## Comp Chen Applied ML for Chemistry 1

Deepchen

description => Deep learning models for drug discovery, quantum chemistry and Use schences.

Applications: chemistry, Brology, Material-science, life-science, dry-discovery

(https://deepcnes.io/) Documentation: https://deepchem.readfridocs.io/en/latest

Source Code: https://github.com/deepchem/deepchem

Instau requirements: Joblib, nungy, pandas, scikit-learn, scipy, redkit

Relevant article: http://arxiv.org/ass/1703.00564

Predictive modeling by machine learning [Cheminformatice]

Quantitative Structure Activity Property Relationship (CLSAR /QSPR)

MolfomMolBlock

MolfromMolfile

MolfromPDBBlock

(85) eXtreme Gradient Boosting. https://github.com/dmlc/xgboost, Accessed: 2017-10-

the second secon

The set of the self-trail and the segment of the

The same of the sa

Della Dade I relieved from the formation a medical

MolfromFASTA

MolfronHELM

Contract the Desirate property of the the the

glycine

phenylalante

histidine

Cysteine

(15) 3 3 3 server Secret Interest in the second in the sec

well thought the Joint year

RDKit => A useful python library wen dealing with chemical data

pip install rakits

from rdkitte import chem

from rdkis. Chen import Draw

from rdkit. Chem import Descriptors

from rdkie. Chun. Draw import Python Console

from rdkit. Chem imports AllChem

from raking import. DataStructus

Import numpy as no

Emporting structures.

or mot ing = Draw. MotTo Image (md)

The chem. Database.

Straure Database.

www. chemisto. org | Chemistry | Name\_to\_structe/

Computing proporties 2

If you have me molecule, men you can complète voirious interesting things

now = Descriptors. Molut (mol)

Looping over list of structures 3

Smiles-list = ['', ', ...]

mod litt = [7]

mol-lise = []

for smiles in smiles\_list: mol = Chen. Molfrom Smiles (smiles)

mol-lise · append (mol) ing = Draw. Mols 10 Grid Image (mol-list, mols Perhow = len (mol-list))

MolfonInchi MoltomPABfil Mol Formol 2 Block MoupomPNGFire Molfrom Mod 2 File MolfromTPLBbouk Molfron prasting MoufromTPLFile Mol from Sequera MoltromXYZBlo Molfonsmarts

MolfromXYZFire Molton Smiles

Developing Neural Neural Neural Force Freeds for Azettoons Emphere Oxido Theofaces are Different Oxidown Levels.
Therface at Miffern Acidoson Lends
Developing Atomición ML force Fields for HEA  pymon, atomic sumuation environment (ase), high performence conjunity (hope)
Searching for partern Sub-structure scarcin (4)
Partern = chem. Moltrom Smiles ('s') suprime elevent present
pattern = Chem. Molfrom Smiles ('c (=0)0') carbonyl group present
parten = Chem. Molfrom Smiles ('c. c (N)(')
partern = Chem. Molton Snarts ('[r]')  partern = Chem. Molton Snarts ('[r5]')  for mol in mol_line:  Search for 9 hing structure  structure  structure
prine (mot. Hags Substructe/Match (patrers)).
aumber of
Fingerprints and Tanimoto Similarly (5)  fp = All Chem. BetMorgan Fingerprint As Bit Vector (glycine, nBites= 1024)  Data Structs. Convert to Numpy Army (fp, fp-arr)  A long list of zeros and 1 (can be converted to a numpy among)  A long list of zeros and 1 (can be converted to a numpy among)  There of the converted to the substructure is there or not.
fp = All Chem. BerMorgan Fingerprint As Bir Vect (glycine, nBirs= 1024)
DataStructs. Converte to Numpy Amay (fp, fp-arr)
DataStructs. Converte Numpy Array (fr, fp-arr)  A long Ust of zeros and I (can be converted to a numpy amay)  The presence of I specifies whether a certain subgroup is more or not.
The presence of I specifies une to see me nonzero element indice
np. nonzero (fp_am).
prints = [ cglycine, x, bi) for x in the fp. Geron Bits ()]
Draws. Vaco Morgan Bits (prints, mals for how = 71 (got - L.
Buting Clusteing [rakit.ML. Cluster. Buting] -> Cluster by Tanimoto
Similaning
Tanimoto Similany. Memod to rank how similar compounds are based on common figerprint bits.
common figerprint bits.
$T(A,B) = \frac{N_c}{N_A + N_B - N_c}$
MA = "On" feature in Structure A
Mg = "on features in Structure B
1) - 41 V Commercia books structure A and B
The second secon
(1 mil-terry) as ) = water telem ( stall-terry) as sure through a suith near

Molecular Fingerprints: Clashing & Clustering Molewlar tingerprints are 9 type of molecular descriptors man encode molecular features/fragments of a molecule in the form of a binary digit (0 or 1). So, we can use molecular fringerprints to perform computations. A bit is of or 1 if a certain fragment is found in a molecule structure. " I was a report and the nil hallower also well a ward a contary weeks Structure of Aspira 1 (St. Marie Land ) Secretary Equation of Secretary Secretary · Search for Simlar Moleury 1 Predictive Models (QSAR \$
QSPR) molecular structure friggerprines: But set by fragments, Structural features
The bit is ON for exter functional group, aromatic ving and composeyling the bit is ON for exter functional group, aromatic ving and composeyling the bit is ON for exter functional group, aromatic ving and composeyling the bits is on for external group, aromatic ving and composeyling the bits is on for external group. functional group. The ON bits represents the structural features present palate property 6

Application of fugurphints:

Strawe, substructure, and similarity search for virtual ligard screening from compound detabase.

>> Machine learning - cheartifative Structure-Actually Relationship (QSAR) and Quantitative Structure - Roperty Relationship (QSPR)"

the the property of the

something outputte

time and the applies the contract and and

Some common fingerpunts available in RDKits.

-> Molecular Access System Keys or MACCS-Keys

-> Avalon Singerprint

-> Atom - pair frigerpint Topological - Torsions frigerpint

Morgan fingerprint or arcular fingerprint

-> ROKis Fingerprint

As molecular fingerpaint gives us a unique en coding of a molecule. It's a representation of a molecule as a series of zeos and ones (0's and 1's). It's basically a one-hot encoding of a molecule. It is done in different ways. There are different algorithms and different types of fingerprints. Morgan or Circular fungerprints is a commonly used. It a specific functional group is present or some pattern is present, more gets encoded in # bit vector as a 1.

A method to represent a molecule as a series of 0s au 1.s

Incodes information on

- 7 Functional groups present (Circular Fls)
- -> Molecular Shapes (arom-pair FPs)
- > Molecular features

Useful for

-> Virtual screening-> used in pharmacoutical world to select for a specific structure

-> Mapping chemical space -> do a dimensionality reduction on the feature vectors (bit reasons) >> Mapping chemical space >> you can Just use this bit-wise fugerpint as properly themselves are

>> Property prediction >> you can Just use this bit-wise fugerpint as properly themselves are

do your NN, random forest algorithm, etc.

The same with the same of the same

for the life of the second

of support topics of the industry of

interpret interpret of the property of

Linitations

-> Difficult to return to original structure

-> Interpretability

Fingerprints can be different leight eg 1024 bits, 2048 bits

BIE Clashing For most fingerparts, the reight of your fingerprint might be 1024 bit vector. The algorithm will generate your vector of such length. Sometimes if your is not long enough, multiple substructures get enough in me some position in your bit vector (bit classing). When mutiple smartmete are represented at me same position in the vector (bit collisions). Me optimally wants to basically have each bit (each position) in met vector, be this on unique substructure— in mat way we can test differences between them and the might contribute differently to unotever you've trying to predict Computing Fingerprints | Avoiding bit clashing | Using in a Statistical model

186908615/ Teol of : 1600

FP Generations.
Bit clashing per 125 (127) was senguler son of the property to clustering for property prediction

moll = Enem. Molfrom Smile (' -... ) mola = Chen. Molfrom Sniles ('a...)

Draw. Mols To Gridhage ([mols, mols], subtry Size = (400, 400))

FpI = All Chem. Get Morgan Finger Print As Bit Vect (mol 1, 3, nBits = 4096, bit Info = Info I, 2)

p2 = All Chem. Get Morgan Finger Print As Bit Vect (mol 2, 3, nBits = 4096, bit Info = info 2, 2)

generating a morgan fingerpino as a bit vector of 4096 bits.

This can be converted to a numpy array of 1's and 0's.

We can also see union functional groups get encoded as 1's and union are encoded as 0's. Molecules with the same "on" index, it the fringerphine very means try have the same similar substructure.

arr = no. zeros ((0,))

arr = np. zeros ((0,)) Chem. Dara Stones. Convert To Nunpy Amay (fpl, am)

an [0:100]

Draw. Draw Morgan Bis (mol 1, 2136, enfol, use svar = Time) Draw Draw Morgan Bit (mol2, 2136, info2, use SVG = True)

Whith some of these figer paths embedding, Maybe we don't have enough bits for all of the structural diversity to be represented. One way to manage this is by increasing the number of bits that we have.

Check how many bits you have in use, across all of our fingerprints It looks the our maximum number of occupied bits plateaux at 163, if we want maximal diversity in our fingerprints win no overlapping.

Fingre print - Based Chrestering i Sci. 1999, 39, 747—750. Buting, D. J. chen info. comput. DOI: 10.1021/ci9803381 circles of tragery and faringed of Visit John Brown John Bridge Line Sent I solling the Water and the sent of the sent of (cost (act) = wilder the Laken then I) spend bind of the ward 1 / I for some const A CONTRACTOR OF THE and the state of t Continue and the continue of t Experience of the formation of the participant of the property The state of the s Markey La State Brown of The Square Landstein Training Sales and all and the first the first that the first will be the first that the state of the first that the first the first that the first the first that t excessive from the sound of the survey of the sound of th W. P. W. SENSE WAY - MANS (may 19th) when he will all of wine of the stand of the first of The following to get a state of the party of the state of the contract of Construction was a strict of the confidence of the to the property of the transfer and the property of the property of the property of the same of the sa Miles and the property of the property of the property of the state of the s is the manual coice iniques to remain morning and the resident of the same of