薬学知識グラフ上のヘテロな情報を利用した文献からの薬物相互作用抽出

浅田 真生　三輪 誠　佐々木 裕
豊田工業大学
{sd19501,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

文献からの薬物相互作用抽出において，近年では事前学習済み深層ニューラルモデルを用いた手法が広く用いられている．これらの手法は薬物メンション周辺の文脈情報のみを考慮しているが，文脈理解に薬物の背景知識が必要とされる薬学文献の読解においては，薬物にまつわるヘテロな情報を深層ニューラルモデルに考慮させることが必要であると考えられる．本研究では，薬物に関する様々な情報が登録されたヘテロ薬学知識グラフの表現を薬物相互作用抽出に援用する手法を提案する．SemEval-2013 Task 9.2 データセットで提案手法を学習・評価し，薬物のヘテロな情報を用いることで世界最高性能となる 85.02%の F 値を達成した．

1　はじめに

薬物相互作用とは，患者に複数の薬物を併用投与した際に，薬物の本来の作用が増強・弱減したり，予期せぬ副作用が生じる現象のことである．「根拠に基づく医療」[1] を実践し，投薬過誤を防ぐためには，薬学論文から薬物相互作用に関する知識を網羅的に抽出しておくことが重要であり，そのために文献からの薬物相互作用の自動抽出技術による支援が期待されている．

著者らはこれまで，深層ニューラルモデルによる文献からの薬物相互作用抽出に取り組む上で，従来の文献のみを対象にした抽出の限界を超えるためには，外部知識の活用が必須であると考え，外部知識を薬物相互作用抽出に取り入れる研究 [2]，様々なドメイン情報を含んだヘテロな薬物知識グラフを作成しリンク予測を行う研究 [3] を行ってきた．本稿ではこれらの研究を発展させ，ヘテロな薬学知識グラフを外部知識として利用した薬物相互作用抽出に取り組む．

本稿では，薬物のヘテロなドメイン情報を考慮した薬物エンティティ表現を，文献からの薬物相互作用抽出に利用する手法を提案する．このエンティティ表現は，薬物データベース DrugBank [4]，タンパク質データベース UniProt [5]，医療用語シソーラス MeSH [6]，パスウェイデータベース Small Molecule Pathway Database (SMPDB) [7] を用いて作成した薬物のヘテロなドメイン情報を含んだヘテロ薬学知識グラフに対して，リンク予測モデルを学習することで獲得した表現である．本研究の貢献は以下の通りである．

- ヘテロなドメイン情報を考慮した薬物エンティティ表現を利用した薬物相互作用抽出モデルを提案した．
- 提案した手法を SemEval-2013 Task 9.2 データセット [8] で評価し，薬物にまつわるヘテロなドメイン情報の活用により抽出性能の向上を達成した．

2　関連研究

2.1　深層ニューラルモデルによる薬物相互作用抽出

文献からの薬物相互作用抽出タスクでは，大規模な薬学文献を用いて事前学習した BioBERT [9], SciBERT [10], PubMedBERT [11] といったモデルを用いた手法が広く使用されている．ただし，これらの研究は入力文の文脈情報のみを利用しており，文書や薬物の背景にある様々なドメイン情報を考慮できていない．

2.2　薬物の分子構造と説明文を用いた薬物相互作用抽出

著者らは，入力文中の薬物メンションについて，薬物データベース DrugBank にアクセスして薬物の分子構造情報と説明文情報を獲得し，分子構造情報，説明文情報をそれぞれ Graph Neural Networks

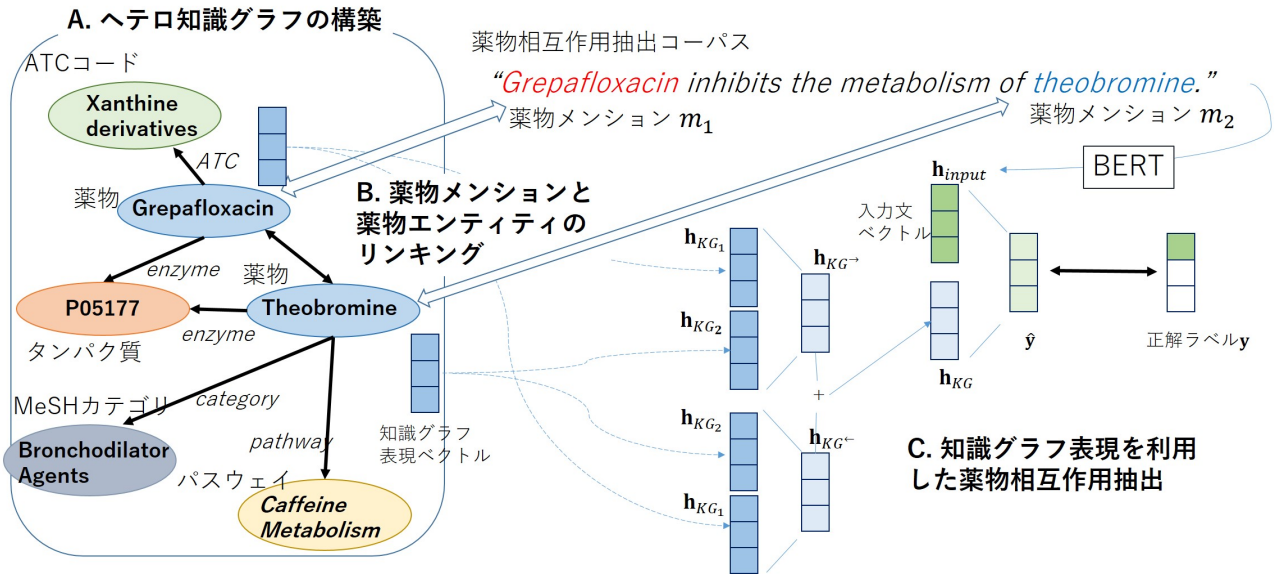

図 1　ヘテロ薬学情報を利用した薬物相互作用抽出モデル

(GNNs) [12], SciBERT [10] で表現し，入力文の表現と組み合わせて薬物相互作用抽出を行う手法 [2] を提案した．薬物の分子構造情報及び説明文情報を利用することでどちらの情報も薬物相互作用の抽出性能を向上できること，さらに，モデルのアンサンブルによって，分子構造情報と説明文情報の両者を考慮することで抽出性能をさらに向上できることを報告した．

2.3　ヘテロ薬学知識グラフの表現ベクトル獲得

著者らは，複数のデータベースを用いて異なる種類の情報を節点としたヘテロ薬学知識グラフを構築し，薬物にまつわるヘテロな情報を考慮したベクトル表現を獲得した [3]．作成したヘテロ薬学知識グラフの概要を図 1A に示す．知識グラフに含まれるノードの種類は以下に示す 5 種類である．

- **薬物**：薬物データベース DrugBank に登録された 10,000 件以上の薬物
- **タンパク質**：タンパク質データベース UniProt のうち，Swiss-Prot に登録されたタンパク質
- **パスウェイ**：パスウェイデータベース SMPDB に登録された 30,000 件以上のパスウェイ
- **MeSH カテゴリ**：薬物のカテゴリ情報を表す医療シソーラス MeSH のエントリ
- **ATC コード**：Anatomical Therapeutic Chemical (ATC) 分類システムによって階層化された薬物のカテゴリ情報

上記，5 種類のヘテロなノードは，以下に示す 8 種類の関係でリンクされる．

- *category*: 薬物ノードと MeSH カテゴリノードの関係．この関係は，薬物が MeSH によって定義された治療カテゴリーまたは一般カテゴリーに分類されることを示す．
- *ATC*: 薬物ノードと ATC コードノードの関係．例えば，薬物 *Morphine* は ATC コードとして A07DA52 が割り当てられているが，全ての親コード (A, A07, A07D, A07DA) のノードも *ATC* 関係でリンクされる．
- *pathway*: 薬物またはタンパク質とそれらが含まれているパスウェイの関係．薬物または酵素タンパク質が代謝，疾患，および生物学的パスウェイに関与している場合，薬物・タンパク質ノードとパスウェイのノードがリンクされる．
- *interact*: 薬物相互作用が報告されている薬物ペア間の関係．
- *target*: 薬物ノードとタンパク質ノードの関係．タンパク質，高分子，拡散，または特定の薬物が結合する小分子であり，結合した分子の正常な機能と望ましい治療効果の変化をもたらす場合，この関係でリンクされる．
- *enzyme*: 薬物ノードとタンパク質ノードの関係．特定の薬物が関与する化学反応を触媒するタンパク質の場合，この関係でリンクされる．
- *carrior*: 薬物ノードとタンパク質ノードの関係．薬物に結合して細胞輸送体に運ばれ，そこで細胞内に移動する分泌タンパク質の場合，この関

係でリンクされる

- *transporter*: 薬物ノードとタンパク質ノードの関係．イオン，小分子，または高分子を膜を越えて細胞内または細胞外にシャトルする膜結合タンパク質の場合，この関係でリンクされる．

作成したヘテロ薬学知識グラフは，有向グラフ $\mathcal{G} = (E, R, F)$ とみなすことができる．ここで，E はエンティティの集合，R は関係の集合，F は関係トリプルの集合である．関係トリプルは (h, r, t) で表現され，h, r, t はそれぞれ先頭エンティティ，関係，末尾エンティティである．グラフ埋め込み表現の学習は，負例サンプリングによって行う．正例の関係トリプル集合を $\mathbb{D}^+$，擬似的に作成した負例の関係トリプル集合を $\mathbb{D}^-$ としたときに，損失関数は以下のように示される．

$$L_{KG} = \sum_{(h,r,t)\in\mathbb{D}^+\cup\mathbb{D})^-} \log\left(1+\exp(y\cdot f(h,r,t))\right) + \lambda||\Theta||_2^2 \tag{1}$$

ここで，$f(h, r, t)$ は関係トリプルのスコア関数である．スコア関数としては，TransE [13], DistMult [14], ComplEx [15], SimplE [16] を用いた．

3 提案手法

本研究では，薬物のヘテロな情報を文献からの薬物相互作用抽出タスクに活用するために，薬学ヘテロ知識グラフ表現を利用した薬物相互作用抽出手法を提案する．提案モデルの全体像を図 1 に示す．

3.1 知識グラフ表現を活用した薬物相互作用抽出

2.3 節で説明した知識グラフ表現を利用して，文献からの薬物相互作用抽出を行う．2 つの薬物メンション m_1, m_2 を含む入力文 $S = (w_1, w_2, \cdots, w_n)$ が与えられたとき，これを BERT エンコーダの入力とし，BERT の [CLS] トークンを入力文の表現ベクトル $\mathbf{h}_{input}$ とする．図 1B に示すように，入力文中の薬物メンション m_1, m_2 を知識グラフの薬物エンティティとリンキングを部分文字列一致によって行い，対応する知識グラフ埋め込みベクトル $\mathbf{h}_{KG_1}, \mathbf{h}_{KG_2}$ を得る．知識グラフの表現ベクトルと，入力文の表現ベクトルを図 1C に示すように組み合わせる．

$$\mathbf{h}_{KG\rightarrow} = W_1([\mathbf{h}_{KG_1}; \mathbf{h}_{KG_2}]) + b_1 \tag{2}$$

$$\mathbf{h}_{KG\leftarrow} = W_1([\mathbf{h}_{KG_2}; \mathbf{h}_{KG_1}]) + b_1 \tag{3}$$

ここで，$[;]$ はベクトルの連結，W^1, b^1 は重み行列とバイアスを表す．$\mathbf{h}_a$ と $\mathbf{h}_b$ の和を入力文表現 $\mathbf{h}_{input}$ と連結し，全結合層の入力とする．

$$\mathbf{h}_{KG} = \mathbf{h}_{KG\rightarrow} + \mathbf{h}_{KG\leftarrow} \tag{4}$$

$$\mathbf{h}_{FC} = W_2([\mathbf{h}_{input}; \mathbf{h}_{KG}]) + b_2 \tag{5}$$

$$\hat{\mathbf{y}} = \mathrm{Softmax}(\mathbf{h}_{FC}) \tag{6}$$

正解ラベル $\mathbf{y}$ との交差エントロピー損失を最小にするように学習を行う．

$$L = -\sum_{x=1}^{X} \mathbf{y}_x \log \hat{\mathbf{y}}_x \tag{7}$$

ここで，X は全インスタンス数を表す．

3.2 モデルのアンサンブルによるヘテロ情報の組み合わせ

薬物エンティティの知識グラフ表現 $\mathbf{h}_{KG_1}, \mathbf{h}_{KG_2}$ を，$\mathbf{h}_{mol_1}, \mathbf{h}_{mol_2}$ または $\mathbf{h}_{desc_1}, \mathbf{h}_{desc_2}$ に置き換えることで，分子構造または説明文を用いた手法 [2] となる．ここで，$\mathbf{h}_{mol_i}$ は薬物メンション $m_i (i = 1, 2)$ の分子構造情報を GNN で表現したベクトル，$\mathbf{h}_{desc_i}$ は説明文情報を BERT でエンコードした [CLS] トークンのベクトル表現である．分子構造表現，説明文表現，知識グラフ表現の組み合わせを深層ニューラルモデルのアンサンブルによって行う．M 個のモデルを用いる場合，それぞれのモデルの出力ベクトルは式 6 により計算され，これを $\hat{\mathbf{y}}_m$ $(m = 1, \ldots, M)$ とする．モデルの予測確率の平均を最終的な出力 $\hat{\mathbf{y}}_{all}$ とする．

$$\hat{\mathbf{y}}_{all} = \frac{1}{M}\sum_{m=1}^{M} \hat{\mathbf{y}}_m \tag{8}$$

4 実験設定

薬物相互作用抽出コーパスとして，SemEval-2013 Task 9.2 (DDIExtraction-2013 shared task) データセットを用いた．このデータセットは，薬物ペアを含んだ文から構成され，薬物ペアそれぞれに対して，以下に示す 4 種類の薬物相互作用が正解付けされている．データセットの内訳を付録 A の表 3 に示す．

- Mechanism 薬物ペアが薬物動態学敵作用を持つ．
- Effect 薬物ペアが薬力学敵作用を持つ．
- Advice 薬物ペアを併用する際の推奨を表す．
- Interaction (Int.) 薬物ペアが薬物相互作用を持つことのみを表す．

表 1 ヘテロ薬学知識グラフ上のリンク予測タスク精度 (MRR)

スコア関数	TransE	DistMult	ComplEx	SimplE
category	0.1978	0.2539	0.0905	0.0461
ATC	0.2929	0.2428	0.3326	0.3278
pathway	0.6741	0.6792	0.6956	0.7531
interact	0.3109	0.7730	0.8678	0.6215
target	0.0802	0.0738	0.0496	0.0815
enzyme	0.3262	0.2501	0.2103	0.1903
carrior	0.4155	0.2023	0.1533	0.1358
transporter	0.3576	0.2293	0.1942	0.2242
Average	0.3319	0.3380	0.3242	0.2873

コーパス中の薬物メンションとヘテロ知識グラフの薬物エンティティとの対応付けは，小文字に変換した後の部分文字列一致により行った．薬物メンションと文字列一致を行う対象として，DrugBank に登録された薬物の見出し語，製品名，製剤名，シノニムを用いた．コーパス中の訓練データの 2,242 種類の薬物メンションのうち，一致が取れたものは 90.5%，評価データの 854 種類の薬物メンションのうち，一致が取れたものは 91.1%であった．

入力文表現のための事前学習済みモデルとして，PubMedBERT [11] を用いた．SemEval-2013 Task 9.2 データセットは公式の開発データセットが提供されていないため，訓練データを 4:1 に分割して開発データセットを作成し，ハイパーパラメータチューニングを行った．ハイパーパラメータの決定後は，分割前の訓練データでモデルを再度学習し，評価データセットで評価を行った．

ヘテロ知識グラフデータセットを訓練・開発・テストに分割し，学習と評価を行った．知識グラフの埋め込み表現獲得のために使用したデータセットの統計を付録 A に示す．

5 結果

表 1 に，ヘテロ薬学知識グラフに対するグラフ埋め込み表現の性能を確認するために，リンク予測タスクの精度を示す．リンク予測の評価指標として，Mean Reciprocal Rank (MRR) を用いた．表 1 より，全ての関係ごとの MRR を平均すると，スコア関数として DisMult を用いた場合が最も性能が高かった．関係ラベルごとに MRR を見ると，スコア関数 TransE を用いた場合は，*interact* の関係において低い MRR を示した．これは，知識グラフ中の薬物相互作用の関係は対称であるが，TransE が対称の関係性を捉えられないからであると考えられる．

続いて，SemEval-2013 Task 9.2 データセット上で評価した文献からの薬物相互作用抽出の性能を表 2 に示す．評価指標として，マイクロ平均 F 値を用いた．SciFive [17] は，入力と出力をどちらもテキストフォーマットで取り扱うモデル Text-to-Text Transfer Transformer (T5) [18] を生物医学ドメイン向けに再学習したモデルである．SciBERT (Mol. + Desc.) [2] は，薬物の分子構造情報と説明文情報を用いた手法である．ヘテロ知識グラフの情報を用いた手法に着目すると，いずれのスコア関数で学習した知識グラフベクトル表現を用いた場合でも，PubMedBERT で入力文情報のみを用いる手法よりも高い F 値を示した．Ensemble (KG + Mol. + Desc.) は，開発データセットで最も高い性能を示したスコア関数によるモデルと，分子構造情報を用いたモデル及び説明文情報を用いたモデルの 3 つのモデルをアンサンブルによって組み合わせた手法である．知識グラフの情報を除いた Ensemble (Mol. + Desc.) と比較すると，分子構造情報と説明文情報に知識グラフの情報を加えることでさらに高い F 値が得られることが分かった．

表 2 薬物相互作用抽出タスクの性能

Method	P	R	F (%)
SciFive [17]	-	-	83.67
SciBERT (Mol. + Desc.) [2]	85.36	82.83	84.08
PubMedBERT (baseline)	82.55	82.63	82.59
+ KG (TransE)	84.03	83.86	83.94
+ KG (DistMult)	**85.38**	82.94	84.14
+ KG (ComplEx)	84.61	82.02	83.29
+ KG (SimplE)	84.26	84.26	84.26
Ensemble (Mol. + Desc.)	85.03	84.16	84.59
Ensemble (KG + Mol. + Desc.)	85.37	**84.67**	**85.02**

6 おわりに

本研究では，文献からの薬物相互作用抽出タスクにおいて，薬物にまつわるヘテロな情報の活用を目的とし，著者らが作成したヘテロ薬学知識グラフの埋め込み表現を利用するモデルを提案した．提案した手法を SemEval-2013 Task 9.2 データセットで学習・評価したところ，薬物に関するヘテロなドメイン情報を利用することで，これまでの最高性能を越え，85.02%の F 値を達成した．

今後は，知識グラフ表現の学習と入力文表現の学習を同時に行い，さらに知識グラフ表現と入力文表現の組み合わせ手法の改善を行う予定である．

謝辞

本研究は JSPS 科研費 JP20K11962 の助成を受けたものです．

参考文献

[1] David L Sackett. Evidence-based medicine. In **Seminars in perinatology**, Vol. 21, pp. 3–5. Elsevier, 1997.

[2] Masaki Asada, Makoto Miwa, and Yutaka Sasaki. Using drug descriptions and molecular structures for drug–drug interaction extraction from literature. **Bioinformatics**, Vol. 37, No. 12, pp. 1739–1746, 10 2020.

[3] Masaki Asada, Nallappan Gunasekaran, Makoto Miwa, and Yutaka Sasaki. Representing a heterogeneous pharmaceutical knowledge-graph with textual information. **Frontiers in Research Metrics and Analytics**, Vol. 6, , 2021.

[4] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. **Nucleic Acids Research**, Vol. 46, No. D1, pp. D1074–D1082, 11 2017.

[5] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. **Nucleic Acids Research**, Vol. 47, No. D1, pp. D506–D515, 11 2018.

[6] C. E. Lipscomb. Medical Subject Headings (MeSH). **Bulletin of the Medical Library Association**, Vol. 88, No. 3, pp. 265–266, Jul 2000. 10928714.

[7] T. Jewison, Y. Su, F. M. Disfany, Y. Liang, C. Knox, A. Maciejewski, J. Poelzer, J. Huynh, Y. Zhou, D. Arndt, Y. Djoumbou, Y. Liu, L. Deng, A. C. Guo, B. Han, A. Pon, M. Wilson, S. Rafatnia, P. Liu, and D. S. Wishart. SMPDB 2.0: big improvements to the Small Molecule Pathway Database. **Nucleic Acids Research**, Vol. 42, No. Database issue, pp. D478–484, Jan 2014.

[8] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In **Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)**, pp. 341–350, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[9] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In **Proceedings of BioNLP 2019**, pp. 58–65, Florence, Italy, August 2019.

[10] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In **Proceedings of EMNLP-IJCNLP 2019**, pp. 3615–3620, Hong Kong, China, November 2019.

[11] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. **ACM Transactions on Computing for Healthcare (HEALTH)**, Vol. 3, No. 1, pp. 1–23, 2021.

[12] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. **Bioinformatics**, Vol. 35, No. 2, pp. 309–318, 2019.

[13] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In **Neural Information Processing Systems (NIPS)**, pp. 1–9, 2013.

[14] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. 2014.

[15] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In **International Conference on Machine Learning**, pp. 2071–2080. PMLR, 2016.

[16] Seyed Mehran Kazemi and David Poole. SimplE embedding for link prediction in knowledge graphs. In **Proceedings of the 32nd International Conference on Neural Information Processing Systems**, pp. 4289–4300, 2018.

[17] Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. SciFive: a text-to-text transformer model for biomedical literature, 2021.

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[19] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In **Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 2019.

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations**, 2019.

[21] Manuel Le Gallo, Abu Sebastian, Roland Mathis, Matteo Manica, Heiner Giefers, Tomas Tuma, Costas Bekas, Alessandro Curioni, and Evangelos Eleftheriou. Mixed-precision in-memory computing. **Nature Electronics**, Vol. 1, No. 4, pp. 246–253, 2018.

A　データセットの統計

各データセットの統計を表 3, 4 5 に示す.

表 3　SemEval-2013 Task 9.2 データセットの統計

	Train	Test
文書数	714	191
文数	6,976	1,299
薬物ペア数	27,792	5,716
正例ペア数	4,021	979
Mechanism	1,319	302
Effect	1,687	360
Advice	826	221
Int.	189	96
負例ペア数	23,771	4,737

表 4　ヘテロ薬学知識グラフのエンティティの統計

Entity type	#
Drug (DrugBank-ID)	11,516
Protein (Uniprot-ID)	5,339
Pathway (SMPDB-ID)	874
Category (MESH-ID)	2,166
ATC (ATC-code)	1,093
Total	20,988

表 5　ヘテロ薬学知識グラフの関係トリプルの統計

Relation type	ALL	train	valid	test
category	60,459	54,419	3,020	3,020
ATC	16,341	14,711	815	815
pathway	18,707	16,847	930	930
interact	2,682,142	2,413,932	134,105	134,105
target	18,467	16,627	920	920
enzyme	5,206	4,686	260	260
carrier	815	735	40	40
transporter	3,093	2,793	150	150
Total	2,750,228	2,525,829	140,240	140,240

B　学習設定

モデルの学習は NVIDIA 社 RTX A6000 が搭載されたマシンを用いて行った．ハイパーパラメータチューニングは Optuna [19] を用いて行った．オプティマイザとして AdamW [20] を使用し，学習率を 5e-05, ミニバッチサイズを 128，ドロップアウト率を 0.1 に設定した．ベクトル $\mathbf{h}_{KG}$ の次元数は，全てのスコア関数に対して 64 で共通とした．メモリ効率化のために mixed-precision 学習 [21] を採用した．