# CASTER: Predicting Drug Interactions with Chemical Substructure Representation

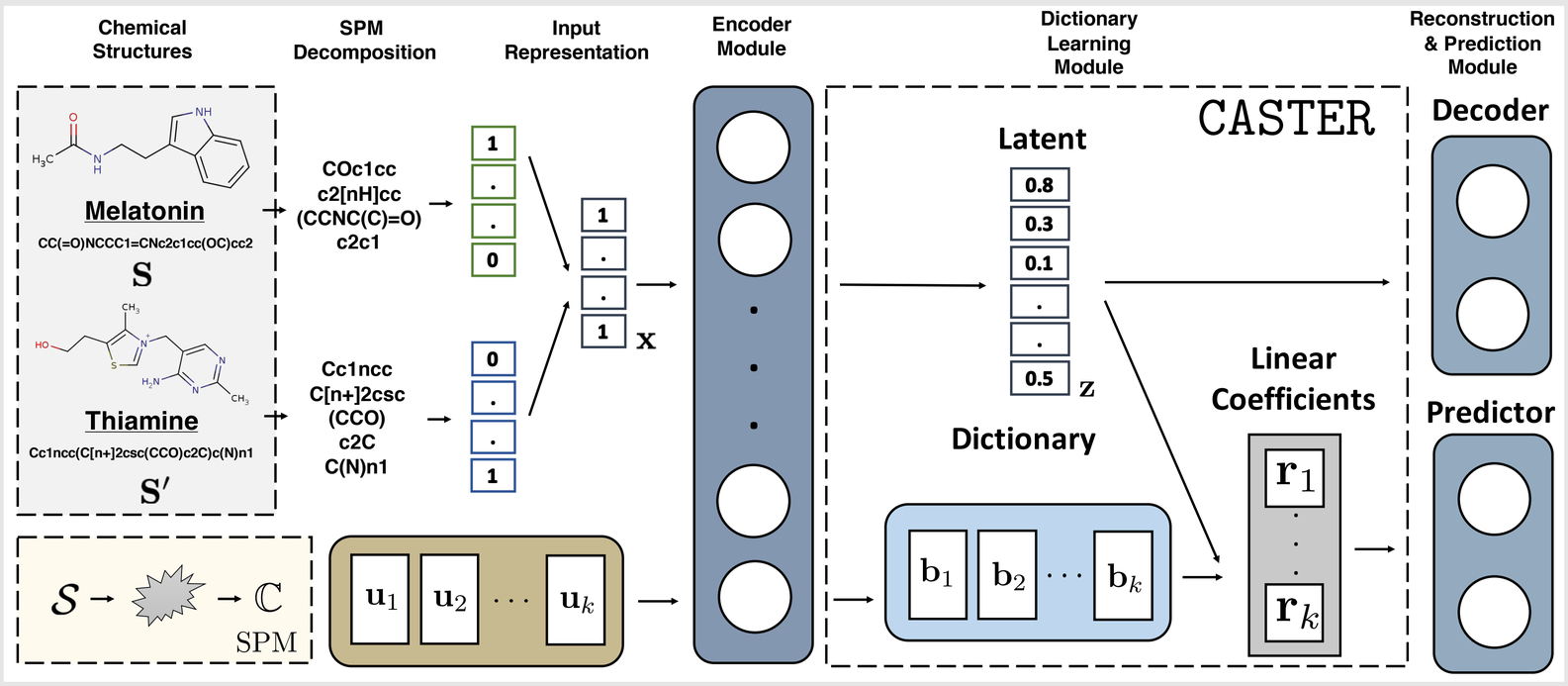
**Kexin Huang[1,2], Cao Xiao[2], Trong Nghia Hoang[3], Lucas M. Glass[2], Jimeng Sun[4,5]**

[1]Harvard [2]IQVIA [3]MIT-IBM Watson AI Lab [4]UIUC CS [5]Georgia Tech

## Motivation

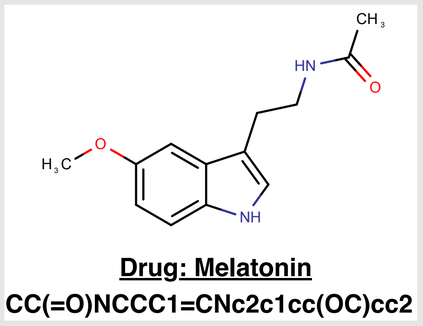
- Drug-Drug Interaction (DDI) is the **delay, decrease, or enhance** absorption of drugs when intaking multiple ones. It can result in **adverse effects** that incur **morbidity & mortality**, and huge **medical costs**.
- Traditional strategies of gaining DDI knowledge includes preclinical *in vitro* safety profiling and clinical safety trials, however they are restricted in terms of **small scales, long duration, and huge costs**.
- Deep learning models that leverage massive biomedical data emerged as a promising direction.
- It assumes that **drugs with similar representations (of chemical structure) will have similar properties.**

## Method




## Background

- **SMILES String** is a sequence of symbols of the chemical atoms & bonds in its depth-first traversal order of its molecular structure graph.



**Drug: Melatonin**

**CC(=O)NCCC1=CNc2c1cc(OC)cc2**

- **Click Chemistry**: One of the major mechanism of drug interactions results from the chemical reactions among only **a few functional sub-structures** of the entire drug's molecular structure, while the remaining substructures are less relevant.

## Challenges

- **Lack of specialized drug representation for DDI prediction (click chemistry).** Previous works often generate drug representations using **the entire chemical representation**, which causes the learned representations to be potentially biased toward irrelevant substructures. This undermines the learned drug similarity and DDI predictions.
- **Limited labels and generalizability.** Some of the previous methods need **external biomedical knowledge** for improved performance and cannot be generalized to drugs in early development phase. Others rely on **a small set of labelled training data**, which impairs their generalizability to new drugs or DDIs.
- **Non-interpretable prediction.** Although deep learning models show good performance in DDI prediction, they often produce predictions that are **characterized by a large number of parameters**, which is hard to interpret.

## Problem & Goal

- **Task:** To predict drug interactions, we need to learn a mapping $\mathcal{G} : \mathcal{S} \times \mathcal{S} \rightarrow [0, 1]$ from a drug-drug pair $(\mathbf{S}, \mathbf{S}') \in \mathcal{S} \times \mathcal{S}$ to a probability that indicates the chance that $\mathbf{S}$ and $\mathbf{S}'$ will have interaction.
- We want the prediction both **accurate** and **interpretable**. It should use the untapped **unlabeled dataset** and leverage the **DDI mechanism**.

## SPM Decomposition

**Algorithm 1:** The Chemical Sequential Pattern Mining Algorithm

Initialize $\mathbb{V}$ to the set of all atoms and bonds, $\mathbb{W}$ as the set of tokenized SMILES strings input
Input $\eta$ as the practitioner-specified frequency threshold, and $\ell$ as the maximum size of $\mathbb{V}$
**for** $t = 1 \ldots \ell$ **do**
  (A, B), FREQ $\leftarrow$ scan $\mathbb{W}$
  //(A,B), FREQ are the frequentest pair and its frequency
  **if** FREQ $< \eta$ **then**
    break // frequency lower than threshold
  **end**
  $\mathbb{W} \leftarrow$ find(A, B) $\in \mathbb{W}$, replace with (AB)
  // update $\mathbb{W}$ with the new token (AB)
  $\mathbb{V} \leftarrow \mathbb{V} \cup$ (AB)
  // add (AB) to the vocabulary set $\mathbb{V}$
**end**

## Dataset

| | Drugbank (DDI) | BIOSNAP |
|---|---|---|
| # Drugs | 1,850 | 1,322 |
| # Positive DDI Labels | 221,523 | 41,520 |
| # Negative Labels | 221,523 | 41,520 |

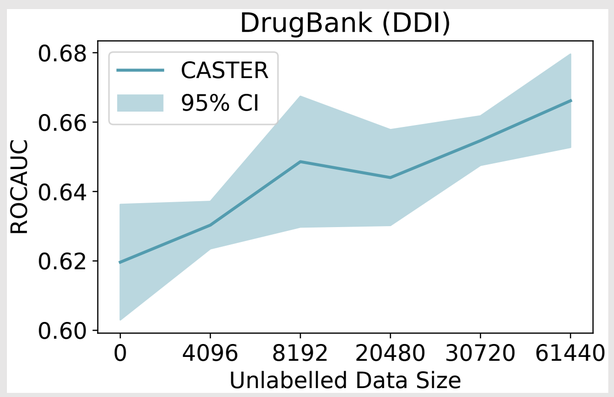
| | Unlabelled |
|---|---|
| # Drugs | 9,675 |
| # Food Compounds | 24,738 |
| # Drug-Drug Pairs | 220,000 |
| # Drug-Food Pairs | 220,000 |

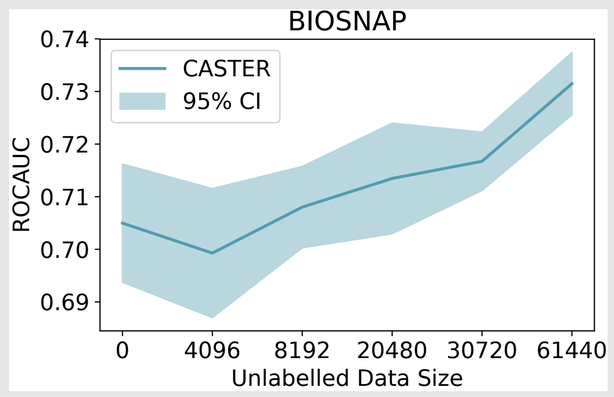
## Predictive Performance

| Model | Dataset | ROC-AUC | PR-AUC | F1 |
|---|---|---|---|---|
| LR | BIOSNAP | $0.802 \pm 0.001$ | $0.779 \pm 0.001$ | $0.741 \pm 0.002$ |
| | DrugBank | $0.774 \pm 0.003$ | $0.745 \pm 0.005$ | $0.719 \pm 0.006$ |
| Nat.Prot | BIOSNAP | $0.853 \pm 0.001$ | $0.848 \pm 0.001$ | $0.714 \pm 0.001$ |
| | DrugBank | $0.786 \pm 0.003$ | $0.753 \pm 0.003$ | $0.709 \pm 0.004$ |
| Mol2Vec | BIOSNAP | $0.879 \pm 0.006$ | $0.861 \pm 0.005$ | $0.798 \pm 0.007$ |
| | DrugBank | $0.849 \pm 0.004$ | $0.828 \pm 0.006$ | $0.775 \pm 0.004$ |
| MolVAE | BIOSNAP | $0.892 \pm 0.009$ | $0.877 \pm 0.009$ | $0.788 \pm 0.033$ |
| | DrugBank | $0.852 \pm 0.006$ | $0.828 \pm 0.009$ | $0.769 \pm 0.031$ |
| DeepDDI | BIOSNAP | $0.886 \pm 0.007$ | $0.871 \pm 0.007$ | $0.817 \pm 0.007$ |
| | DrugBank | $0.844 \pm 0.003$ | $0.828 \pm 0.002$ | $0.772 \pm 0.006$ |
| CASTER | BIOSNAP | $0.910 \pm 0.005$ | $0.887 \pm 0.008$ | $0.843 \pm 0.005$ |
| | DrugBank | $0.861 \pm 0.005$ | $0.829 \pm 0.003$ | $0.796 \pm 0.007$ |

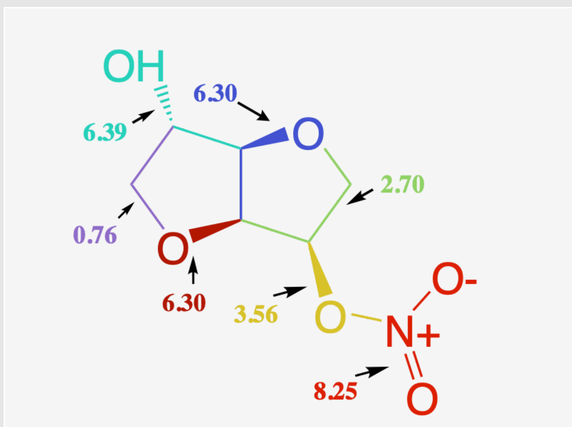
CASTER achieves the **best** predictive performance!

## Usage of Unlabelled Data




## Interpretability

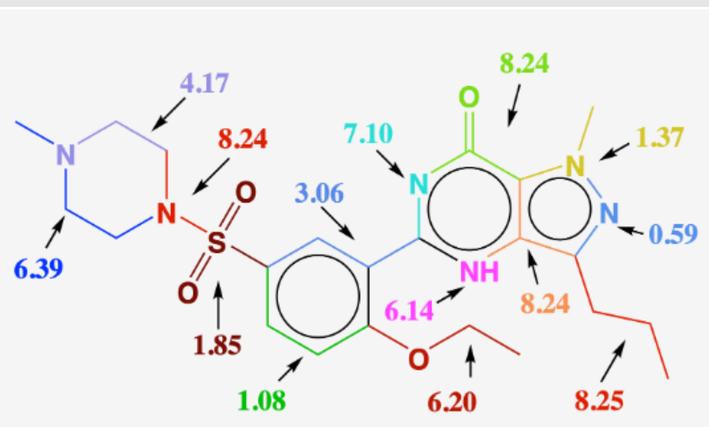



**Isosorbide Mononitrate**

**O=[N+]([O-])O[C@@H]1CO[C@@H]2[C@@H](O)CO[C@H]12**

CASTER

$p(\mathbf{x}) = 0.793$




**Sildenafil**

**CCCc1nn(C)c2c(=O)nc(-c3cc(S(=O)(=O)N4CCN(C)CC4)ccc3OCC)[nH]c12**

- Random initialization **does not affect** our prediction! CASTER achieves **0.7673** average correlation score across five models with different random seeds. Also, we find all nitrate-based drugs and CASTER assigns on average **50% higher coefficient** to nitrate than the mean of coefficients of other sub-structures existed in the input pair, which is not a coincidence.

**Code**: https://github.com/kexinhuang12345/CASTER **Paper:** https://arxiv.org/pdf/1911.06446.pdf **Contact:** kexinhuang@hsph.harvard.edu