Estimating the uncertainty of AI predictions of the Kováts retention index

Lewis Y. Geer, Douglas J. Slotta, and Stephen E. Stein

National Institute of Standards and Technology
U.S. Department of Commerce
Mass Spectrometry Data Center

INTRODUCTION

The Kováts Retention Index (RI) is a molecular property widely used to help identify chemicals.

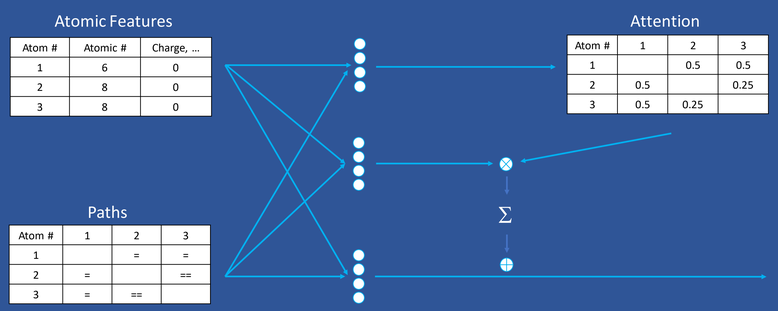
The NIST/EPA/NIH EI-MS spectral library includes measured and estimated RIs. Estimated RIs are predicted using a deep neural network called ***AIRI*** (Geer, Stein & Wallace, ASMS 2019).

AIRI takes a molecular structure as input and outputs a predicted RI. Previous AIRI predictions used a single fixed uncertainty calculated over the entire test set.

We investigate the use of an ensemble of networks to estimate a per structure uncertainty with the goal of improving the overall average confidence limits of the ***AIRI*** predictions.

METHOD

Path Augmented Transformer Model



The models were trained with an AdamW optimizer and a mean average error loss. The software libraries used are Masskit (https://github.com/usnistgov/masskit) for computation on spectra and structures, and Masskit AI (https://github.com/usnistgov/masskit_ai), used for AI computation on spectra and structures.

Data sets

The models were trained on structures from the NIST 2020 library with experimentally measured RIs for standard semi-nonpolar columns.

The test set consisted of structures from the NIST 2023 library with experimentally measured RIs and with a different connectivity than any structure in the training set.

Uncertainty calculation

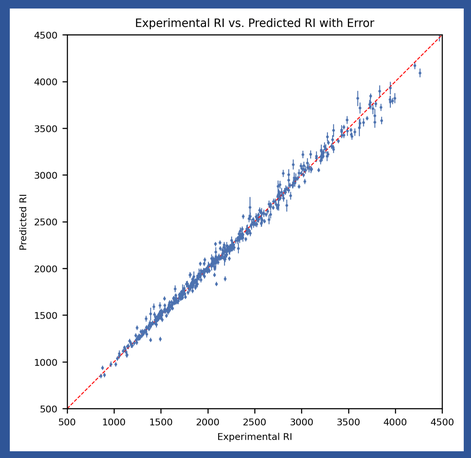
To create uncertainties for each prediction, RI values were predicted for each structure using the 8 models trained. These values were used to generate an average and standard deviation for each RI value.

It was observed that the uncertainties for some predictions were too small. To compensate for this, a minimum error of 22 was used as it resulted in the lowest average value at the 90% confidence interval.

RESULTS

The prediction of per RI value uncertainties resulted in an ***improvement in the mean 50% confidence limit from 27 to 22 and of the mean 90% confidence limit from 102 to 70***.

A selection of test structures with experimentally measured RIs and predicted RIs with error bars:



CONCLUSIONS

The use of model ensembles results in a better estimation of uncertainty.

Incorporating AIRI predictions in the NIST MS Search reduces false positives by the following method:

- All library structures have a measured or predicted RI.
- If the query spectrum of an unknown chemical has an RI, a match factor is computed for each hit that incorporates the difference of the query RI and the library RI, which can be measured or predicted.
- The resulting match factor can help suppress false positives. For example, the following two searches are performed on a spectrum of ***o-Xylene,*** one using RI and one not:

No AIRI

#	Lib	MF	RMF	Prob	RI	DBs	Name	
1	M	941	942	29.7	865	32 …	p-Xylene
2	M	937	939	26.5	866	32 …	Benzene, 1,3-dimethyl-
3	M	924	925	18.0	888	36 …	o-Xylene
4	M	888	895	6.09	855	35 …	Ethylbenzene
5	M	878	881	4.55	121…		Benzeneethanol, α,β-dimethyl-

With AIRI

#	Lib	MF	RMF	Prob	RI	origMF	DBs	Name	
1	M	924	925	90.6	888	924	36 …	o-Xylene
2	M	826	877	6.70	878*	876		Cyclopentene, 1-ethenyl-3-methylene-
3	M	767	939	1.22	866	937	32 …	Benzene, 1,3-dimethyl-
4	M	761	942	1.03	865	941	32 …	p-Xylene
5	M	699	750	0.17	898	749	1 G	1,3,5-Cyclooctatriene