Creating a merged dataset and investigation of correlations in the data with SOM and VAE models

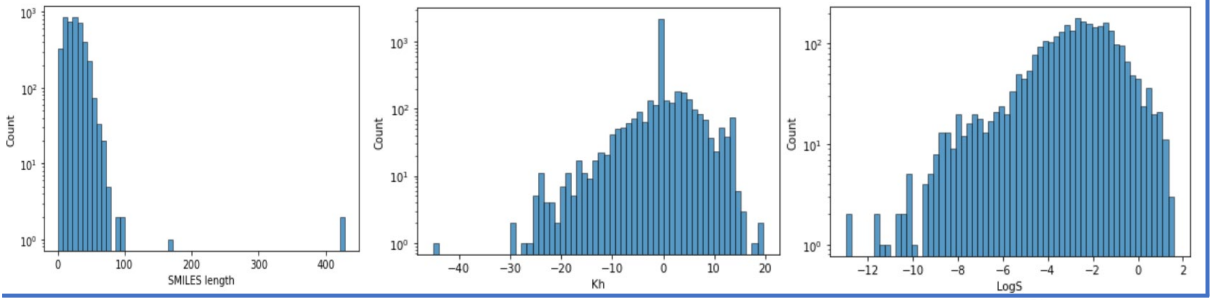
Maximilian Hoffmann, Freie Universität Berlin, hoffmam98@zedat.fu-berlin.de

Skills 4 Scientists

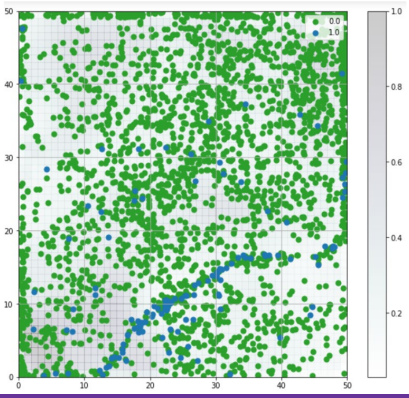
Sponsored by: optibrium, dotmatics knowledge solutions

	SMILES	InChI_Key	LogS	LogKh	MW	MLOGP
0	[CH]1CCCCC1.[CH]1CCCCC1.[CH]1CCCCC1.[OH]	DSQCYFOORCMTBH-UHFFFAOYSA-N	-5.590	0.0	266.26	NaN
1	[O-]N=C(C(=NO)c1ccccc1)c1ccccc1	JJZONEUCDUQVGR-UHFFFAOYSA-M	-5.899	0.0	239.27	2.873

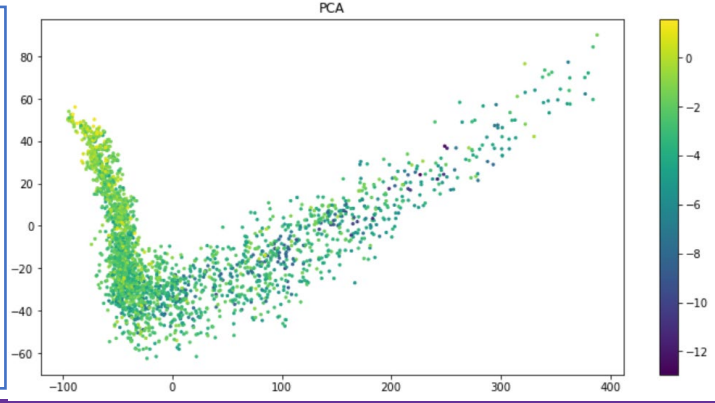
The dataset, containing 4237 species



The use of layerwise relevance propagation in a VAE-Classifier model could provide insight into which substructures or atoms of a molecule are relevant for its properties.



U-Matrix from SOM on Lipinsky descriptors and Henry's law constant as features;
blue: logS > 0,
green: logS < 0



Principal Component Analysis on the variational autoencoder's (VAE) latent vectors from the dataset's compounds;
color labels logS

AI³ Science Network+

PSDS PHYSICAL SCIENCES DATA-SCIENCE SERVICE

RSC INTEREST GROUP CHEMICAL INFORMATION AND COMPUTER APPLICATIONS GROUP