CheMixNet: Mixed DNN Architectures for Predicting Chemical Properties using Multiple Molecular Representations

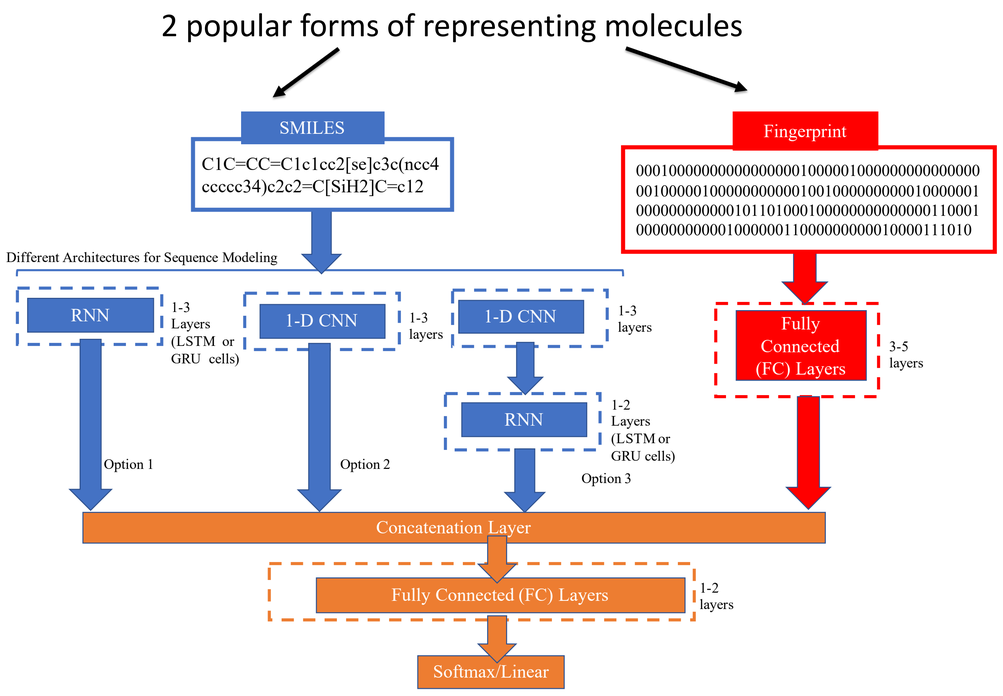
Arindam Paul, Dipendra Jha, Reda Al-Bahrani, Wei-keng Liao, Alok Choudhary, Ankit Agrawal
Department of Electrical Engineering and Computer Science, Northwestern University

INTRODUCTION

- CheMixNet: Set of general-purpose architectures for chemical property prediction.
- There exists multiple molecular representations for molecules: SMILES/InChI, fingerprints, graphs etc.
- Different networks exist that are successful on certain representations and/or size of molecules.
- There is motivation for a multi-input architecture that can harness multiple representations.
- CheMixNet models outperform other architectures such as fully connected networks, SMILES2vec, Chemception & Molecular Graph Convolutions across multiple datasets.

EXPERIMENTS

METHOD



We use 3 neural network candidate options: CNNs, RNNs, CNNs followed by RNNs for learning from SMILES. For better interpretability, we use 166 bit MACCS fingerprints instead of compressed 1024 bit representations (Atom Pair, Morgan etc.)

DATA

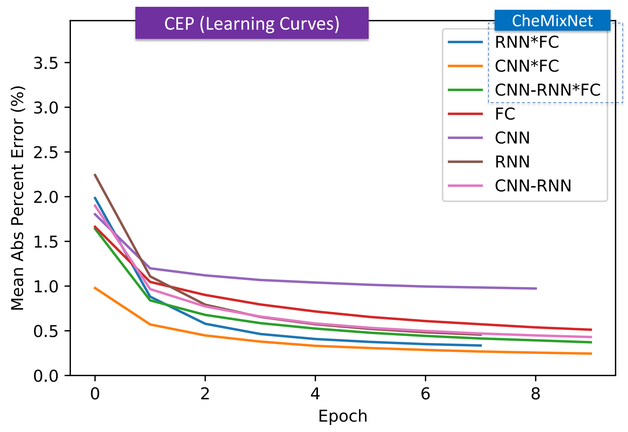
We demonstrate the effectiveness of CheMixNet for predicting chemical properties from SMILES and fingerprints using six different datasets. For the SMILES sequence, we used 1-hot encoding to convert the SMILES into a fixed length representation. The length of the sequence was determined by the length of the longest SMILES sequence in each dataset. The vocabulary size was determined by finding the number of unique characters in each dataset.

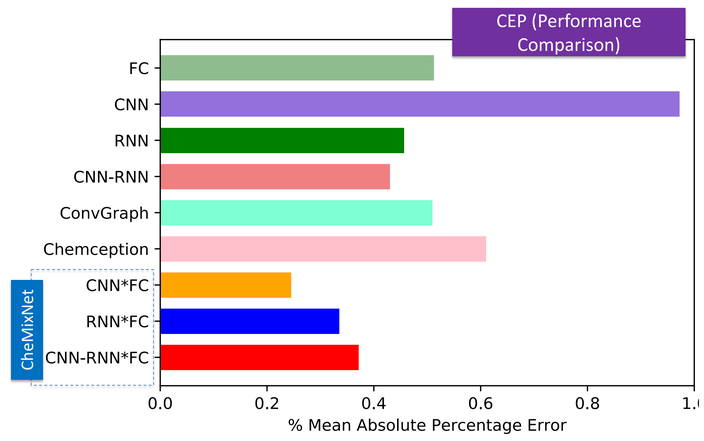
Dataset	Property	Task	Size
Clean Energy Project (CEP)	Highest Occupied Molecular Orbital Energy	Regression	2.3 million
ESOL	Activity	Regression	1,128
FreeSolv (Experimental)	Solvation Energy	Regression	643
FreeSolv (Computed)	Solvation Energy	Regression	643
HIV	Activity	Classification	2,886
Tox21	Toxicity	Classification	8,981

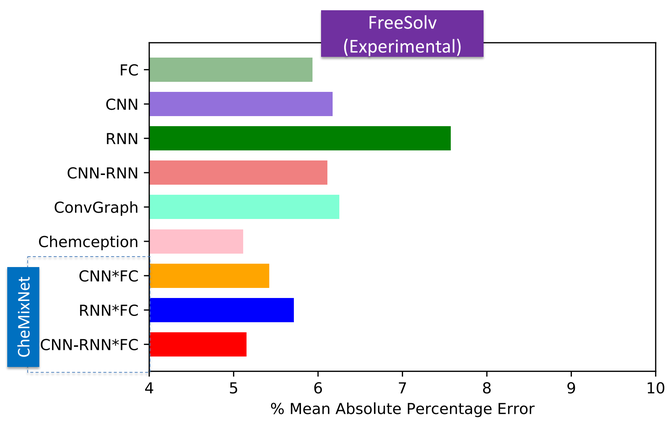
Dataset	Size of Vocabulary	Maximum Input Sequence Length
CEP	23	83
ESOL	33	98
FreeSolv (Experimental)	32	83
FreeSolv (Computed)	32	83
HIV	54	400
Tox21	42	940

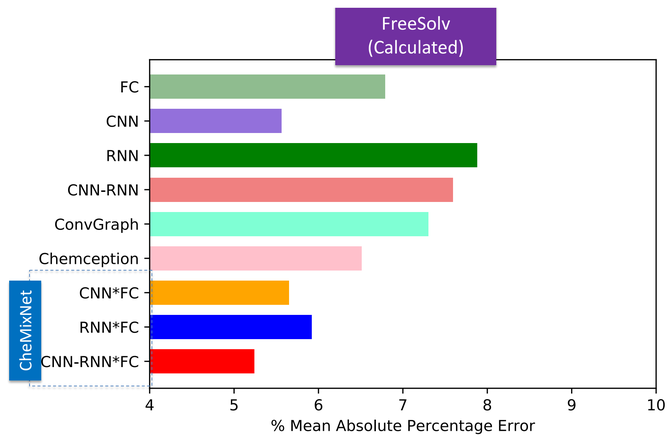
RESULTS

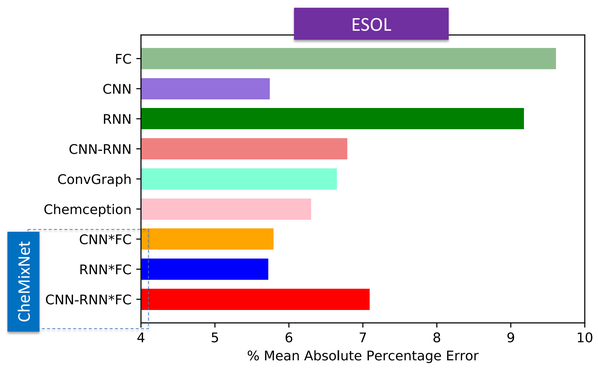
For our experiments, we used Adam as the optimizer with initial learning rate of 0.001. Early stopping was used during training to avoid over-fitting. For the CEP dataset, we used a batch size of 64; for the two classification datasets (HIV and Tox21), we used a batch size of 32; for the ESOL and FreeSolv, we used a batch size of 16.





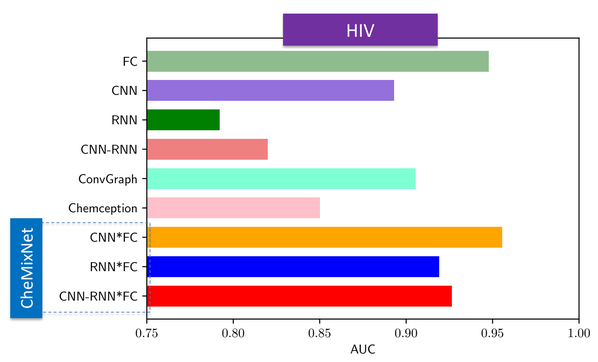


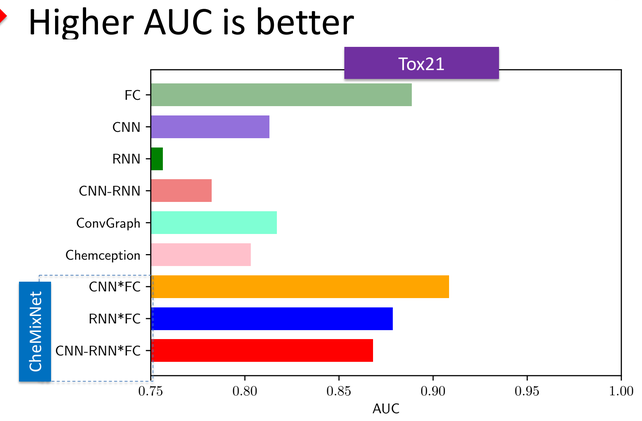




For *regression* tasks, Lower MAPE is better

For *classification* tasks, Higher AUC is better





FUTURE WORK

- Extend CheMixNet with other candidate representations such as molecular graphs and 2D molecular drawings
- Interpretability of candidate neural networks w.r.t. inputs
- Hierarchical Attention Networks with attention on individual atoms and sub-groups

Code available at: *tinyurl.com/chemixnet*

apaul@u.northwestern.edu

Supported by U.S. Department of Commerce, National Institute of Standards and Technology as part of Center for Hierarchical Materials Design (CHiMaD) under financial assistance award 70NANB14H012.