Machine Learning Guided Cocrystal Prediction Based on Large Data Base

Dingyan Wang[a,c, ‡], Zeen Yang[b,c, ‡], Bingqing Zhu[b], Xuefeng Mei[*,b,c], Xiaomin Luo[*,a,c]

[a] *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China*

[b] *Pharmaceutical Analytical & Solid-State Chemistry Research Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China*

[c] *University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China*

Corresponding Authors

[*](Xuefeng Mei) E-mail: xuefengmei@simm.ac.cn

[*](Xiaomin Luo) E-mail: xmluo@simm.ac.cn

Author Contributions

[‡]Dingyan Wang and Zeen Yang contributed equally to this work.

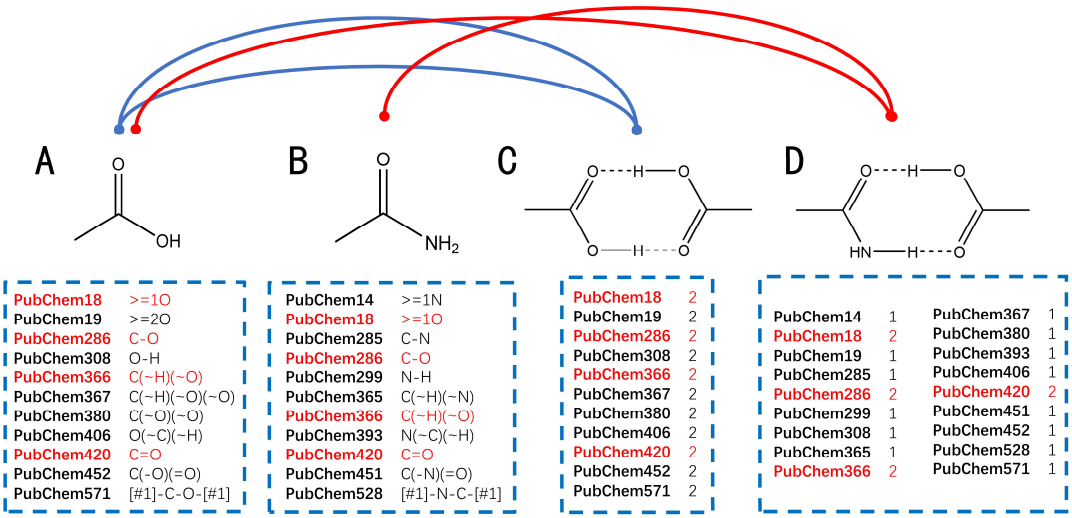


Figure S1. In this study we use molecular fingerprints (FP) to encode the structural information of molecules and further construct the features of chemical pairs. A bit in a fingerprint is a pre-defined substructure. Here we use two usual functional groups (carboxyl group and amide group) and PubChem fingerprint to make a further explanation. There are 881 bits in PubChem fingerprint and each bit represents a substructure or structural feature. For example, PubChem18 means more than one oxygen atom in the molecule and PubChem420 refers to the existence of carbonyl group (**A**). A simple carboxyl group will activate (set the value of a bit in fingerprint to 1 instead of 0) at least 11 bits in PubChem FP (**A**). Also, amide group will activate at least 11 bits in PubChem FP but partly different from which carboxyl group activates (**B**). In this way a functional group is represented by a set of (not one) bits. Then, according to **Equation (2)**, input feature of a chemical pair is generated by adding the fingerprints of two molecules element-wisely. Element '0' in the output vector means both molecules do not have this substructure, '1' means one of them has and '2' means both of them have. Unique fingerprints which can be used for modeling will be generated for different compositions of functional groups in chemical pairs. As shown in the figure, basic feature vectors of carboxylic dimer synthon (**C**) and amide-carboxylic acid synthon (**D**) are different. In this way the output of **Equation (2)** can be used to represent a chemical pair and taken as the input of ML models.

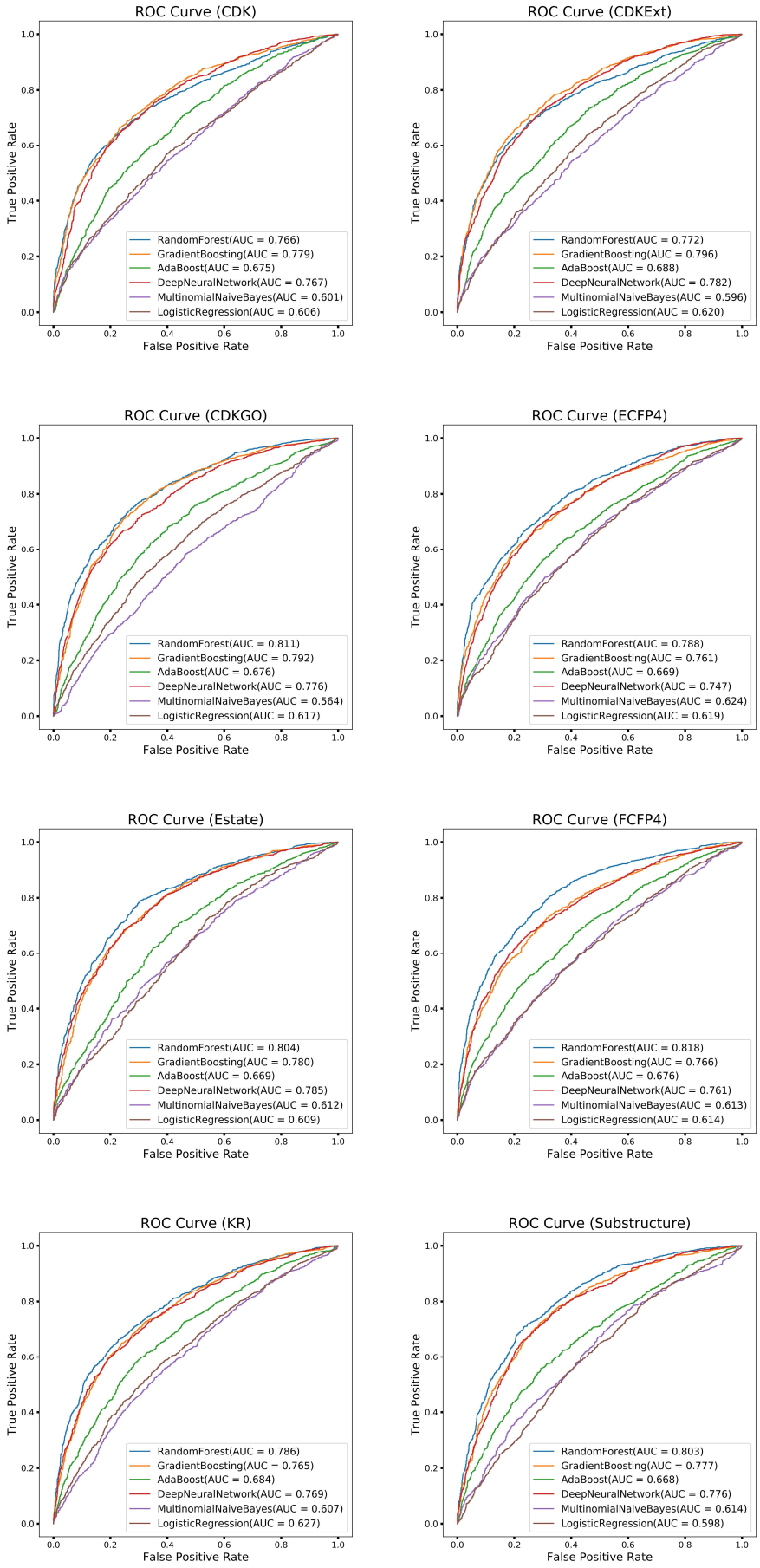


Figure S2. ROC curves of different ML models using eight fingerprints on the test set.

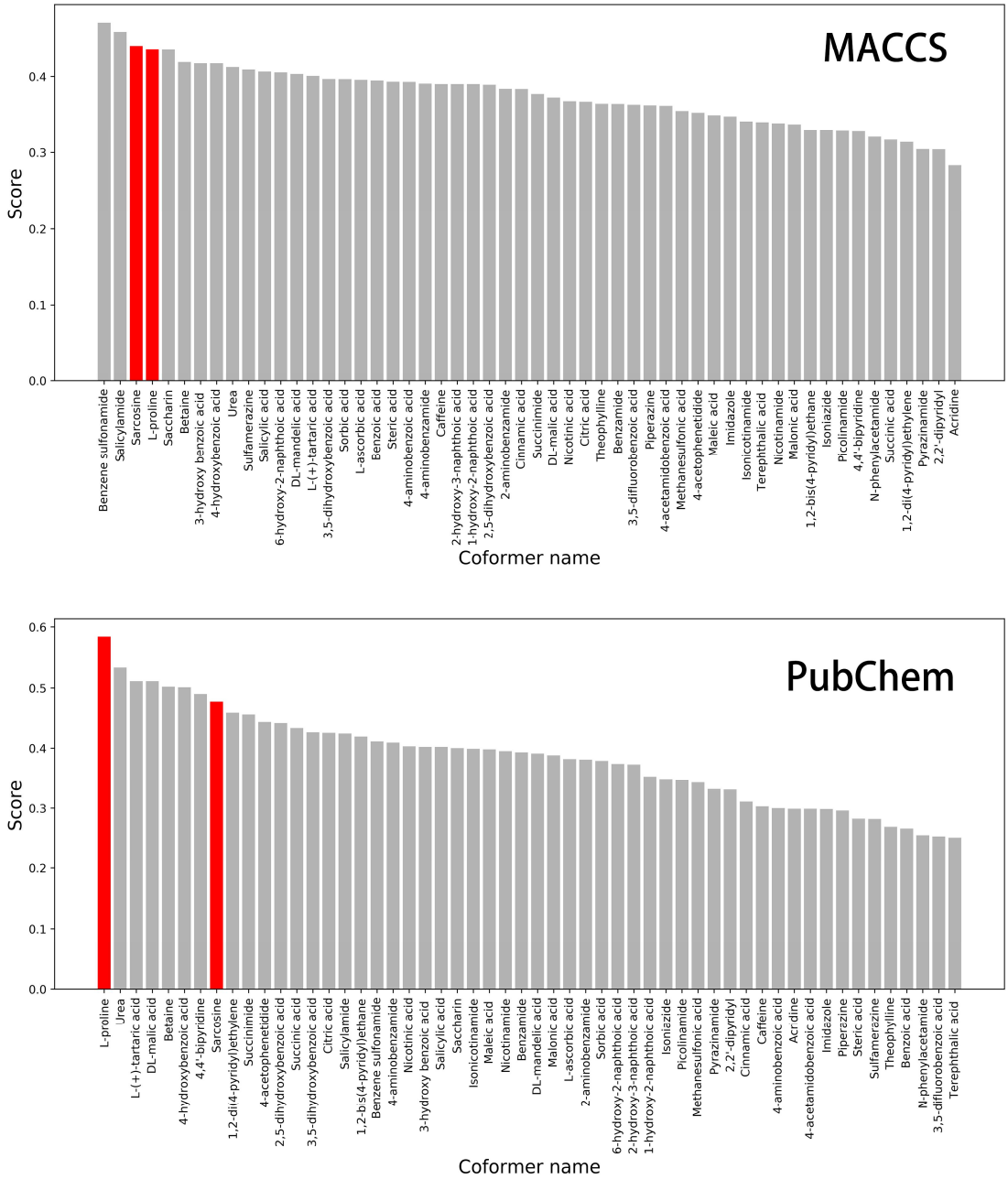
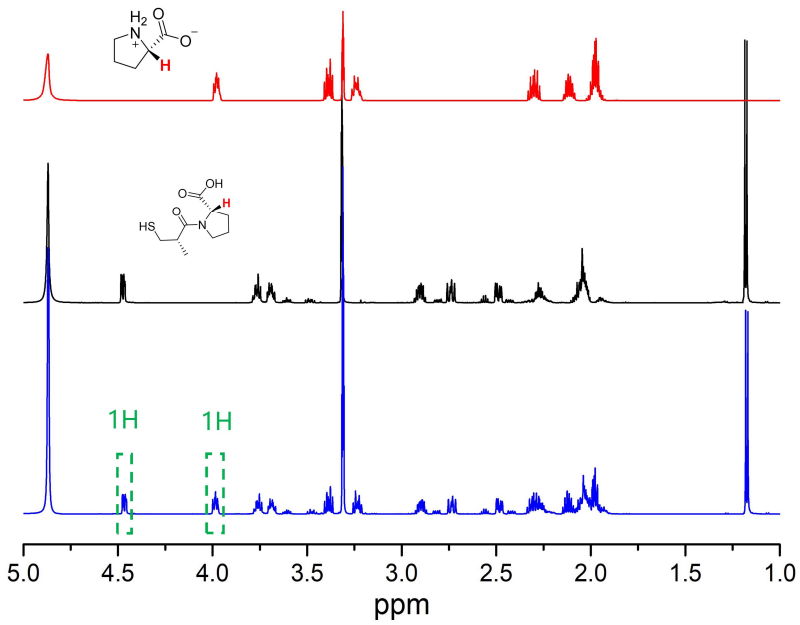
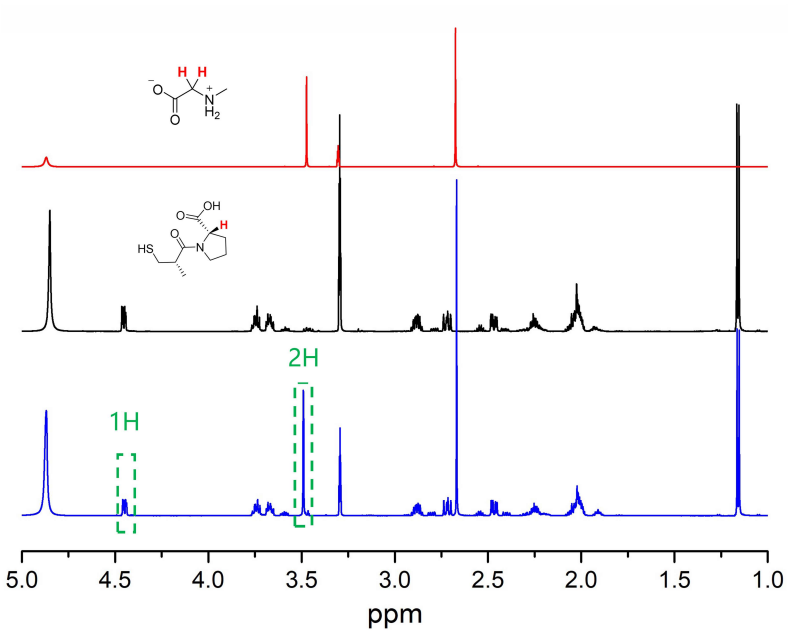


Figure S3. Cocrystal probability ranking by the MACCS-based RF and PubChem-based RF for 54 coformers against captopril. Pro and Sar are colored red.



(A)



(B)

Figure S4. ^{1}H-NMR (Methanol-D4, 600MHz) of **(A)** CapPro and **(B)** CapSar compared with API and the corresponding coformers. The intergration values suggested that the stoichiometric ratio of Cap : Pro = 1:1, and Cap : Sar = 1:1.

Table S1. Hyperparamter searching results for different combinations of ML models and molecular fingerprints

Fingerprint	Model	Mean of ROC-AUC [a]	Best hyperparameter
MACCS	RF	0.871 (0.022)	min_samples_split=3 n_estimators=1000
	GB	0.854 (0.034)	learning_rate=0.1 n_estimators=1000 sub_sample=0.6
	DNN	0.839 (0.029)	dropout_rate=0.1 hidden_num=256 learning_rate=0.01
	LR	0.641 (0.021)	C=0.01 penalty='none'
	MNB	0.609 (0.025)	-
	AB	0.740 (0.046)	learning_rate=1 n_estimators=1000
PubChem	RF	0.867 (0.026)	min_samples_split=2 n_estimators=1000
	GB	0.853 (0.024)	learning_rate=0.1 n_estimators=1000 sub_sample=0.8
	DNN	0.838 (0.028)	dropout_rate=0.1 hidden_num=256 learning_rate=0.0001
	LR	0.651 (0.015)	C=0.01 penalty='l2'
	MNB	0.594 (0.026)	-
	AB	0.749 (0.026)	learning_rate=1 n_estimators=1000
Substructure	RF	0.840 (0.021)	min_samples_split=5 n_estimators=1000
	GB	0.815 (0.030)	learning_rate=0.1 n_estimators=1000 sub_sample=0.6
	DNN	0.819 (0.026)	dropout_rate=0.1 hidden_num=256 learning_rate=0.0001
	LR	0.639 (0.013)	C=0.01 penalty='none'
	MNB	0.612 (0.020)	-
	AB	0.701 (0.050)	learning_rate=1 n_estimators=500

Estate	RF	0.831 (0.020)	min_samples_split=5 n_estimators=1000
	GB	0.808 (0.026)	learning_rate=0.1 n_estimators=1000 sub_sample=0.7
	DNN	0.803 (0.023)	dropout_rate=0.1 hidden_num=256 learning_rate=0.001
	LR	0.633 (0.022)	C=0.01 penalty='none'
	MNB	0.603 (0.008)	-
	AB	0.705 (0.030)	learning_rate=1 n_estimators=1000
CDK	RF	0.815 (0.018)	min_samples_split=5 n_estimators=1000
	GB	0.826 (0.019)	learning_rate=0.1 n_estimators=1000 sub_sample=0.7
	DNN	0.802 (0.016)	dropout_rate=0.1 hidden_num=256 learning_rate=0.01
	LR	0.653 (0.012)	C=0.01 penalty='none'
	MNB	0.613 (0.014)	-
	AB	0.732 (0.013)	learning_rate=1 n_estimators=1000
KR	RF	0.833 (0.022)	min_samples_split=3 n_estimators=500
	GB	0.814 (0.029)	learning_rate=0.1 n_estimators=1000 sub_sample=0.7
	DNN	0.804 (0.023)	dropout_rate=0.1 hidden_num=256 learning_rate=0.01
	LR	0.658 (0.013)	C=0.01 penalty='none'
	MNB	0.629 (0.010)	-
	AB	0.714 (0.028)	learning_rate=1 n_estimators=500
CDKGO	RF	0.834 (0.027)	min_samples_split=3 n_estimators=1000

	GB	0.822 (0.024)	learning_rate=0.1 n_estimators=1000 sub_sample=0.7
	DNN	0.799 (0.020)	dropout_rate=0.3 hidden_num=256 learning_rate=0.01
	LR	0.651 (0.017)	C=0.01 penalty='none'
	MNB	0.601 (0.027)	-
	AB	0.714 (0.036)	learning_rate=1 n_estimators=500
CDKExt	RF	0.821 (0.019)	min_samples_split=5 n_estimators=1000
	GB	0.835 (0.021)	learning_rate=0.1 n_estimators=1000 sub_sample=0.8
	DNN	0.804 (0.032)	dropout_rate=0.1 hidden_num=256 learning_rate=0.01
	LR	0.651 (0.014)	C=0.01 penalty='l2'
	MNB	0.617 (0.020)	-
	AB	0.718 (0.022)	learning_rate=1 n_estimators=1000
ECFP4	RF	0.830 (0.023)	min_samples_split=3 n_estimators=500
	GB	0.800 (0.028)	learning_rate=0.1 n_estimators=1000 sub_sample=0.7
	DNN	0.796 (0.026)	dropout_rate=0.1 hidden_num=256 learning_rate=0.0001
	LR	0.662 (0.016)	C=0.01 penalty='l2'
	MNB	0.647 (0.008)	-
	AB	0.703 (0.029)	learning_rate=1 n_estimators=500
FCFP4	RF	0.844 (0.015)	min_samples_split=2 n_estimators=1000
	GB	0.800 (0.029)	learning_rate=0.1 n_estimators=1000 sub_sample=0.7

DNN	0.801 (0.020)	dropout_rate=0.3 hidden_num=256 learning_rate=0.0001
LR	0.662 (0.021)	C=0.01 penalty='l2'
MNB	0.642 (0.010)	-
AB	0.710 (0.040)	learning_rate=1 n_estimators=500

[a] Reported in the form of mean value (standard derivation)

Table S2. Probabilities of formation of a cocrystal between diclofenac and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
XUVSIR	**Positive**	**62.9**	**0.361**
XUVQUB	**Positive**	**54.8**	**0.443**
4,4'-bipyridine	**Positive**	**41.8**	**0.688**
3-aminopyridine	**Positive**	**12.1**	**0.475**
2-aminopyridine	**Positive**	**9.9**	**0.559**
2-amino-4,6-dimethylpyrimidine	**Positive**	**9.1**	**0.565**
3,5-dimethylpyrazole	Negative	8.9	0.333
2-amino-3,5-dibromopyridine	**Positive**	**8.7**	**0.458**
2-chloropyrimidine	Negative	8.5	0.589
Isonicotinamide	**Positive**	**8.5**	**0.604**
2-amino-5-chloropyridine	**Positive**	**6.4**	**0.562**
XUVQEL	**Positive**	**6.4**	**0.451**
4-chloro-2,6-diaminopyrimidine	**Positive**	**6.1**	**0.518**
Pyrazole	Negative	6.0	0.391
2-aminopyrimidine	**Positive**	**5.4**	**0.600**
4-iodopyrazole	Negative	4.8	0.388
2-amino-4-hydroxy-6-methylpyrimidine	**Positive**	**4.3**	**0.516**
2-amino-4-chloro-6-methylpyrimidine	**Positive**	**4.3**	**0.554**
3,5-dimethyl-4-chloropyrazole	Negative	3.8	0.348
4-bromopyrazole	Negative	3.6	0.372
4-iodo-3,5-dimethylpyrazole	Negative	3.2	0.336
3-hydroxypyridine	**Positive**	**3.0**	**0.483**

Table S3. Probabilities of formation of a cocrystal between piracetam and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
3,4-dihydroxybenzoic	**positive**	**98.1**	**0.818**
3,5-dihydroxybenzoic	**positive**	**97.8**	**0.802**
Hydroquinone	**positive**	**92.1**	**0.517**
Gentisic acid	**positive**	**89.2**	**0.806**
4-hydroxybenzoic acid	**positive**	**78.9**	**0.811**
Fumaric acid	negative	72.7	0.789
2,3-dihydroxybenzoic	**positive**	**72.7**	**0.804**
Urea	negative	56.8	0.468
2,4-dihydroxybenzoic	**positive**	**56.3**	**0.847**
Succinic acid	negative	54.5	0.794
S-mandelic acid	**positive**	**51.6**	**0.733**
Tartaric acid	**positive**	**40.2**	**0.823**
Glucuronic acid	negative	38.5	0.662
Saccharin	negative	30.3	0.513
Citric acid	**positive**	**28.8**	**0.691**
Camphoric acid	negative	22.3	0.510
Maleic acid	negative	20.1	0.789
4-acetamidobenzoic acid	negative	14.4	0.386
2-amino-5-methylbenzoic	negative	13.0	0.394
Methyl-p-	negative	12.9	0.631
Glutaric acid	negative	11.0	0.716
Glycine	negative	10.6	0.453
2,6-dihydroxybenzoic acid	negative	10.3	0.809

mannitol	negative	7.5	0.522
piperazine	negative	7.2	0.372
Hippuric acid	negative	5.6	0.290
sulfaproxyline	negative	4.0	0.432
L-proline	negative	3.6	0.419
Imidazole	negative	3.1	0.368

Table S4. Probabilities of formation of a cocrystal between pyrazine carboxamide and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
3,5-dihydroxybenzoic acid	**positive**	**87.0**	**0.810**
Gallic acid	**positive**	**84.1**	**0.785**
Gentisic acid	**positive**	**54.9**	**0.807**
4-hydroxybenzoic acid	**positive**	**49.5**	**0.798**
Oxalic acid	**positive**	**47.5**	**0.863**
3-hydroxybenzoic acid	**positive**	**41.9**	**0.798**
1-hydroxy-2-naphthoic acid	**positive**	**39.3**	**0.636**
Fumaric acid	**positive**	**31.5**	**0.860**
Indole-2-carboxylic acid	**positive**	**29.3**	**0.631**
Salicylic acid	**positive**	**28.1**	**0.787**
α-Ketoglutaric acid	negative	22.4	0.798
Malonic acid	**positive**	**20.6**	**0.862**
4-aminobenzoic acid	**positive**	**19.7**	**0.804**
Adamantanecarboylic acid	negative	18.3	0.594
Succinic acid	**positive**	**17.2**	**0.880**
Indole-3-carboxylic acid	negative	16.8	0.614
Undecylenic acid	negative	16.7	0.690
2-aminobenzoic acid	**positive**	**14.5**	**0.789**
Pyruvic acid	negative	13.6	0.509
Tartaric acid	negative	12.8	0.748
Camphoric acid	negative	12.2	0.559
4-aminosalicylic acid	**positive**	**11.4**	**0.792**
Adipic acid	**positive**	**11.3**	**0.890**
Isonicotinic acid	negative	11.2	0.713
L-tyrosine	negative	10.9	0.553
Phenylalanine	negative	10.8	0.548
Pyrazine-2-carboxylic acid	negative	10.4	0.575

Azelaic acid	negative	10.4	0.847
Vanillic acid	**positive**	**9.4**	**0.780**
Hydrocinnamic acid	negative	9.1	0.757
Saccharin	negative	8.5	0.521
Nicotinic acid	negative	8.2	0.661
4-nitrobenzamide	**positive**	**8.2**	**0.394**
Glutaric acid	**positive**	**8.1**	**0.885**
Hippuric acid	negative	7.6	0.528
3-aminobenzoic acid	**positive**	**7.3**	**0.807**
Pyrazine-2,3-carboxylic acid	negative	7.2	0.627
Benzoic acid	negative	6.5	0.784
Glycolic acid	negative	4.9	0.693
L-pyroglutammic acid	negative	4.7	0.410
Isonicotinamide	negative	4.2	0.415
Cinnamic acid	negative	3.8	0.726
Nicotinamide	negative	3.5	0.434
L-histidine	negative	3.4	0.511
Benzamide	negative	3.0	0.211

Table S5. Probabilities of formation of a cocrystal between acetazolamide and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
4-aminobenzamide	negative	29.8	0.438
3,5-dihydroxybenzoic acid	negative	25.7	0.340
3-aminobenzamide	negative	21.4	0.430
Caffeine	negative	21.2	0.546
3,4-dihydroxybenzoic acid	negative	20.9	0.332
2-aminobenzamide	**positive**	**20.8**	**0.417**
2-pyridinecarboxamide	**positive**	**16.1**	**0.445**
2-hydroxybenzamide	**positive**	**13.4**	**0.419**
4-aminobenzoic acid	negative	12.1	0.467
2,3-hydroxybenzoic acid	**positive**	**11.9**	**0.315**
4-hydroxybenzoic acid	**positive**	**8.4**	**0.350**
3-hydroxybenzamide	negative	8.1	0.431
4-hydroxybenzamide	negative	7.7	0.452
2-pyridinecarboxylic acid	negative	7.1	0.496
Salicylic acid	negative	6.9	0.303
3-hydroxybenzoic acid	negative	6.9	0.328
Nicotinamide	**positive**	**6.9**	**0.437**
Gentisic acid	negative	6.7	0.323
2,4-dihydroxybenzoic acid	negative	6.6	0.336
2-aminobenzoic acid	negative	5.6	0.407
3-aminobenzoic acid	negative	4.1	0.435
4-pyridinecarboxylic acid	negative	4.1	0.496
Isonicotinic acid	negative	4.0	0.496
Succinic acid	negative	3.8	0.517
Nicotinic acid	negative	3.7	0.480
Oxalic acid	negative	3.7	0.471
3-pyridinecarboxylic acid	negative	3.6	0.480

Adipic acid	negative	3.5	0.518
Isonicotinamide	negative	3.4	0.440
Camphoric acid	negative	3.4	0.298
2-hydroxynicotinic acid	negative	3.2	0.481
Malonic acid	negative	3.1	0.487
Saccharin	negative	3.1	0.589
Citric acid	negative	3.0	0.444
Glutaric acid	negative	2.9	0.502
Benzoic acid	negative	2.8	0.285

Table S6. Probabilities of formation of a cocrystal between indomethacin and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
Ethyl maltol	negative	77.20	0.288
Urea	negative	44.60	0.443
Lidocaine	**positive**	**37.40**	**0.350**
1,2-ethanedisulfonic acid	negative	35.60	0.288
lysine	negative	29.20	0.378
4-aminobenzamide	negative	27.40	0.453
4-aminobenzoic acid	negative	21.60	0.601
Lactamide	**positive**	**19.50**	**0.440**
Aspartic acid	negative	19.40	0.357
Maleic acid	negative	18.60	0.289
N-methyl-D-glucamine	negative	15.00	0.352
4-hydroxybenzoic acid	negative	14.30	0.337
4,4-bipyridine	**positive**	**13.00**	**0.136**
4-hydroxybenzamide	negative	12.80	0.468
1-hydroxy-2-naphthoic acid	negative	10.40	0.250
Glycolamide	negative	10.40	0.437
Hippuric acid	negative	10.10	0.436
Gentisic acid	negative	9.90	0.337
Tromethamine	negative	9.60	0.303
Tartaric acid	negative	9.20	0.267
Mannitol	negative	8.50	0.187
Maltose	negative	8.10	0.230
Nicotinamide	**positive**	**8.00**	**0.651**
Benzamide	**positive**	**7.90**	**0.426**
Vanillic acid	negative	7.80	0.359
Glycine	negative	7.30	0.411
Stearic acid	negative	6.60	0.231

Lactose	negative	6.50	0.230
Neotame	negative	6.40	0.358
DL-Lactic acid	negative	5.90	0.283
Sucrose	negative	5.50	0.226
Fructose	negative	4.80	0.208
Malic acid	negative	4.80	0.249
Mannose	negative	4.70	0.230
Saccharin	**positive**	**4.40**	**0.414**
Tryptophan	negative	4.40	0.460
Isonicotinamide	negative	4.40	0.690
DL-Mandelic acid	**positive**	**4.10**	**0.285**
Sorbic acid	negative	3.70	0.303
Malonic acid	negative	3.60	0.262
Cyclamic acid	negative	3.60	0.307
Ascorbic acid	negative	3.60	0.240
Arabinose	negative	3.50	0.218
4-Toluenesulfonic acid	negative	3.50	0.291
Naphthalenesulfonic acid	negative	3.30	0.234
Glycolic acid	negative	3.30	0.245
Cinnamic acid	**positive**	**3.30**	**0.302**
Fumaric acid	negative	3.20	0.289
Succinic acid	negative	3.00	0.228
Glucose	negative	3.00	0.230
Adipic acid	negative	2.90	0.170
Lactic acid	negative	2.90	0.283
Leucine	negative	2.90	0.352
Oxalic acid	negative	2.90	0.260
Glutaric acid	negative	2.90	0.197
Citric acid	negative	2.70	0.244
Benzoic acid	negative	2.70	0.338

Table S7. Probabilities of formation of cocrystal between drug candidate and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
Orotic acid	negative	55.6	0.347
Gentisic acid	**positive**	**39.4**	**0.458**
Fumaric acid	**positive**	**27.7**	**0.480**
Succinic acid	**positive**	**20.9**	**0.503**
Salicylic acid	negative	19.4	0.444
Uracil	negative	18.0	0.356
Malonic acid	**positive**	**17.4**	**0.491**
Piperazine	negative	14.7	0.285
Tartaric acid	negative	10.5	0.475
Saccharin	negative	9.5	0.404
Adipic acid	**positive**	**9.2**	**0.544**
Citric acid	negative	7.9	0.468
Succinamide	negative	7.5	0.382
Maleic acid	**positive**	**5.9**	**0.480**
Lysine	negative	5.3	0.385
L-arginine	negative	4.8	0.379
Glutaric acid	**positive**	**4.6**	**0.526**
Ascorbic acid	negative	4.6	0.374
Nicotinic acid	negative	4.5	0.395
Glycine	negative	3.6	0.359
L-valine	negative	3.4	0.351
Tromethamine	negative	3.4	0.286
N-methyl-D-glucamine	negative	3.3	0.323
Hippuric acid	negative	3.0	0.350
L-histidine	negative	2.9	0.376
L-pyroglutammic acid	negative	2.9	0.391
Isonicotinamide	negative	2.8	0.297
Nicotinamide	negative	2.7	0.327

Table S8. Probabilities of formation of cocrystal between furosemide and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
Gallic acid	negative	25.3	0.135
Cytosine	**positive**	**19.8**	**0.446**
Urea	**positive**	**16.8**	**0.419**
Asparagine	negative	14.7	0.385
Acetylsalicylic acid	negative	11.2	0.212
Caffeine	**positive**	**10.2**	**0.471**
Salicylic acid	negative	9.1	0.111
4-aminobenzoic acid	**positive**	**8.9**	**0.368**
Fumaric acid	negative	8.8	0.234
Aspartic acid	negative	8.4	0.351
Tartaric acid	negative	7.9	0.264
Acetamide	**positive**	**6.9**	**0.420**
Saccharin	negative	6.9	0.382
Adenine	**positive**	**6.7**	**0.312**
Succinic acid	negative	6.2	0.260
Glutamic acid	negative	5.8	0.365
Citric acid	negative	5.8	0.268
Nicotinamide	**positive**	**5.3**	**0.491**
Benzamide	negative	4.7	0.357
Glutamine	negative	4.7	0.397
Nicotinic acid	negative	4.5	0.468
Adipic acid	negative	4.4	0.258
4-hydroxybenzamide	negative	4.3	0.330
Leucine	negative	3.9	0.343
Phenylalanine	negative	3.8	0.281
Benzoic acid	negative	3.6	0.146
Ascorbic acid	negative	3.4	0.316
Isonicotinamide	**positive**	**3.1**	**0.509**

Table S9. Probabilities of formation of cocrystal between nalidixic acid and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
Phloroglucinol	**positive**	**100.0**	**0.504**
Resorcinol	**positive**	**100.0**	**0.522**
Orcinol	**positive**	**100.0**	**0.501**
Hydroquinone	**positive**	**100.0**	**0.518**
Pyrogallol	**positive**	**99.9**	**0.501**
Catechol	**positive**	**98.7**	**0.504**
Ferulic acid	negative	95.1	0.474
Urea	negative	92.5	0.407
Salicylic acid	negative	91.8	0.518
Adenine	negative	90.8	0.321
L-glutamine	negative	77.2	0.334
Thymine	negative	60.0	0.392
Isonicotinamide	negative	51.7	0.386
Nicotinic acid	negative	48.5	0.443
Benzoic acid	negative	48.4	0.471
L-threonine	negative	43.2	0.392
Cytosine	negative	42.8	0.382
Nicotinamide	negative	32.5	0.381
L-proline	negative	23.6	0.367
L-lysine	negative	19.0	0.342
L-histidine	negative	18.8	0.401
Acetamide	negative	7.6	0.397

Table S10. Probabilities of formation of cocrystal between paracetamol and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
N,N-dimethyl piperazine	**positive**	**100.0**	**0.405**
DABCO	**positive**	**99.7**	**0.474**
Cyclam	**positive**	**96.6**	**0.368**
N-methylmorpholine	**positive**	**84.6**	**0.410**
Morpholine	**positive**	**57.8**	**0.438**
Piperazine	**positive**	**54.2**	**0.453**
trans-1,4-diaminocyclohexane	**positive**	**50.9**	**0.418**
3-isochromanone	negative	44.2	0.425
4,4-trimethylenedipyridine	negative	41.0	0.563
1,2-bisdipyridylethane	**positive**	**34.0**	**0.563**
4,4'-bipyridine	**positive**	**21.5**	**0.560**
Resorcinol	negative	15.0	0.435
Theophylline	**positive**	**14.2**	**0.605**
Caffeine	negative	13.8	0.584
trans-1,4-dipyridylethylene	**positive**	**13.3**	0.558
Maleic acid	negative	12.6	0.481
Pyrazine	**positive**	**10.7**	**0.551**
Theobromine	negative	10.2	0.601
1-naphthol	negative	9.4	0.385
Phenazine	**positive**	**8.2**	**0.468**
Gentisic acid	negative	8.0	0.455
cis-3,6-dimethy-1,4-dioxane-2,5-dione	negative	8.0	0.463
Melamine	negative	6.6	0.490
Malonic acid	negative	6.5	0.447
Nicotinamide	negative	6.0	0.429

Malic acid	negative	5.8	0.388
Succinic acid	negative	5.2	0.418
Saccharin	negative	4.9	0.514
Oxalic acid	**positive**	**4.6**	**0.469**
Adipic acid	negative	4.4	0.397
Isonicotinamide	negative	4.2	0.459
Imidazole	negative	4.0	0.446
Ascorbic acid	negative	3.9	0.354
Benzoic acid	negative	3.6	0.415
Citric acid	**positive**	**3.2**	**0.352**
Naphthalene	**positive**	**2.6**	**0.387**
Anthracene	negative	2.6	0.357

Table S11. Probabilities of formation of cocrystal between 3-cyanophenol and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
4-phenylpyridine	**positive**	**54.8**	**0.560**
1,2-bis(4-dipyridyl)ethane	**positive**	**37.5**	**0.677**
1-hexadecanol	negative	34.9	0.223
trans-1,2-bis(4-dipyridyl)ethylene	**positive**	**17.5**	**0.687**
4,4'-bipyridine	negative	17.4	0.699
1,3,5-trihydroxybenzene	negative	14.3	0.264
4,4'-biphenol	negative	13.8	0.325
1,3-dihydroxybenzene	negative	5.3	0.295
1-cyanonaphthalene	negative	5.1	0.355
1-naphthol	negative	4.9	0.239
5-hydroxyisoquinoline	negative	4.7	0.513
3-hydroxypyridine	negative	4.0	0.546
1,3-dicyanobenzene	negative	4.0	0.445
1,4-dicyanobenzene	negative	4.0	0.453
Bicalutamide	negative	3.2	0.302
4-cyanopyridine	negative	3.2	0.560
3-cyanopyridine	negaitve	2.9	0.523
4-cyanophenol	negative	2.6	0.431

Table S12. Probabilities of formation of cocrystal between 4-cyanophenol and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
4-phenylpyridine	**positive**	**49.5**	**0.550**
1,2-bis(4-dipyridyl)ethane	**positive**	**31.5**	**0.658**
1-hexadecanol	negative	28.1	0.206
1,3,5-trihydroxybenzene	negative	16.8	0.281
4,4'-biphenol	negative	16.0	0.381
trans-1,2-bis(4-dipyridyl)ethylene	**positive**	**14.3**	**0.674**
4,4'-bipyridine	**positive**	**12.8**	**0.682**
1-naphthol	negative	6.1	0.233
5-hydroxyisoquinoline	negative	5.8	0.515
1,3-dihydroxybenzene	negative	5.3	0.312
1-cyanonaphthalene	negative	4.8	0.343
1,3-dicyanobenzene	negative	4.1	0.433
1,4-dicyanobenzene	negative	4.1	0.438
3-hydroxypyridine	negative	3.8	0.557
4-cyanopyridine	negative	3.4	0.558
3-cyanopyridine	negative	3.2	0.519
Bicalutamide	negative	3.1	0.301
3-cyanophenol	negative	2.6	0.431

Table S13. Probabilities of formation of cocrystal between 3-cyanopyridine and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
1,3,5-trihydroxybenzene	negative	39.2	0.492
4,4'-biphenol	**positive**	**36.0**	**0.610**
1-naphthol	negative	16.7	0.540
1-hexadecanol	negative	16.2	0.242
1,3-dihydroxybenzene	negative	11.6	0.530
4-phenylpyridine	negative	8.0	0.069
1,2-bis(4-dipyridyl)ethane	negative	6.4	0.045
4,4'-bipyridine	negative	4.7	0.043
1,4-dicyanobenzene	negative	4.6	0.454
1,3-dicyanobenzene	negative	4.1	0.418
5-hydroxyisoquinoline	negative	3.6	0.215
trans-1,2-bis(4-dipyridyl)ethylene	negative	3.0	0.049
bicalutamide	negative	3.4	0.288
4-cyanophenol	negative	3.2	0.519
1-cyanonaphthalene	negative	3.0	0.463
3-hydroxypyridine	negative	3.0	0.293
3-cyanophenol	negative	2.9	0.523
4-cyanopyridine	negative	2.6	0.180

Table S14. Probabilities of formation of cocrystal between 4-cyanopyridine and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
1,3,5-trihydroxybenzene	**positive**	**41.5**	**0.499**
4,4'-biphenol	**positive**	**38.7**	**0.659**
1-naphthol	negative	21.5	0.567
1-hexadecanol	negative	16.8	0.248
1,3-dihydroxybenzene	**positive**	**10.4**	**0.565**
4-phenylpyridine	negative	7.7	0.059
1,4-dicyanobenzene	negative	6.0	0.473
1,2-bis(4-dipyridyl)ethane	negative	6.0	0.038
1,3-dicyanobenzene	negative	5.1	0.437
5-hydroxyisoquinoline	negative	4.4	0.222
Bicalutamide	negative	4.2	0.298
4,4'-bipyridine	negative	4.2	0.033
4-cyanophenol	negative	3.4	0.558
3-cyanophenol	negative	3.2	0.560
3-hydroxypyridine	negative	3.2	0.303
trans-1,2-bis(4-dipyridyl)ethylene	negative	3.1	0.044
1-cyanonaphthanlene	negative	3.0	0.467
3-cyanopyridine	negative	2.6	0.180

Table S15. Probabilities of formation of cocrystal between benzamide and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
3,5-dinitrobenzoic acid	**positive**	**75.3**	**0.930**
2,4,6-trihydroxybenzoic acid	negative	52.0	0.827
4-nitrobenzoic acid	**positive**	**30.6**	**0.767**
4-hydroxybenzoic acid	negative	24.7	0.926
3-nitrobenzoic acid	**positive**	**24.6**	**0.854**
4-amino-3-nitrobenzoic acid	negative	24.3	0.749
4-hydroxy-3-nitrobenzoic acid	**positive**	**19.5**	**0.829**
Succinic acid	**positive**	**18.8**	**0.837**
Salicylic acid	**positive**	**15.9**	**0.839**
Pentafluorobenzoic acid	**positive**	**13.9**	**0.675**
4-aminobenzoic acid	negative	12.8	0.739
Camphoric acid	negative	11.3	0.451
Cinnamic acid	negative	3.4	0.764

Table S16. Probabilities of formation of cocrystal between itraconazole and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
Fumaric acid	**positive**	**94.3**	**0.340**
Malonic acid	negative	89.5	0.338
Succinic acid	**positive**	**86.4**	**0.322**
Malic acid	**positive**	**78.1**	**0.370**
Tartaric acid	**positive**	**74.9**	**0.376**
Maleic acid	negative	59.7	0.340
Adipic acid	negative	57.4	0.318
Glutaric acid	negative	37.8	0.318

Table S17. Probabilities of formation of cocrystal between bicalutamide and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
1,2-bis(4-dipyridyl)ethane	negative	86.3	0.310
4,4'-biphenol	negative	81.6	0.280
4-phenylpyridine	negative	81.1	0.287
1,3,5-trihydroxybenzene	negative	63.7	0.301
1-hexadecanol	negative	58.6	0.327
4,4'-bipyridine	**positive**	**58.3**	**0.307**
trans-1,2-bis(4-dipyridyl)ethylene	**positive**	**53.0**	**0.302**
1-naphthol	negative	39.4	0.303
1,3-dihydroxybenzene	negative	20.0	0.306
5-hydroxyisoquinoline	negative	12.6	0.315
1-cyanonaphthalene	negative	10.7	0.295
3-hydroxypyridine	negative	7.7	0.368
4-cyanopyridine	negative	4.2	0.298
1,4-dicyanobenzene	negative	3.4	0.288
3-cyanopyridine	negative	3.4	0.288
3-cyanophenol	negative	3.2	0.302
1,3-dicyanobenzene	negative	3.2	0.279
4-cyanophenol	negative	3.1	0.301

Table S18. Probabilities of formation of cocrystal between meloxicam and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
Gentisic acid	**positive**	**94.0**	**0.629**
4-hydroxybenzoic acid	**positive**	**87.6**	**0.605**
Fumaric acid	**positive**	**85.6**	**0.455**
Malonic acid	**positive**	**76.4**	**0.538**
Succinic acid	**positive**	**72.4**	**0.537**
Malic acid	**positive**	**63.7**	**0.521**
1-hydroxy-2-naphthoic acid	**positive**	**59.1**	**0.553**
Tartaric acid	negative	53.8	0.509
Salicylic acid	**positive**	**49.7**	**0.653**
Camphoric acid	**positive**	**44.1**	**0.135**
Adipic acid	**positive**	**44.0**	**0.508**
Maleic acid	**positive**	**26.1**	**0.455**
Glutaric acid	**positive**	**25.5**	**0.504**
Ascorbic acid	negative	20.5	0.416
Hydrocinnamic acid	**positive**	**19.4**	**0.496**
Benzoic acid	**positive**	**14.4**	**0.601**
Glycolic acid	**positive**	**10.8**	**0.475**

Table S19. Probabilities of formation of cocrystal between nicotinamide and coformers predicted by MEPS-based method and the consensus model. Positive samples are highlighted in bold.

Coformer	Label	MEPS-based Method	Consensus
Salicylic acid	**positive**	**18.3**	**0.811**
Ibuprofen	**positive**	**10.1**	**0.835**
Flurbiprofen	**positive**	**8.9**	**0.760**
Fenbufen	**positive**	**6.7**	**0.749**
Paracetamol	negative	6.0	0.429
Ketoprofen	negative	5.6	0.615
Piracetam	negative	3.7	0.215

Table S20. Brief explanations for symbols used in Figure 4

Symbol	Explanation
-	single bond (aliphatic)
=	double bond
#	triple bond
%	aromatic bond
$	ring bond
@	ring linkage
@1	linked back to the first atom in the list
~	any bond
>= 1N	more than one nitrogen atom
[#1]	hydrogen atom
Q	heteroatoms; any non-C or non-H atom
A	any valid periodic table element symbol

Table S21. Counts of values of top 20 important fingerprint bits on the whole data set

Value	0		1		2	
Classification	Positive	Negative	Positive	Negative	Positive	Negative
PubChem(374)	3820	2825	2712	3536	708	879
PubChem(308)	2504	3183	4304	3229	432	828
PubChem(406)	2710	3371	4114	3123	416	746
PubChem(617)	3054	3648	3830	2967	356	625
PubChem (2)	5142	4405	1856	2522	242	313
PubChem(14)	737	1089	4411	3403	2092	2748
PubChem(1)	1047	836	3942	3397	2251	3007
PubChem(440)	4098	4474	2890	2443	252	323
PubChem(285)	764	1113	4413	3420	2063	2707
PubChem(12)	4994	4341	2019	2546	227	353
PubChem(11)	1393	905	3543	3422	2304	2913
PubChem(186)	4437	3839	2377	2900	426	501
PubChem(181)	3352	3908	3645	2835	243	497
PubChem(380)	3980	4303	2968	2565	292	372
PubChem(452)	4064	4409	2913	2489	263	342
PubChem(667)	3929	4454	3114	2426	197	360
PubChem(351)	960	1238	4424	3493	1856	2509
PubChem(346)	5972	5498	1093	1631	175	111
PubChem(180)	3426	4016	3596	2770	218	454
PubChem(366)	5561	4938	1434	2111	245	191
MACCS(160)	3820	2825	2712	3536	708	879
MACCS(139)	2504	3183	4304	3229	432	828
MACCS(98)	2994	3390	3834	3133	412	717
MACCS(161)	737	1089	4411	3403	2092	2748
MACCS(65)	3276	3692	3701	2986	263	562
MACCS(137)	1895	2075	4335	3624	1010	1541
MACCS(123)	3980	4303	2968	2565	292	372
MACCS(149)	5073	4346	1860	2566	307	328
MACCS(134)	4931	4735	2132	2243	177	262
MACCS(121)	2417	2630	4075	3457	748	1153
MACCS(159)	1333	1594	4211	3612	1696	2034
MACCS(142)	1879	2239	4342	3541	1019	1460
MACCS(145)	2710	2432	3615	3568	915	1240
MACCS(146)	3330	3650	3523	2963	387	627
MACCS(96)	4876	4294	2066	2572	298	374
MACCS(131)	3028	3310	3527	3199	685	731
MACCS(154)	2426	2172	3068	3599	1746	1469
MACCS(151)	3718	3335	2817	3176	705	729
MACCS(105)	4176	3735	2589	2967	475	538
MACCS(157)	2024	2259	4107	3574	1109	1407

Table S22. DSC data of Cap and cocrystals

Phase	Cap	Pro	Sar	CapPro	CapSar
T_m (°C) [a]	108.8 ± 1.5	232.3 ± 2.0	216.1 ± 1.9	118.8 ± 1.6	126.1 ± 1.1
ΔH_{fusion} (J/g)	105.2	116.1	113.2	93.1	84.4

[a] T_m: melting point

Table S23. pKa_1 of Cap, Pro and Sar[1-3]

Compound	pKa_1
Cap	3.90 ± 0.05
Pro	1.95 ± 0.1
Sar	2.21 ± 0.1

1. Gupta, M.;da Silva, E. F.;Svendsen, H. F., Explicit Solvation Shell Model and Continuum Solvation Models for Solvation Energy and pKa Determination of Amino Acids. *J. Chem. Theory Comput.* **2013,** *9*, 5021-37.
2. Niazi, A.;Ghasemi, N., PARAFAC and PLS applied to determination of captopril in pharmaceutical preparation and biological fluids by ultraviolet spectrophotometry. *Annali Di Chimica* **2007,** *97*, 845-858.
3. Ratner, S.;Clarke, H. T., The action of formaldehyde upon cysteine. *J. Am. Chem. Soc.* **1937,** *59*, 200-206.