PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Prediction models of retention indices: application to gas chromatography coupled to high-resolution mass spectrometry for two column types: DB-624 and HP-5ms

A. Haiduc*, E. Dossin, P. Diana, P.A. Guy, N.V. Ivanov, M.C. Peitsch
PMI R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland

* Presenting author

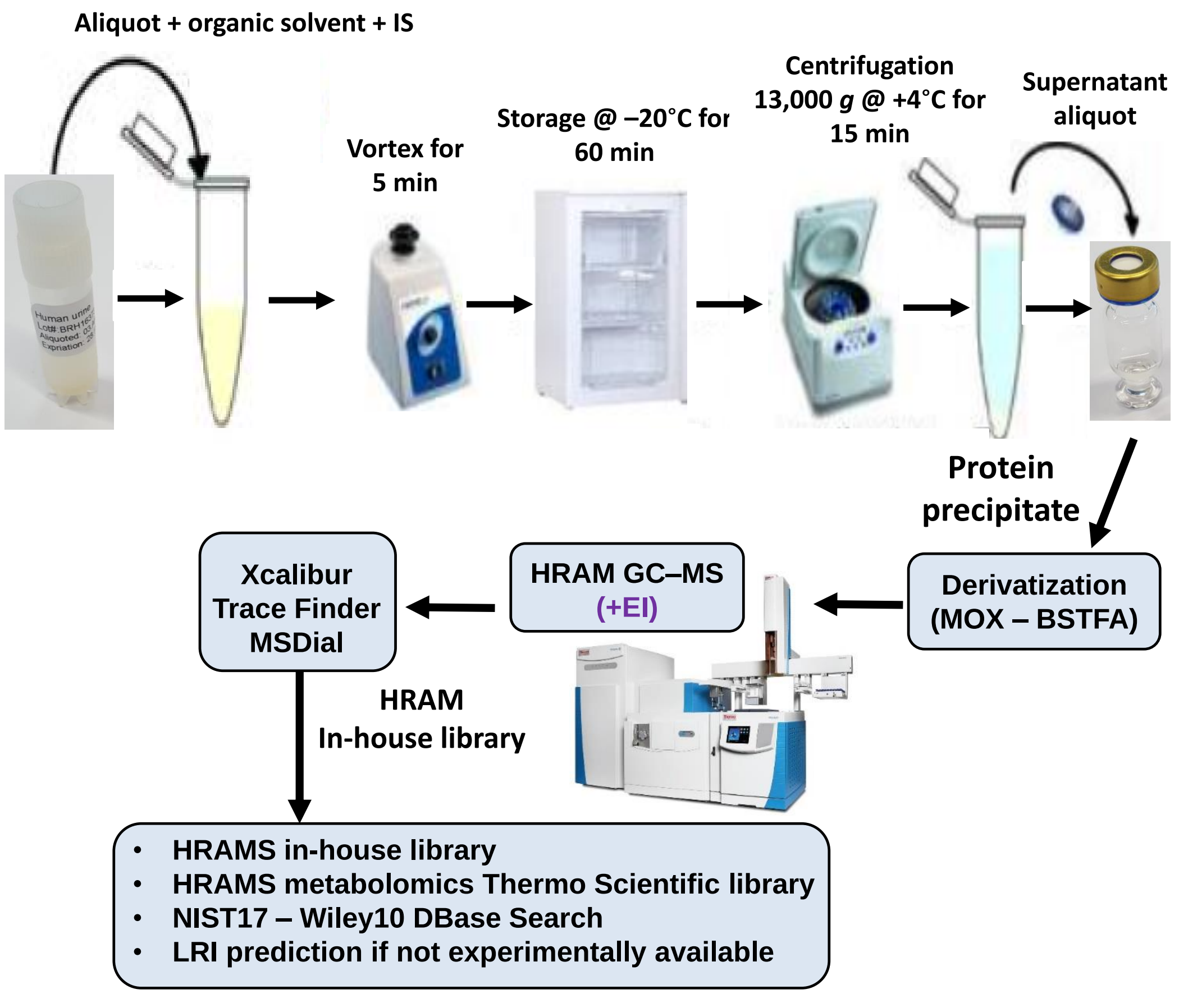
Introduction and objectives

- ➢ The scientific community has largely reported and accepted that high-resolution accurate mass spectrometry (HRAMS) combined with advanced chemoinformatic tools enhance the confidence level for compound identification **[1,2,3]**.
- ➢ Using chemoinformatic tools and software openly available, a prediction model for linear retention indices (LRI) was developed based on the structural 2D molecular data to help in the tedious task of small molecule identification for untargeted metabolomics application.
- ➢ The accuracy of our LRI prediction is assessed from the Mahalanobis distance and used as a quality check to estimate the capacity of the model when predicting all urine and blood metabolites currently registered in t)he Human Metabolome Database (HMDB) **[4]**.
- ➢ LRI prediction was assessed for two gas chromatography (GC) columns from reported metabolites to be present in urine or plasma after oximation (MOX) and silylation (TMS) derivatization.

Methods

Workflow GC-Q Exactive MS

- ➢ Reference standards (biological sample) dried under nitrogen.
- ➢ Derivatization (MOX and TMS) using TriPlus RSH autosampler.
- ➢ EI full scan acquisition using GC-Q Exactive™ MS (60k resolution @ *m/z* 200).
- ➢ Analyses were realized on two GC columns (DB-624 and HP-5ms with different physicochemical properties).
- ➢ **DB-624** covers LRI values from 500 to 1,900 (using C5 to C19 alkanes).
- ➢ **HP-5ms** covers LRI values from 800 to 3,000 (using C8 to C30 alkanes).
- ➢ Curated EI accurate mass spectra with experimental LRI values registered in our HRAMS in-house library (**Figure 1**).


Aliquot + organic solvent + IS
Vortex for 5 min
Storage @ –20°C for 60 min
Centrifugation 13,000 *g* @ +4°C for 15 min
Supernatant aliquot
Protein precipitate
Derivatization (MOX – BSTFA)
HRAM GC–MS (+EI)
Xcalibur Trace Finder MSDial
HRAM In-house library

- HRAMS in-house library
- HRAMS metabolomics Thermo Scientific library
- NIST17 – Wiley10 DBase Search
- LRI prediction if not experimentally available

Figure 1: *Workflow used for metabolomics application. Similar protocol used for building HRAMS in-house library.*

In silico LRI prediction

A subset of these reference standards was randomly split into training (**HP-5ms** n=407; **DB-624** n=549) and test (**HP5-ms** n=431; **DB-624** n=346) sets and was used to optimize the LRI prediction model.

1) **Molecular descriptor extraction.** Based on the CAS number, the «MOL» files were downloaded from the CIR **[5]**. *In silico* MOX-TMS derivatization was performed using Metaboderivatizer **[6]**, and molecular descriptors were calculated with the Mordred Python package **[7]**. The resulting descriptor tables were filtered for variables having non-empty fields and/or showing consistent values across all training molecules. The resulting data was used as the base for modelling LRI from molecular properties (**Figure 2**).

2) **Model development retaining Lasso and PLS**. The prediction equation between calculated physicochemical parameters and structural similarity with experimental LRI values was calculated in Matlab **[8]** using multiple algorithms: Lasso, PLS, genetic algorithm variable selection, and MLR and neural networks. Lasso, PLS, and neural network provided the best LRI model with the smallest validation test set error. In terms of prediction ease of implementation in Python, Lasso and PLS were retained for their portability.

3) **In-model calculation for new predicted molecules with the Mahalanobis distance**. To verify the model applicability, the Mahalanobis distance, –d1, between the molecular descriptors of the training set and the new molecules to be predicted is calculated. The predicted LRI is retained only if the d1 value is less than 3 times the average Mahalanobis distance of the training set.

4) **Model validation with test set & NIST17 database.** LRI prediction models were tested with n=431 (**HP-5ms**) and n=346 (**DB-624**) reference compounds. In addition, experimental LRI values were extracted from NIST14 for non-polar and semi-polar GC columns (similar to **HP-5ms**). Experimental LRI values for 1,104 molecules were compared with predicted LRI values from the PLS and Lasso models.

Results

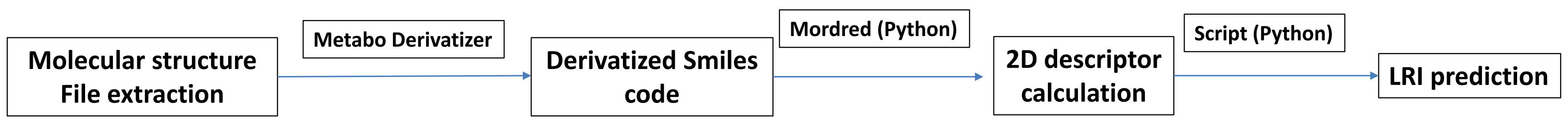

Molecular structure File extraction
Metabo Derivatizer
Derivatized Smiles code
Mordred (Python)
2D descriptor calculation
Script (Python)
LRI prediction

Figure 2: *Workflow to predict LRI values.*

- ➢ **<u>Model performance assessment</u>**: LRI prediction values were plotted vs. LRI experimental internal and external values using Lasso models (**Figure 3 for both training and test sets**).
- ➢ **<u>To identify possible outliers</u>**: application of the maximum Mahalanobis distance (**Figure 3 insets**).
- ➢ **<u>Confidence interval</u>** for the predicted LRI as compared to model error allows setting of retention index window in searchable GC-MS libraries.
- ➢ **<u>Regression coefficients</u>**: out of 1,681 descriptors, Lasso regression retained 276 significant predictors (**Table 1**).
- ➢ **<u>Lasso model</u>** resulting in a R^2 of 0.9922 (training HP-5ms) and 0.9740 (training DB-624), thus confirming the large applicability to a wide set of molecular structures (**Table 2**).

Table 1: *Selected descriptors significantly impacting LRI prediction.*

Selected Descriptors	
Increasing LRI	Decreasing LRI
AATSC7c	6,2, and 5 mean Topological Charge (JGI)
Atom-bond connectivity (molecular branching)	Geary coeffcient
Number of double bonds	BalabanJ
Sp2 carbons bound to three carbon atoms	number of ssO
Amide N	number of aaO
number of sssN	n9 fused and aromatic ring count
number of hydrogen bond donor	
molecular distance edge between primary N and tertiary N	
molecular distance edge between primary N and primary N	
n6 ring and n10 fused aromatic ring count	

Table 2: *LRI model regression performance.*

	HP5		DB624	
Method	R2 training	R2 test	R2 training	R2 test
Lasso	0.99	0.98	0.97	0.94
PLS	0.99	0.97	0.97	0.94
NN	0.99	0.96	0.97	0.91
GA-MLR	0.96	0.91	0.89	0.85
Regression Trees	0.87	0.81	0.84	0.79

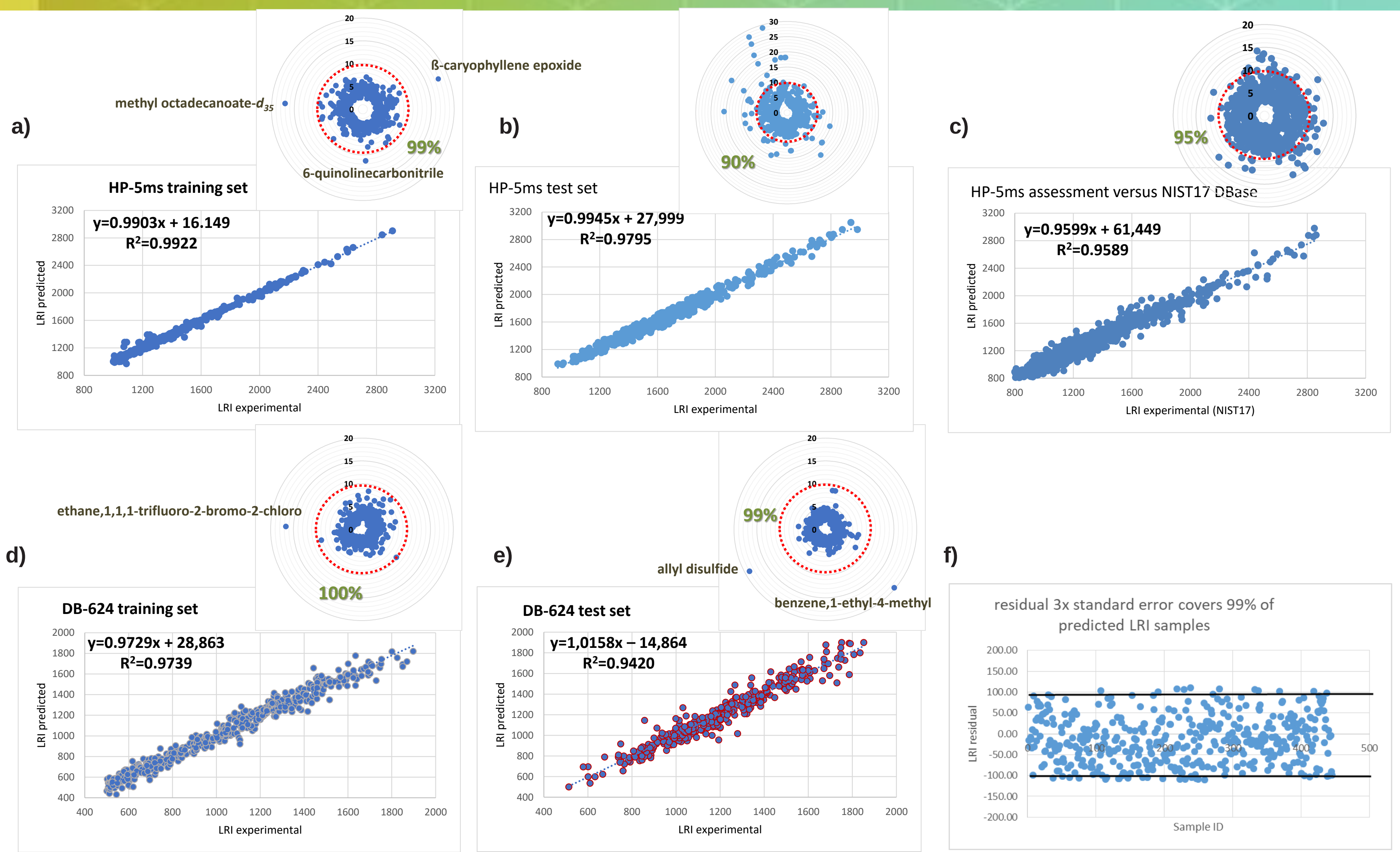


Figure 3: *Correlation plots, accuracy, and Mahalanobis distance (insets) of the compounds used a) HP-5ms training (**n=407**), b) HP-5ms test (**n=431**), c) HP-5ms model validation with nonpolar-semi-polar NIST14 (**n=1,104**), d) DB-624 training (**n=549**), e) DB-624 test (**n=346**), f) HP-5ms residual of mean LRI values.*

Application to metabolomics (e.g., blood & urine samples)

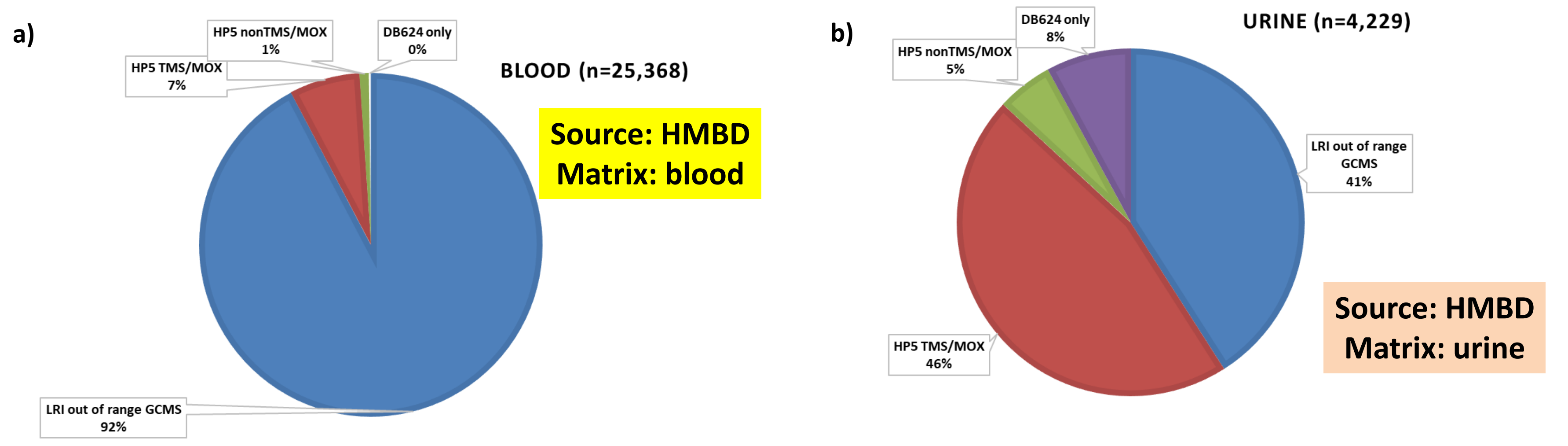


Figure 4: *Metabolites percentage coverage (taken from metabolites registered in HMDB [**4**]) expected on a GC-MS system based on predicted LRI values from a) blood and b) urine matrices. In this plot, the percentage value given for DB-624 GC column considered only LRI predicted values from 500–800, whereas HP-5ms from LRI prediction of 800–2,900.*

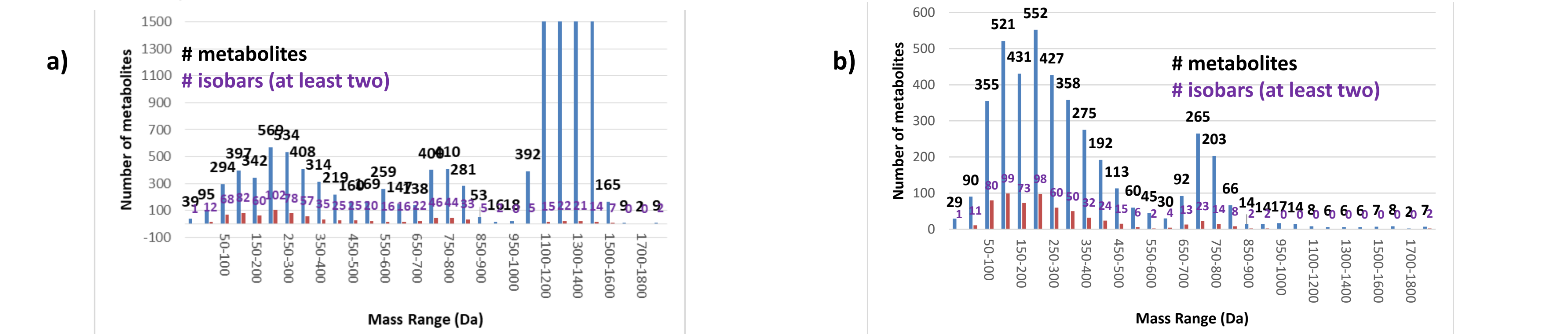


Figure 5: *Mass range distribution of the metabolites for which the GC-MS LRI was predicted in a) blood and b) urine matrices.*

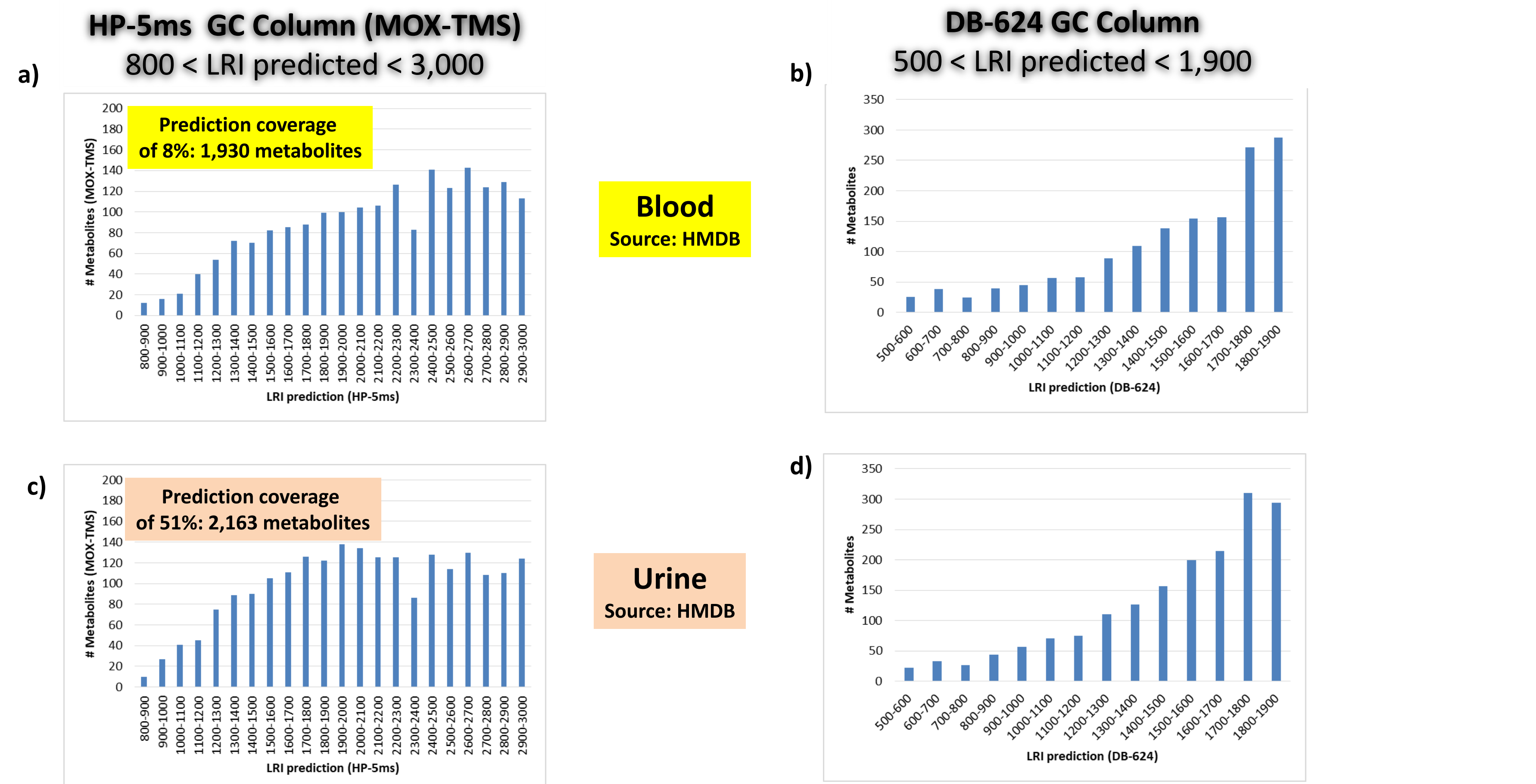


Figure 6: *Predicted LRI distribution for GC-MS in blood a) HP-5ms, b) DB-624, and in urine c) HP-5ms, d) DB-624 GC columns. It is worth mentioning that MOX-TMS was not taken into consideration for the DB-624 GC column, and these metabolites will elute later than their native species.*

Conclusions

- LRI prediction models have been developed for GC-MS using two columns: HP-5ms and DB-624.
- Several algorithms were evaluated, and the Lasso model provided optimal results for both GC columns (open-source software).
- These models have been cross-validated using leave-one-out methodology.
- LRI values are of great interest, as a strong correlation with NIST14 data have been shown (reproducible LRI data across laboratories).
- Application to metabolomics shows that 51% and 9% of metabolites reported in HMDB in urine and blood, respectively, could be monitored with GC-MS (using HP-5ms GC column) upon their intrinsic endogenous concentration levels.
- The Mahalanobis distance provides a critical confidence level of LRI predicted values with the possibility to select new molecules to be included in the training model if necessary.

References

[1] Dossin E. *et al.* **2016** *Anal. Chem.* *88*, 15, 7539-7547.

[2] Allen F. *et al.* **2016** *Anal. Chem.* *88*, 15, 7689-7697.

[3] Schymanski E.L. *et al.* **2017** *J. Chemoinform.* *9*, 1, 22.

[4] Wishart D.S. *et al.* **2018** *Nucleic Acids Res.* *46*, (D1):D608-17.

[5] https://cactus.nci.nih.gov/chemical/structure

[6] Matsuo T. *et al.* **2017** *Anal. Chem.* *89*, 6766-6773.

[7] Moriwaki H. *et al.* **2018** *J. Cheminform.* *10*, 4.

Metabolomics 2019, Den Haag

Competing Financial Interest – The research described in this poster was sponsored by Philip Morris International