Property Prediction Briefing

Mol2vec & RNN Models

田野&赵伟豪

Property Prediction Briefing

Contents

Property Prediction Briefing

SMILES string

- SMILES: Simplified Molecular Input Line Entry Specification
  - Example: CCOC(=Oc1c(C)cc(O)c(C=O)c1O, Cl.N=C(N)n1cccn1, ...
  - Atoms: ‘C’, ‘[Au]’
  - Chemical bond: ‘=’, ‘#’
  - Ring: ‘C1CCCCC1’
- Proceccing SMILS string: rdkit.Chem.MolFromSmiles

Property Prediction Briefing

Embedding

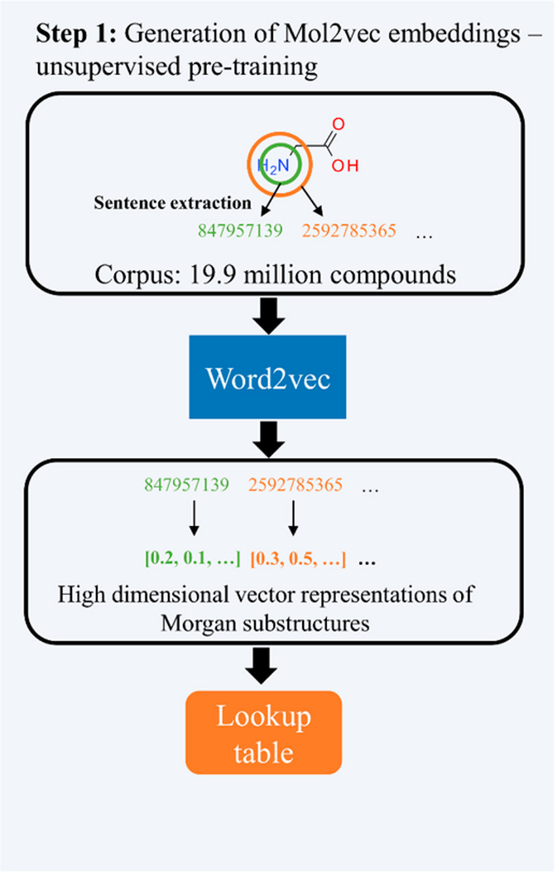
- Mol2vec: https://github.com/samoturk/mol2vec
- Workflow of embedding SMILES string in vectors:
  - Convert tokens in SMILES string to words
  - Embed words in vectors
  - Sum up

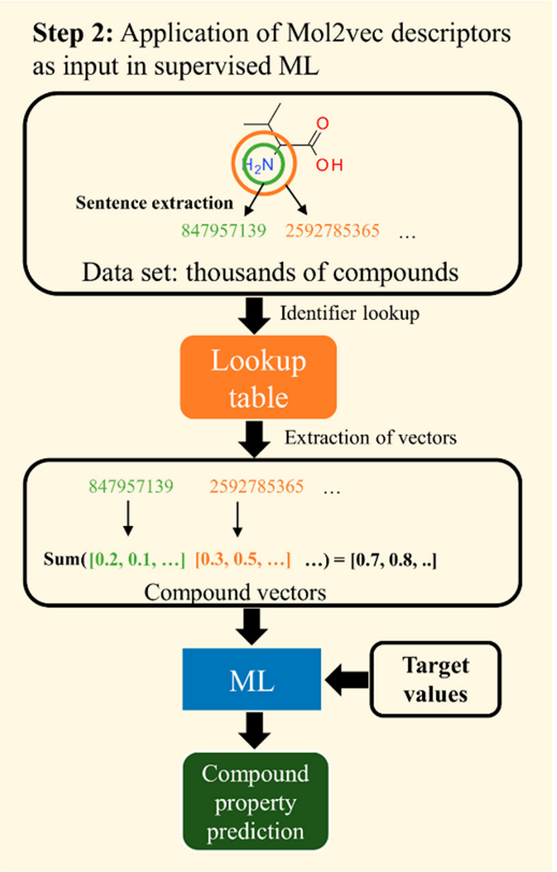
Property Prediction Briefing

Embedding

- Workflow of training mol2vec models and application of mol2vec in property prediction tasks: https://pubs.acs.org/doi/10.1021/acs.jcim.7b00616

- Pretrained mol2vec model: model_300dim.pkl





Property Prediction Briefing

Feature selection

- Pretrained mol2vec model generates a 300 dimension vector for a single molecule
  - Use all features to predict property
  - Use specified features to predict property
  - Apply PCA to embeddings and choose some major features

Property Prediction Briefing

RNN models

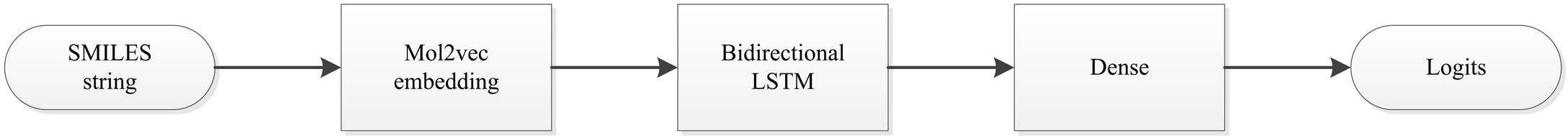
- Feed selected features to simple RNN model

- Poor performance
- Fall into local optimal solution

Property Prediction Briefing

RNN models

- Feed selected features to Bidirectional LSTM model



- Also poor performance and fall into local optimal solution
- Various structure of molecules neglected by both RNN models

Property Prediction Briefing

RNN models

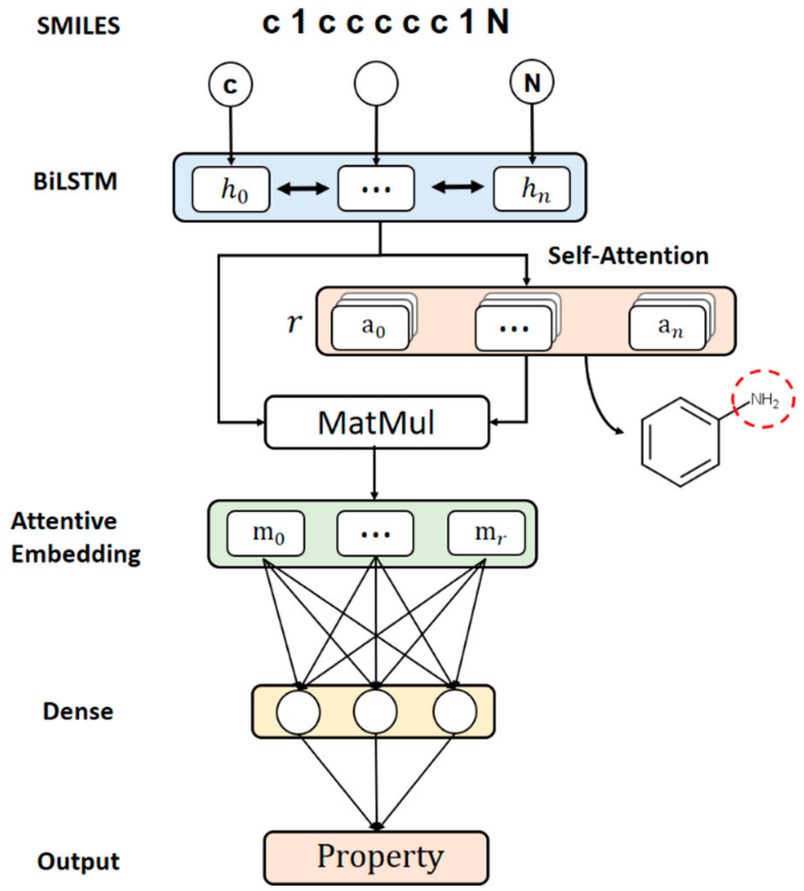
- Instead of embedding SMILES strings in vectors, embed them in *tensors*:
  - Convert tokens in SMILES string to words
  - Embed words in vectors
  - Feature selection
  - Stack up
- Construct a model which can utilize the structural information

Property Prediction Briefing

RNN models

- Consider introducing self attention mechanism into RNN models: https://pubs.acs.org/doi/10.1021/acs.jcim.8b00803

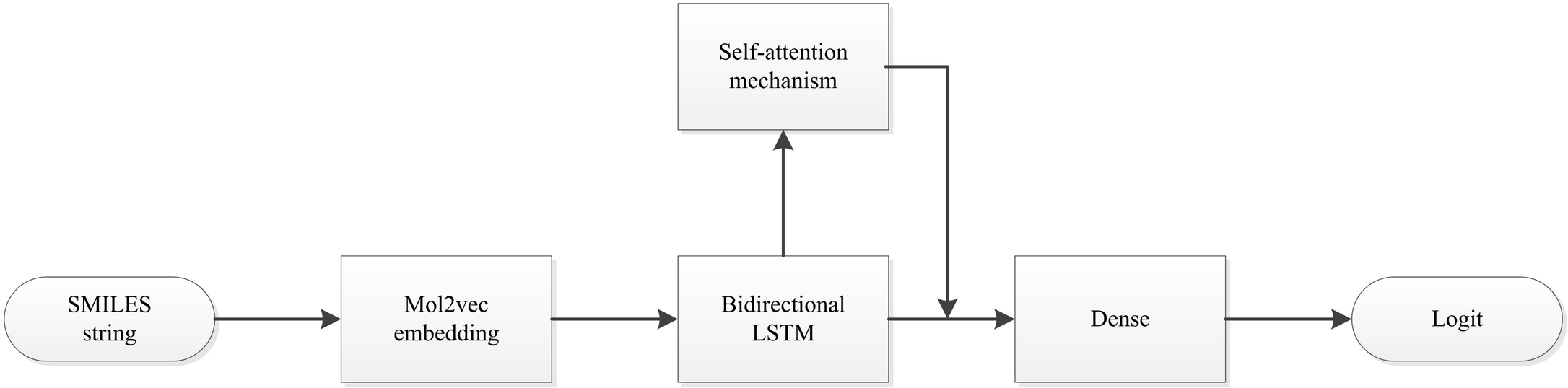
- Adapt to tensor inputs in order to apply mol2vec and utilize structural information



Property Prediction Briefing

RNN models

- Feed selected features to Self attention BiLSTM model



- Get improvement in model performance

Property Prediction Briefing

RNN models

- 20 tests with 501 molecules:

ROC-AUC	PRC-AUC
0.9618 ± 0.0129	0.6258 ± 0.0907

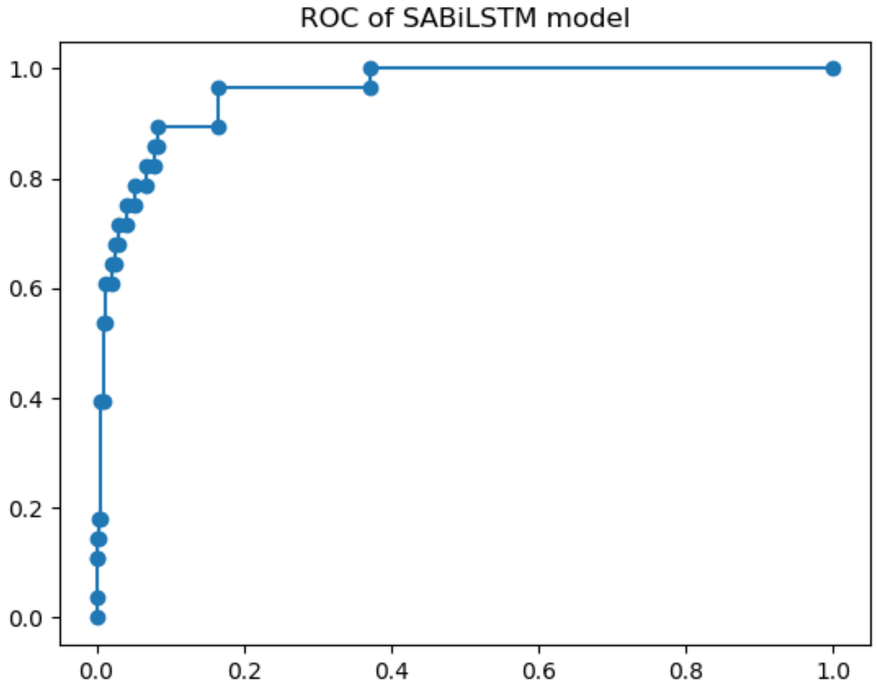
- 10 fold cross validation:

	ACC	ROC-AUC	PRC-AUC		ACC	ROC-AUC	PRC-AUC
0	0.9104	0.4959	0.1275	5	0.8756	0.8586	0.3811
1	0.8168	0.6391	0.4530	6	0.9748	0.5849	0.0680
2	0.8706	0.7563	0.5150	7	0.8558	0.4010	0.0040
3	0.8119	0.3750	0.0146	8	0.8507	0.9670	0.2156
4	0.7214	0.8915	0.3878	9	0.7624	0.9062	0.4017

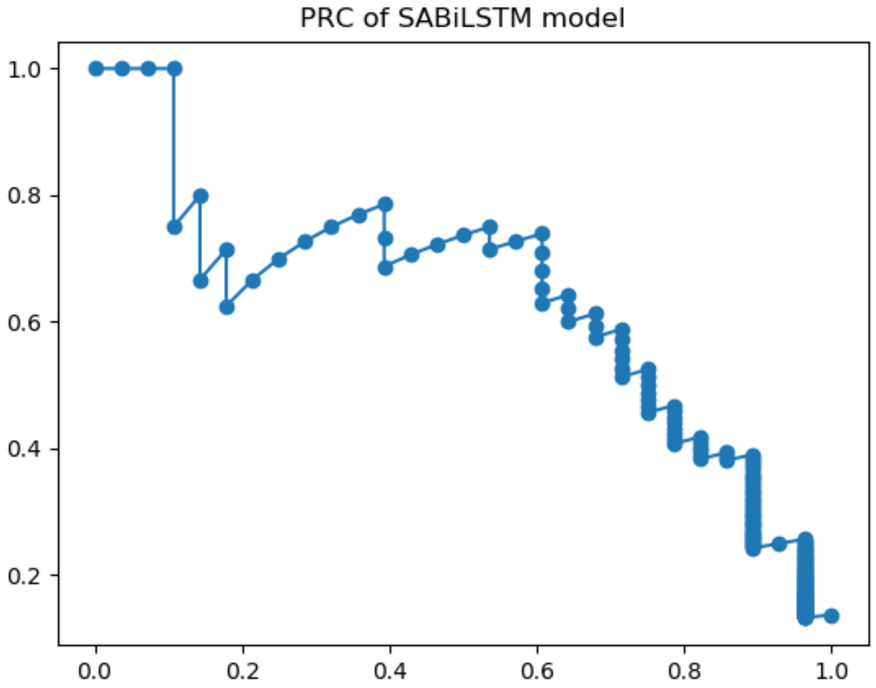
Property Prediction Briefing

RNN models

- ROC



- PRC



Property Prediction Briefing

Unbalanced data

- Ecoli.csv: 120 valid molecules, 2215 invalid molecules
  - Replicate valid molecules and add to dataset
  - Fastest training: train SABiLSTM model with a half valid and half invalid dataset
  - Other ratio: SABiLSTM model will always converge, but slower
  - Use balanced dataset at the beginning of training, then lower the ratio

Property Prediction Briefing

Appendix

- RDKit: https://github.com/rdkit/rdkit
- Mol2vec: https://github.com/samoturk/mol2vec
- Mol2vec workflow introduction: https://pubs.acs.org/doi/10.1021/acs.jcim.7b00616
- Self attention BiLSTM model: https://pubs.acs.org/doi/10.1021/acs.jcim.8b00803