Supplementary Information

1. Other classification methods for comparison

1.1 Logistic Regression

Logistic Regression (LR) is a classical classification method and can also be seen as a simple neural network without hidden layers. LR fits the parameters from the training set, fitting the target to [0, 1], and then discretizing the target to achieve classification. We used the 'LogisticRegression' module of the sklearn package in this study. SGD was used to learn weights of the model. The regularization penalty was 'l1', and the regularization parameters C was 1.00.

1.2 k-Nearest Neighbor

k-Nearest Neighbor (kNN) is a simple and well-known classification method. It determines the class of a new sample by looking at the majority of the classes of the nearest k samples. The metrics we used for evaluating the distance between samples is the Euclidean distance. The number of neighbors k∈[2, 5] has been considered and 3 has been selected for our model.

1.3 Random Forest

Random Forest (RF) is an ensemble learning method based on decision trees. It has the advantages of simplicity, ease of implementation, and low computational overhead, and powerful performance in many tasks. We use the 'RandomForestClassifier' module of the sklearn package in this study. The number of trees has been considered in the range [100, 500], and the number of features randomly sampled for each split is optimized in the range [50, 200]. Finally, 500 trees and 90 features were selected.

1.4 Back propagation artificial neural network

Artificial neural network is a complex network structure formed by a large number of neurons. It is a simplification and simulation of the structure and operation of the human brain. In this study, two types of ANN models were used for comparison with DeepCID. The first category, called BP-ANN, are Conventional Back Propagation neural networks. After optimized in range from 25 to 70, the number of nodes in BPs are determined to be 40, respectively. The initial bias is -0.5 to 0.5, and the batch size is 500. The Sigmoid was used as activation function and the Stochastic Gradient Descent was used as the Optimizer. The second category, FCNN models, were established by simply removing the convolutional layers of CNN models. They are used to further verify the effectiveness of the convolutional layers.

Table S1. Prediction accuracy of component identification models on simulated test set.

NO.	Model for compound	Accuracy（%）
1	1,4-Benzoquinone	100.0
2	2-Propanol	99.7
3	30% Hydrogen Peroxide	99.2
4	36% Acetic Acid	99.7
5	4-Hydroxybenzoic Acid	100.0
6	4-Isobutylacetophenone	100.0
7	a-Tocopherol	100.0
8	a-Tocopheryl Acetate	99.9
9	Acetacetic Ester	100.0
10	Acetaminophen	100.0
11	Acetanilide Melting Point Standard	100.0
12	Acetic Anhydride	100.0
13	Acetic Ether	99.8
14	Acetone	99.8
15	Acetonitrile	99.6
16	Adenosine	100.0
17	Albuterol Sulfate	100.0
18	Aniline	100.0
19	Ascorbic Acid	100.0
20	Avobenzone	100.0
21	Azithromycin	100.0
22	Benzoic Acid	99.9
23	Benzyl Alcohol	99.7
24	Bis(trimethylsilyl)amine	99.8
25	Boric Acid	100.0
26	Bupivacaine HCl	99.8
27	Butanone	100.0
28	Butylparaben	99.9
29	Caffiene	99.8
30	Chloral Hydrate	100.0
31	Chlorpheniramine	100.0
32	Cimetidine HCl	100.0
33	Cimetidine	100.0
34	Ciprofloxacin HCl	100.0
35	Citric Acid	100.0
36	Citric Acid	100.0
37	Clarithromycin	100.0
38	Clindamycin Phosphate	100.0
39	Clotrimazole	100.0
40	Cyclohexanone	99.8

NO.	Model for compound	Accuracy（%）
41	Cyclosporine (Ciclosporin)	99.9
42	Dextrose (D-Glucose)	100.0
43	Dichloromethane	99.7
44	Diethylene Glycol	100.0
45	Dimethyl Benzene	100.0
46	Dimethyl Sulfoxide	99.5
47	Diphenhydramine	100.0
48	Dopamine HCl	100.0
49	Edetate Disodium	99.9
50	EDTA-2Na	99.9
51	Erythromycin	100.0
52	Ethanol	99.5
53	Ether	99.5
54	Ethylene Glycol	99.9
55	Ethylenediaminetetraacetic Acid	99.8
56	Ethylparaben	100.0
57	Famotidine	100.0
58	Formic acid	99.6
59	Fructose	100.0
60	Furosemide	100.0
61	Gabapentin	100.0
62	Glycerin	99.8
63	Glycerin	99.9
64	Glycine	100.0
65	Guaiacol	100.0
66	Guaifenesin	100.0
67	Homosalate	100.0
68	Hydrazine Hydrate	99.8
69	Hydrochlorothiazide	100.0
70	Hydrocortisone	100.0
71	Ibuprofen	100.0
72	Isopropyl Amine	98.8
73	Isopropyl Ether	99.8
74	L-Alanine	99.9
75	L-Arginine	100.0
76	L-Aspartic Acid	100.0
77	L-Cysteine HCl	99.9
78	L-Glutamic Acid	99.9
79	L-Glutamine	100.0
80	L-Histidine	100.0
81	L-Isoleucine	100.0

NO.	Model for compound	Accuracy（%）
82	L-Leucine	99.9
83	L-Lysine Acetate	100.0
84	L-Lysine HCl	99.9
85	L-Phenylalanine	99.9
86	L-Serine	100.0
87	L-Tyrosine	99.8
88	Lactic Acid	99.5
89	Lactose, Anhydrous	99.7
90	Lactose, Monohydrate	100.0
91	Lidocaine	100.0
92	Magnesium Sulphate	99.9
93	Mannitol	99.9
94	Meslamine (Mesalazine)	99.8
95	Metformin HCl	100.0
96	Methanol	99.2
97	Methyl Silicone	99.5
98	Methylparaben	100.0
99	Metoprolol	100.0
100	Metronidazole	100.0
101	N, N-Dimethyl Formamide	99.9
102	N-Acetyl-L-Cysteine	100.0
103	N-Methyl Pyrrolidone	99.9
104	Naproxen	100.0
105	Niacinamide (Nicotinamine)	99.9
106	Nitric Acid	99.4
107	Octinoxate	100.0
108	Octisalate	100.0
109	Octocrylene	100.0
110	Omeprazole	100.0
111	Oxybenzone	99.9
112	p-Toluenesulfonic Acid	99.7
113	Paraformaldehyde	99.5
114	Phenacetin Melting Point Standard	100.0
115	Phenoxyethanol	100.0
116	Phenyl Salicyclate Melting Point Standard	100.0
117	Phenylephrine HCl	99.4
118	Phenylethyl Alcohol	99.8
119	Phosphorus Pentoxide	99.7
120	Phthalic Anhydride	100.0
121	Phytonadione	100.0
122	Polyacrylamide	99.7

NO.	Model for compound	Accuracy（%）
123	Polyethylene Glycol	99.8
124	Polyethylene Glycol	99.6
125	Polyvinylpyrrolidone	99.9
126	Potassium Bichromate	99.7
127	Potassium Gluconate	99.9
128	Potassium Pyrophosphate	99.9
129	Prednisolone	99.9
130	Propylene Glycol	99.9
131	Propylparaben	100.0
132	Pyridoxine HCl	99.9
133	Ranitidine HCl	100.0
134	Salicylic Acid	100.0
135	Silica Gel for TLC	99.4
136	Sodium Acetate Trihydrate	99.9
137	Sodium Bicarbonate	99.8
138	Sodium Bisulfite	99.6
139	Sodium Carbonate	99.9
140	Sodium Chlorate	99.8
141	Sodium Lactate	100.0
142	Sodium Metabisulfite	99.7
143	Sodium Phosphate	99.2
144	Sodium Sulfite	100.0
145	Sodium Tetraborate	99.9
146	Sorbitol	99.9
147	Sucrose	100.0
148	Sulfamethoxazole	100.0
149	Sulfanilamide Melting Point Standard	99.9
150	Sulfapyridine Melting Point Standard	100.0
151	Taurine	100.0
152	Tetracaine HCl	100.0
153	Tetracycline HCl	99.6
154	Tetrahydrofuran	99.8
155	Theophylline	100.0
156	Thiamine HCl	99.9
157	Titanium Dioxide	99.4
158	Trichloromethane	100.0
159	Trihydroxymethyl Aminomethane	100.0
160	Trimethoprim	100.0
161	Trisodium Citrate	99.7
162	Trolamine	99.6
163	Valproic Acid	100.0

NO.	Model for compound	Accuracy（%）
164	Vanillin Melting Point Standard	99.9
165	Vanillin	99.9
166	Verapamil HCl	100.0
167	Zinc Oxide	99.8

Table S2. The prediction accuracy (ACC%) of the simulated test set by different methods (after pre-processed).

Model	DeepCID	RF	LR	kNN	BP-ANN	FCNN
Methanol	100.0	99.5	99.8	88.0	99.7	99.9
Ethanol	99.7	99.7	99.7	78.0	99.3	99.6
Acetonitrile	99.6	99.3	99.8	78.1	99.0	99.8
Polyacrylamide	99.7	99.5	99.8	78.8	99.2	99.7
Sodium Acetate	100.0	99.6	99.9	76.8	99.3	99.8
Sodium Carbonate	99.9	99.5	99.8	80.9	99.6	99.8

Table S3. The information of liquid and powder dataset and the corresponding prediction results of DeepCID (after pre-processed)

NO.	Status	Real rate	Component	Prediction results	TP	TN	FP	FN	TPR (%)	FPR (%)
1	Liquid	5: 3: 2	Methanol Ethanol Acetonitrile	Methanol Ethanol Acetonitrile Trolamine	3	164	1	0	100.0	0.6
2	Liquid	7: 2: 1	Methanol Ethanol Acetonitrile	Methanol Ethanol Acetonitrile	3	164	0	0	100.0	0.0
3	Liquid	4: 3: 3	Methanol Ethanol Acetonitrile	Methanol Ethanol Acetonitrile Trolamine	3	163	1	0	100.0	0.6
4	Powder	1: 1	Polyacrylamide Sodium Acetate Trihydrate	Polyacrylamide Sodium Acetate Trihydrate	2	165	0	0	100.0	0.0
5	Powder	1: 1: 1	Polyacrylamide Sodium Acetate Trihydrate Sodium Carbonate	Polyacrylamide Sodium Acetate Trihydrate Sodium Carbonate	3	164	0	0	100.0	0.0
6	Powder	5: 3: 2	Polyacrylamide Sodium Acetate Trihydrate Sodium Carbonate	Polyacrylamide Sodium Acetate Trihydrate □	2	164	0	1	66.7	0.0