measured***		***Predicted-Predicted***	
		CoM_1	***CoM_2***	***CoM_1***	***CoM_2***	***CoM_1***	***CoM_2***	***CoM_1***	***CoM_2***
General statistics									
Min	0.939	0.367	0.370	0.905	0.914	0.417	0.254	0.855	0.781
1st Q.	0.999	0.711	0.874	0.946	0.971	0.965	0.926	0.970	0.966
Median	1.000	0.839	0.908	0.990	0.994	0.982	0.986	0.981	0.994
3th Q.	1.000	0.918	0.937	0.996	0.997	0.993	0.994	0.990	0.996
Max	1.000	0.974	0.972	1.000	1.000	0.999	0.998	0.999	0.999
Medians values for the single ammunition types									
Ge357	0.995	0.933	0.927	0.994	0.985	0.988	0.986	0.978	0.972
Ge45	1.000	0.658	0.954	0.992	0.972	0.975	0.986	0.973	0.996
Ma357	0.995	0.841	0.952	0.937	0.994	0.977	0.992	0.935	0.966
Ma45	0.970	0.458	0.470	0.945	0.917	0.597	0.473	0.897	0.856
Pm45	1.000	0.740	0.878	0.933	0.994	0.974	0.994	0.981	0.994
Re45	0.999	0.808	0.927	0.996	0.970	0.992	0.976	0.988	0.982
Sa357	1.000	0.857	0.840	0.995	0.996	0.986	0.872	0.987	0.995
Se357	1.000	0.874	0.902	0.948	0.999	0.994	0.996	0.988	0.996
Se45	0.999	0.971	0.958	0.969	0.958	0.995	0.993	0.989	0.935

M-M: measured-measured comparisons

Table S7 – Observed PCCs after between-ammunition comparisons of profiles predicted using the two tested combinations of models (CoM_1 and CoM_2) and both training schemes (extrapolation and interpolation).

	M-M	***Extrapolation***				***Interpolation***			
		Predicted-Measured		***Predicted-Predicted***		***Predicted-Measured***		***Predicted-Predicted***	
		CoM_1	***CoM_2***	***CoM_1***	***CoM_2***	***CoM_1***	***CoM_2***	***CoM_1***	***CoM_2***
General statistics									
Min	-0.387	-0.232	-0.387	-0.025	0.191	-0.379	-0.437	0.185	0.188
1st Q.	0.457	0.585	0.550	0.707	0.619	0.533	0.522	0.597	0.554
Median	0.659	0.715	0.708	0.817	0.755	0.707	0.679	0.757	0.721
3th Q.	0.785	0.867	0.845	0.893	0.879	0.816	0.817	0.853	0.830
Max	0.990	0.993	0.997	0.989	0.986	0.997	0.992	0.994	0.994
Medians values for the single ammunition types									
Ge357	0.483	0.665	0.643	0.803	0.619	0.604	0.664	0.674	0.621
Ge45	0.653	0.698	0.679	0.574	0.657	0.750	0.669	0.733	0.679
Ma357	0.776	0.812	0.806	0.775	0.845	0.820	0.808	0.815	0.776
Ma45	0.280	0.651	0.611	0.844	0.826	0.329	0.285	0.767	0.665
Pm45	0.528	0.728	0.736	0.851	0.838	0.588	0.591	0.653	0.561
Re45	0.679	0.693	0.667	0.829	0.712	0.682	0.643	0.748	0.711
Sa357	0.657	0.692	0.755	0.857	0.825	0.714	0.701	0.766	0.797
Se357	0.701	0.822	0.808	0.880	0.843	0.755	0.775	0.797	0.769
Se45	0.686	0.759	0.734	0.786	0.656	0.743	0.782	0.739	0.718

M-M: measured-measured comparisons

Numbers of pairwise comparisons and observed AUCs in ROC analysis

Table S8 – Numbers of pairwise comparisons carried out under each tested scenario (n) and observed areas under curves (AUCs) after receiver operating characteristics (ROC) analysis of PCC data.

	M-M	***Extrapolation***				***Interpolation***			
		Predicted-Measured		***Predicted-Predicted***		***Predicted-Measured***		***Predicted-Predicted***	
		CoM_1	***CoM_2***	***CoM_1***	***CoM_2***	***CoM_1***	***CoM_2***	***CoM_1***	***CoM_2***
n (within)	54	108	108	27	27	108	108	27	27
n (between)	576	864	864	324	324	864	864	324	324
AUC	0.998	0.633	0.824	0.940	0.976	0.916	0.894	0.966	0.962

M-M: measured-measured comparisons