Green simultaneous quantification of levodopa, carbidopa, and benserazide in anti-Parkinson tablets by ATR-FTIR spectroscopy combined with hurdle modelling and machine learning

Manh Huy Nguyen[a,b], Thanh Dam Nguyen[a,b], Hong Anh Duong[a,b], Hung Viet Pham[*b]

[a.] Key Laboratory of Analytical Technology for Environmental Quality and Food Safety Control (KLATEFOS), VNU University of Science, Vietnam National University, Hanoi, 334 Nguyen Trai, Thanh Xuan, Hanoi 100000, Vietnam

[b.] Research Centre for Environmental Technology and Sustainable Development (CETASD), Faculty of Chemistry, VNU University of Science, Vietnam National University, Hanoi, 334 Nguyen Trai, Thanh Xuan, Hanoi 100000, Vietnam

Email: vietph@vnu.edu.vn

Supplementary Information

- Text S1 to S3
- Table S1 to S6
- Figure S1

Supplementary Text

Text S1. Ranking procedure used in module 2

Selection of the optimal pipeline was based on a composite multi-criteria ranking scheme rather than on any single performance metric. For each pipeline p, analyte c, metric m, and outer split s, a rank $r_{pcms}$was assigned. For BAcc, F1 score, and AUC, higher values were considered preferable, whereas for RMSE-based metrics, lower values were considered preferable. The composite rank for each pipeline–analyte–split combination was calculated as:

$$R_{pcs} = \frac{1}{M}\sum_{m=1}^{M} r_{pcms}$$

where Mdenotes the number of metrics included in the ranking procedure. The overall rank of pipeline $p$was then obtained by averaging R_{pcs}across all analytes and outer splits:

$$GlobalRank_p = \frac{1}{CS}\sum_{c=1}^{C}\sum_{s=1}^{S} R_{pcs}$$

where $C = 3$ is the number of analytes and S is the total number of outer evaluations. Lower values of $GlobalRank_p$indicated better overall pipeline performance.

Pipeline stability was quantified as the standard deviation of R_{pcs}across all analytes and outer splits:

$$Stability_p = SD(R_{pcs})$$

Lower $Stability_p$ values indicated lower sensitivity to data partitioning.

Text S2. Ranking procedure used in step 3.1 of module 3

Classifier performance was evaluated on the outer test folds using balanced accuracy (BAcc) and F1 score, with AUC used as a supplementary descriptor. Model selection for each analyte–region combination was based on a weighted composite rank:

$$Rank_{total} = 0.7\ Rank_{BAcc} + 0.3\ Rank_{F1}$$

A greater weight was assigned to BAcc because, within the hurdle-model framework, balanced performance on both the positive and negative classes was required. The F1 score was retained to ensure adequate performance on the positive class, since false-negative predictions would completely disable the downstream regression branch.

Text S3. Ranking procedure used in step 3.2 of module 3

To identify the optimal regression model within the hurdle-model framework, each regressor configuration was evaluated using three metrics: $RMSE_{pos}$, $RMSE_{pos,TP}$, and $RMSE_0$. These metrics were converted into ranks in ascending order, such that lower RMSE values corresponded to better ranks, and were combined into a single composite score as follows:

$$Rank_{total} = 0.6\ Rank_{RMSE_{pos}} + 0.25\ Rank_{RMSE_{pos,TP}} + 0.15\ Rank_{RMSE_0}$$

The greater weight assigned to $RMSE_{pos}$was intended to emphasize overall predictive performance under practical conditions, whereas $RMSE_{pos,TP}$and $RMSE_0$were included to capture intrinsic regression accuracy and control error on the negative class, respectively. The composite rank was calculated separately for each analyte and then averaged across analytes to determine the globally optimal regressor configuration. In

addition, configuration stability was assessed from the standard deviation of $RMSE_{pos}$across cross-validation iterations, such that models with more consistent performance were preferentially selected.

Supplementary Tables

Table S1. Designed API contents and HPLC-DAD measured actual contents of the 103 pellets in the calibration set

Pellet ID	Designed content (%w/w)			HPLC-DAD measured content (%w/w)		
	LD	BZ	CD	LD	BZ	CD
1	20	5	0	18.51	5.09	0
2	20	10	0	19.21	11.03	0
3	20	15	0	19.31	17.42	0
4	20	20	0	19.42	23.14	0
5	40	5	0	40.53	5.52	0
6	40	10	0	41.78	10.57	0
7	40	15	0	39.7	16.96	0
8	40	20	0	32.92	20.76	0
9	60	5	0	60.49	6.04	0
10	60	10	0	59.04	13.64	0
11	60	15	0	70.37	17.67	0
12	60	20	0	59.84	21.44	0
13	80	5	0	72.41	4.54	0
14	80	10	0	74.81	9.33	0
15	80	15	0	71.74	13.62	0
16	80	20	0	74.8	19.49	0
17	20	0	2	18.85	0	1.86
18	20	0	4	18.76	0	3.57
19	20	0	6	18.89	0	5.33
20	20	0	8	19.37	0	7.2
21	40	0	2	43.01	0	2.57
22	40	0	4	38.91	0	3.38
23	40	0	6	40.69	0	6.41
24	40	0	8	42.84	0	7.38
25	60	0	2	60.3	0	2.87
26	60	0	4	41.6	0	2.49
27	60	0	6	52.74	0	5.75
28	60	0	8	58.21	0	7.18
29	80	0	2	72.55	0	2.43
30	80	0	4	79.02	0	4.43
31	80	0	6	72.77	0	5.85
32	80	0	8	56.73	0	5.74
33	0	5	0	0	5.44	0
34	0	10	0	0	12.99	0
35	0	15	0	0	19.36	0
36	0	20	0	0	19.89	0
37	0	0	2	0	0	2.3
38	0	0	4	0	0	3.45
39	0	0	6	0	0	5.03

40	0	0	8	0	0	6.59
41	0	0	0	0	0	0
42	0	0	2.5	0	0	1.97
43	0	0	5	0	0	4.65
44	0	0	7.5	0	0	6.48
45	0	6.25	0	0	7.55	0
46	0	6.25	2.5	0	6.42	4.42
47	0	6.25	5	0	6.64	6.19
48	0	6.25	7.5	0	6.15	8.29
49	0	12.5	0	0	12.78	0
50	0	12.5	2.5	0	12.3	5.64
51	0	12.5	5	0	11.45	6.86
52	0	12.5	7.5	0	12.02	9.14
53	0	18.75	0	0	19.48	0
54	0	18.75	2.5	0	17.69	4.69
55	0	18.75	5	0	17.98	7.37
56	0	18.75	7.5	0	16.02	8.45
57	25	0	0	24.15	0	0
58	25	0	2.5	24.85	0	2.41
59	25	0	5	25.38	0	3.32
60	25	0	7.5	22.8	0	6.49
61	25	6.25	0	23.26	6.91	0
62	25	6.25	2.5	23.69	6.23	3.71
63	25	6.25	5	21.86	5.72	5.52
64	25	6.25	7.5	22.97	6.13	0
65	25	12.5	0	21.58	12.51	7.34
66	25	12.5	2.5	23.39	12.8	5.24
67	25	12.5	5	23.67	11.78	7.4
68	25	12.5	7.5	24.67	12.87	9.79
69	25	18.75	0	19.17	16.22	0
70	25	18.75	2.5	18.64	14.85	5.15
71	25	18.75	5	23.28	17.88	9
72	25	18.75	7.5	21.78	17.91	11.18
73	50	0	0	47.96	0	0
74	50	0	2.5	45.7	0	2.44
75	50	0	5	48.91	0	4.73
76	50	0	7.5	46.52	0	6.65
77	50	6.25	0	49.94	5.86	0
78	50	6.25	2.5	49.17	6.67	3.24
79	50	6.25	5	49.8	5.52	5.14
80	50	6.25	7.5	46.99	5.42	7.09
81	50	12.5	0	44.46	12.43	0
82	50	12.5	2.5	49.33	12.07	5.85
83	50	12.5	5	47.91	11.88	8.2
84	50	12.5	7.5	48.06	11.19	9.64
85	50	18.75	0	46.84	18.63	0
86	50	18.75	2.5	49.61	17.03	6.25

87	50	18.75	5	49.68	17.67	9.1
88	50	18.75	7.5	48.13	14.99	10.52
89	75	0	0	69.77	0	0
90	75	0	2.5	68.37	0	2.45
91	75	0	5	67.75	0	4.59
92	75	0	7.5	67.74	0	6.72
93	75	6.25	0	66.36	6.94	0
94	75	6.25	2.5	73.87	7.14	3.09
95	75	6.25	5	70.8	5.24	4.49
96	75	6.25	7.5	72.41	5.21	7.24
97	75	12.5	0	66.29	11.02	0
98	75	12.5	2.5	68.76	11.27	3.88
99	75	12.5	5	71.38	13.31	7.66
100	75	12.5	7.5	68.24	12.16	10.34
101	75	18.75	0	71.18	19.8	0
102	75	18.75	2.5	69.14	16.31	6.92
103	75	18.75	5	66.02	17.41	7.88

Table S2. Range of hyperparameters used as GridSearchCV in module 2

Classifier/Regressor	Hyperparameters	Range
Logistic regression	C	0.01, 0.1, 1, 10, 100, 1000
PLS	n_components	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25

scoring = 'balanced_accuracy' for classifier, and 'neg_root_mean_squared_error' for regressor

Table S3. Range of hyperparameters used for RandomizedSearchCV in step 3.1 and step 3.3 of module 3

Classifier	Hyperparameters	Range
Logistic regression	C	loguniform(1e-3, 1e3)
SVC	C	loguniform(1e-2, 1e3)
	gamma	loguniform(1e-4, 1e0)
RF	n_estimators	randint(300, 1500),
	max_depth	None, 8, 12, 18, 25, 35
	min_samples_leaf	randint(1, 6),
	max_features	sqrt', 'log2', 0.3, 0.5, 0.8

scoring = 'balanced_accuracy', n_iter = 100 and 300 for the step 3.1 and 3.3, respectively

Table S4. Range of hyperparameters used as GridSearchCV for PLS/PCR and RandomizedSearchCV for SVR/RF in step 3.2 and step 3.3 of module 3

Classifier	Hyperparameters	Range
PLS	n_components	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25
PCR	n_components	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25
SVR	C	loguniform(1e-2, 1e3)
	gamma	loguniform(1e-4, 1e0)
	Epsilon	loguniform(1e-3, 1e0)
RF	n_estimators	randint(300, 1500),
	max_depth	None, 8, 12, 18, 25, 35
	min_samples_leaf	randint(1, 6),
	max_features	sqrt', 'log2', 0.3, 0.5, 0.8

scoring = 'neg_root_mean_squared_error', n_iter = 100 and 300 for the step 3.2 and 3.3, respectively

Table S5. Rank and stability of 16 combinations of preprocessing pipeline and spectral type in module 2

Spectral type	Preprocessing pipeline	Global rank	Stability
mean	SG first derivative + SNV	3.80	1.25
mean	RAW + MSC	4.55	1.17
mean	SG smoothing + MSC	5.18	1.19
mean	RAW + SNV	5.73	1.18
mean	SG second derivative + SNV	7.63	1.22
mean	SG first derivative + MSC	8.00	1.18
median	RAW + MSC	8.00	1.51
median	SG smoothing + MSC	8.20	1.48
mean	SG smoothing + SNV	8.60	2.46
median	SG first derivative + MSC	10.17	0.97

mean	SG second derivative + MSC	10.17	1.48
median	RAW + SNV	10.23	1.35
median	SG smoothing + SNV	10.80	1.43
median	SG second derivative + MSC	11.07	2.10
median	SG second derivative + SNV	11.83	1.42
median	SG first derivative + SNV	12.03	1.35

Table S6. The ordered rankings of regression profiles in the step 3.2 of module 3

No.	Spectral region	Mode	Regressor	Total rank	Stability
1	half	single	RF	6.40	1.25
2	fingerprint	single	RF	7.40	1.17
3	half	single	PCR	8.45	0.93
4	full	single	RF	8.63	0.97
5	full	multi	RF_multi	9.43	1.19
6	full	single	PCR	9.72	0.60
7	fingerprint	single	PLS	9.95	1.10
8	half	multi	RF_multi	10.23	1.30
9	fingerprint	single	PCR	10.37	0.90
10	half	single	PLS	10.97	1.11
11	fingerprint	multi	RF_multi	11.60	1.09
12	half	multi	PCR_multi	12.43	0.89
13	fingerprint	multi	PLS2	12.80	1.03
14	full	multi	PCR_multi	12.87	0.98
15	fingerprint	single	SVR	12.98	1.13
16	full	single	PLS	13.17	0.75
17	fingerprint	multi	PCR_multi	13.30	1.04

18	half	multi	SVR_multi	15.80	1.30
19	fingerprint	multi	SVR_multi	16.10	1.09
20	full	multi	SVR_multi	16.30	1.15
21	full	single	SVR	16.40	1.09
22	half	single	SVR	16.73	1.27
23	half	multi	PLS2	18.40	0.75
24	full	multi	PLS2	19.57	0.87

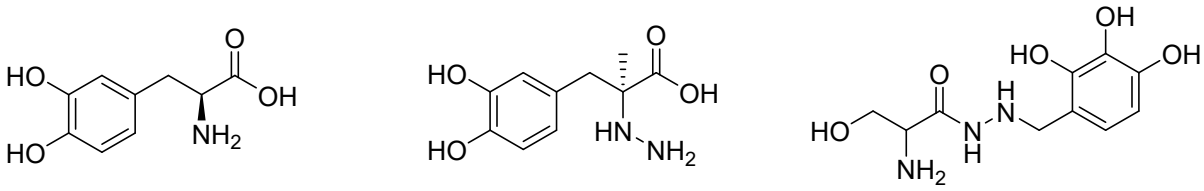
Table S7. Recovery of LD, BZ and CD from synthetic excipient pellets at three spiking levels (n = 3 for low and high levels, n = 6 for medium level)

Synthetic mixture type	API	Spiking level	Added content (% w/w)	HPLC-DAD found (% w/w)	ATR-FTIR found (% w/w)	Recovery (%)
LD+CD	LD	Low	40	34.34 ± 4.93	35.48 ± 3.21	103.9 ± 7.0
LD+CD	LD	Medium	50	48.02 ± 10.18	48.40 ± 11.71	101.1 ± 12.7
LD+CD	LD	High	60	61.12 ± 2.74	64.04 ± 6.89	104.6 ± 7.0
LD+CD	CD	Low	4	4.17 ± 0.98	4.79 ± 0.82	116.3 ± 11.6
LD+CD	CD	Medium	5	4.88 ± 0.99	4.78 ± 0.93	98.7 ± 10.8
LD+CD	CD	High	6	6.31 ± 0.22	5.90 ± 0.57	93.6 ± 9.6
LD+BZ	LD	Low	28.8	28.37 ± 4.46	29.91 ± 7.92	104.3 ± 11.2
LD+BZ	LD	Medium	36	35.69 ± 5.92	33.73 ± 8.03	93.9 ± 11.4
LD+BZ	LD	High	43.2	41.95 ± 2.11	40.11 ± 4.67	95.5 ± 7.4
LD+BZ	BZ	Low	7.2	6.12 ± 0.22	6.44 ± 0.20	105.2 ± 5.9
LD+BZ	BZ	Medium	9	9.28 ± 2.09	9.66 ± 1.85	105.3 ± 10.8
LD+BZ	BZ	High	10.8	10.67 ± 0.59	10.11 ± 0.98	94.8 ± 9.9

Table S8. Detailed AGREE scoring parameters for ATR-FTIR/hurdling model

No.	Criteria	Condition
1	Direct analytical techniques should be applied to avoid sample treatment	In-field sampling and direct analysis
2	Minimal sample size and minimal number of samples are goals	1.12 g (= 30 mg * 112 total samples / 3 samples)
3	In situ measurements should be performed	At-line
4	Integration of analytical processes and operations saves energy and reduces the use of reagents	3 or fewer
5	Automated and miniaturized methods should be selected	Degree of automation: Semi-automatic Sample preparation: none or miniaturized
6	Derivatization should be avoided	None
7	Generation of a large volume of analytical waste should be avoided, and proper management of analytical waste should be provided	1.12 g
8	Multi-analyte or multi-parameter methods are preferred versus methods using one analyte at a time	Number of analytes determined in a single run: 3 Sample throughput (samples analysed per hour): 0.4
9	The use of energy should be minimized	FTIR
10	Reagents obtained from renewable sources should be preferred	No reagents
11	Toxic reagents should be eliminated or replaced	Does the method involve the use of toxic reagents or solvents: No
12	The safety of the operator should be increased	None

Supplementary Figure



levodopa carbidopa benserazide

Figure S1. The chemical structures of LD, CD, and BZ.