United States Environmental Protection Agency

A comparison of three chromatographic retention time prediction models

Andrew D. McEachran[1,3]*, Kamel Mansouri[1,3], Seth R. Newton[2], Brandiese E.J. Beverly[1,2],
Jon R. Sobus[2], and Antony J. Williams[3]

[1]Oak Ridge Institute of Science and Education (ORISE) Research Participant, Research Triangle Park, NC
[2]U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Research Triangle Park, NC
[3]U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology, Research Triangle Park, NC

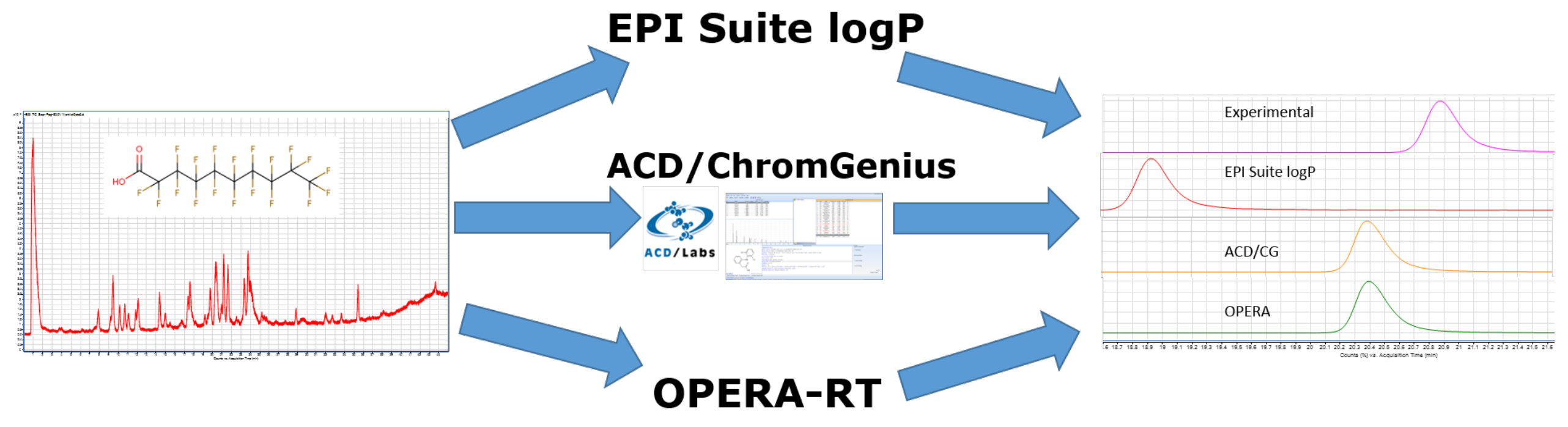
* mceachran.andrew@epa.gov I ORCiD: 0000-0003-1423-330X

CINF 28
ACS Fall 2017
Washington, DC
August 20-24, 2017

Problem Definition and Goals

Problem: Non-targeted analysis (NTA) using high resolution mass spectrometry (HRMS) has revolutionized the identification of environmental contaminants. However, chemical identification remains challenging due to the vast number of unknown molecular features observed. This requires the implementation of advanced data processing techniques to improve workflows. The ideal workflow brings together harmonized data and tools from a variety of sources to increase certainty of identification. One such tool is chromatographic retention time (RT) modelling. By comparing predicted RTs of candidate structures to observed RTs of unknowns analysts can improve identification. Here we evaluate three RT prediction models using High Performance LC (HPLC)-Time of Flight (TOF)/MS data on 97 chemicals: a logP-based RT model using EPI Suite^{TM} property predictions, ACD/ChromGenius, and an in-house QSRR model, termed "OPERA-RT."

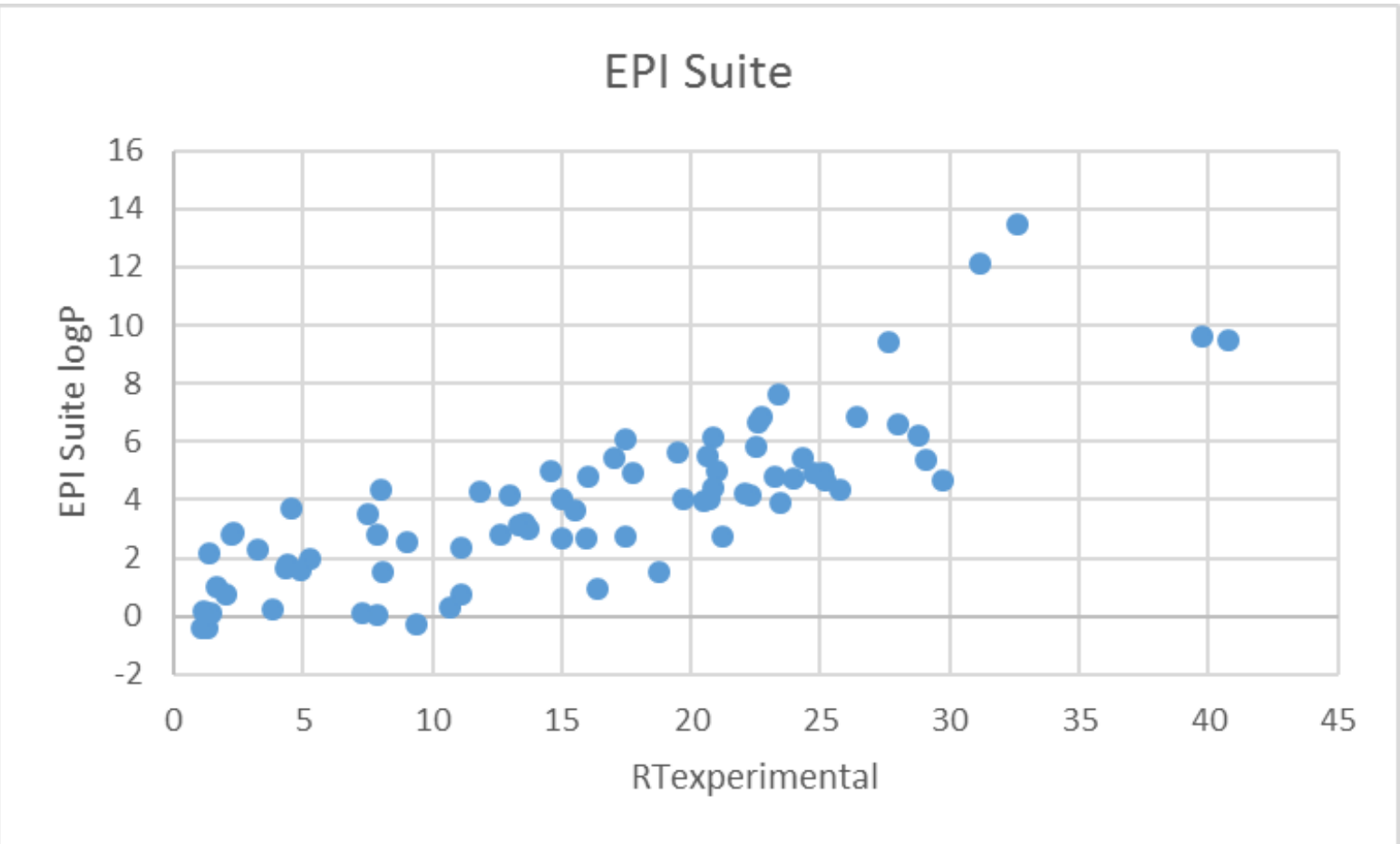
Goals: The goal of this research is to identify an efficient, accessible, and adaptable tool capable of supporting a comprehensive NTA workflow. We aim to demonstrate the applicability of three separate RT prediction models for use in NTA by comparing the relative predictive abilities and applicability to NTA.



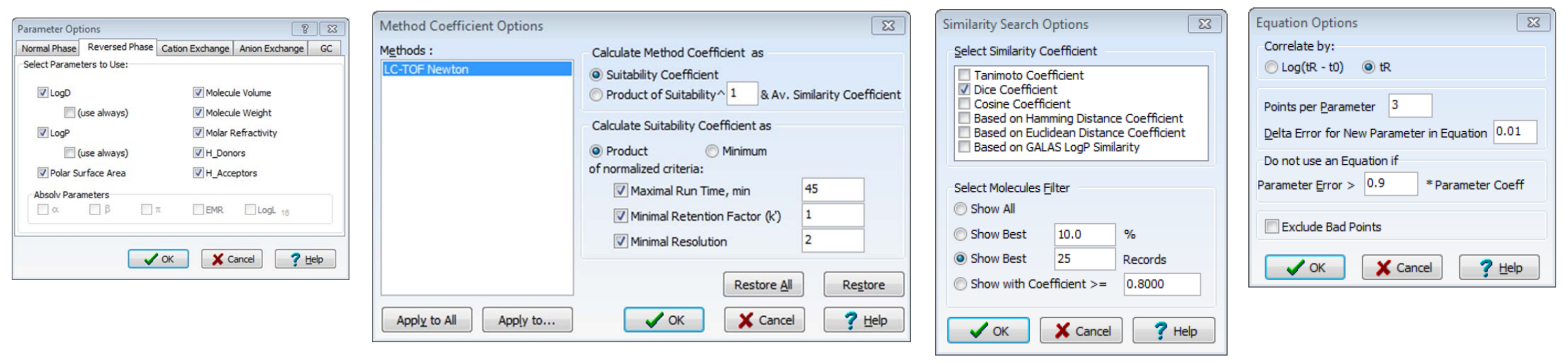
Modeling Approaches

Structure Sets: Experimental RTs obtained from standard mixtures analyzed via LC-TOFMS between 0-45 min (1). Chemicals were split into a Training Set (n=78) and Test/Validation Set (n=19)

1. EPI Suite logP Retention Time Prediction Model
   - EPI Suite^{TM} (2) was used to generate logP values
   - logP regressed against experimental RT (RTexp)



2. ACD/ChromGenius (Advanced Chemistry Development, Toronto, Canada)
   - A proprietary algorithm using physicochemical parameters including logP, logD, molecular weight, molecular volume, polar surface area, etc.



3. OPERA-RT
   - Quantitative Structure-Retention Relationship (QSRR) Model (3)
   - Genetic Algorithms (GAs) coupled to partial least squares (PLS) using PaDeL structural descriptors (4)

Predicted vs Experimental RT

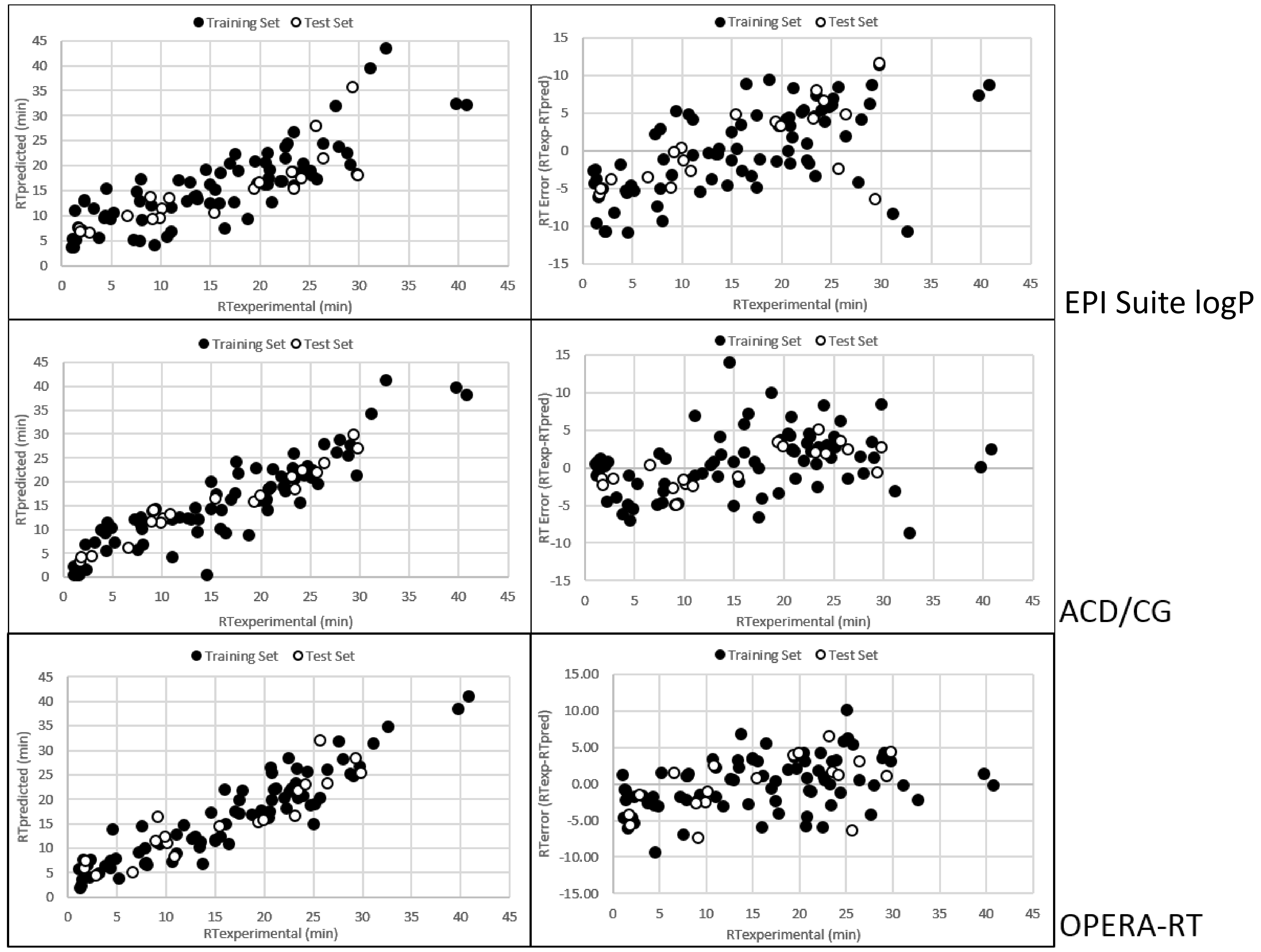


Figure 1 (ABOVE). Experimental versus predicted retention times (left) and RT prediction error (right) of the combined training and test sets for all three models: EPI Suite^{TM} logP (top), ACD/ChromGenius (middle), and OPERA-RT (bottom). Total run time was 45 minutes.

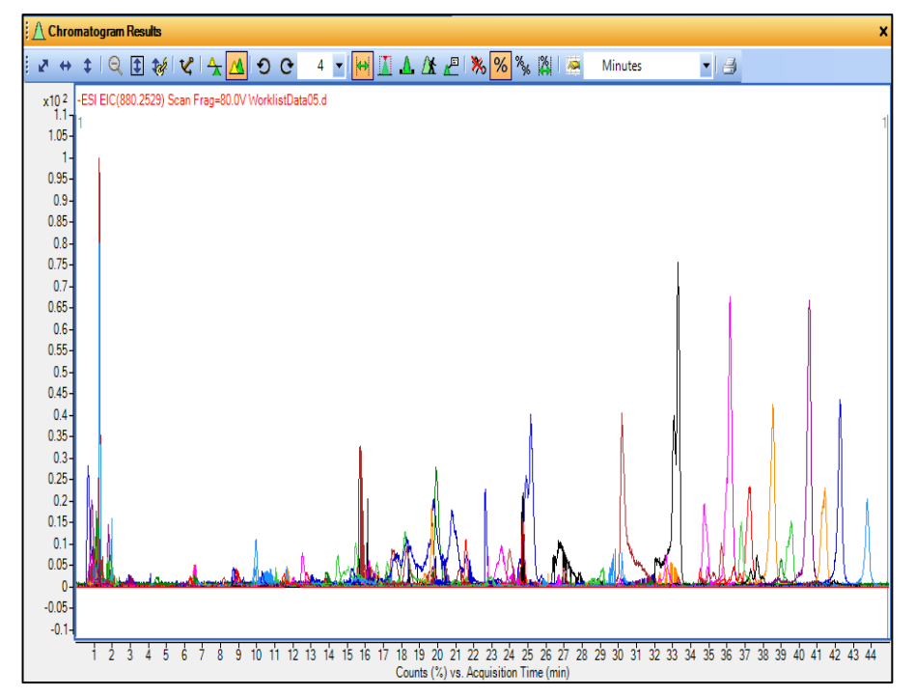
	logP	ChromGenius	OPERA-RT
Training Set (n=78)			
R^2	0.66	0.81	0.86
RMSE (min)	5.58	4.18	3.56
Absolute Mean Error (min)	4.71	3.25	2.88
Test Set (n=19)			
R^2	0.69	0.92	0.83
RMSE (min)	5.14	2.66	3.86
Absolute Mean Error (min)	4.41	2.36	3.28
Combined (n=97)			
R^2	0.66	0.83	0.86
RMSE (min)	5.50	3.93	3.60
Absolute Mean Error (min)	4.65	3.03	2.93

Table 1. Model performance summary statistics for all three models.

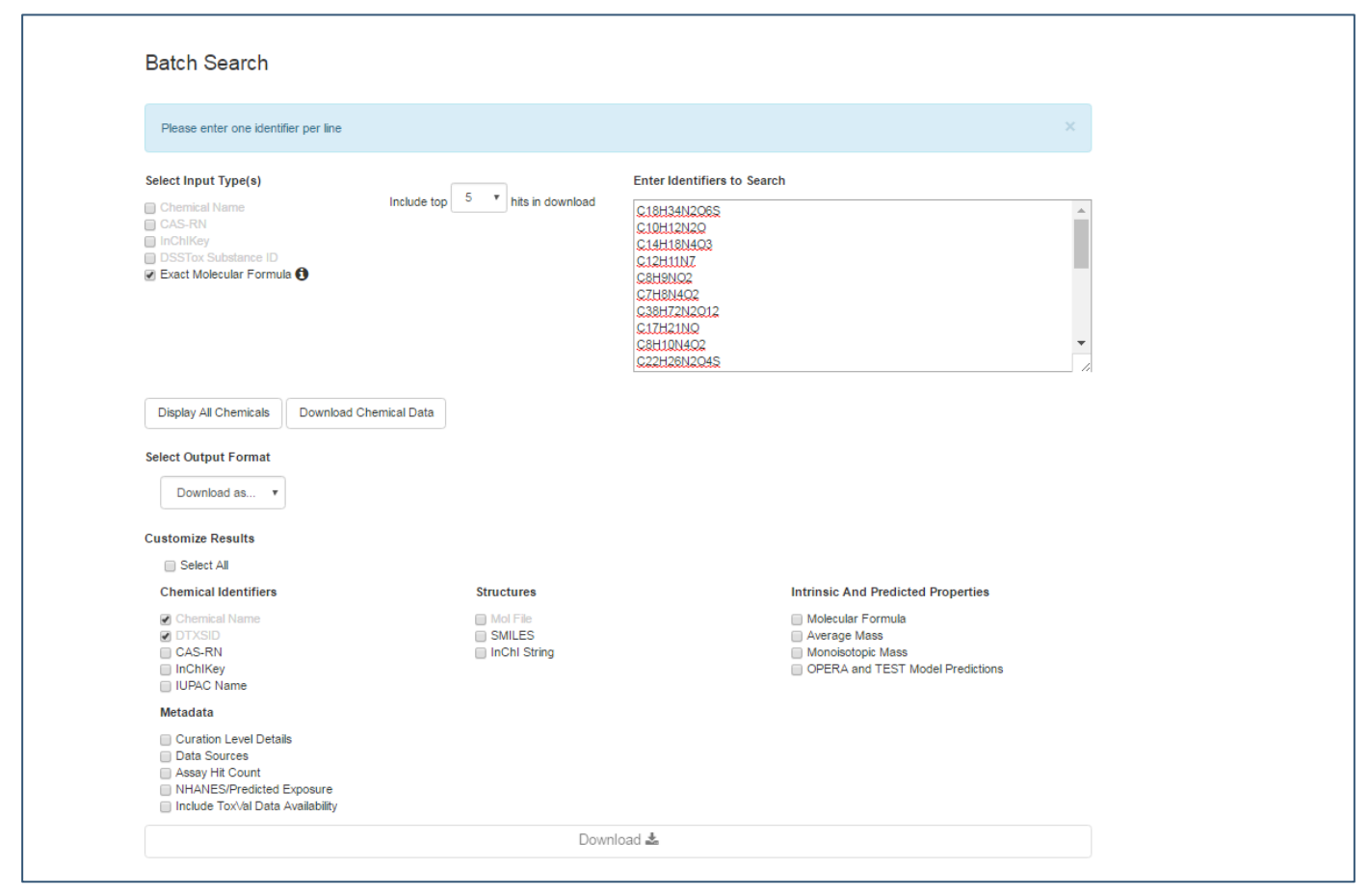
	Number of predicted RTs found within window of experimental RTs		
RT window (± % of total run, ± min)	EPI Suite^{TM} logP	ChromGenius	OPERA-RT
Training Set (n=78)			
± 5% (2.25 min)	19	36	36
± 10% (4.50 min)	39	56	63
± 15% (6.75 min)	59	70	74
± 20% (9.00 min)	70	76	76
Test Set (n=19)			
± 5% (2.25 min)	3	9	7
± 10% (4.50 min)	10	17	15
± 15% (6.75 min)	17	19	18
± 20% (9.00 min)	18	19	19

Table 2. Number of predicted RTs that fell within specified windows surrounding experimental RTs for the training and test set compounds.

RT Prediction in NTA



- Unknown features
- DB Matching for formula(e)



- Search formula(e) in the CompTox Chemistry Dashboard to retrieve likely candidates (5)

	OPERA-RT		ACD/ChromGenius	
RT Window	% Screened Out	% Knowns Kept	% Screened Out	% Knowns Kept
±5 min	10%	92%	20%	100%
±3 min	60%	42%	40%	83%
±2 min	80%	33%	75%	25%

Table 3. Predicted RTs of top 10 most likely structures, results displayed as percentage of candidate structures screened out within RT window of experimental and percentage of the known candidates kept

Conclusions

- OPERA-RT and ACD/ChromGenius outperform EPI Suite^{TM} logP RT prediction model
- OPERA-RT and ACD/ChromGenius predict >90% of RTs within ±15% time window of experimental RTs
- OPERA-RT, generated using Open Data, performed as well as ACD/ChromGenius, a commercial software tool

Future Work

- Incorporate RT Prediction into combined structure identification workflows
- Use RT prediction to make assessments of data quality
- Apply OPERA-RT to different chromatographic runs and implement on large scale

References

1. Rager J., et al. 2016. Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring, Environment International 88 (2016) 269-280. https://doi.org/10.1016/j.envint.2015.12.008
2. U.S. Environmental Protection Agency, Estimation programs interface suite for microsoft windows, Washington, DC, USA (2016).
3. Mansouri, K. et al. 2017. OPERA. In prep.
4. C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, Journal of Computational Chemistry 32(7) (2011) 1466-1474.
5. McEachran AD, Sobus JR, Williams AJ. 2017. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. *Anal. Bioanal. Chem.* 409(7): 1729-1735. doi:10.1007/s00216-016-0139-z
6. McEachran AD, et al. 2017. A comparison of three chromatographic retention time prediction models. *Submitted*.

Acknowledgements

The authors would like to thank Drake Phelps and Tamara Tal of the US EPA for providing standards, and Mark Strynar and Julia Rager for analytical assistance. The authors would like to acknowledge ACD/Labs for use of ACD/ChromGenius and guidance regarding the software. This work was supported in part by an appointment to the ORISE Research Participation Program at the Office of Research and Development, U.S. EPA, through an interagency agreement between the US EPA and DOE.

Innovative Research for a Sustainable Future