ChemicalX: A Deep Learning Library for Drug Pair Scoring

Benedek Rozemberczki*, Charles Tapley Hoyt**, Anna Gogleva*, Piotr Grabowski*,
Klas Karis**, Andrej Lamov*, Andriy Nikolov*, Sebastian Nilsson,* Michael Ughetto*,
Yu Wang*, Tyler Derr***, Benjamin M. Gyori**

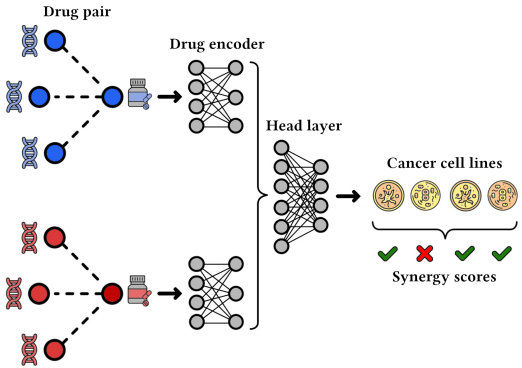
* AstraZeneca
** Harvard Medical School
*** Vanderbilt University

KDD 2022 Applied Data Science Track

1. Drug Pair Scoring
2. A System for Repurposing
3. Experimental Results
4. Conclusions

Drug Pair Scoring

What is drug pair scoring?



- The drug pair scoring task
- Motivation
  - Timelines
  - Material costs (assays)
  - Labour costs
  - Tractability
- Application domains - tasks
  - Interaction
  - Polypharmacy side effect
  - Synergy
- Multi-objective optimization

How do we represent drugs in drug pair scoring?

Let $G = (\mathcal{V}, \mathcal{R}, \mathcal{E})$ be a heterogeneous biological graph with drug entities $\mathcal{D} \subset \mathcal{V}$.

1. Molecular - low level encoders:

$$\mathbf{h}^d = f_{\Theta_D}(\mathcal{M}^d)$$

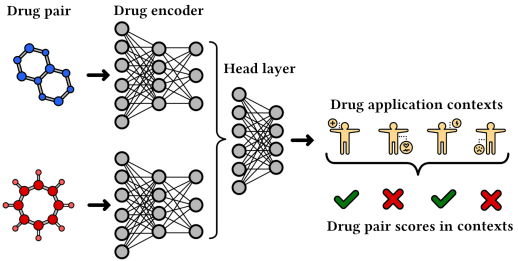
2. Systems biology based - high level encoders:

$$\mathbf{h}^d = \text{AGGREGATE}(\{\Theta^u, \forall u \in \mathcal{N}(d)\})$$

3. Hierarchical encoders (structure and systems view):

$$\mathbf{h}^d = \text{AGGREGATE}(\{f_{\Theta_D}(\mathcal{M}^u), \forall u \in \mathcal{N}(d)\})$$

Why molecule level models?



- Induction
- Attribution (explanation)
- Pre-training
- Transfer learning
- Structural ablation

How do we represent biological contexts and score pairs?

Given $\mathcal{C}$ a set of biological contexts the representation of $c \in \mathcal{C}$ is:

$$\mathbf{h}_c = f_{\Theta_C}(\mathbf{x}_c).$$

We score the pair $d, d' \in \mathcal{D}$ in the context $c \in \mathcal{C}$ with:

$$\widehat{\mathbf{y}}^{d,d',c} = f_{\Theta_H}(\mathbf{h}^d, \mathbf{h}^{d'}, \mathbf{h}^c).$$

The loss for the pair in the context is defined as:

$$\ell_{d,d',c} = \ell(\widehat{\mathbf{y}}^{d,d',c}, \mathbf{y}^{d,d',c}).$$

The parametric functions $f_{\Theta_D}(\cdot)$, $f_{\Theta_C}(\cdot)$, and $f_{\Theta_H}(\cdot)$ can be trained jointly by minimizing the cost accumulated from data point level losses.

How do we conceptualize these models?

Data: $\mathcal{X}_{\mathcal{D}}$ - Drug feature set.
$\mathcal{X}_{\mathcal{C}}$ - Context feature set.
$\mathcal{B}$ - Labeled drug pair - context batch.
Result: $\mathcal{L}$ - The cost for the batch.

$\mathcal{L} \leftarrow 0$
for $(d, d', c, y^{d,d',c}) \in \mathcal{B}$ **do**
  $\mathbf{h}^{d} \leftarrow f_{\Theta_D}(\mathbf{x}^{d}, \mathcal{G}^{d}, \mathbf{X}_N^{d}, \mathbf{X}_E^{d})$ // Compute drug representation for $d \in \mathcal{D}$.
  $\mathbf{h}^{d'} \leftarrow f_{\Theta_D}(\mathbf{x}^{d'}, \mathcal{G}^{d'}, \mathbf{X}_N^{d'}, \mathbf{X}_E^{d'})$ // Compute drug representation for $d' \in \mathcal{D}$.
  $\mathbf{h}^{c} \leftarrow f_{\Theta_C}(\mathbf{x}^{c})$ // Compute context representation for $c \in \mathcal{C}$.
  $\widehat{y}^{d,d',c} \leftarrow f_{\Theta_H}(\mathbf{h}^{d}, \mathbf{h}^{d'}, \mathbf{h}^{c})$ // Score based on the representations.
  $\mathcal{L} \leftarrow \mathcal{L} + \ell(y^{d,d',c}, \widehat{y}^{d,d',c})$ // Add loss to the accumulated cost.
end

A System for Repurposing

What is ChemicalX?

README.md

pypi package 0.1.0 | docs passing | codecov 95% | CI passing | ArXiv 2202.05240

Documentation | External Resources | Datasets | Examples

ChemicalX is a deep learning library for drug-drug interaction, polypharmacy side effect, and synergy prediction. The library consists of data loaders and integrated benchmark datasets. It also includes state-of-the-art deep neural network architectures that solve the drug pair scoring task. Implemented methods cover traditional SMILES string based techniques and neural message passing based models.

https://github.com/AstraZeneca/chemicalx/

How did you engineer it?

- Documentation
- Unit and integration tests with coverage reports
- Tutorials
- Example datasets
- Continuous integration
- Linting, type hinting and docstrings

Library	Year	Backend	Drug Domain	Pair Scoring
PyG [7]	2018	PT	✘	✘
DGL [24]	2019	PT/TF/MX	✘	✘
StellarGraph [5]	2019	TF	✘	✘
DeepChem [19]	2019	TF	✔	✘
CHChem [16]	2019	CH	✔	✘
Jraph [8]	2020	JAX	✘	✘
Spektral [9]	2020	TF	✘	✘
DIG [15]	2021	PT	✘	✘
TorchDrug [28]	2021	PT	✔	✘
CogDL [3]	2021	PT	✘	✘
TFG [10]	2021	TF	✘	✘
DGL-LS [13]	2021	PT	✔	✘
Our Work	2022	PT	✔	✔

What is included in ChemicalX?

Model	Year	Domain	Encoder
DeepDDI [20]	2018	Interaction	Feedforward
DeepSynergy [18]	2018	Synergy	Feedforward
MHCADDI [6]	2019	Polypharmacy	GAT
MR-GNN [25]	2019	Interaction	GCN
CASTER [12]	2019	Interaction	Feedforward
SSI-DDI [17]	2020	Interaction	GAT
EPGCN-DS [21]	2020	Interaction	GCN
DeepDrug [2]	2020	Interaction	GCN
GCN-BMP [4]	2020	Interaction	GCN
DeepDDS [23]	2021	Synergy	GCN or GAT
MatchMaker [1]	2021	Synergy	Feedforward

How do we load the dataset?

```python
from chemicalx.data import DrugCombDB, BatchGenerator

loader = DrugCombDB()

context_set = loader.get_context_features()
drug_set = loader.get_drug_features()
triples = loader.get_labeled_triples()

generator = BatchGenerator(batch_size=1024,
                           context_features=True,
                           drug_features=True,
                           drug_molecules=False,
                           context_feature_set=context_set,
                           drug_feature_set=drug_set,
                           labeled_triples=triples)
```

How do we train a model?

```python
import torch
from chemicalx.models import DeepSynergy

model = DeepSynergy(context_channels=112,
                    drug_channels=256)

optimizer = torch.optim.Adam(model.parameters())
model.train()
loss = torch.nn.BCELoss()

for epoch in range(200):
    for batch in generator:
        optimizer.zero_grad()
        prediction = model(batch.context_features,
                           batch.drug_features_left,
                           batch.drug_features_right)
        loss_value = loss(prediction, batch.labels)
        loss_value.backward()
        optimizer.step()
```

How do we score a dataset with the model?

```python
import pandas as pd
from chemicalx.data import HAEM

model.eval()

loader = HAEM()

generator.labeled_triples = loader.get_labeled_triples()

predictions = []
for batch in generator:
    prediction = model(batch.context_features,
                       batch.drug_features_left,
                       batch.drug_features_right)
    prediction = prediction.detach().cpu().numpy()
    identifiers = batch.identifiers
    identifiers["prediction"] = prediction
    predictions.append(identifiers)
predictions = pd.concat(predictions)
```

How can you BYOD (Bring Your Own Data)?

Training and scoring for specific pairs would only need these things:

- Drug pairs with biological/chemical contexts and labels.
- Context set with context identifier keys and context feature vector values.
- Drug set with SMILES strings and molecule level features.

Checkout the following link for an example:

`https://github.com/AstraZeneca/chemicalx/tree/main/dataset/drugbankddi`

Experimental Results

What are the datasets integrated?

Table 1: Datasets available in ChemicalX in the domain of the pair scoring task and the number of drugs ($|\mathcal{D}|$), administration contexts ($|\mathcal{C}|$), and labeled triples ($|\mathcal{Y}|$).

Dataset	Task	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{C}\rvert$	$\lvert\mathcal{Y}\rvert$
TWOSIDES [22]	Polypharmacy	644	10	499,582
Drugbank DDI [20]	Interaction	1,706	86	383,496
DrugComb [26, 27]	Synergy	4,146	288	659,333
DrugCombDB [14]	Synergy	2,956	112	191,391
OncolyPharm [11]	Synergy	38	39	23,052

How about predictive performance?

Table 2: The predictive performance of (some) models in *ChemicalX* on TWOSIDES [22]. Feed-forward encoder based architectures are noted with a ■.

		AUROC	AUPR	F_1
DeepDDI [20]	■	$.929 \pm .001$	$.907 \pm .001$	$.848 \pm .009$
DeepSynergy [18]	■	$.940 \pm .001$	$.919 \pm .001$	$.887 \pm .001$
MR-GNN [25]		$.937 \pm .002$	$.917 \pm .001$	$.875 \pm .002$
SSI-DDI [17]		$.823 \pm .002$	$.800 \pm .003$	$.756 \pm .001$
EPGCN-DS [21]		$.855 \pm .003$	$.834 \pm .002$	$.785 \pm .004$
DeepDrug [2]	■	$.923 \pm .004$	$.904 \pm .002$	$.857 \pm .002$
GCN-BMP [4]		$.709 \pm .003$	$.694 \pm .002$	$.592 \pm .003$
DeepDDS [23]		$.915 \pm .002$	$.898 \pm .002$	$.839 \pm .003$
MatchMaker [1]	■	$.912 \pm .002$	$.892 \pm .001$	$.849 \pm .001$

How about training time?

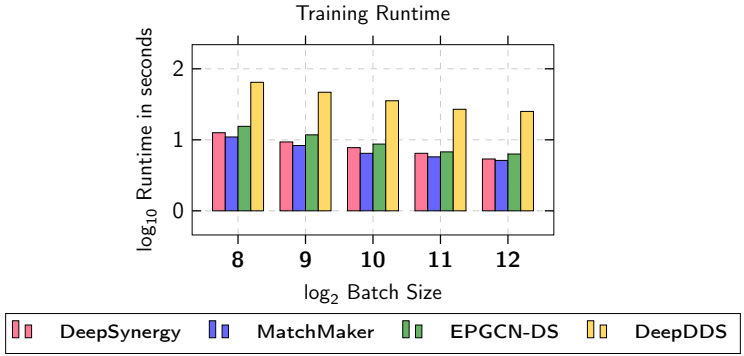


Figure 1: The average runtime of doing an epoch on DrugBankDDI [11].

How about inference time?

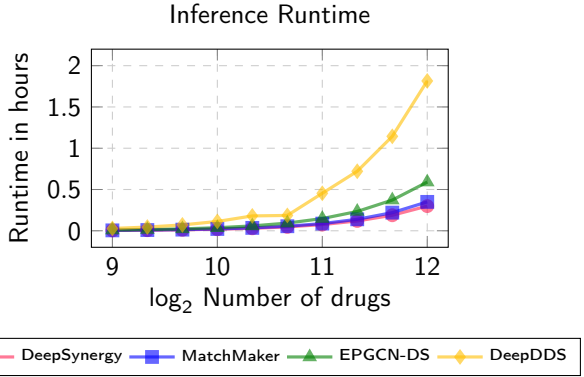


Figure 2: The average runtime of doing a scoring pass for all combinations in DrugBankDDI [11].

Conclusions

What are the main takeaways?

Having impact!

- ChemicalX is used for targeting haematological malignancies.
- Early oncology scientists in AZ are using ChemicalX self-service.

Having fun!

- Internal and external collaborations.
- Skills elevated for people in AZ who are not core machine learning.

What could be improved?

- Splits that take scaffolds into account.
- Geometric graph encoders.
- Scaling with locality sensitive hashing.

Thank you for the kind attention!

[1] Kuru Halil Brahim, Oznur Tastan, and Ercument Cicek. 2021. MatchMaker: A Deep Learning Framework for Drug Synergy Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2021).

[2] Xusheng Cao, Rui Fan, and Wanwen Zeng. 2020. DeepDrug: A General Graph-Based Deep Learning Framework for Drug Relation Prediction. *bioRxiv* (2020).

[3] Yukuo Cen, Zhenyu Hou, Yan Wang, Qibin Chen, et al. 2021. CogDL: An Extensive Toolkit for Deep Learning on Graphs. (2021).

[4] Xin Chen, Xien Liu, and Ji Wu. 2020. GCN-BMP: Investigating Graph Representation Learning for DDI Prediction Task. *Methods* 179 (2020), 47–54. Interpretable machine learning in bioinformatics.

[5] CSIRO's Data61. 2018. StellarGraph Machine Learning Library. `https://github.com/stellargraph/stellargraph`.

[6] Andreea Deac, Yu-Hsiang Huang, Petar Velickovic, Pietro Liò, and Jian Tang. 2019. Drug-Drug Adverse Effect Prediction with Graph Co-Attention. *ICML Workshop on Computational Biology* (2019).

[7] Matthias Fey and Jan E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

[8] Jonathan Godwin, Thomas Keck, Peter Battaglia, Victor Bapst, Thomas Kipf, et al. 2020. *Jraph: A Library for Graph Neural Networks in Jax*.

[9] Daniele Grattarola and Cesare Alippi. 2021. Graph Neural Networks in TensorFlow and Keras with Spektral. *IEEE Computational Intelligence Magazine* 16, 1 (2021), 99–106.

[10] Jun Hu, Shengsheng Qian, Quan Fang, et al. 2021. Efficient Graph Deep Learning in TensorFlow with TF Geometric. *arXiv preprint 2101.11552* (2021).

[11] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, et al. 2021. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. In *35th Conference on Neural Information Processing Systems*.

[12] Kexin Huang, Cao Xiao, Trong Nghia Hoang, Lucas M Glass, and Jimeng Sun. 2020. CASTER: Predicting Drug Interactions with Chemical Substructure Representation. *AAAI* (2020).

[13] Mufei Li, Jinjing Zhou, Jiajing Hu, Wenxuan Fan, Yangkang Zhang, Yaxin Gu, and George Karypis. 2021. DGL-LifeSci: An Open-Source Toolkit for Deep Learning on Graphs in Life Science. *ACS Omega* 6, 41 (2021), 27233–27238.

[14] Hui Liu, Wenhao Zhang, Bo Zou, Jinxian Wang, and Yuanyuan Deng. 2020. DrugCombDB: A Comprehensive Database of Drug Combinations Toward the Discovery of Combinatorial Therapy. *Nucleic acids research* 48 (2020), 871–881.

[15] Meng Liu, Youzhi Luo, Limei Wang, et al. 2021. DIG: A Turnkey Library for Diving into Graph Deep Learning Research. *Journal of Machine Learning Research* 22, 240 (2021), 1–9.

[16] Abe Motoki, Mihai Mororiu, Tomoya Otabi, Kenshin Abe, and Others. 2017. *Chainer Chemistry: A Library for Deep Learning in Biology and Chemistry*. `https://github.com/chainer/chainer-chemistry`

[17] Arnold K Nyamabo, Hui Yu, and Jian-Yu Shi. 2021. SSI–DDI: Substructure–Substructure Interactions for Drug–Drug Interaction Prediction. *Briefings in Bioinformatics* (2021).

[18] Kristina Preuer, Richard PI Lewis, Sepp Hochreiter, Andreas Bender, Krishna C Bulusu, and Günter Klambauer. 2018. DeepSynergy: Predicting Anti-Cancer Drug Synergy with Deep Learning. *Bioinformatics* 34, 9 (2018), 1538–1546.

[19] Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, et al. 2019. *Deep Learning for the Life Sciences*. O'Reilly Media.

[20] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. 2018. Deep Learning Improves Prediction of Drug–Drug and Drug–Food Interactions. *Proceedings of the National Academy of Sciences* 115, 18 (2018), E4304–E4311.

[21] Mengying Sun, Fei Wang, Olivier Elemento, and Jiayu Zhou. 2020. Structure-Based Drug-Drug Interaction Detection via Expressive Graph Convolutional Networks and Deep Sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13927–13928.

[22] Nicholas P Tatonetti, P Ye Patrick, Roxana Daneshjou, and Russ B Altman. 2012. Data-Driven Prediction of Drug Effects and Interactions. *Science translational medicine* 4, 125 (2012).

[23] Jinxian Wang, Wenhao Zhang, Siyuan Shen, Lei Deng, and Hui Liu. 2021. DeepDDS: Deep Graph Neural Network with Attention Mechanism to Predict Synergistic Drug Combinations. *bioRxiv* (2021).

[24] Minjie Wang, Lingfan Yu, Da Zheng, et al. 2019. Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs. (2019).

[25] Nuo Xu, Pinghui Wang, Long Chen, Jing Tao, and Junzhou Zhao. 2019. MR-GNN: Multi-Resolution and Dual Graph Neural Network for Predicting Structured Entity Interactions. *Proceedings of IJCAI* (2019).

[26] Bulat Zagidullin, Jehad Aldahdooh, Shuyu Zheng, Wenyu Wang, Yinyin Wang, et al. 2019. DrugComb: An Integrative Cancer Drug Combination Data Portal. *Nucleic acids research* 47, W1 (2019), W43–W51.

[27] Shuyu Zheng, Jehad Aldahdooh, Tolou Shadbahr, Yinyin Wang, Dalal Aldahdooh, et al. 2021. DrugComb Update: A More Comprehensive Drug Sensitivity Data Repository and Analysis Portal. *Nucleic Acids Research* (2021).

[28] Zhaocheng Zhu, Shengchao Liu, Chence Shi, et al. 2021. *TorchDrug: A Powerful and Flexible Machine Learning Platform for Drug Discovery*.