Deep Supramolecular Language Processing For Co-crystal Prediction

Rebecca Birolo[1,2], Rıza Özçelik[1], Andrea Aramini[3], Roberto Gobetto[2], Michele Chierotti[2], Francesca Grisoni[1*]

Check this out!

Co-crystals

✓ Multicomponent systems formed by supramolecular interactions between an active pharmaceutical ingredient (**API**) and a second organic molecule (**coformer**).

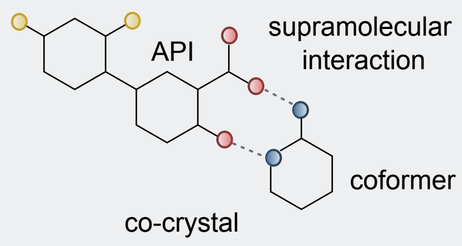
✓ Phamaceutical formulations to increase API **solubility** and **bioavailability**.

? Coformer selection is **time-consuming** and **resources-intensive** due to the vast number of possible combinations. While thousands of coformer are available, only a few are able to establish supramolecular interactions with a specific API.

DeepCocrystal

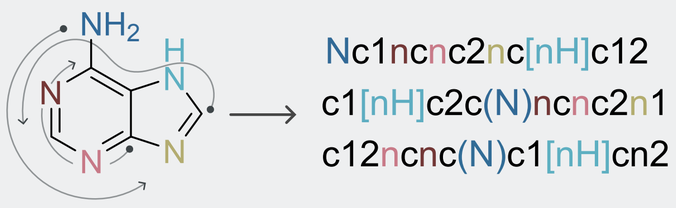
selects promising **API-coformer** pairs, limiting unsucessful lab experiments.

Unbalance dataset for model training
5240 (79%) 'positives' and 1392 (21%) 'negatives' data
↓
SMILES AUGMENTION

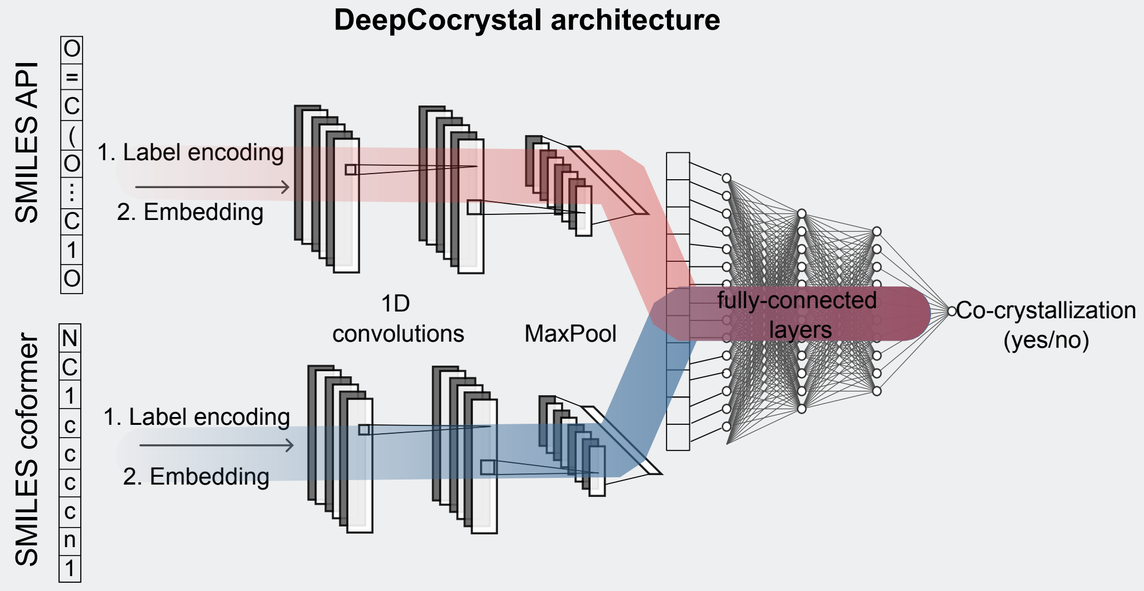


DeepCocrystal training, validation & benchmarking

SMILES strings



DeepCocrystal architecture



Performance of DeepCocrystal, tested on internal and external test sets.
Internal test set: 664 pairs sampled by stratified splits of the collected dataset.
External test set: 364 pairs more structurally diverse edit distance, and Tanimoto similarity computed) to the training set, also used to benchmark DeepCocrystal with existing literature models.

Test set	Model	BAcc	Recall	Specificity
Internal	DeepCocrystal - canonical	88% ± 2%	96% ± 1%	79% ± 6%
	DeepCocrystal - augmented (1:4)	88% ± 2%	91% ± 2%	86% ± 3%
	DeepCocrystal - augmented (2:7)	**89% ± 2%**	92% ± 2%	**87% ± 3%**
External	DeepCocrystal - canonical	59%	**93%**	26%
	DeepCocrystal - augmented (1:4)	69%	71%	66%
	DeepCocrystal - augmented (2:7)	**78%**	75%	**81%**
	CCGNet	60%	51%	69%
	CC-Descriptor-ML	63%	79%	48%
	Descriptor-DNN	63%	84%	41%
	Fingerprint-DNN	57%	90%	25%

Chemical language processing, traditionally employes a 'one-molecule-one-property' approach. → '**Supramolecular language**' processing, simultaneosly learns from the SMILES strings of pairs of molecules.

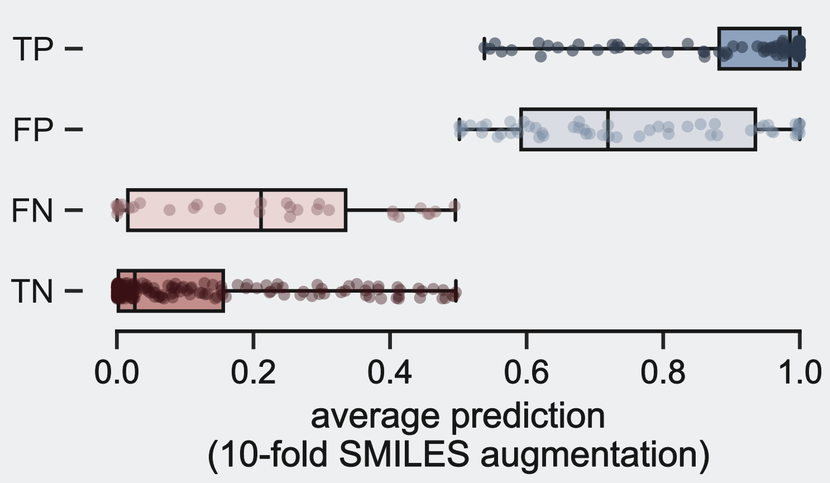
canonical SMILES, univocal standarized string per molecule. → *'randomized'* SMILES, differents strings based on the starting atom and graph trasveral route.

↓

Different levels of augmentation
specific number of randomization for positive and negatives optimized to balance the two classes [positive:negative = 1:4 or 2:7]

DeepCocrystal - augmented (2:7)
✓ outperforms benchmarks
✓ better trade-off between positive and negative predictions
✓ higher generalization potential

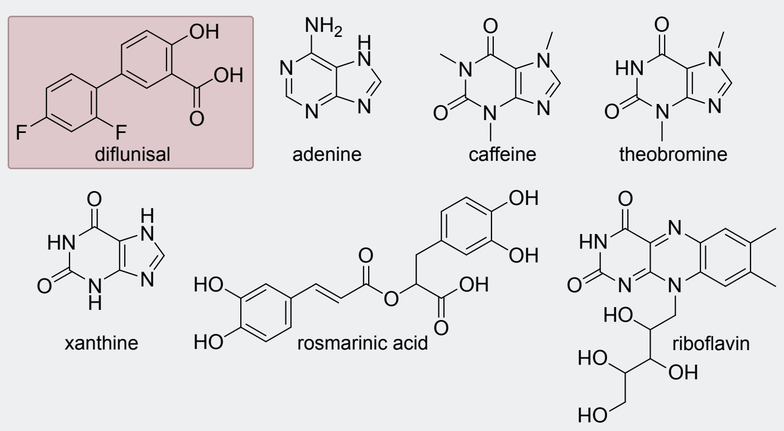
Uncertainty estimation



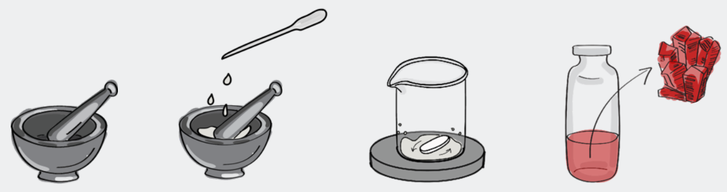
Uncertainty estimation with DeepCocrystal. External test set molecules were represented as **10 SMILES strings** each before prediction (using DeepCocrystal 2:7).
Majority voting and **standard deviation** were used to estimate uncertainty.

SMILES input	Method	Thr.	No. Pairs (%)	BAcc	Recall	Specificity
Canonical	-	-	364 (100%)	78%	75%	81%
Augmented (10-fold)	Major.	≥ 50%	364 (100%)	76%	75%	77%
	Major.	≥ 60%	348 (96%)	77%	75%	79%
	Major.	≥ 70%	313 (86%)	79%	77%	82%
	Major.	≥ 80%	287 (79%)	82%	79%	84%
	Major.	≥ 90%	254 (70%)	84%	82%	86%
	Major.	= 100%	218 (60%)	87%	**86%**	89%
	St. dev.	≤ 0.50	364 (100%)	76%	75%	77%
	St. dev.	≤ 0.40	351 (96%)	77%	76%	78%
	St. dev.	≤ 0.30	275 (76%)	82%	80%	83%
	St. dev.	≤ 0.30	227 (62%)	86%	85%	87%
	St. dev.	≤ 0.10	191 (52%)	**88%**	**86%**	90%
	St. dev.	= 0.05	161 (44%)	**88%**	84%	**91%**

Prospective experimental application



12 natural products containing polyphenolic or purine moieties selected as **coformer candidates** and **predicted** with DeepCocrystal 2:7 → **Experimental screening** of:
- top-two *high-certainty positive* predictions
- top-two *high-certainty negative* predictions
- two *most uncertainty* predictions



Samples were analyzed by **IR spectroscopy** and **solid-state NMR** to discriminate between co-crystal and non-cocrystal formation.

Tested coformer	DeepCocrystal Prediction	DeepCocrystal Outcome	Experimental Outcome
Adenine	0.99 ± 0.00	✓	✓
Caffeine	0.99 ± 0.01	✓	✓
Theobromine	0.66 ± 0.35	?	x
Xanthine	0.63 ± 0.38	?	x
Rosmarinic acid	0.02 ± 0.02	x	x
Riboflavin	0.00 ± 0.00	x	x

[1] Institute for Complex and Molecular Systems, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands.
[2] Department of Chemistry and NIS Centre, University of Torino, Torino, Italy.
[3] Research and early Development, Dompé Farmaceutici S.p.A, L'Aquila, Italy.

ICML International Conference On Machine Learning

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

UNIVERSITÀ DI TORINO