A Strategy for the Best Candidate Selection from an Ensemble of Plausible Structures

Alexander Waked, Dimitris Argyropoulos, Maxim Kisko, Mikhail Elyashberg and Sergey Golotvin

[1]Advanced Chemistry Development, Inc. (ACD/Labs), 8 King Street East, Toronto, ON. M5C 1B5. Canada

ACD/Labs

Introduction

In the typical Computer Assisted Structure Elucidation (CASE) workflow [1,2] the last step is to select the best structure from the ones generated. It is the same in Automated Structure Verification (ASV) systems when there are more than 1 proposed structures (Combined and Concurrent Verification, CCV) and in Unbiased Verification (UBV) [3]. Usually either a Match Factor (MF) and/or the mean deviation between predicted and experimental chemical shifts are calculated. Despite the metrics used for the ranking it is not uncommon to have two or more structures with similar validities.

The DP4 Approach

The DP4 methodology has been developed for stereochemistry determination [4]. It can be very valuable in resolving cases where 2 or more structures have similar, high validities as determined by other means. DP4 has been used with DFT calculated NMR spectra and here we are using it with Neural Network (NN) and HOSE-codes predicted spectra. It requires knowledge of the prediction accuracy and error distribution, which we estimated using a list of 36,000 ^{1}H and 52,000 ^{13}C chemical shifts of 3100 fully assigned chemical structures that are not present in the predictor training databases (Fig. 1).

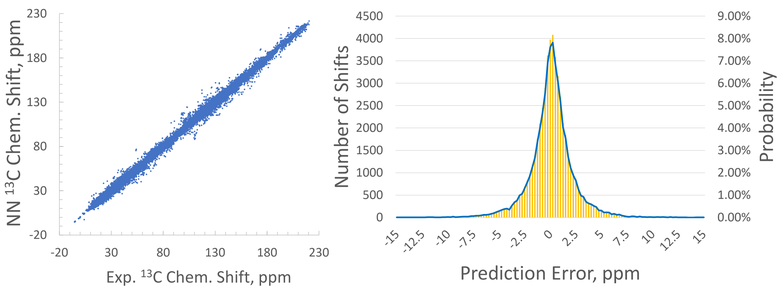


Figure 1. The ^{13}C NN prediction accuracy (left) and error distribution (right) as calculated using 52,000 chemical shifts of 3100 structures.

The final DP4 style probabilities are calculated using the method described in the original DP4 reference [4]. A comparison of the use cases for the three possible metrics that can be used is shown in Table 1.

Table 1: Comparing the possible metrics for structure ranking

Metrics	Works w/ single structure	Requires fully assigned structure	Correct structure must be in set for accurate result	Accounts for observed/missing correlations	Accounts for observed integrals
Match Factor	✓	×	×	✓	✓
Mean Deviations	✓	✓	×	×	×
DP4	×	✓	✓	×	×

Application in UBV: Artemisinine

We applied this to a dataset of 1D ^{1}H and ^{13}C, and 2D HSQC, COSY, and HMBC spectra of artemisinine, run through CCV and UBV initially with a wrong proposed structure. Isomeric structures to it were generated (CCV) and MFs and DP4 probabilities calculated (Table 2). The DP4 results for the fully assigned isomeric structures incorrectly point to the wrong structure, despite the MF being low. If a full UBV run is performed, then the correct structure is generated, and it gets high DP4 and MF scores as well as low mean deviation, d_N (Table 3).

Table 2: CCV result

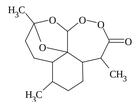
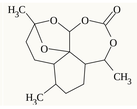
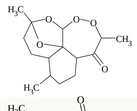
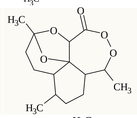
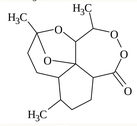
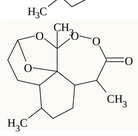
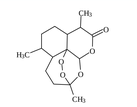
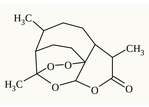
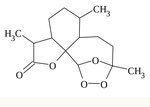
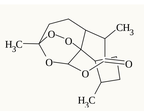
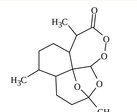
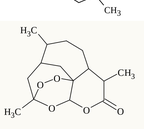
	Structure	MF	DP4$_N$ (^{13}C)
1		0.69	99.97
2		0.00	
3		0.00	
4		0.00	
5		0.00	
6		0.58	0.03

Table 3: UBV result

	Structure	MF	DP4$_N$ (^{13}C)	d_N ^{13}C
1		0.82	99.83	3.820
2		0.73	0.00	5.736
3		0.63	0.00	5.893
4		0.67	0.17	6.074
5		0.69	0.00	6.232
6		0.49	0.00	6.529

Application in CASE

We calculated the DP4 probabilities for up to the top 10 ranked generated structures, in more than 200 previous CASE problems. The datasets contained 1D ^{1}H and ^{13}C, and 2D HSQC and HMBC spectra. The DP4 probabilities calculated were the highest for the correct structure in >90% of cases. Table 4 lists a subset of these datasets, containing the ones with the most ambiguous results as ranked by the average deviations between experimental and predicted ^{13}C chemical shifts. Since sometimes the deviation ranking was different for the three prediction methods used (HOSE codes, Neural Networks, and Incremental) we list the DP4 probabilities as calculated for the prediction method that gave the lowest deviation. We see that on average, in 8 out of 10 such cases the DP4 probability correctly identified the right structure, while in only 2 out of 10 the result was not correct.

Table 4: Selected NMR datasets with deviations difference between 1st and 2nd structures less than 30%

Set	1st Structure Deviation	2nd Structure Deviation	Prediction Method	Correct Structure Rank	DP4 Probability
1	1.282	1.312	NN	1	62.17
2	1.988	2.280	NN	1	99.89
3	1.139	1.287	HOSE Codes	2	3.23
4	1.652	1.952	NN	1	98.2
5	2.306	2.460	HOSE Codes	3	92.77
6	5.811	6.452	NN	1	72.23
7	1.497	1.528	NN	2	45.76
8	2.534	2.577	HOSE Codes	1	86.95
9	2.755	3.000	HOSE Codes	1	98.90
10	1.654	1.940	HOSE Codes	1	93.07

Conclusions

DP4 probability metrics can be very valuable in discriminating structural candidates in ASV, UBV, and CAS workflows even when used with empirical NMR shift predictions.

References

1. Elyashberg, M. E.; Blinov, K. A.; Molodtsov, S. G.; Williams, A. J.; Martin, G. E. (2004). *J. Chem. Inf. Comput. Sci.*, *44*, 771-792.
2. Elyashberg, M. E., Argyropoulos, D. (2019). *eMagRes.*, *8*, 239-254.
3. Golotvin, S. S., Pol, R., Sasaki, R. R., Nikitina, A., Keys, P. (2012). *Mag. Res. Chem.*, *6*, 429-435.
4. Smith, S., Goodman, J. (2010). *J. Am. Chem. Soc.*, *132*, 12946-12959.
5. Bremser, W. (1978). *Anal. Chim. Acta*, *103*, 355-365.