ĐẠI HỌC BÁCH KHOA HÀ NỘI

NGHIÊN CỨU TỐT NGHIỆP 1

Ứng dụng trí tuệ nhân tạo diễn giải trong tin-sinh học

PHAN HOÀNG HẢI
hai.ph225715@sis.hust.edu.vn

Công nghệ thông tin Việt - Nhật

Giảng viên hướng dẫn:	TS. Nguyễn Đức Anh
Chương trình đào tạo:	Công nghệ thông tin Việt - Nhật
Trường:	Công nghệ Thông tin và Truyền thông

HÀ NỘI, 06/2025

Chapter 1

GIỚI THIỆU VỀ CƠ SỞ DỮ LIỆU MOLECULENET, SIDER VÀ TOX21

1.1 Giới thiệu về MoleculeNet

MoleculeNet là một tập hợp các tập dữ liệu điểm chuẩn quy mô lớn, được thiết kế để đánh giá các phương pháp học máy trong việc dự đoán thuộc tính phân tử [10, 20]. Phát hành vào năm 2017 bởi nhóm Pande tại Đại học Stanford, MoleculeNet đã trở thành một điểm chuẩn quan trọng trong lĩnh vực khám phá thuốc bằng học máy [4]. Mục tiêu chính của nó là thúc đẩy sự phát triển của học máy phân tử thông qua việc tuyển chọn dữ liệu, cung cấp phần mềm đặc trưng hóa phân tử, và triển khai các thuật toán đã biết [9, 10].

MoleculeNet tích hợp dữ liệu từ nhiều cơ sở dữ liệu công cộng, bao gồm thông tin về hơn 700.000 hợp chất với đa dạng thuộc tính [8, 10]. Toàn bộ dữ liệu được tích hợp vào gói mã nguồn mở DeepChem để dễ dàng truy cập [8]. Các tập dữ liệu được phân loại thành bốn nhóm chính:

- **Cơ học lượng tử (Quantum Mechanics):** Cấu trúc 3D và thuộc tính tính toán bằng hóa học lượng tử [4, 8].
- **Hóa lý (Physical Chemistry):** Độ hòa tan trong nước, năng lượng tự do solvation, độ ưa béo [4, 25].
- **Sinh lý học (Physiology):** Khả năng xuyên hàng rào máu não (BBB) và dữ liệu độc tính như Tox21, ToxCast, SIDER [4, 25].

- **Lý sinh (Biophysics):** Liên kết protein-ligand, bao gồm PCBA, MUV, HIV, PDBbind, BACE [4, 25].

MoleculeNet cũng cung cấp các triển khai thuật toán học máy chất lượng cao, chứng minh hiệu quả của các biểu diễn học được, đặc biệt là mạng tích chập đồ thị [7]. Các tập dữ liệu bao gồm cả nhiệm vụ hồi quy (ví dụ: QM7, ESOL) và phân loại (ví dụ: Tox21, SIDER), đòi hỏi các số liệu hiệu suất khác nhau [7, 9].

1.2 Cơ sở dữ liệu SIDER

Cơ sở dữ liệu SIDER (Side Effect Resource) là một nguồn tài nguyên quan trọng chứa thông tin về các phản ứng bất lợi của thuốc (ADR) từ các loại thuốc đã được phê duyệt [7, 13]. Thông tin được trích xuất từ tài liệu công khai và tờ hướng dẫn sử dụng thuốc [13]. SIDER cung cấp chi tiết về tần suất tác dụng phụ, phân loại thuốc và tác dụng phụ, cùng với các liên kết đến thông tin thuốc-mục tiêu [13].

Phiên bản hiện tại là SIDER 4.1, phát hành ngày 21 tháng 10 năm 2015 [13]. Phiên bản này sử dụng từ điển MedDRA (phiên bản 16.1) cho các thuật ngữ ưu tiên và cấp thấp hơn [13]. So với SIDER 2, số lượng thuốc tăng từ 996 lên 1430, và các tác dụng phụ được truy xuất tốt hơn; SIDER 4.1 đã loại bỏ các tác dụng phụ chỉ được đề cập là tiềm năng hoặc không xảy ra [13].

SIDER phân loại tác dụng phụ thành 27 lớp hệ thống cơ quan [25]. Tuy nhiên, dữ liệu được cập nhật đến năm 2015 và không còn được cấp kinh phí phát triển, do đó thông tin có thể lỗi thời [13]. Mặc dù vậy, SIDER vẫn là một nguồn tài nguyên giá trị cho nghiên cứu và giáo dục về tác dụng phụ của thuốc [13]. Trong học máy, SIDER là một tập dữ liệu phân loại trong bộ điểm chuẩn MoleculeNet, hỗ trợ dự đoán các phản ứng có hại của thuốc [7, 25].

1.3 Cơ sở dữ liệu Tox21

Tox21 là một chương trình hợp tác liên bang Hoa Kỳ (EPA, FDA, NIH) nhằm phát triển các phương pháp đánh giá nhanh chóng và hiệu quả sự an toàn của hóa chất thương mại và sản phẩm y tế [18].

Tập dữ liệu Tox21 bao gồm 12.060 mẫu huấn luyện và 647 mẫu kiểm tra, đại diện cho các hợp chất hóa học [3, 5, 17]. Mỗi mẫu có 12 nhãn nhị phân tương ứng với kết quả của 12 thí nghiệm độc tính khác nhau [3, 17]. Các phép đo độc tính định tính này bao gồm các thụ thể hạt nhân và con đường phản ứng căng thẳng [18, 25]. Cụ thể, các thí nghiệm tập trung vào các thụ thể hạt nhân (AhR, AR, ERα, FXR, GR,

PPARδ, PPARγ, TR, VDR) và các con đường phản ứng căng thẳng (p53, NF-κB, pH2AX, căng thẳng lưới nội chất, tiềm năng màng ty thể, ARE/Nrf-2, phản ứng sốc nhiệt, tổn thương DNA) [18].

Dữ liệu hóa học được biểu diễn bằng 801 "dense features" (khối lượng phân tử, độ hòa tan) và 272.776 "sparse features" (cấu trúc con hóa học như ECFP10, DFS6, DFS8) [3, 17]. Ma trận nhãn chứa nhiều giá trị bị thiếu (NA) [3].

Dữ liệu Tox21 được công bố rộng rãi qua PubChem (NLM) và ToxCast (EPA) [18, 11]. NCATS cũng cung cấp trình duyệt kho hóa chất Tox21 miễn phí [18, 6]. ToxCast là một tập dữ liệu độc tính khác từ cùng sáng kiến, cung cấp dữ liệu từ sàng lọc *in vitro* thông lượng cao [7, 8].

Tox21, là một phần của MoleculeNet, thường được sử dụng làm điểm chuẩn trong các nghiên cứu học máy để dự đoán độc tính hóa học, góp phần rút ngắn thời gian đưa thuốc ra thị trường và giảm thử nghiệm trên động vật [14, 11].

Chapter 2

PHÂN TÍCH THỰC NGHIỆM VÀ KẾT QUẢ

2.1 Phương pháp dự đoán tác dụng phụ của thuốc

Mô hình học máy là XGBoost (eXtreme Gradient Boosting) kết hợp với các đặc trưng phân tử (molecular descriptors) để dự đoán các tác dụng phụ của thuốc dựa trên dữ liệu từ cơ sở dữ liệu SIDER.

2.1.1 XGBoost

XGBoost thuộc nhóm gradient boosting dựa trên cây quyết định. Kết hợp của các cây quyết định yếu (weak learners) được huấn luyện tuần tự, mỗi cây cố gắng khắc phục lỗi của các cây trước đó, giúp XGBoost đạt được độ chính xác cao.

2.1.2 Đặc trưng phân tử (Molecular Descriptors)

Để mô hình học máy có thể hiểu và xử lý thông tin về cấu trúc hóa học của các hợp chất, chúng ta cần chuyển đổi các chuỗi SMILES (Simplified Molecular Input Line Entry System) thành các biểu diễn số học, gọi là đặc trưng phân tử. Trong nghiên cứu này, chúng tôi đã sử dụng một tập hợp các đặc trưng phân tử toàn diện, bao gồm:

- **Morgan Fingerprints (kích thước 1024 bit, bán kính 2):** structural fingerprint dựa trên thuật toán Morgan, được sử dụng rộng rãi để mã hóa

thông tin về các môi trường nguyên tử xung quanh từng nguyên tử trong phân tử.

- **MACCS Keys (166 bit):** MACCS Keys là một tập hợp 166 khóa cấu trúc, mỗi khóa tương ứng với sự hiện diện hoặc vắng mặt của một motif hóa học.
- **Các đặc trưng hóa lý thiết yếu (14 đặc trưng):** 14 đặc trưng hóa lý quan trọng, thường được sử dụng trong đánh giá ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity). Các đặc trưng này bao gồm:
  - Khối lượng phân tử (Molecular Weight)
  - LogP (hệ số phân bố octanol-nước, biểu thị độ ưa béo)
  - Số liên kết hydro cho (NumHDonors)
  - Số liên kết hydro nhận (NumHAcceptors)
  - Số liên kết quay được (NumRotatableBonds)
  - Diện tích bề mặt cực topo (TPSA - Topological Polar Surface Area)
  - Số vòng thơm (NumAromaticRings)
  - Số vòng bão hòa (NumSaturatedRings)
  - Tổng số vòng (RingCount)
  - Số dị nguyên tử (NumHeteroatoms)
  - Chỉ số phức tạp phân tử BertzCT (BertzCT)
  - Chỉ số linh hoạt HallKierAlpha (HallKierAlpha)
  - Diện tích bề mặt tiếp cận LabuteASA (LabuteASA)
  - Chỉ số trạng thái điện topo lớn nhất (MaxAbsEStateIndex)

Tổng cộng, mỗi hợp chất được biểu diễn bằng một vector đặc trưng có kích thước 1204 (1024 từ Morgan Fingerprints, 166 từ MACCS Keys, và 14 từ các đặc trưng hóa lý). Các vector đặc trưng này sau đó được sử dụng làm đầu vào cho mô hình XGBoost để dự đoán khả năng gây ra từng tác dụng phụ.

2.1.3 Kết quả dự đoán tác dụng phụ

Hiệu suất của mỗi mô hình được đánh giá bằng chỉ số ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) và báo cáo phân loại chi tiết (precision, recall, f1-score, accuracy). Kết quả được tổng hợp trong Bảng 2.1.

adjustbox

Table 2.1: Kết quả đánh giá mô hình XGBoost trên các lớp tác dụng phụ của SIDER

Tác dụng phụ	**ROC-AUC**	**Precision**	**Recall**	**F1-score**	**Ac**
Hepatobiliary disorders	0.7257	0.66	0.66	0.66	
Metabolism and nutrition disorders	0.5656	0.57	0.55	0.55	
Product issues	0.5100	0.49	0.50	0.49	
Eye disorders	0.7136	0.65	0.65	0.65	
Investigations	0.6462	0.55	0.53	0.53	
Musculoskeletal and connective tissue disorders	0.6302	0.60	0.58	0.59	
Gastrointestinal disorders	0.6592	0.67	0.59	0.61	
Social circumstances	0.6447	0.64	0.57	0.58	
Immune system disorders	0.6157	0.57	0.55	0.55	
Reproductive system and breast disorders	0.7505	0.68	0.68	0.67	
Neoplasms benign, malignant and unspecified	0.7153	0.70	0.64	0.65	
General disorders and administration site conditions	0.6308	0.54	0.52	0.52	
Endocrine disorders	0.7315	0.70	0.62	0.63	
Surgical and medical procedures	0.5517	0.58	0.52	0.51	
Vascular disorders	0.6501	0.54	0.52	0.52	
Blood and lymphatic system disorders	0.7223	0.67	0.64	0.65	
Skin and subcutaneous tissue disorders	0.6104	0.72	0.56	0.59	
Congenital, familial and genetic disorders	0.6638	0.68	0.60	0.61	
Infections and infestations	0.6001	0.53	0.52	0.52	
Respiratory, thoracic and mediastinal disorders	0.5306	0.50	0.50	0.50	
Psychiatric disorders	0.6947	0.60	0.57	0.57	
Renal and urinary disorders	0.6307	0.58	0.57	0.57	
Pregnancy, puerperium and perinatal conditions	0.6309	0.58	0.54	0.55	
Ear and labyrinth disorders	0.6687	0.63	0.63	0.63	
Cardiac disorders	0.6724	0.56	0.55	0.55	
Nervous system disorders	0.6535	0.52	0.51	0.51	
Injury, poisoning and procedural complications	0.5913	0.56	0.54	0.54	
Trung bình	**0.6534**	**0.61**	**0.58**	**0.59**	
Trung vị	**0.6462**	**0.60**	**0.57**	**0.57**	
Min	**0.5100** (Product issues)	**0.49**	**0.50**	**0.49**	
Max	**0.7505** (Reproductive system)	**0.72**	**0.69**	**0.67**	

2.1.4 Phân tích và So sánh

Dựa trên kết quả thực nghiệm, mô hình XGBoost của chúng tôi đạt điểm ROC-AUC trung bình là 0.6534 trên tập dữ liệu SIDER. Điểm ROC-AUC cao nhất đạt được là

0.7505 cho lớp "Reproductive system and breast disorders", trong khi thấp nhất là 0.5100 cho lớp "Product issues". Sự biến động này cho thấy hiệu suất của mô hình khác nhau đáng kể giữa các lớp tác dụng phụ, có thể do sự mất cân bằng lớp (class imbalance) hoặc độ phức tạp vốn có của từng loại tác dụng phụ.

Các mô hình tiên tiến (State-of-the-Art - SOTA) trên tập dữ liệu SIDER:

- BioAct-Het: 91.11 [1]
- Deep-CBN: 78.2 [1]
- MolXPT: 71.7 [1]
- IterRefLSTM: 70.40 [1]
- ChemRL-GEM: 67.2 [1]

Chapter 3

Tiếp cận elEmBERT: Biểu diễn và Embedding Hợp chất Hóa học

Mô hình elEmBERT [**?**] ứng dụng kiến trúc bộ mã hóa Transformer của BERT để tạo các biểu diễn vector (embeddings) cho hợp chất hóa học, từ đó dự đoán các thuộc tính trên các tập dữ liệu như SIDER và Tox21.

3.0.1 Biểu diễn Phân tử thành Chuỗi Token Hóa học

Quá trình chuyển đổi cấu trúc hóa học thành chuỗi token đầu vào cho BERT gồm các bước:

1. **Từ SMILES đến Cấu trúc 3D và Hàm Phân bố Cặp (PDFs):** Chuỗi SMILES được chuyển đổi sang cấu trúc 3D. Từ đó, Hàm Phân bố Cặp Nguyên tử (PDFs) được tính cho từng nguyên tử, mô tả môi trường hóa học cục bộ của chúng.

2. **Token hóa "Tiểu Nguyên tố" (Sub-elements) - Mô hình V1:** Đây là điểm khác biệt chính. Thay vì chỉ sử dụng loại nguyên tố (ví dụ: C), mỗi nguyên tố được phân cụm thành các "tiểu nguyên tố" dựa trên PDF và trạng thái oxy hóa của chúng (ví dụ: C_{methyl}, C_{aromatic}). Điều này tạo ra một từ điển token giàu thông tin hơn (tới 565 token), phản ánh sự đa dạng về môi trường hóa học của cùng một loại nguyên tố.

3. **Token Đặc biệt:** Giống BERT chuẩn, token `[CLS]` được thêm vào đầu chuỗi để thu thập biểu diễn tổng hợp của phân tử.

Kết quả là một chuỗi các token hóa học (tiểu nguyên tố và token đặc biệt), sẵn sàng cho BERT.

3.0.2 BERT Embedding: Học Biểu diễn Phân tử

Chuỗi token được đưa vào các lớp mã hóa Transformer của BERT để tạo ra các embedding ngữ cảnh:

1. **Lớp Embedding Ban đầu:** Mỗi token t_i được ánh xạ tới một vector embedding ban đầu e_i.

2. **Cơ chế Tự Chú ý Đa đầu (Multi-Head Self-Attention):** Là cốt lõi của BERT, cơ chế này cho phép mô hình học các mối quan hệ tương tác hóa học giữa các "tiểu nguyên tố" trong phân tử. Đối với mỗi "đầu" chú ý j:

   - Các vector Query (Q_j), Key (K_j), Value (V_j) được tạo từ ma trận embedding E (hoặc đầu ra lớp trước):

   $$Q_j = EW_j^Q, \quad K_j = EW_j^K, \quad V_j = EW_j^V$$

   - Attention weights:

   $$\text{AttentionScores}_j = \text{softmax}\left(\frac{Q_j K_j^T}{\sqrt{d_k}}\right)$$

   - Attention weights là tổng có trọng số của Value:

   $$\text{Head}_j = \text{AttentionScores}_j \cdot V_j$$

   Các đầu ra từ h đầu được ghép nối và chiếu qua ma trận trọng số W^O:

   $$\text{MultiHead}(E) = \text{Concat}(\text{Head}_1, \ldots, \text{Head}_h)W^O$$

   Residual connection và chuẩn hóa lớp được áp dụng sau mỗi bước:

   $$E' = \text{LayerNorm}(E + \text{MultiHead}(E))$$

3. **Feed-Forward Network - FFN:** E' được xử lý qua FFN (hai lớp tuyến tính với hàm kích hoạt):

   $$\text{FFN}(E') = \max(0, E'W_1 + b_1)W_2 + b_2$$

   Tiếp tục thực hiện lại LayerNorm và Residual Connection:

   $$E_{out} = \text{LayerNorm}(E' + \text{FFN}(E'))$$

Đặc điểm hóa học quan trọng:

- **Bỏ qua Positional Embeddings:** elEmBERT không sử dụng embedding vị trí. Điều này phản ánh tính hoán vị bất biến của phân tử (thứ tự nguyên tử không làm thay đổi thuộc tính hóa học).
- **Embedding Phân tử từ `[CLS]` Token:** Vector đầu ra của token `[CLS]` từ lớp mã hóa cuối cùng ($E_{\texttt{[CLS]}}$) được dùng làm biểu diễn embedding tổng hợp cho toàn bộ phân tử, chứa thông tin cấu trúc và hóa học đã được mã hóa. $E_{\texttt{[CLS]}}$ này sau đó được đưa vào lớp phân loại để dự đoán.

3.0.3 Kết quả trên SIDER và Tox21

Hiệu suất được đánh giá bằng ROC-AUC.

Table 3.1: Kết quả ROC-AUC trung bình của elEmBERT trên tập dữ liệu SIDER và Tox21 . V0: element embeddings, V1: sub-element embeddings.

Tập dữ liệu	**elEmBERT V0**	**elEmBERT V1**	**SOTA Model**
SIDER	0.778	0.773	0.659 (Li et al. [2022])
Tox21	0.965	**0.967**	0.860 (Li et al. [2021])

3.1 Tiếp cận Meta-MGNN: Học đồ thị Few-Shot để dự đoán thuộc tính phân tử

Trong bối cảnh khám phá thuốc hiện đại, các mô hình học sâu thường yêu cầu lượng lớn dữ liệu gán nhãn, điều này hiếm khi có sẵn cho các thuộc tính phân tử mới hoặc ít được nghiên cứu. Để giải quyết vấn đề này, Guo và cộng sự (2021) [**?**] đã đề xuất **Meta-MGNN** (Meta-Molecular Graph Neural Network), một mô hình học đồ thị few-shot để dự đoán thuộc tính phân tử. Meta-MGNN kết hợp mạng nơ-ron đồ thị (GNN) với khung meta-learning, mô-đun tự giám sát và cơ chế chú ý nhận biết tác vụ.

3.1.1 Kiến trúc và Cơ chế hoạt động của Meta-MGNN

Biểu diễn Đồ thị Phân tử

Mỗi phân tử G được biểu diễn dưới dạng đồ thị $G = (V, E)$, trong đó V là tập hợp các nút (nguyên tử hóa học) và E là tập hợp các cạnh (liên kết hóa học). Mỗi nút

và cạnh có các thuộc tính ban đầu (ví dụ: số nguyên tử, thẻ tính chirality cho nút; loại liên kết, hướng liên kết cho cạnh), như được trình bày chi tiết trong Bảng 1 của bài báo gốc.

Học Biểu diễn Phân tử bằng GNN

Meta-MGNN sử dụng Mạng Isomorphism Graph (GIN) làm GNN cơ sở để học biểu diễn của các nút và đồ thị.

- **Cập nhật biểu diễn nút:** Biểu diễn $h_v^{(l)}$ của nút v ở lớp l được cập nhật bằng cách tổng hợp thông tin từ các nút lân cận $N(v)$ và các cạnh liên quan.

  $$h_v^{(l)} = \text{AGG}^{(l)}(\{h_u^{(l-1)} : u \in N(v)\}, \{h_e^{(l-1)} : e = (v, u)\})$$

  trong đó AGG là hàm tổng hợp và σ là hàm kích hoạt phi tuyến (ví dụ: LeakyReLU).

- **Biểu diễn đồ thị cấp độ:** Sau khi thông tin được truyền qua L lớp GNN, biểu diễn cấp độ đồ thị h_G cho toàn bộ phân tử được tính bằng cách lấy trung bình các biểu diễn nút cuối cùng:

  $$h_G = \text{MEAN}(\{h_v^{(L)} : v \in V\})$$

  Biểu diễn h_G này sau đó được đưa vào một bộ phân loại (ví dụ: MLP) để dự đoán thuộc tính phân tử.

Học Meta (Meta-learning) và Tự giám sát (Self-supervised learning)

Meta-MGNN tích hợp khung meta-learning dựa trên MAML [**?**] để học các tham số mô hình θ sao cho chúng có thể nhanh chóng thích nghi với các tác vụ dự đoán thuộc tính phân tử mới chỉ với một số ít mẫu.

- **Cập nhật tham số bên trong (Inner-loop update):** Đối với một tác vụ T_τ và tập support S_τ, các tham số GNN được cập nhật để tối ưu hóa lỗi tác vụ:

  $$\theta' = \theta - \alpha \nabla L_T(\theta)$$

  trong đó α là tốc độ học và $L_T(\theta)$ là hàm lỗi tổng hợp cho tác vụ T_τ.

- **Học tự giám sát:** Để khai thác dữ liệu không gán nhãn và cải thiện biểu diễn phân tử, Meta-MGNN tích hợp hai tác vụ tự giám sát phụ trợ:

– **Tái tạo liên kết (Bond Reconstruction):** Dự đoán sự tồn tại của liên kết giữa các cặp nút. Hàm lỗi được định nghĩa là cross-entropy nhị phân:

$$L_{\text{edge}}(\theta) = -\frac{1}{|\mathcal{E}_s|} \sum_{(u,v)\in\mathcal{E}_s} \text{BINARYCROSSENTROPY}(l_{uv}, \hat{c}_{uv})$$

với $\hat{c}_{uv} = h_u \cdot h_v$ là điểm tái tạo liên kết giữa nút u và v.

– **Dự đoán loại nguyên tử (Atom Type Prediction):** Dự đoán loại của một nguyên tử dựa trên ngữ cảnh cục bộ của nó. Hàm lỗi được định nghĩa là cross-entropy:

$$L_{\text{node}}(\theta) = -\frac{1}{|\mathcal{V}_{\text{sel}}|} \sum_{v\in\mathcal{V}_{\text{sel}}} \text{CROSSENTROPY}(\hat{t}_v, t_v)$$

với $\hat{t}_v$ là dự đoán loại nguyên tử v, được tính từ một MLP trên trung bình của các embedding lân cận của v.

- **Hàm lỗi tổng hợp (Joint Loss):** Hàm lỗi tổng hợp cho tác vụ T_τ trong quá trình meta-training là:

$$L_T(\theta) = L_{\text{node}}(\theta) + \lambda_1 L_{\text{edge}}(\theta) + \lambda_2 L_{\text{label}}(\theta)$$

trong đó L_{label} là lỗi dự đoán thuộc tính chính (cross-entropy), và λ_1, λ_2 là các tham số cân bằng.

- **Cập nhật tham số bên ngoài (Outer-loop update):** Các tham số mô hình chính θ được cập nhật dựa trên lỗi của tập query từ tất cả các tác vụ trong batch, có tính đến trọng số tác vụ (task-aware attention).

Cơ chế chú ý nhận biết tác vụ (Task-aware Attention)

Để phản ánh tầm quan trọng khác nhau của các tác vụ, Meta-MGNN giới thiệu cơ chế chú ý nhận biết tác vụ để gán trọng số $\eta(T)$ cho mỗi tác vụ T:

$$\eta(T) = \frac{\exp(\text{MLP}(H_T))}{\sum_{T'\in T} \exp(\text{MLP}(H_{T'}))}$$

trong đó H_T là embedding của tác vụ T, được tính bằng cách lấy trung bình các embedding phân tử của tập query thuộc tác vụ đó.

3.1.2 Kết quả và Phân tích trên SIDER và Tox21

Meta-MGNN được đánh giá trên hai tập dữ liệu công khai Tox21 và SIDER, trong đó hiệu suất được đo bằng chỉ số ROC-AUC. Bảng 3.2 tổng hợp kết quả ROC-AUC trung bình cho cả hai tập dữ liệu trong các kịch bản few-shot (1-shot và 5-shot).

Table 3.2: Hiệu suất ROC-AUC trung bình của Meta-MGNN so với các phương pháp khác trên Tox21 và SIDER [**?**].

Tập dữ liệu	**Kịch bản**	**Meta-MGNN (ROC-AUC)**	**SOTA trước đó (ROC-AUC)**	**Cả**
Tox21	1-shot	**0.7687** (so với EGNN 0.7581)	0.7581 (EGNN)	
	5-shots	**0.7802** (so với Seq3seq 0.7718)	0.7718 (Seq3seq)	
SIDER	1-shot	**0.7334** (so với PreGNN 0.7154)	0.7154 (PreGNN)	
	5-shots	**0.7472** (so với PreGNN 0.7285)	0.7285 (PreGNN)	

Phân tích kết quả:

- **Hiệu suất vượt trội:** Meta-MGNN thể hiện hiệu suất vượt trội so với tất cả các phương pháp cơ sở trên cả hai tập dữ liệu Tox21 và SIDER, đặc biệt là trong các kịch bản few-shot. Điều này chứng tỏ hiệu quả của việc kết hợp GNN với meta-learning để giải quyết vấn đề thiếu dữ liệu.
- **Cải thiện đáng kể:**
  - Trên **Tox21**, Meta-MGNN đạt mức cải thiện trung bình +1.04% cho 1-shot và +0.84% cho 5-shots so với các phương pháp tốt nhất trước đó.
  - Trên **SIDER**, mức cải thiện thậm chí còn ấn tượng hơn với +1.80% cho 1-shot và +1.87% cho 5-shots.
- **Tính ổn định và hiệu quả của các thành phần:** Các nghiên cứu loại bỏ (ablation studies) trong bài báo gốc xác nhận rằng mỗi thành phần của Meta-MGNN (bao gồm pre-training GNN, mô-đun tự giám sát và cơ chế chú ý nhận biết tác vụ) đều đóng góp đáng kể vào việc cải thiện hiệu suất mô hình. Việc pre-training GNN giúp khởi tạo tham số tốt hơn, trong khi mô-đun tự giám sát và chú ý nhận biết tác vụ giúp mô hình khai thác hiệu quả hơn thông tin không gán nhãn và sự đa dạng giữa các tác vụ.