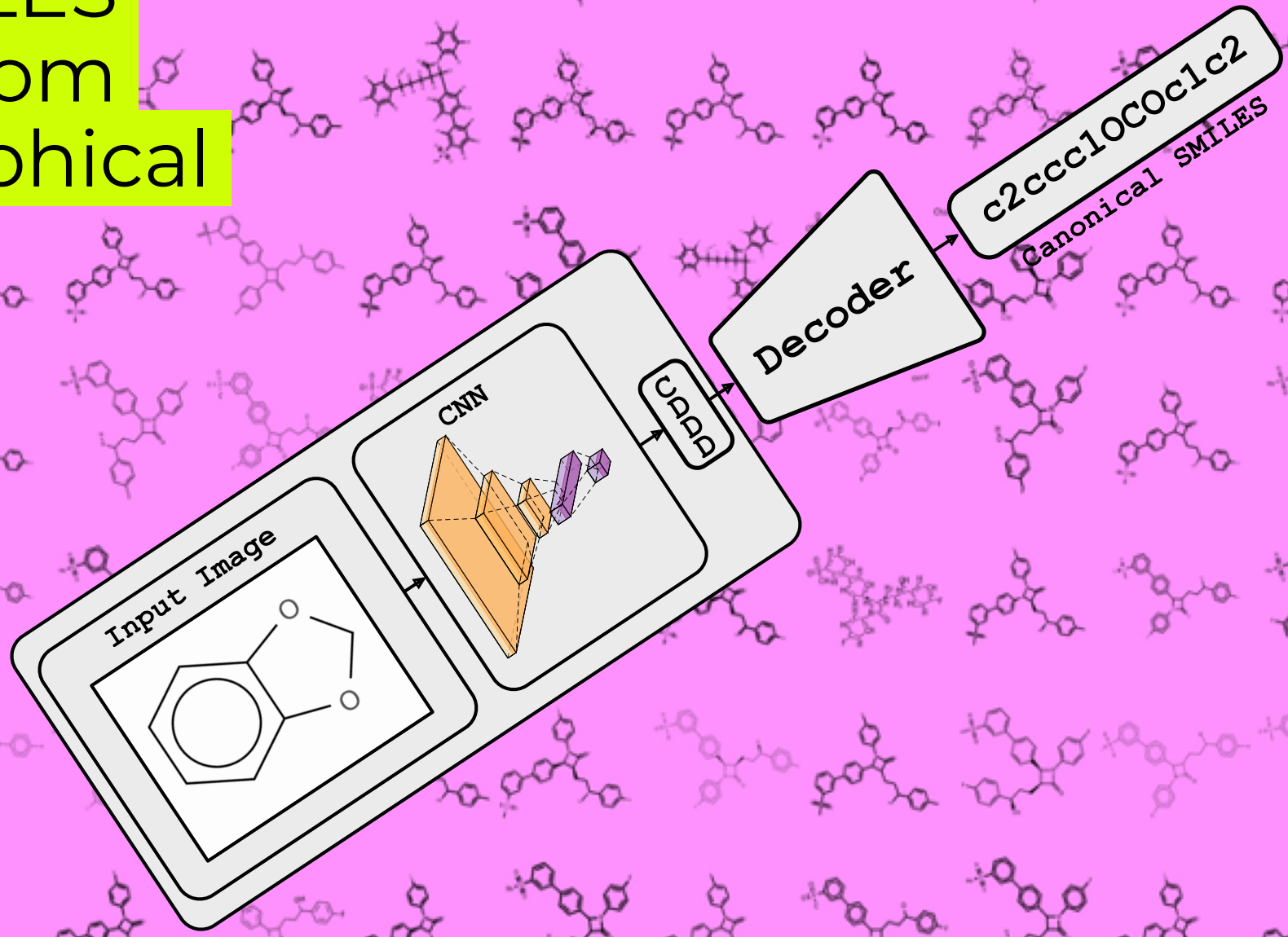
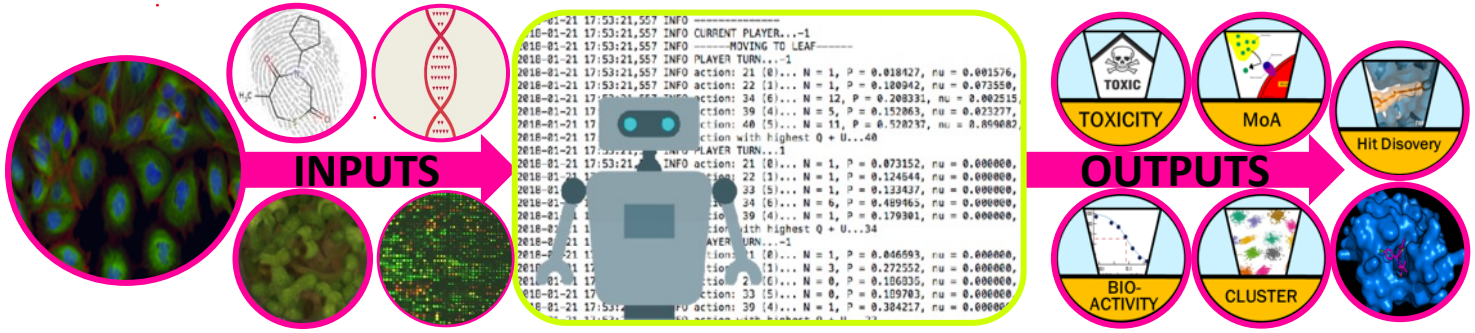
Img2Mol - SMILES Recognition from Molecular Graphical Depictions

Djork-Arné Clevert,
Tuan Le,
Robin Winter,
Floriane Montanari

RDKit UGM 2021
14.10.2021

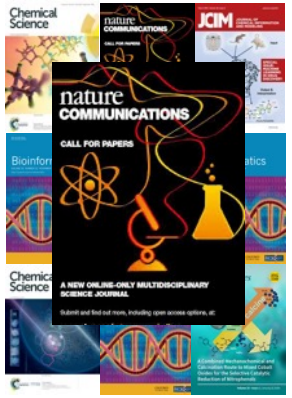


Machine Learning Research @ Bayer



MLR develops learning systems to translate data that is ‘relatively easy’ to collect (inputs) into information that is ‘expensive’ to acquire (outputs)

MLR@Bayer has a recognized reputation in the field of machine learning for drug discovery, methods developed by our group are used in academia and pharmaceutical industry



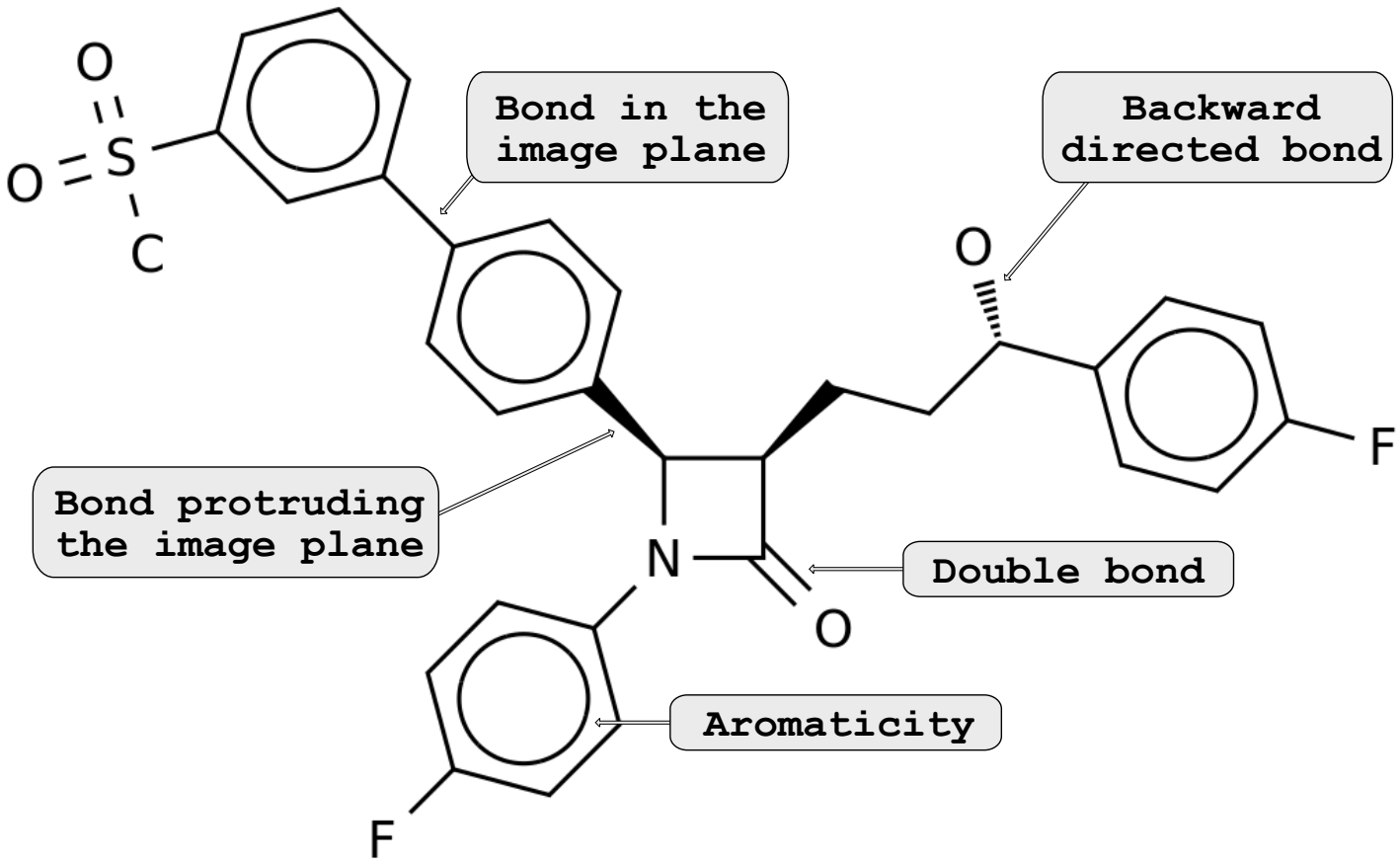
During the last years MLR have published more than 16 papers (Nature Communication, Chemical Science, ICML, ICANN, Bioinformatics, NeurIPS, Molecule)

MLR is currently involved in six IMI EU consortia, one Massachusetts Life Sciences grant and one BMBF grant

Build with love by the Machine Learning Research group at Bayer

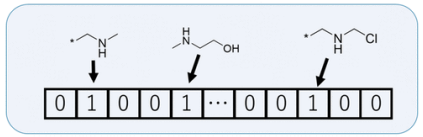
SMILES

Problem: A picture is worth a 1000 ~~words~~

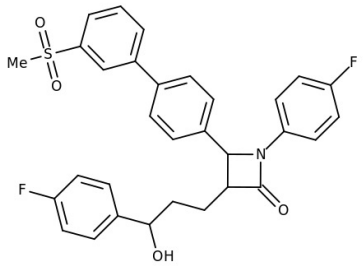


- Initial step in developing new therapeutics is the collection, analysis, and utilization of previously published data
- Most chemistry journals still do not have clear guidelines on computer-readable supplementary information
- Most drug discovery-relevant publications contain chemical in the form of pictures
- Currently more than 2000 new life science papers are published per day
- There is strong demand for tools to curate chemistry literature
- Sketching molecules by hand is time-consuming and error-prone

Inverse molecular problems



hashed ECFP vector



molecular depiction

forward

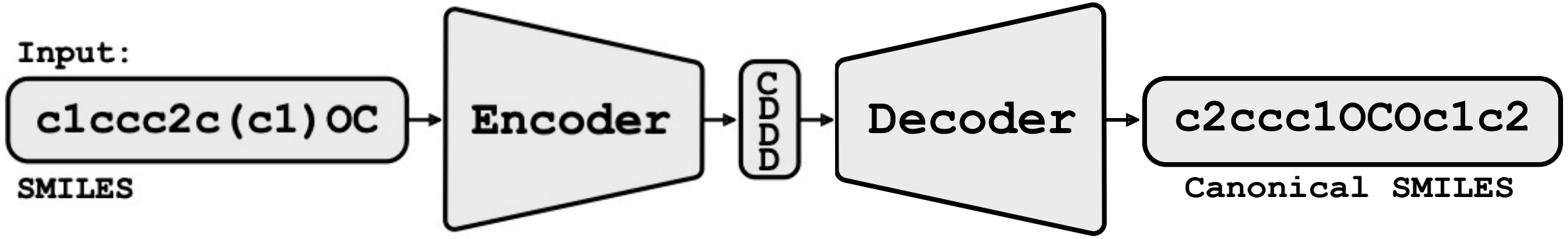
inverse

CS(=O)(=O)c1cccc(-c2ccc(C3C(CCC(O)c4ccc(F)...

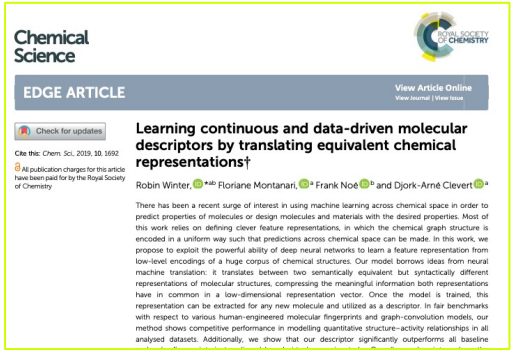
SMILES, InChI,...

- Most molecular problems have a non-linear forward model
- Most are structured inverse problems that impose additional structural constraints on the recovery task (graph isomorphism)

Reverse-engineering depictions



- Two-step approach:
  - Given a molecular depiction, predict the latent CDDD representation.
  - From the predicted CDDD, utilize the fixed decoder network to retrieve the SMILES representation (and hence the molecular structure)

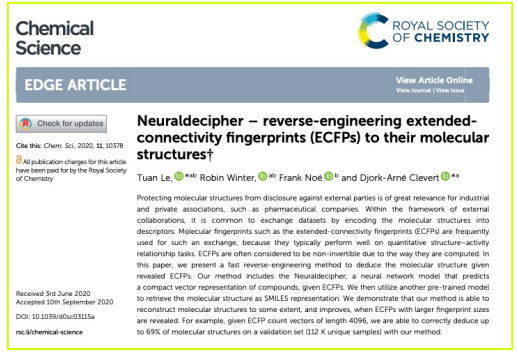

Chemical Science

EDGE ARTICLE

View Article Online

Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations†

Robin Winter, Floriane Montanari, Frank Noé and Djork-Arné Clevert


Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online

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

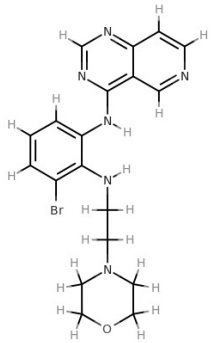
Tuan Le, Robin Winter, Frank Noé and Djork-Arné Clevert

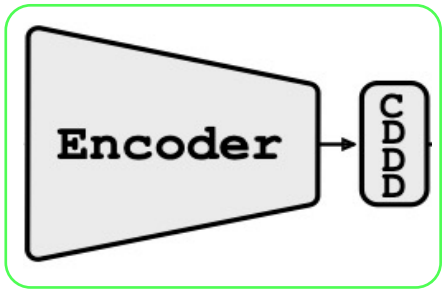
Img2Mol - workflow

Brc1cccc(Nc2ncnc3ccncc23)c1NCCN1CCOCC1

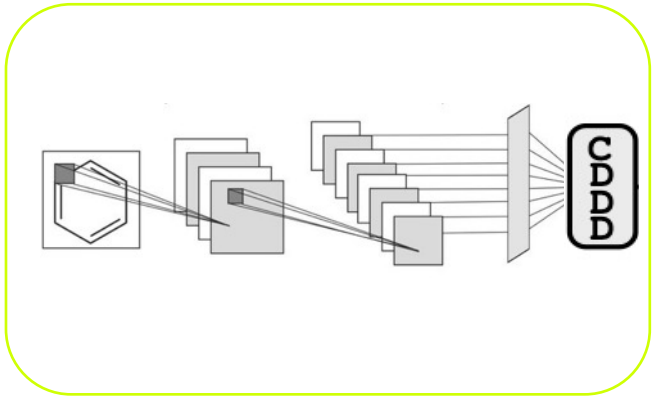
ENCODERS

encode

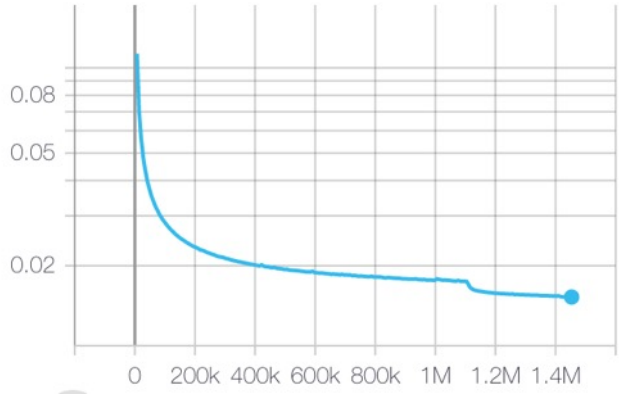


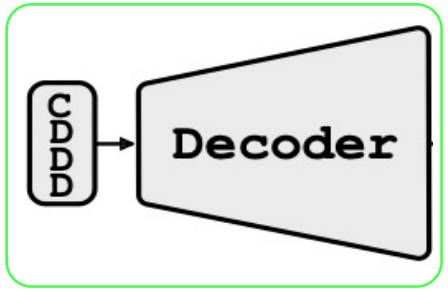


$f_\theta: [224x224x1] \rightarrow$ CDDD

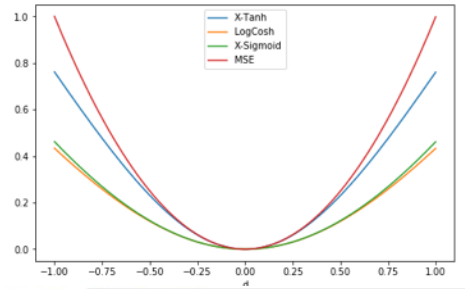


train_loss_epoch





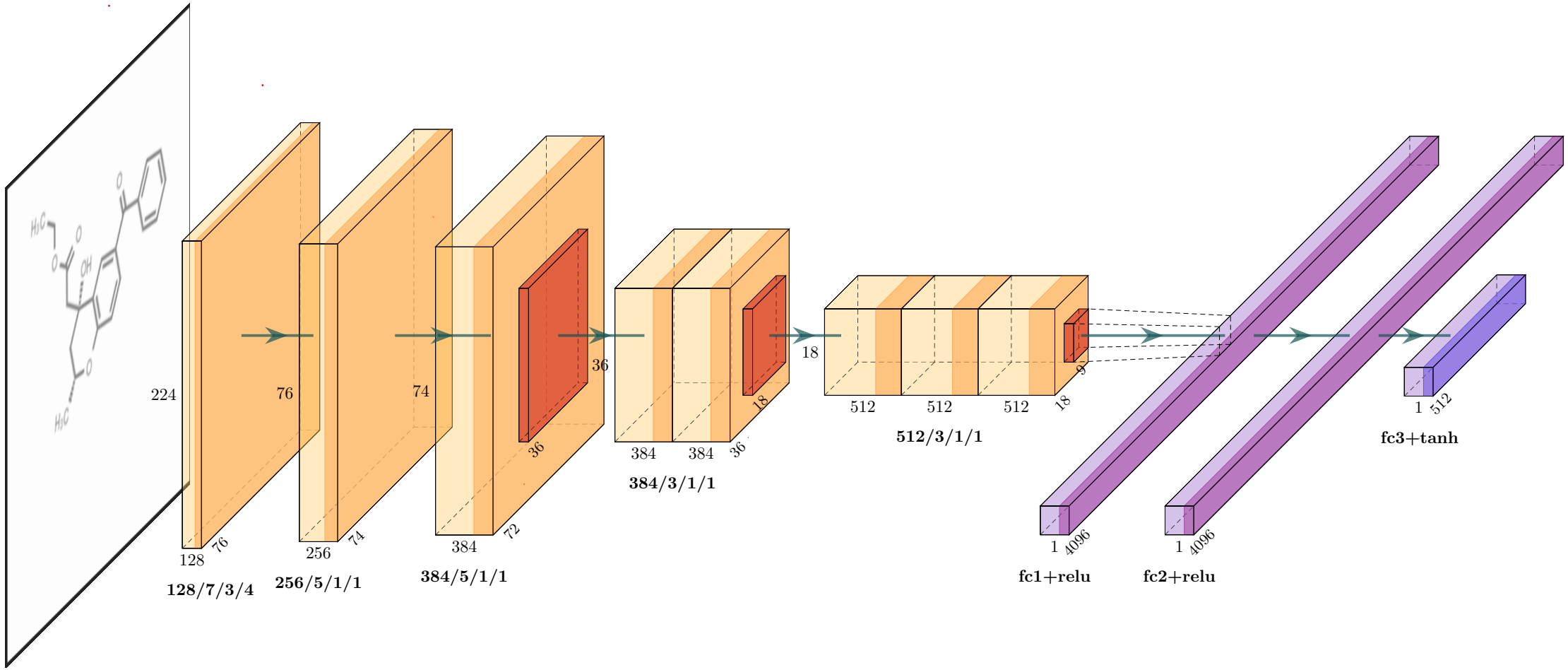
decode



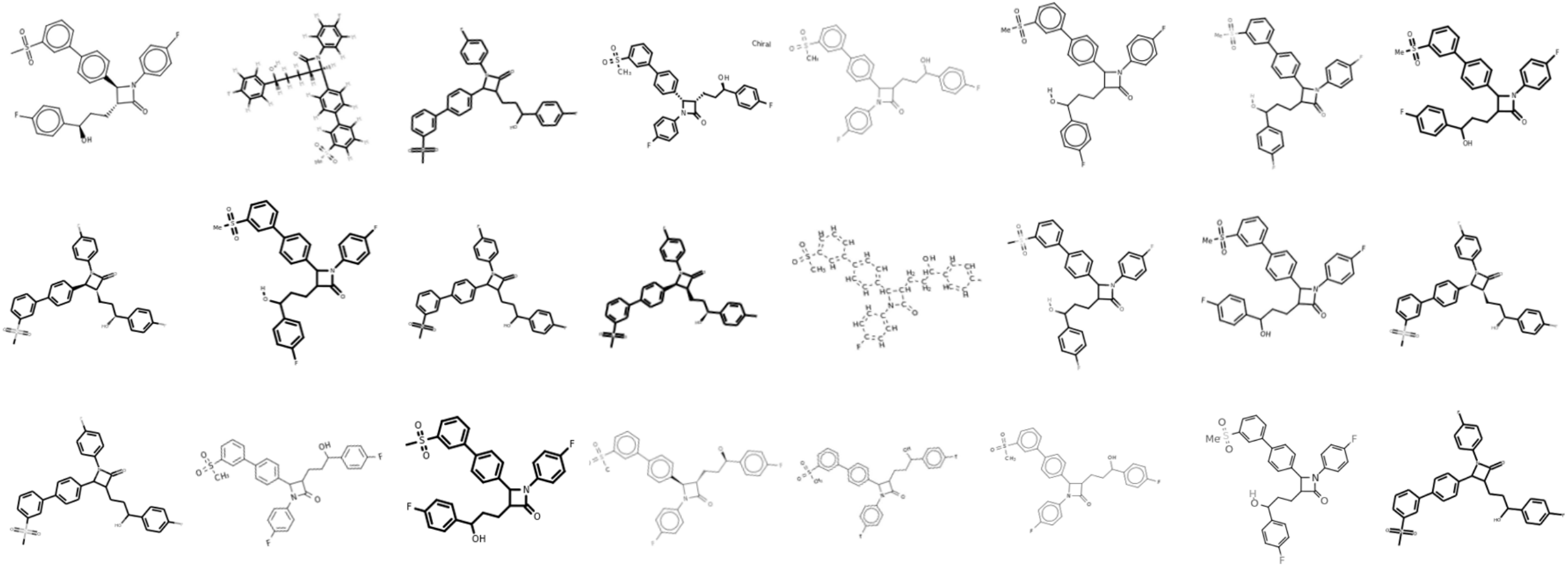

```
In [2]: def LogCosh(d):
            return np.log(np.cosh(d))

In [3]: def sigmoid(x):
            return 1/(1+np.exp(-x))
```

Brc1cccc(Nc2ncnc3ccncc23)c1NCCN1CCOCC1

Img2Mol - CNN encoder architecture



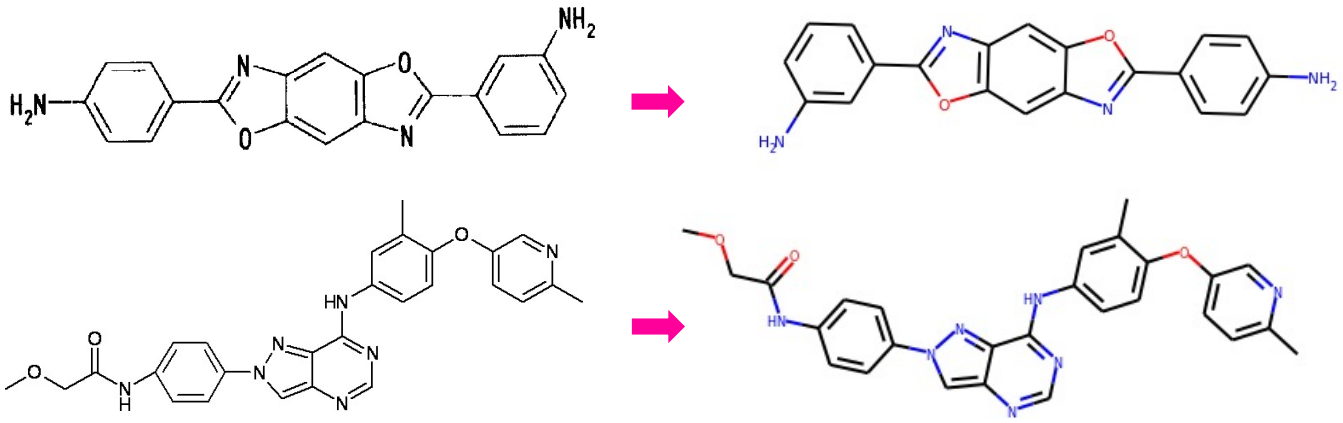
Molecular depiction generation

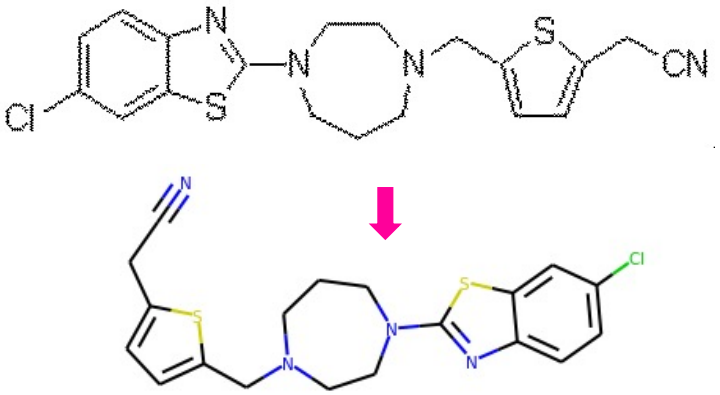


Results benchmark data

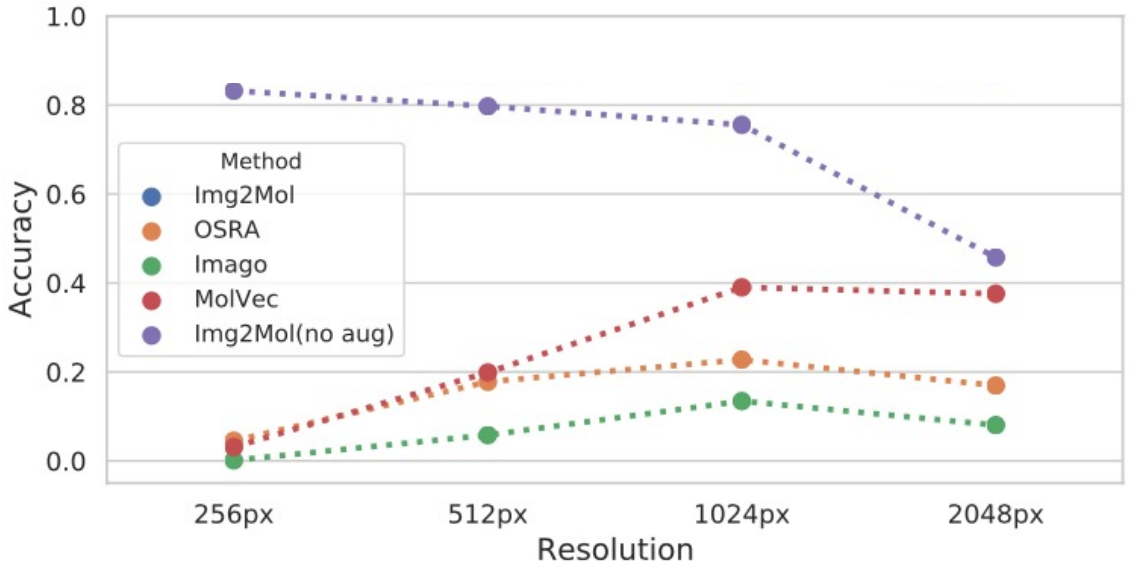
	Img2Mol		MolVec 0.9.8		Imago 2.0		OSRA 2.1	
	Accuracy	Tanimoto	Accuracy	Tanimoto	Accuracy	Tanimoto	Accuracy	Tanimoto
Benchmark								
Img2Mol	**88.25**	**95.27**	2.59	13.03	0.02	4.74	2.59	13.03
STAKER	**64.33**	**83.76**	5.32	31.78	0.07	5.06	5.23	26.98
USPTO	**42.29**	**73.07**	30.68	65.50	5.07	7.28	6.37	44.21
UoB	**78.18**	**88.51**	75.01	86.88	5.12	7.19	70.89	85.27
CLEF	**48.84**	**78.04**	44.48	76.61	26.72	41.29	17.04	58.84
JPO	45.14	**69.43**	**49.48**	66.46	23.18	37.47	33.04	49.62

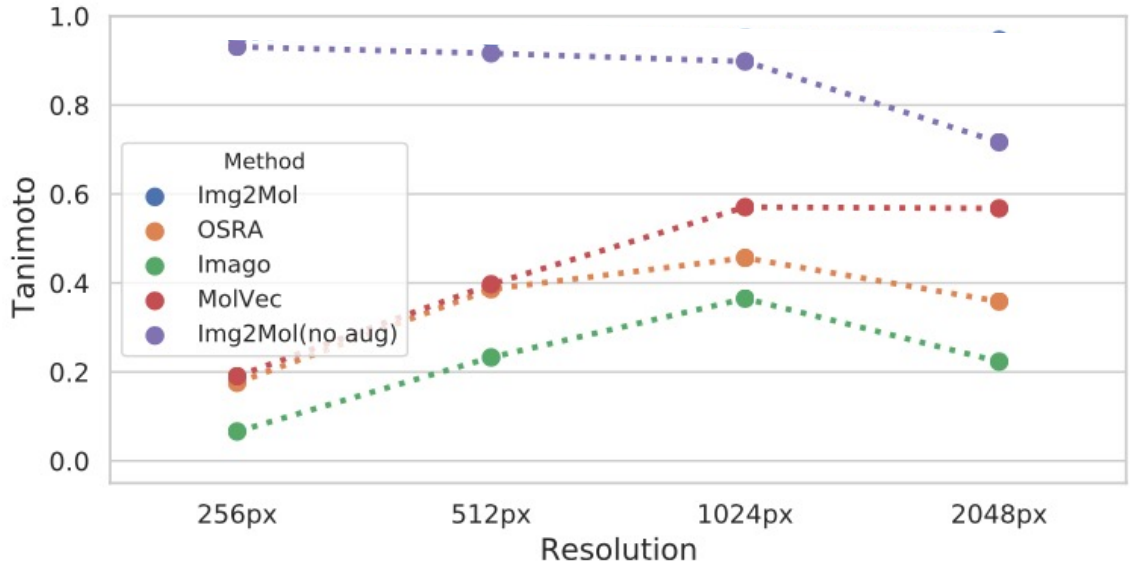
2,6-(3,4'-ジアミノジフェニル)ベンゾ[1,2-d:4,5-d']ビスオキサゾール





Results: Influence of the image resolution





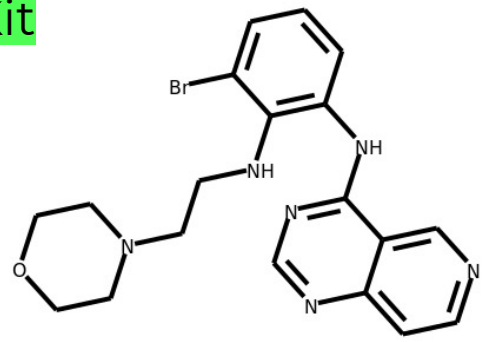
- Reconstruction accuracy decreases with resolution for Img2Mol (no aug)
  - First Img2Mol (no aug) wasn't scaling invariant
  - Final Img2Mol is scaling invariant
- Reconstruction accuracy increases with resolution for competitors

Results: Influence of the depiction library

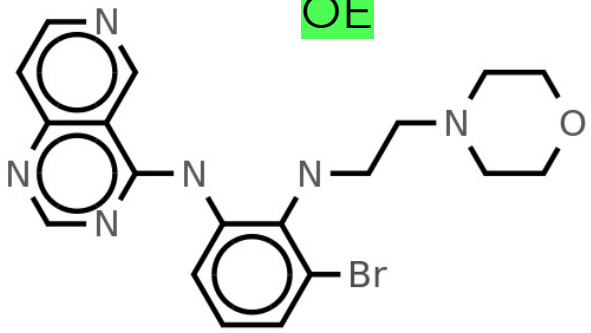
	Img2Mol		MolVec 0.9.8		Imago 2.0		OSRA 2.1	
	Accuracy	Tanimoto	Accuracy	Tanimoto	Accuracy	Tanimoto	Accuracy	Tanimoto
Depiction								
RDKit	**93.4**±0.2	**97.4**±0.1	3.7±0.3	24.7±0.1	0.3±0.1	17.9±0.3	4.4±0.4	17.5±0.5
OE	**89.5**±0.2	**95.8**±0.1	33.4±0.4	57.4±0.3	12.3±0.2	32.0±0.2	26.3±0.4	50.0±0.4
Indigo	**79.0**±0.3	**91.5**±0.1	22.2±0.5	37.0±0.5	4.2±0.2	19.7±0.2	22.6±0.2	41.0±0.2

- Reconstruction accuracy stable across depiction libraries for Img2Mol
- Competitors struggle to reconstruct RDKit depictions

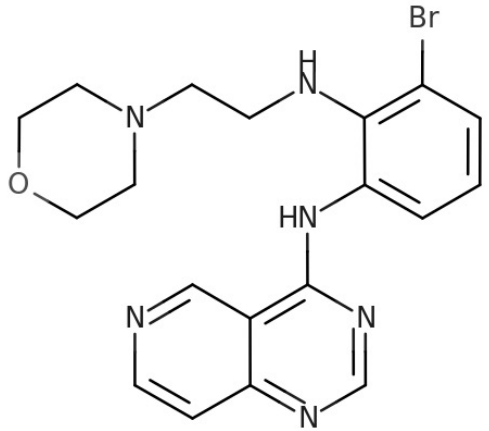
RDKit



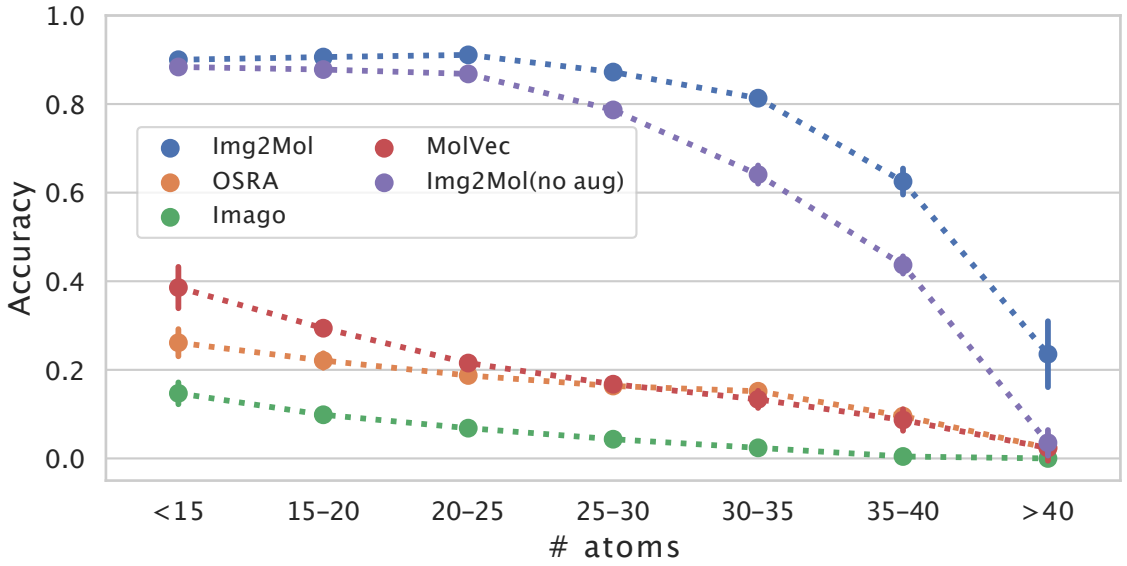
OE

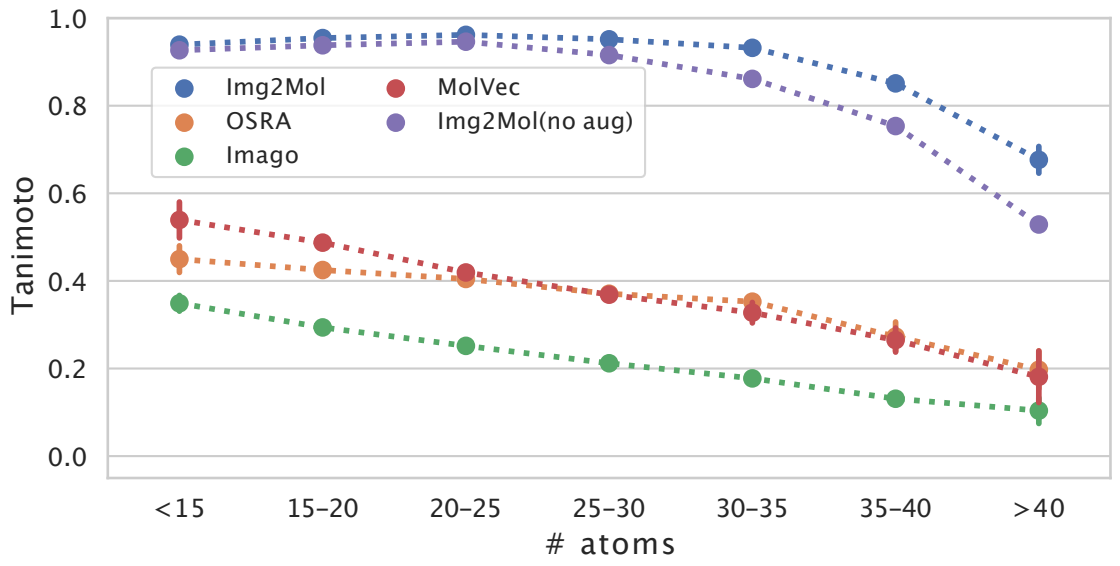


Indigo



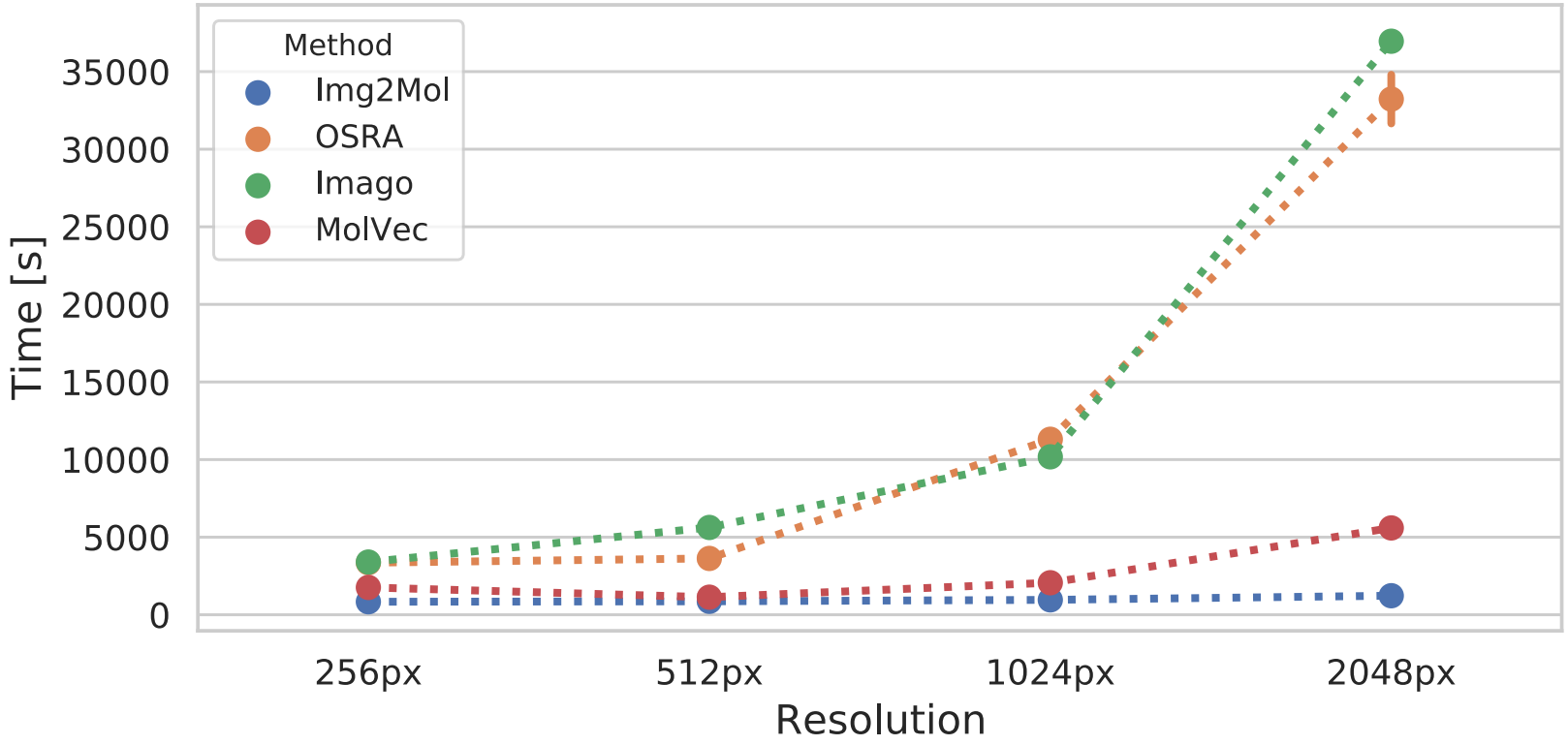
Results: Influence of the molecular size



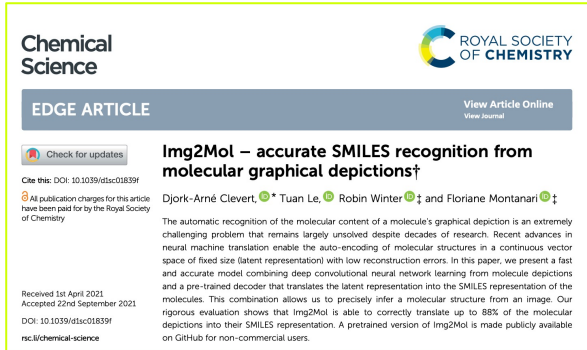


- Reconstruction accuracy and Tanimoto similarity decrease with molecular size
- The tasks become more difficult as more atoms and bonds have to be correctly identified
- Curriculum learning applied to improve predictive performance for large moleclues

Computational cost - wall-clock time



- Wall-clock time in [s] for processing 5,000 images as a function of image resolution.


Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d1sc01839f

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

Received 1st April 2021
Accepted 22nd September 2021

DOI: 10.1039/d1sc01839f

rsc.li/chemical-science

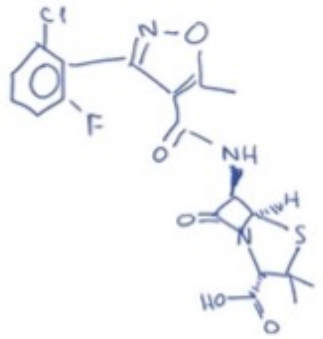
Img2Mol – accurate SMILES recognition from molecular graphical depictions†

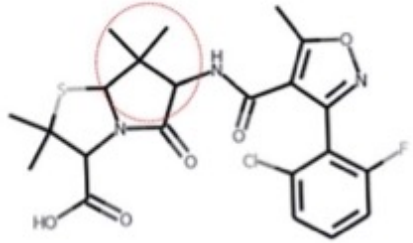
Djork-Arné Clevert, * Tuan Le, Robin Winter ‡ and Floriane Montanari ‡

The automatic recognition of the molecular content of a molecule's graphical depiction is an extremely challenging problem that remains largely unsolved despite decades of research. Recent advances in neural machine translation enable the auto-encoding of molecular structures in a continuous vector space of fixed size (latent representation) with low reconstruction errors. In this paper, we present a fast and accurate model combining deep convolutional neural network learning from molecule depictions and a pre-trained decoder that translates the latent representation into the SMILES representation of the molecules. This combination allows us to precisely infer a molecular structure from an image. Our rigorous evaluation shows that Img2Mol is able to correctly translate up to 88% of the molecular depictions into their SMILES representation. A pretrained version of Img2Mol is made publicly available on GitHub for non-commercial users.

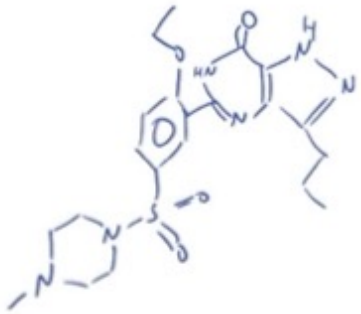
Does Img2Mol generalizes to the wild?

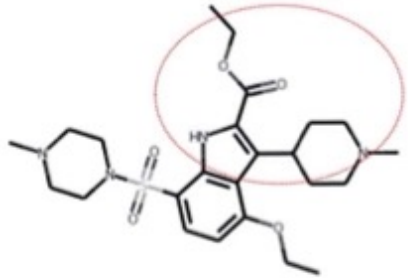
a)



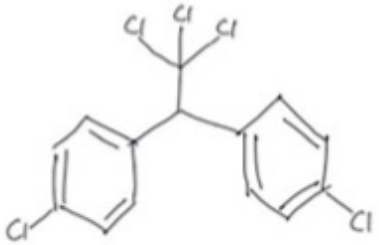


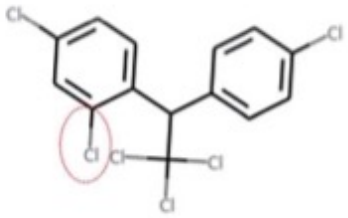
b)



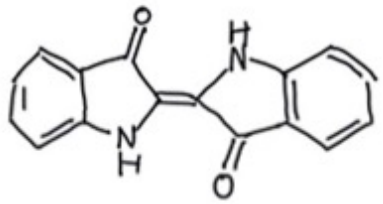


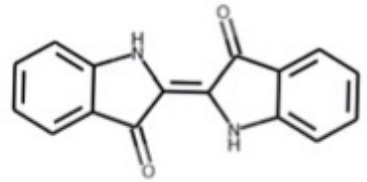
c)



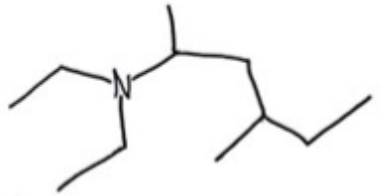


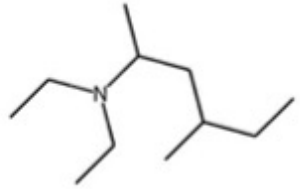
d)



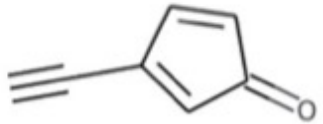


e)





f)



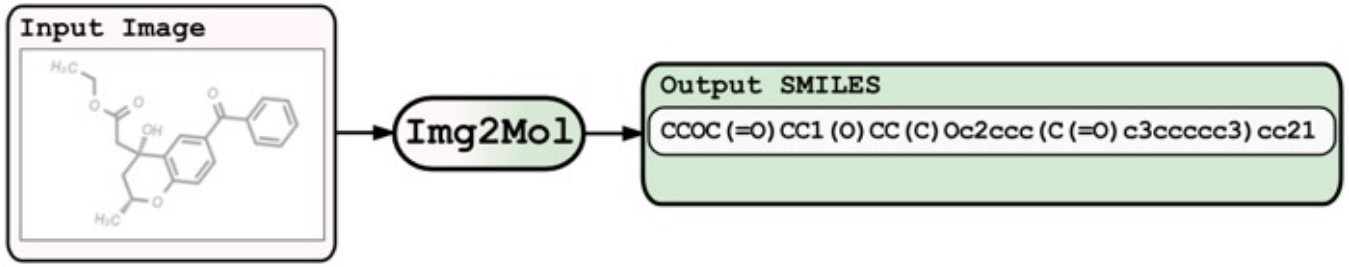
demo

Img2Mol - Demo Version

Img2Mol is a fast algorithm for automatic recognition of the molecular content of a molecule's graphical depiction. To read more about how the model works, click here.

This proof of concept web application converts images containing 2D structural representations of a molecule to the corresponding SMILES representation. In its current state not all special cases are covered, i.e. try to avoid uploading images that containt text artifacts due to cropping.

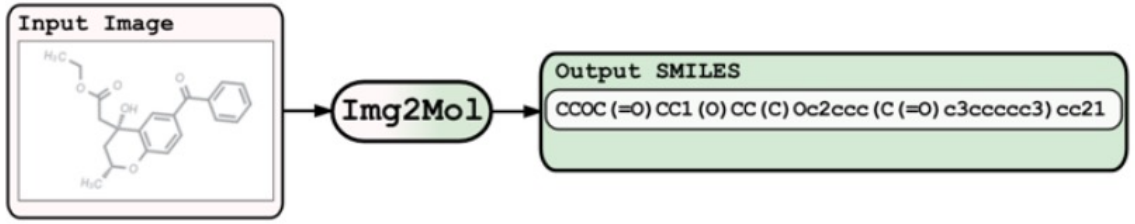
Please note that the current deployment of this web application is temporary and we will provide a more permanent instance in the near future.



Choose an image to process

Drag and drop file here
Limit 10MB per file
Browse files

Get SMILES

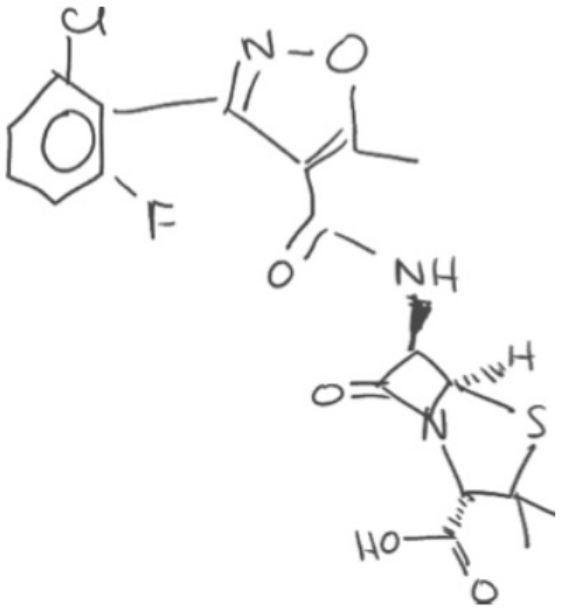


Choose an image to process

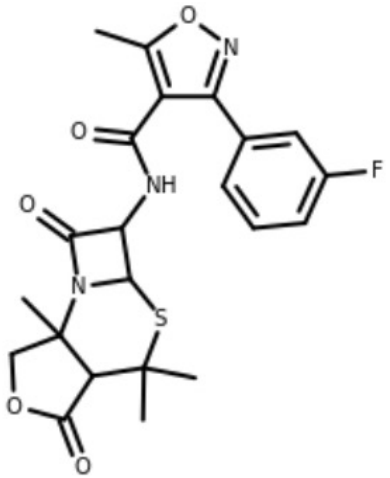
Drag and drop file here
Limit 200MB per file
Browse files

Bildschirmfoto 2021-05-06 um 23.23.28.png 41.6KB

Get SMILES



Input image



Molecule from predicted SMILES

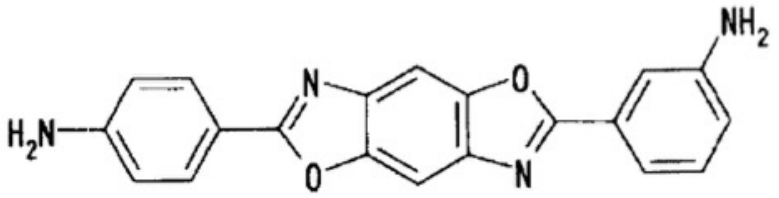
Predicted SMILES

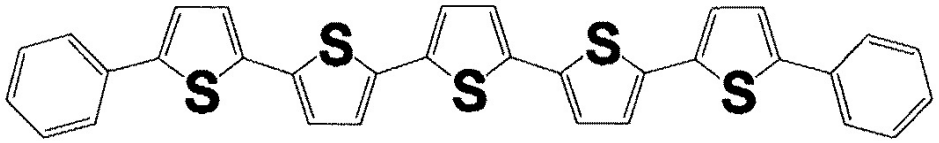
Cc1onc(-c2cccc(F)c2)c1C(=O)NC1C(=O)N2C1SC(C)(C)C1C(=O)OCC12C

Error analysis

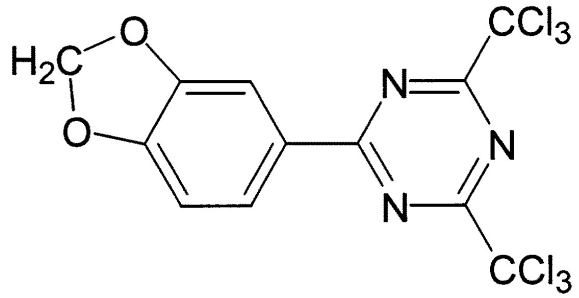
	Img2Mol	
	Accuracy	Tanimoto
Benchmark		
Img2Mol	**88.25**	**95.27**
STAKER	**64.33**	**83.76**
USPTO	**42.29**	**73.07**
UoB	**78.18**	**88.51**
CLEF	**48.84**	**78.04**
JPO	45.14	**69.43**

2,6-(3,4'-ジアミノジフェニル)ベンゾ[1,2-d:4,5-d']ビスオキサゾール





P5T



(16)

一般式(3)

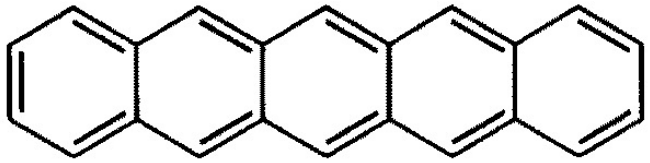
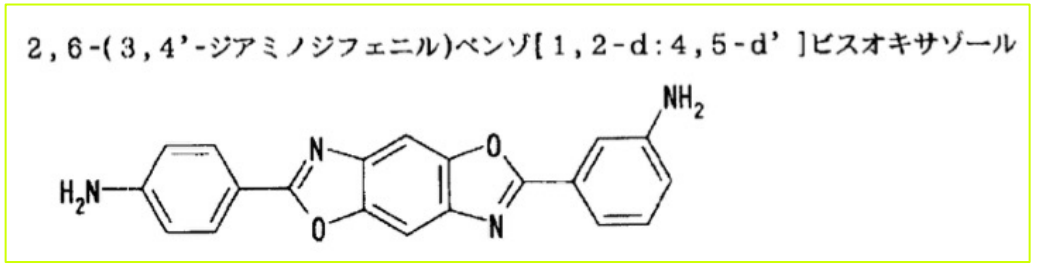
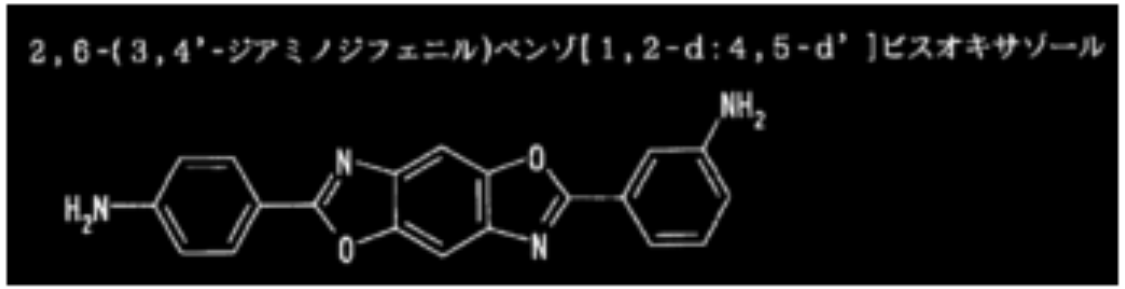


Image segmentation



input molecule



convex hull molecule

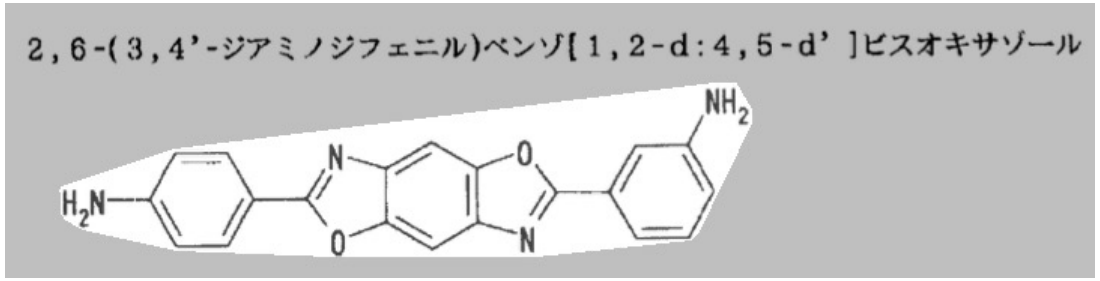
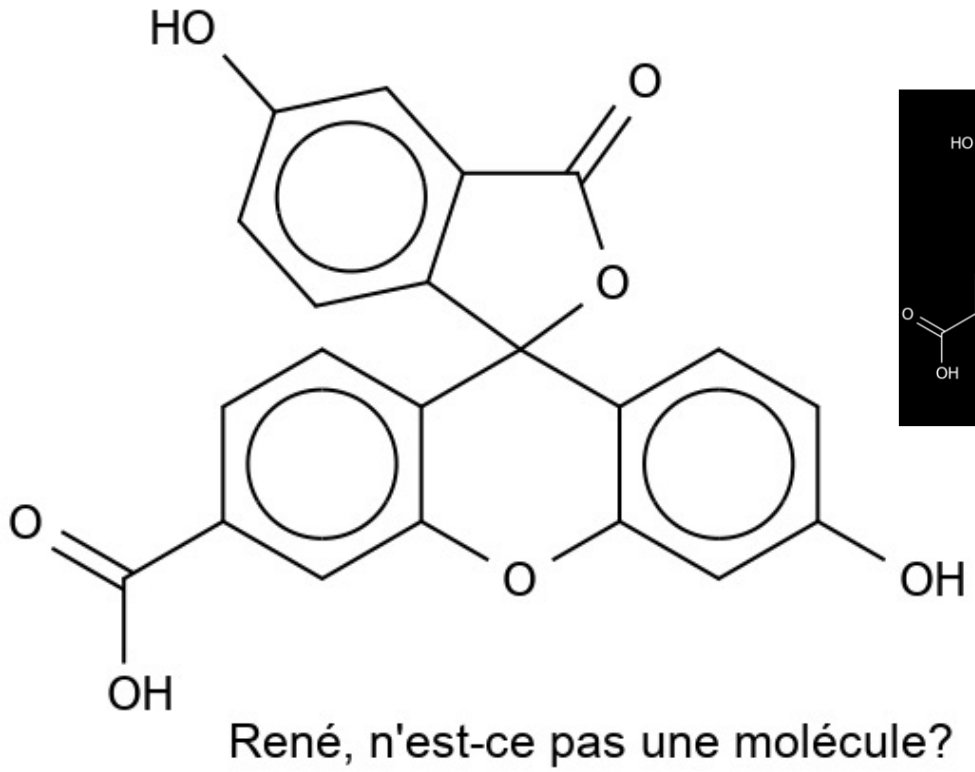
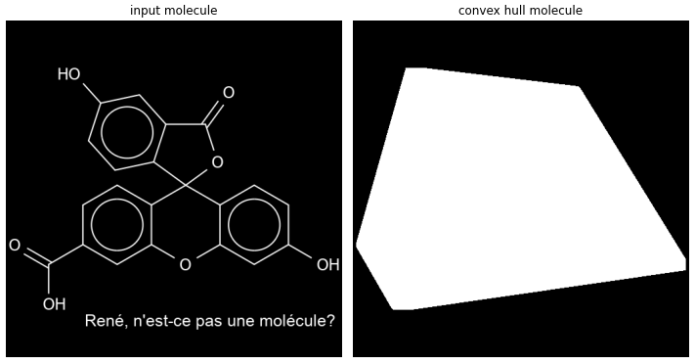
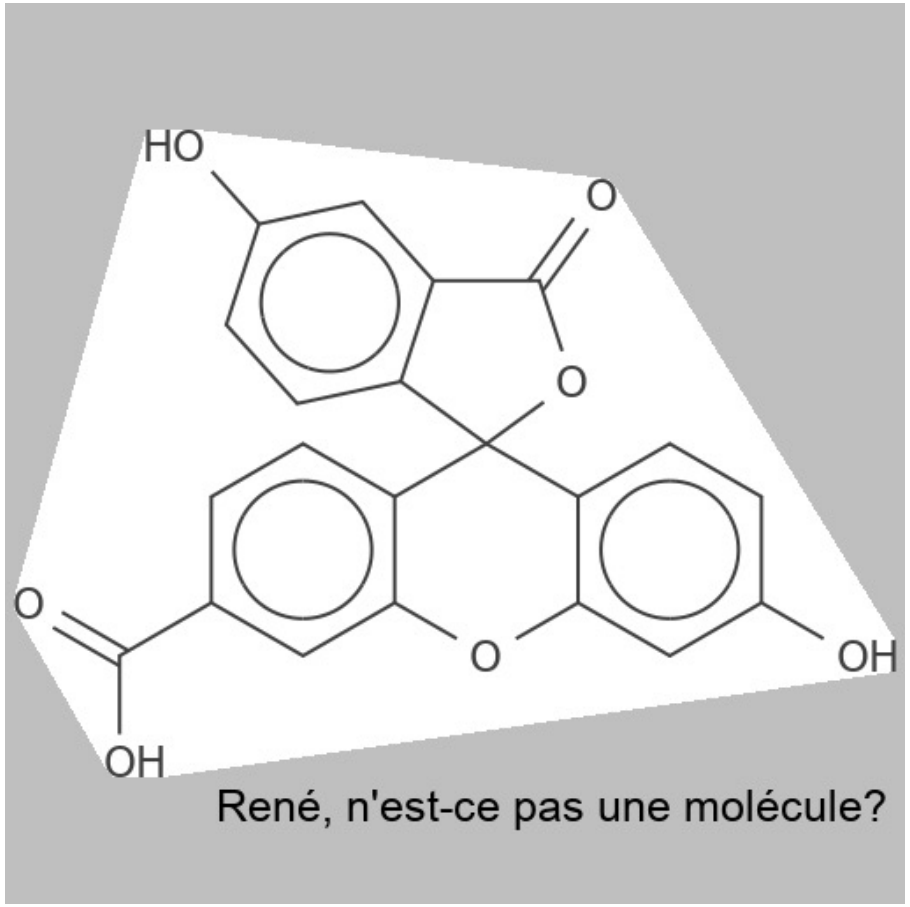


Image segmentation II







Conclusion / Outlook

- Unsupervised learned representation of molecules can be used to tackle inverse problems
- Img2Mol can quickly and accurately reconstruct molecular structures from images with up to 88% reconstruction accuracy
- Next steps: Img2Mol will be extended for structure segmentation and to identify stereochemistry and Markush structures
- Img2Mol code & model available at:
  - https://github.com/bayer-science-for- a-better-life/Img2Mol

MLR@Bayer

digital pathology

RJ Winter

J Retel

natural language processing

A Poehlmann

Tuan Le

P Reis

federated learning

A Pentina

DA Clevert

active learning

M Osterland

machine learning research

R Henderson

M Bertolini

molecular de-novo design

F Montanari

PM Zapata

antibody design

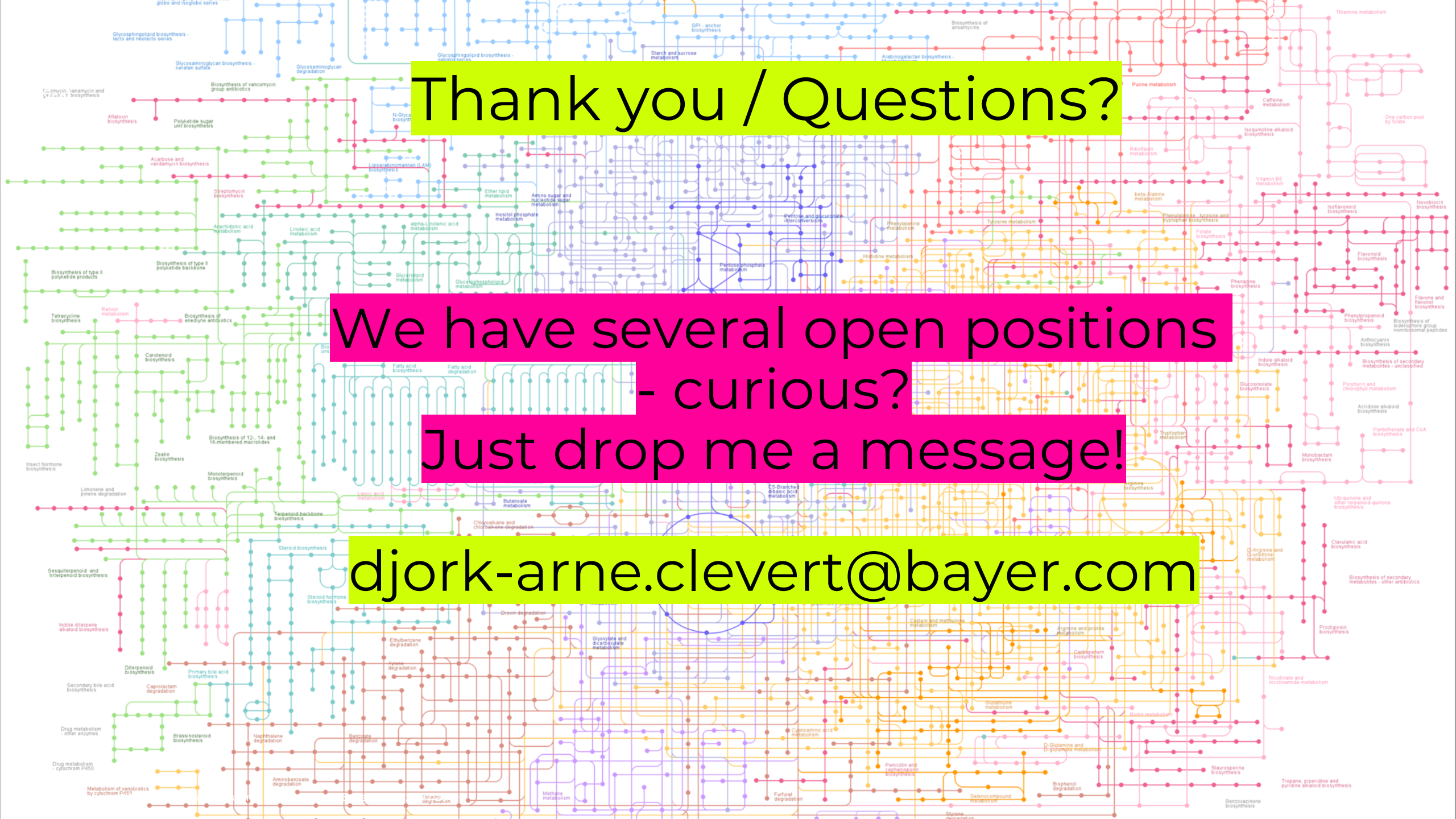
S Villalba

conformer embeddings

explainable AI

high-content imaging

bioactivity modelling


Thank you / Questions?
We have several open positions
- curious?
Just drop me a message!
djork-arne.clevert@bayer.com