Classification of Adverse Drug Reactions

Term Project

Julian Krumm | Juraj Trappl

Table of Contents

Table of Contents

General Information / Adverse Drug Reactions

- ADRs are harmful or unpleasant reactions to a medication or other therapeutic intervention that are based on a variety of factors:

1. Patient-related factors:
   age, gender, weight, genetics,

2. Route of administration:
   intravenous vs oral administration etc.

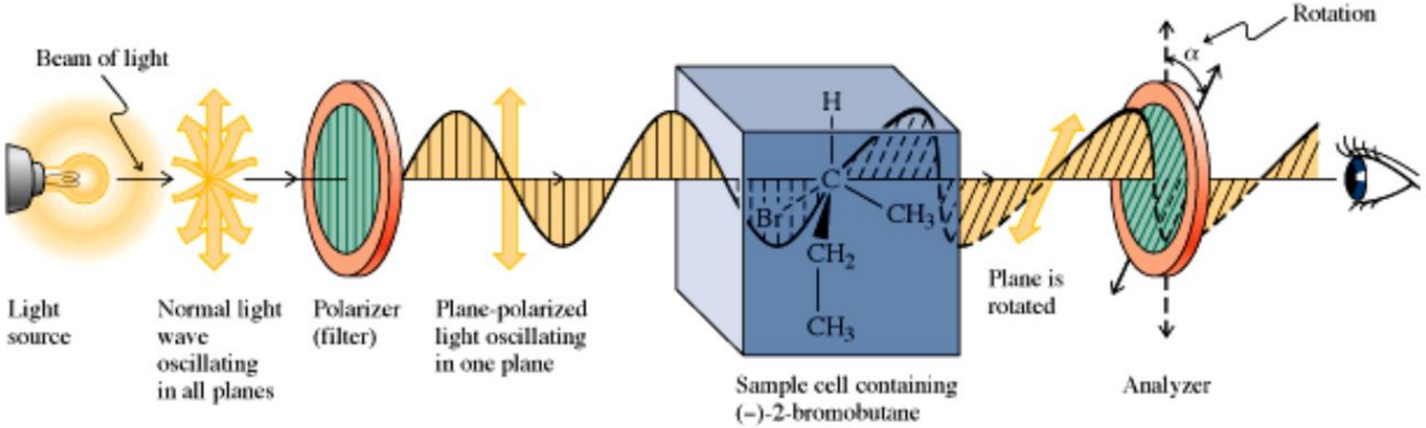
3. Environmental factors:
   exposure to toxins, pollutants, or other chemicals

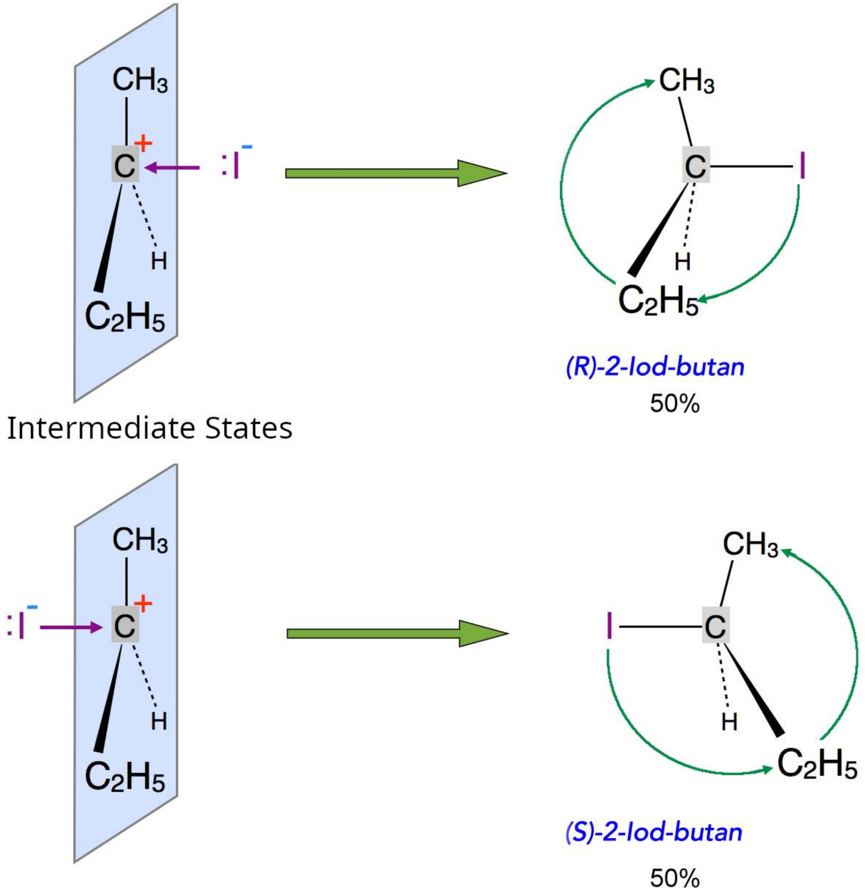
4. Drug-related factors:
   drug dosage, drug interactions, duration of use, **stereochemistry**

- Carbon that is bound to 4 different atoms is considered as chirality center

- (R)/(S) forms of compounds are called enatiomers

> Different chemical properties



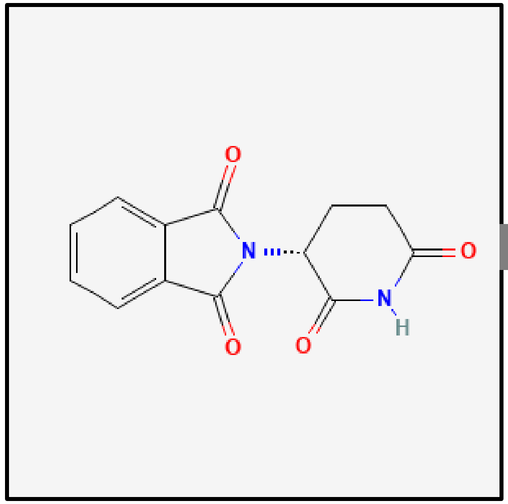


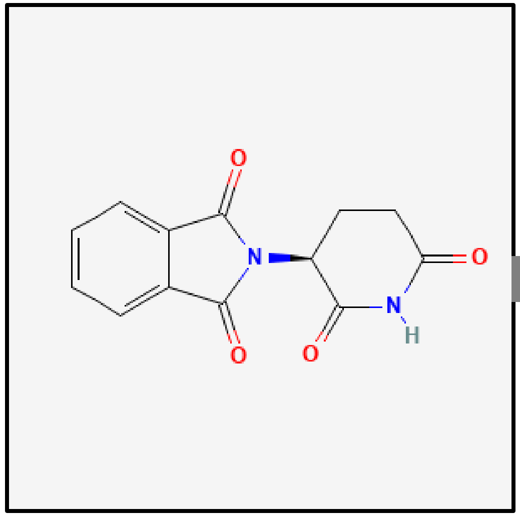
Racemat

General Information / Stereochemistry

The Thalidomide Scandal:

- Medical compound developed for treating anxiety trouble sleeping, tension and morning sickness
- Considered as safe for pregnant women even though not clinically tested





(R)-Thalidomide

50:50
Racemat

(S)-Thalidomide

General Information / SMILES

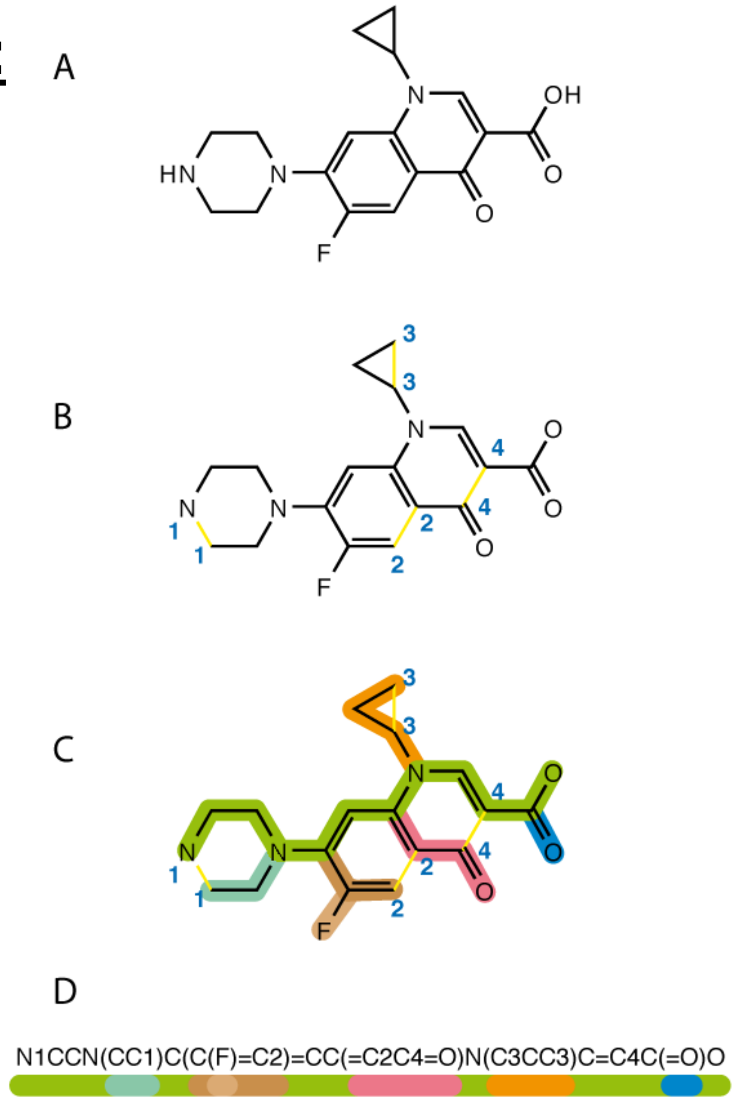
<u>Simplified molecular-input line-entry system (SMILES):</u>

<u>(+)-Thalidomide:</u>

C1CC(=O)NC(=O)C1N2C(=O)C3=CC=CC=C3C2=O

<u>(-)-Thalidomide:</u>

C1CC(=O)NC(=O)C1N2C(=O)C3=CC=CC=C3C2=O



General Information / SMILES

Simplified molecular-input line-entry system (SMILES):

(+)-Thalidomide:

C1CC(=O)NC(=O)C1N2C(=O)C3=CC=CC=C3C2=O

C1CC(=O)NC(=O)[C@@H]1N2C(=O)C3=CC=CC=C3C2=O

(-)-Thalidomide:

C1CC(=O)NC(=O)C1N2C(=O)C3=CC=CC=C3C2=O

C1CC(=O)NC(=O)[C@H]1N2C(=O)C3=CC=CC=C3C2=O

> Isomeric SMILES

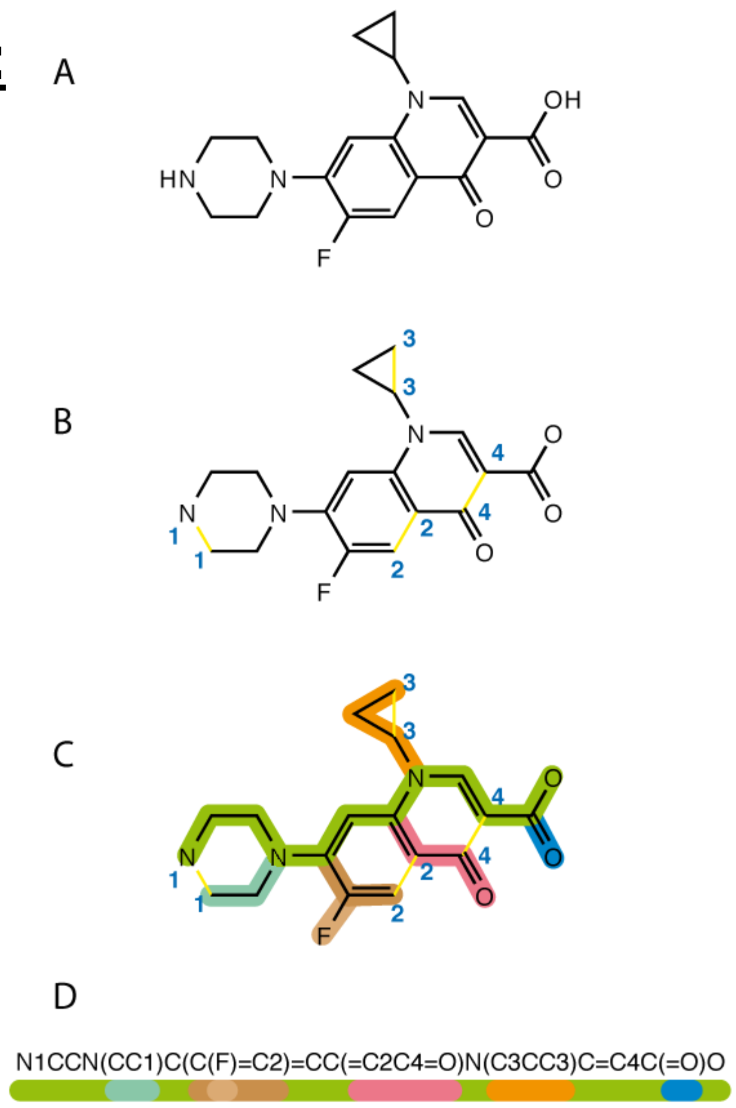
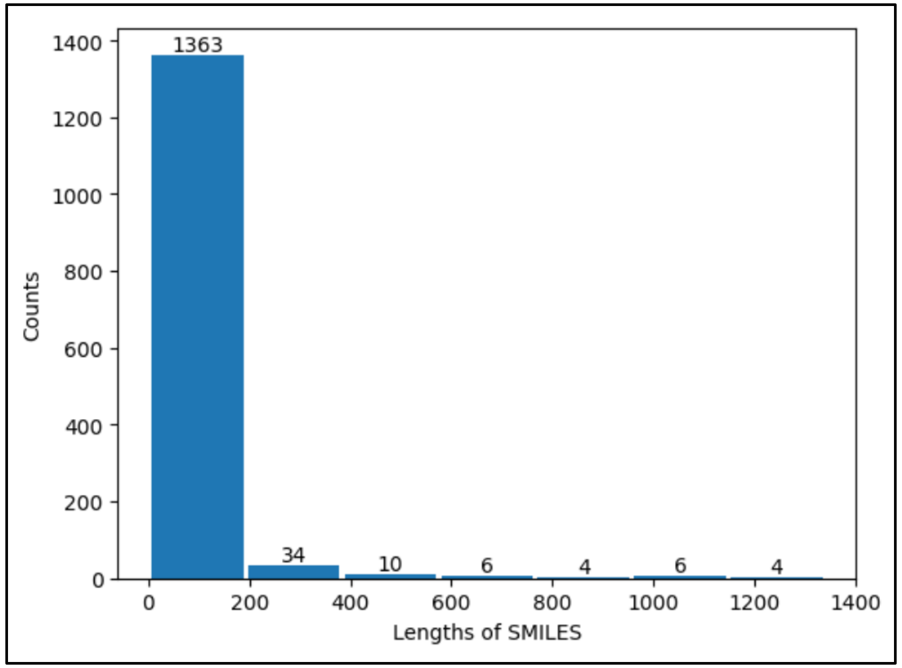


Table of Contents

Data / SIDER data set

- SIDER contains information on marketed medicines and their recorded adverse drug reactions
- More comprehensive data set available via DeepChem:

>1427 medical compounds with 27 distinct binary labeled ADRs



smiles	Hepatobiliary disorders	Metabolism and nutrition disorders	Product issues	Eye disorders	Investigations	Musculoskeletal and connective tissue disorders	Gastrointestinal disorders
C(CNCCNCCNCCN)N	1	1	0	0	1	1	1
CC(C)(C)C1=CC(=C(C=C1NC(=O)C2=CNC3=CC=CC=C3C2=...	0	1	0	0	1	1	1
CC[C@]12CC(=C)[C@H]3[C@H]([C@@H]1CC[C@]2(C#C)O...	0	1	0	1	1	0	1
CCC12CC(=C)C3C(C1CC[C@]2(C#C)O)CCC4=CC(=O)CCC34	1	1	0	1	1	1	1
C1C(C2=CC=CC=C2N(C3=CC=CC=C31)C(=O)N)O	1	1	0	1	1	1	1

Table of Contents

Aims & Problems

Aim: train the best possible classifier to predict ADRs

Problems:

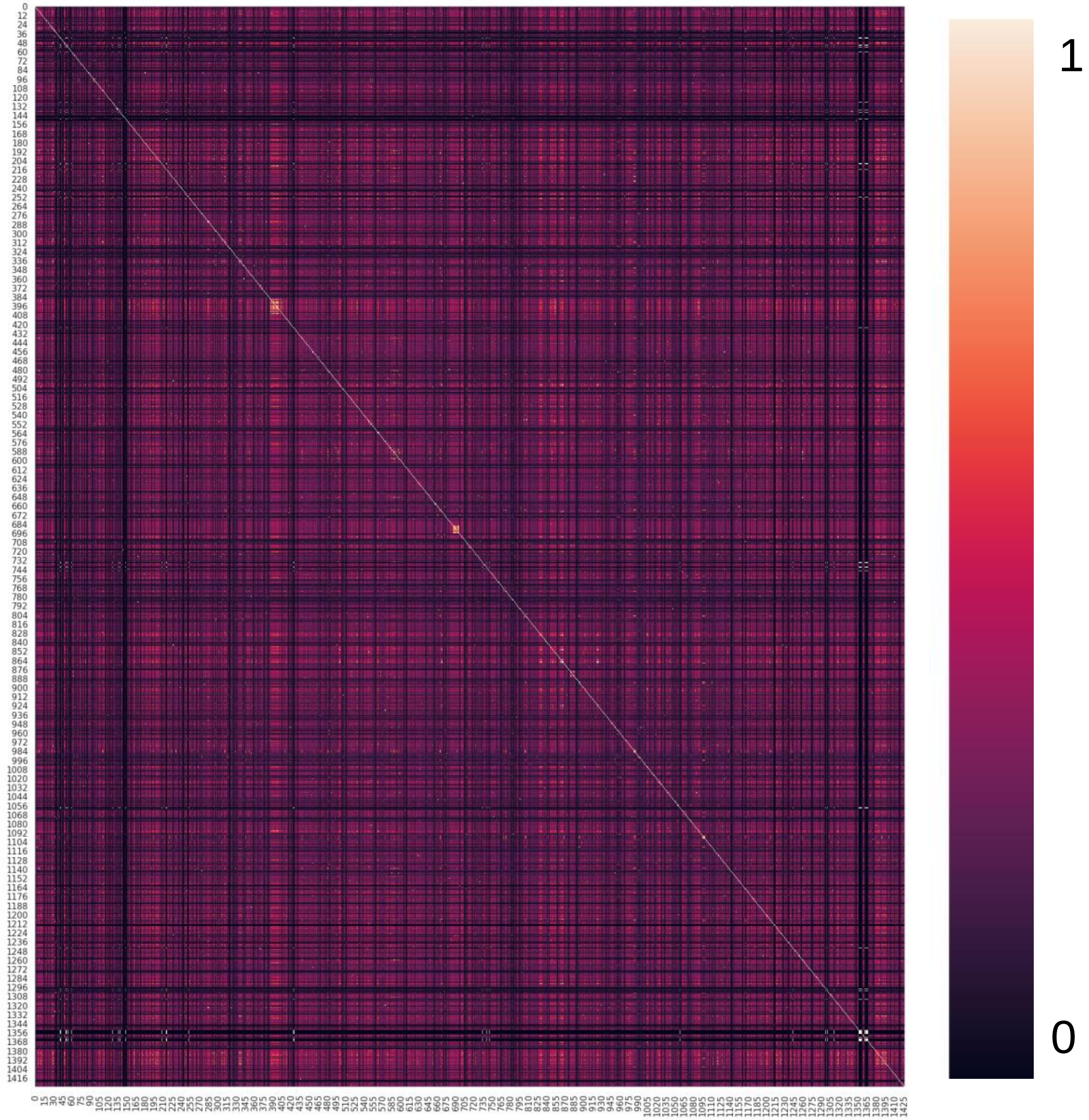
- Numerical representation of SMILES
- Class Imbalance

Aims & Problems

- Tanimoto coefficient: Measures the structural similarity between molecules

- SIDER molecules show generally a strong structural correlation

> One-hot-encoding not suitable



Aims & Problems

- NLP technique
- Word embedding = representation of a word
- Result is typically a real valued vector
- Captures similarity

Aims & Problems

- Class imbalance if the ratio of samples in the minority class is much lower than the ratio of majority class
  > Synthetic Minority Oversampling Technique (SMOTE)
  > Duplicate samples from the minority class

ADR	Negative count	Positive count	Count difference	Minority ratio	Majority ratio	Ratio difference
Hepatobiliary disorders	684	743	59	48	52	4
Metabolism and nutrition disorders	431	996	565	30	70	40
Product issues	1405	22	1383	2	98	96
Eye disorders	551	876	325	39	61	22
Investigations	276	1151	875	19	81	62
Musculoskeletal and connective tissue disorders	430	997	567	30	70	40

Table of Contents

WE Network / Preprocessing

SMILES (string)

['C1=CC(C(C(=C1)C(=O)O)O)O']

> input

SMILES Pair Encoding (list of strings)

['C1=', 'CC(', 'C(', 'C(=', 'C1)', 'C(=O)O)', 'O)', 'O']

> structural groups

Keras Tokenizer

{'c1=':1, 'cc(':2, 'c(':3, 'c(=':4, 'c1)':5, 'c(=o)o)':6, 'o)':7, 'o':8}

> vocabulary based on word frequency

[[1, 2, 3, 4, 5, 6, 7, 8]]

> **padded numerical SMILES representation based on structural groups**

WE Network / Training Data

- Train a word embedding network for finding structural patterns / correlations

Training Data

~ 440k isomeric SMILES from PubChem

Preprocessing

SMILES (string)

['C1=CC(C(C(=C1)C(=O)O)O)O']

> input

SMILES Pair Encoding (list of strings)

['C1=', 'CC(', 'C(', 'C(=', 'C1)', 'C(=O)O)', 'O)', 'O']

> structural groups

Keras Tokenizer

{'c1=':1, 'cc(':2, 'c(':3, 'c(=':4, 'c1)':5, 'c(=o)o)':6, 'o)':7, 'o':8}

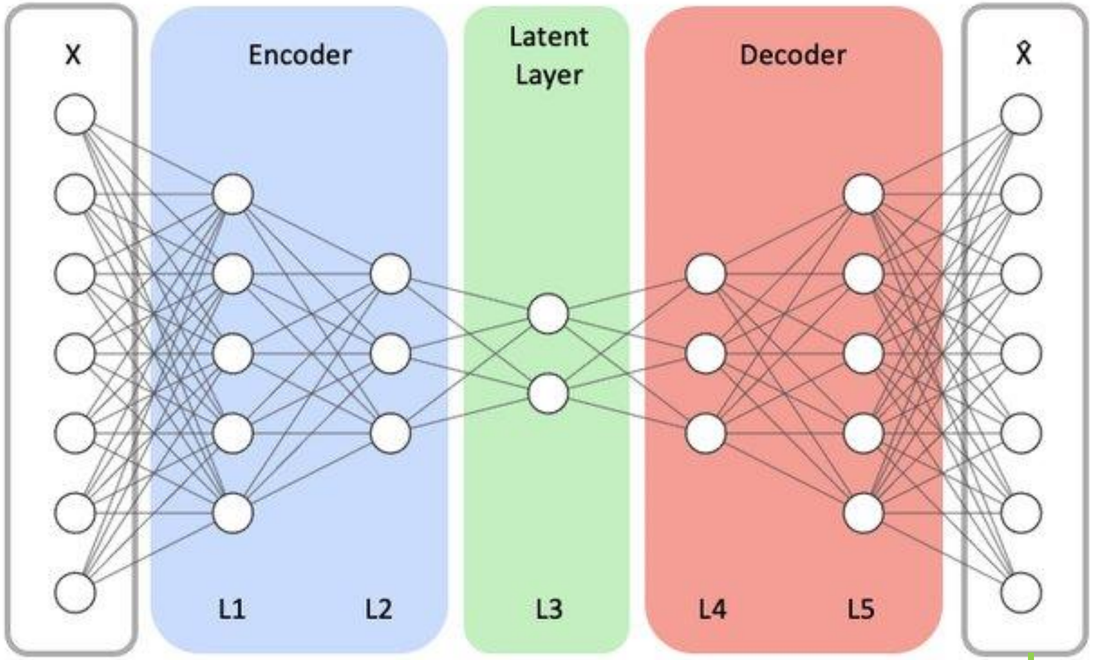
> vocabulary based on word frequency

[[1, 2, 3, 4, 5, 6, 7, 8]]

> padded numerical SMILES representation based on structural groups

Vectors (358,)

Word Embedding (WE) Network



Feedback

> floating vector with 50 entries for each SMILE

WE Network / Encoder

```
Model: "we_encoder"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 encoder_input (InputLayer)  [(None, 358)]             0

 encoder_embedding (Embeddin  (None, 358, 50)          102400
 g)

 encoder_rnn (Bidirectional)  (None, 358, 256)         628736

 encoder_avg_pool (GlobalAve  (None, 256)              0
 ragePooling1D)

 encoder_fc_dense_1 (Dense)  (None, 256)               65792

 encoder_fc_dense_2 (Dense)  (None, 256)               65792

 encoder_output (Dense)      (None, 50)                12850

=================================================================
Total params: 875,570
Trainable params: 875,570
Non-trainable params: 0
_________________________________________________________________
```

Encoder

> generates a WE from preprocessed SMILE

WE Network / Decoder

```
Model: "we_decoder"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 decoder_input (InputLayer)  [(None, 50)]              0

 decoder_fc_dense_1 (Dense)  (None, 50)                2550

 decoder_fc_dense_2 (Dense)  (None, 256)               13056

 decoder_fc_dense_3 (Dense)  (None, 256)               65792

 decoder_output (Dense)      (None, 358)               92006

=================================================================
Total params: 173,404
Trainable params: 173,404
Non-trainable params: 0
_________________________________________________________________
```

Decoder

> Takes word embedding created by the encoder

> Tries to reconstruct back the initial SMILE

WE Network / Architecture

```
Model: "we_network"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 encoder_input (InputLayer)  [(None, 358)]             0

 encoder_embedding (Embeddin  (None, 358, 50)          102400
 g)

 encoder_rnn (Bidirectional)  (None, 358, 256)         628736

 encoder_avg_pool (GlobalAve  (None, 256)              0
 ragePooling1D)

 encoder_fc_dense_1 (Dense)  (None, 256)               65792

 encoder_fc_dense_2 (Dense)  (None, 256)               65792

 encoder_output (Dense)      (None, 50)                12850

 we_decoder (Functional)     (None, 358)               173404

=================================================================
Total params: 1,048,974
Trainable params: 1,048,974
Non-trainable params: 0
_________________________________________________________________
```

Autoencoder

> Comprise both models into one network

> capture semantics of SMILE and return word embeddings

> Only Encoder part of the Autoencoder model is needed for generation of Word Embeddings

Table of Contents

Classifier

Vectors (358,)

Vectors (50,)

Training Data

SIDER Dataset (1427,28)

Preprocessing

SMILES (string)

['C1=CC(C(C(=C1)C(=O)O)O)O']

> input

SMILES Pair Encoding (list of strings)

['C1=', 'CC(', 'C(', 'C(=', 'C1)', 'C(=O)O)', 'O)', 'O']

> structural groups

Keras Tokenizer

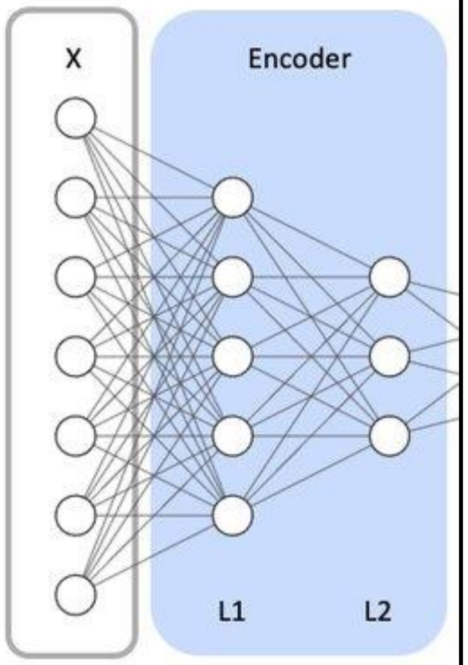
{'c1=':1, 'cc(':2, 'c(':3, 'c(=':4, 'c1)':5, 'c(=o)o)':6, 'o)':7, 'o':8}

> vocabulary based on word frequency

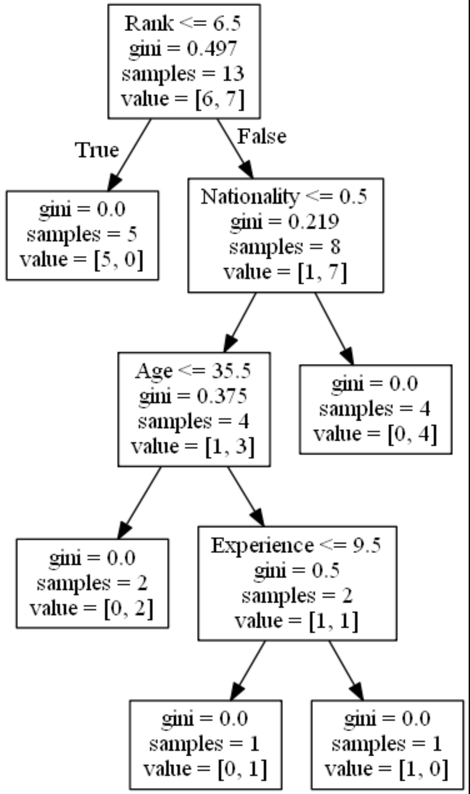
[[1, 2, 3, 4, 5, 6, 7, 8]]

> padded numerical SMILES representation based on structural groups

Pretrained Encoder



Decision Tree



AdaBoost

ADR Prediction

Classifier / Results

Decision Tree

ADR	Accuracy score	F1 score
Hepatobiliary disorders	64 %	58 %
Metabolism and nutrition disorders	98 %	98 %
Product issues	59 %	53 %
Eye disorders	75 %	75 %
Investigations	62 %	62 %
Musculoskeletal and connective tissue disorders	78 %	78 %
Gastrointestinal disorders	73 %	77 %
Social circumstances	64 %	70 %
Immune system disorders	68 %	75 %
Reproductive system and breast disorders	87 %	85 %
Neoplasms benign, malignant and unspecified (i...	70 %	72 %
General disorders and administration site cond...	70 %	72 %
Endocrine disorders	68 %	73 %
Surgical and medical procedures	56 %	67 %
Vascular disorders	87 %	86 %
Blood and lymphatic system disorders	64 %	61 %
Skin and subcutaneous tissue disorders	70 %	73 %
Congenital, familial and genetic disorders	71 %	67 %
Infections and infestations	62 %	64 %
Respiratory, thoracic and mediastinal disorders	63 %	66 %
Psychiatric disorders	86 %	87 %
Renal and urinary disorders	66 %	61 %
Pregnancy, puerperium and perinatal conditions	81 %	82 %
Ear and labyrinth disorders	58 %	57 %

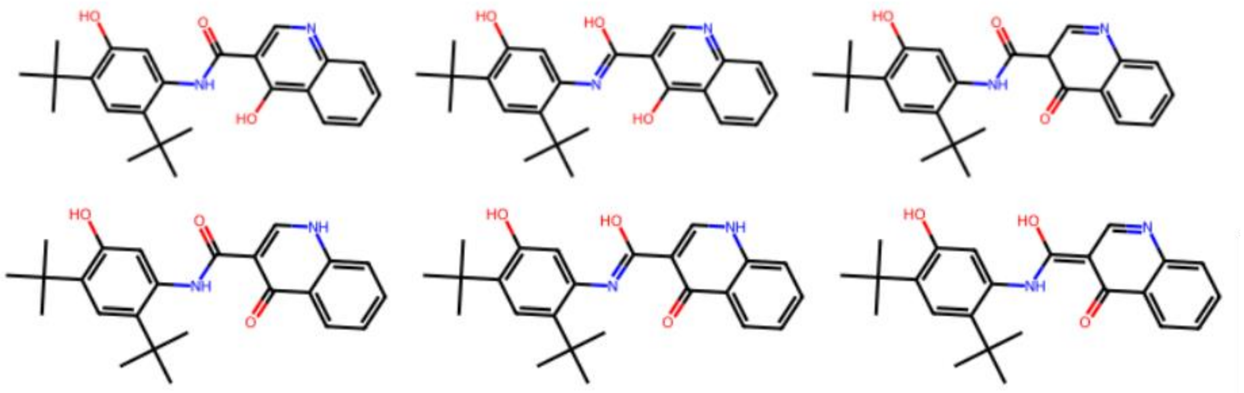
Decision Tree + AdaBoost

ADR	Accuracy score	F1 score
Hepatobiliary disorders	76 %	71 %
Metabolism and nutrition disorders	100 %	100 %
Product issues	66 %	64 %
Eye disorders	85 %	84 %
Investigations	77 %	73 %
Musculoskeletal and connective tissue disorders	95 %	95 %
Gastrointestinal disorders	88 %	90 %
Social circumstances	80 %	78 %
Immune system disorders	83 %	85 %
Reproductive system and breast disorders	94 %	93 %
Neoplasms benign, malignant and unspecified (i...	83 %	85 %
General disorders and administration site cond...	90 %	91 %
Endocrine disorders	81 %	80 %
Surgical and medical procedures	69 %	68 %
Vascular disorders	95 %	95 %
Blood and lymphatic system disorders	88 %	90 %
Skin and subcutaneous tissue disorders	77 %	75 %
Congenital, familial and genetic disorders	83 %	82 %
Infections and infestations	75 %	73 %
Respiratory, thoracic and mediastinal disorders	72 %	69 %
Psychiatric disorders	93 %	93 %
Renal and urinary disorders	78 %	75 %
Pregnancy, puerperium and perinatal conditions	94 %	94 %
Ear and labyrinth disorders	72 %	68 %

Table of Contents

Discussion

- Data Augmentation:
  > Creating tautomeric structures
  from SIDER SMILES



- Better Word Embedding Network

- Using Graph Embeddings

- More comprehensive dataset with better covering of the dynamics of ADRs

Thank you for your Attention!

&

Don’t f**o**rget t**o** SMILE

C1[C@H](SC(=C(C#N)N2C=CN=C2)S1)C3=C(C=C(C=C3)Cl)Cl

Sources

- **Pictures:**

https://www.welt.de/geschichte/article169174727/Wie-die-Atombombe-gegen-Contergan-blind-machte.html
https://www.welt.de/kultur/history/article108632744/Der-Kampf-der-Contergan-Firma-gegen-die-Opfer.html
https://www.freepik.com/free-photo/portrait-happy-pregnant-woman-touching-her-belly_7336825.htm
https://en.wikipedia.org/wiki/File:SMILES.png
https://www.u-helmich.de/che/Sek2/Organik/Mechanismen/SN/Bilder/SN1-02.jpg
https://www.uibk.ac.at/organic/micura/teaching/grundlagen-organische-chemie/download/ocphkap6.pdf
https://www.w3schools.com/python/python_ml_decision_tree.asp

- **Data Sources:**

Drug components with annotated ADRs
https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/sider.csv.gz

Isomeric SMILES for WE-network training
https://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full/SDF/

- **Software**

https://pypi.org/project/SmilesPE/

https://github.com/jurajtrappl/adverse-drug-reactions