Neuraldecipher -
Reverse-engineering extended-connectivity fingerprints (ECFPs)

9th RDKit User Group Meeting (UGM)
06 October – 08 October 2020

Tuan Le
Machine Learning Research
06.10.2020

Agenda

// Motivation

// Workflow

// Results

Motivation

- Protecting molecular structures from disclosure is of great relevance for pharmaceutical companies
- Within external collaborations, descriptors are often exchanged to improve QSAR or ADMET models
- The folded extended-connectivity fingerprints (ECFPs) are frequently shared, as they are specifically developed for QSAR modeling[1]
- Unfolded ECFPs are often considered as non-invertible due to the usage of a hashing function to map input arrays to integers of the 2^{32}-space
- Folding ECFPs into fix-sized vectors is required for downstream tasks such as clustering or predictive modeling → Information is lost due to folding, also known as *hash/bit collision*

(1) D. Rogers and M. Hahn , *J. Chem. Inf. Model.*, 2010, **50** , 742 —754

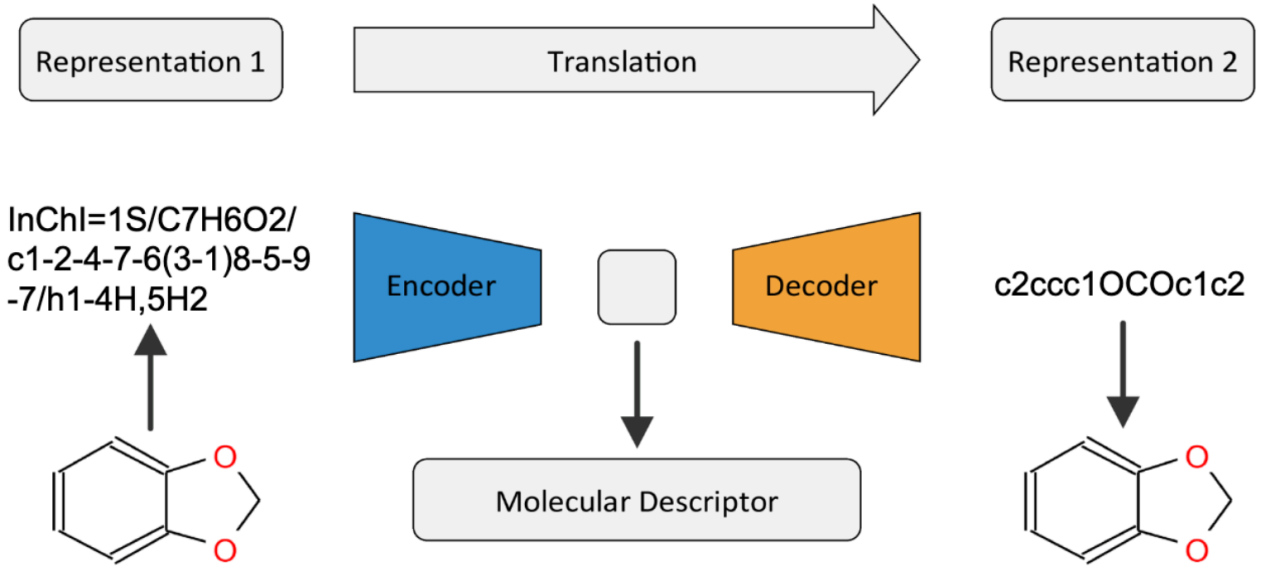
Motivation

- ECFPs have been exchanged between AstraZeneca (1.41M samples) and Bayer AG (2.75M samples) in 2013 to analyze the chemical similarity of the two collections[2]
- Within the IMI funded Joint European Compound Library (JECL, 2013-2018), 312K ECFPs were shared among seven pharmaceuticals as well as academic research groups to accelerate drug discovery on a precompetitive stage[3]
- The exchange of ECFPs was mainly reference by means of structure-free comparison without disclosure of the compound structure
- **Question: Is it possible to deduce molecular structures of folded ECFPs ?**
  - Related work in de novo drug design by Kotsias *et al.*[4] and in the GuacaMol benchmark for drug discovery generative models by Brown *et al.*[5] → Our intention differs, as we want to learn the mapping between ECFP and corresponding compound.

(2) T. Kogej , N. Blomberg , P. J. Greasley , S. Mundt , M. J. Vainio , J. Schamberger , G. Schmidt and J. Hüser , *Drug Discovery Today*, 2013, **18** , 1014 —1024
(3) J. Besnard , P. S. Jones , A. L. Hopkins and A. D. Pannifer , *Drug Discovery Today*, 2015, **20** , 181 —186
(4) P.-C. Kotsias , J. Arús-Pous , H. Chen , O. Engkvist , C. Tyrchan and E. J. Bjerrum , *Nat. Mach. Intell.*, 2019, **2** , 254 —265
(5) N. Brown , M. Fiscato , M. H. Segler and A. C. Vaucher , *J. Chem. Inf. Model.*, 2019, **59** , 1096 —1108
(6) RDKit mailing list discussion from 2018: Re: [Rdkit-discuss] Any known papers on reverse engineering fingerprints into structures?

Workflow

Learning **C**ontinuous and **D**ata-**D**riven Molecular **D**escriptors by translating equivalent chemical representations*

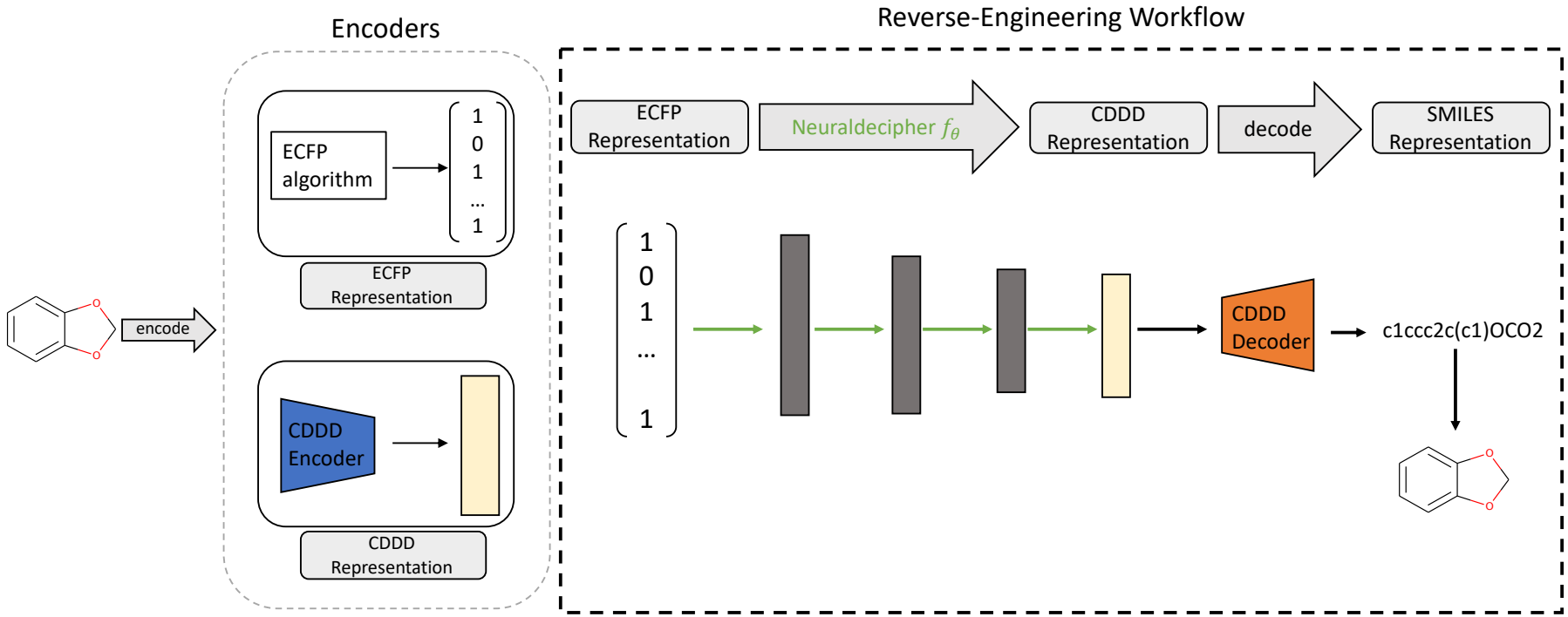


- Train on ~70M compounds.
- Model learns to extract the essence that both representations have in common
- The CDDD translation-model has learned its own representation of chemical structure and can reconstruct SMILES string with the correct syntax

*Winter *et al*. Chemical Science (2019)

Workflow

Reverse-engineering pipeline of Neuraldecipher*



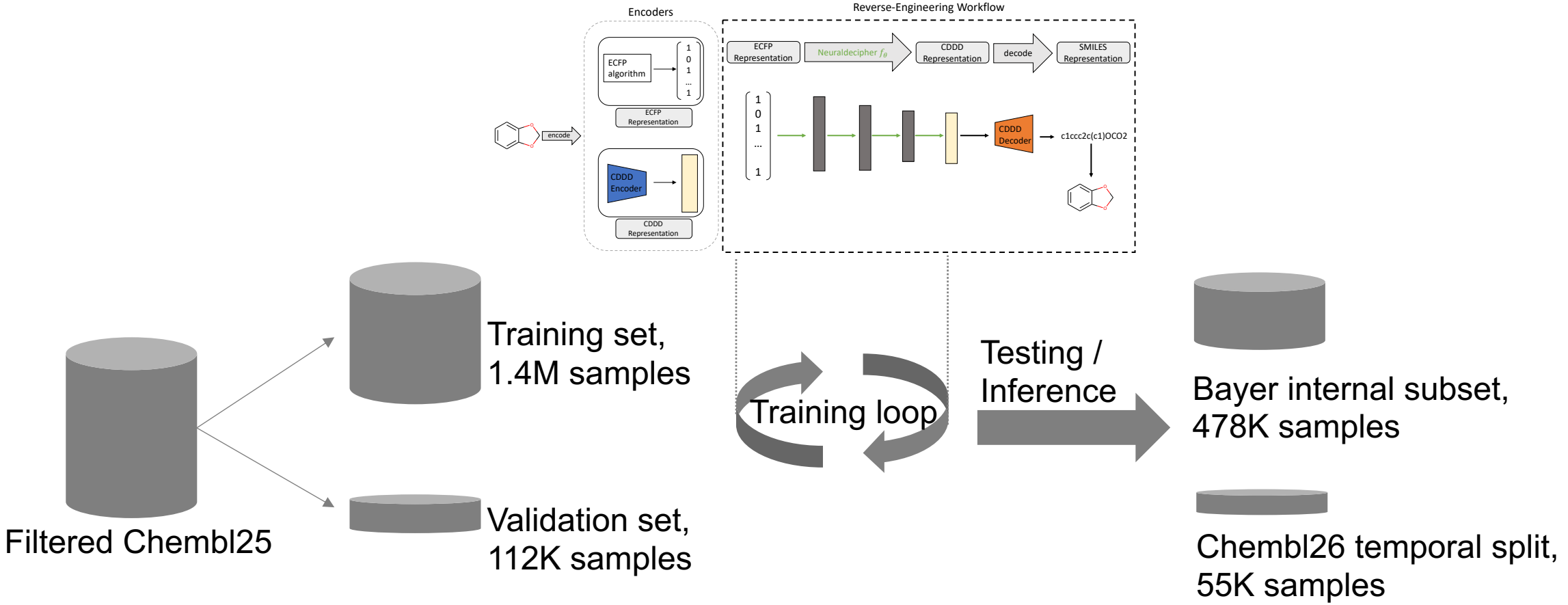
Le *et al.* Chemical Science (2020)

- $\theta^* = \arg\min_\theta \frac{1}{|B|} \sum_{b \in \mathrm{B}} l(cddd_{true,b}\ ,\ \ f_\theta(ECFP_b))$

- As the folded ECFP might contain information loss, the reconstruction accuracy depends on the fingerprint size k and bond-diameter d
- Experiments of deducing molecular structure on various settings

Workflow

Datasets for training and testing



Results

Fixing bond-diameter $d = 6$ and increasing fingerprint size k.

- Reconstruction accuracy increases with larger fingerprint size
- Reconstruction performs better if count ECFPs are used to train the Neuraldecipher
- No performance difference for cluster and random split on the internal and temporal datasets

Table 1. Results on binary $ECFP_6$

	Cluster split						Random split					
	Reconstruction [%]			Tanimoto [%]			Reconstruction [%]			Tanimoto [%]		
k	Valid.	Inter.	Temp.	Valid.	Inter.	Temp.	Valid.	Inter.	Temp.	Valid.	Inter.	Temp.
1, 024	12.14	11.32	13.34	47.08	45.31	46.84	28.70	12.11	14.14	60.64	40.30	47.60
2, 048	18.85	15.85	18.04	53.65	49.68	51.17	37.87	16.34	18.81	67.11	50.26	51.87
4, 096	32.90	25.08	28.12	63.02	57.06	59.11	57.35	25.30	28.43	79.36	57.39	59.55
8, 192	48.83	37.14	39.98	74.25	66.45	68.24	72.91	36.84	39.81	88.01	66.57	68.33
16, 384	57.85	44.64	47.38	79.80	71.86	73.46	79.79	46.22	48.86	91.30	72.96	74.34
32, 768	59.04	45.81	48.31	80.77	72.84	74.21	80.02	46.92	49.66	91.40	73.35	74.76

Le *et al.* Chemical Science (2020)

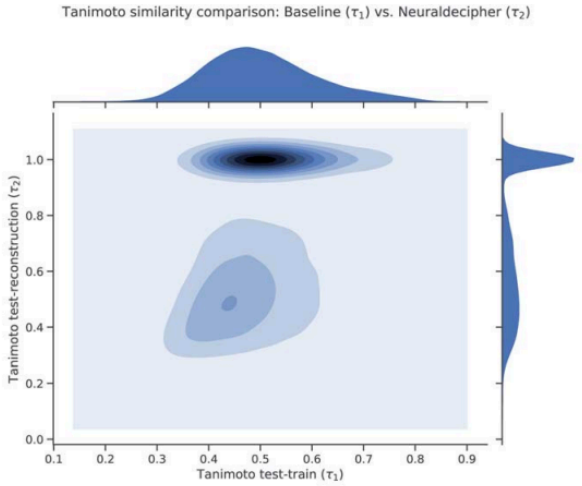
Table 2. Results on count $ECFP_6$

	Cluster split						Random split					
	Reconstruction [%]			Tanimoto [%]			Reconstruction [%]			Tanimoto [%]		
k	Valid.	Inter.	Temp.	Valid.	Inter.	Temp.	Valid.	Inter.	Temp.	Valid.	Inter.	Temp.
1, 024	22.27	16.92	19.41	61.39	57.59	59.06	38.29	17.85	20.90	71.11	58.13	59.42
2, 048	30.45	22.35	25.94	66.25	61.32	62.90	47.73	22.22	25.77	76.36	61.34	62.99
4, 096	41.02	29.98	34.61	72.58	66.43	68.52	66.61	31.73	36.22	85.98	67.61	69.59
8, 192	55.01	39.63	44.56	80.49	72.77	74.85	77.07	40.89	44.97	90.98	73.60	75.29
16, 384	62.42	46.47	50.61	84.30	76.83	78.44	80.02	46.02	49.48	92.45	76.69	78.05
32, 768	64.03	48.52	52.32	85.07	78.01	79.30	83.52	50.35	54.25	93.85	79.09	80.44

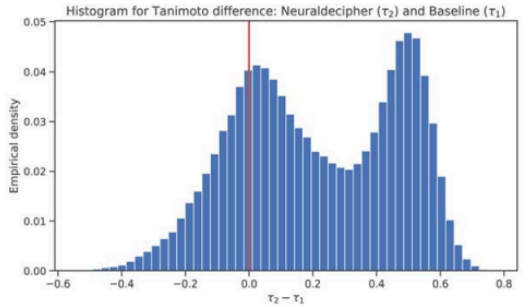
Le *et al.* Chemical Science (2020)

Comparison against a Baseline

- Baseline Model is a purely computational approach from virtual screening for the validation set of 112K samples
- For each validation sample, compute all pairwise Tanimoto similarities to a reference set, here: Training set
- “Weak” baseline because the reference set is rather small with 1.4M samples
- In “real”-life scenarios, it is likely that there is no overlap between the reference set and target set e.g. as shown in Kogej *et al.* and Besnard *et al.*



(a) Tanimoto similarity from Baseline (τ_1) vs. Neuraldecipher (τ_2).



(b) Distribution for Tanimoto similarity difference between Neuraldecipher (τ_2) and Baseline model (τ_1).

Le *et al.* Chemical Science (2020)

Conclusion

- The folded ECFP exhibits a lossy compression method to represent molecular structures that are often exchanged along collaborations to improve QSAR or ADMET models
- With the Neuraldecipher model, we were able to reconstruct molecular structures with better performance, when count-ECFPs and larger fingerprints are shared
  - Regarding count-$\mathrm{ECFP}_{6,4096}$, the trained Neuraldecipher can reconstruct 30% of Bayer internal data (478K samples) and 35% of novel compounds from ChEMBL26 (55K samples)
  - Considering that we only used publicly available data from ChEMBL25 to train the Neuraldecipher model, increasing the training data can lead to better generalization
  - Exchanging ECFPs with partners should be done with caution, as molecular (druglike) compounds can be reconstructed with our method
  - We showcast an example of inverse-problem based on ECFP representations. Are there other representations we could you as input to deduce the molecular structure?
    - NMR spectra of compounds: Pre-liminary results on NMR-ShiftDB: 18% Top-5 Reconstruction

Acknowledgement

- To all contributers in the RDKit mailing discussion thread:
  - https://www.mail-archive.com/rdkit-discuss@lists.sourceforge.net/msg07851.html
- The anonymous reviewers from Chemical Science for their suggestions to improve the paper
- Robin Winter and Djork-Arné Clevert

Thank you for your time!
Any questions?

Tuan Le
9th RDKit UGM 2020

Machine Learning Research
Djork-Arné Clevert

Anastasia Pentina
Floriane Montanari
Paula Marin Zapata
Andreas Pohlmann
Joren Retel
Marc Arne Boef
Marc Osterland
Marco Bertolini
Pedro Reis
Robin Winter
Ryan Henderson
Santiago Villalba
Van Khoa Le

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d0sc03115a

All publication charges for this article have been paid for by the Royal Society of Chemistry

Neuraldecipher – reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures†

Tuan Le, *[ab] Robin Winter, [ab] Frank Noé [b] and Djork-Arné Clevert *[a]

Paper: https://doi.org/10.1039/D0SC03115A
Open-source code on GitHub:
https://github.com/bayer-science-for-a-better-life/neuraldecipher