Project Final Report

Drug repurposing based on network

เสนอ

อ.ดร. ดวงดาว วิชาดากุล

รายชื่อผู้จัดทำ

1. มารีนญา	ตะโจปะรัง	6231352421
2. ณิชกานต์	ชัยพจนา	6231322621
3. ดรากรณ์	ผดุงพัฒโนดม	6231323221

โครงการนี้เป็นส่วนหนึ่งของรายวิชา 2110581 ชีวสารสนเทศ

ภาคการศึกษาต้น ปีการศึกษา 2564

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

Table of contents

List of tables

Introduction

ปัจจุบันการพัฒนายาใด ๆ ขึ้นมาใหม่เป็นเรื่องที่ต้องใช้เวลาและทุนจำนวนมาก จึงได้มีแนวคิดในการประยุกต์ใช้ยาเดิมที่มีอยู่เพื่อพัฒนาประสิทธิภาพให้มากขึ้น ซึ่งเราทราบว่ายาใด ๆ จะมีผลต่อ protein ต่าง ๆ ทำให้คิดว่าเราสามารถนำยามาใช้ในโรคต่าง ๆ ได้มากขึ้น จึงเกิดเป็น drug repositioning หรือ drug repurposing ขึ้นมา ซึ่งเป็นการหาแนวโน้มใหม่ในการใช้ยาเดิม โดยในการหาความสัมพันธ์ต่าง ๆ ระหว่างยาชนิดใด ๆ และโรคต่าง ๆ ที่อาจรักษาได้ด้วยยานั้น สามารถใช้ความรู้ของด้านอื่น ๆ นอกจากความรู้ทางการแพทย์มาช่วย เพื่อเป็นเครื่องมือในการลดเวลาและต้นทุนได้ ความรู้ส่วนหนึ่งที่สามารถนำมาแก้ปัญหาส่วนนี้ได้คือเรื่องกราฟในทางคณิตศาสตร์ หรืออาจเรียกว่า network ก็ได้ จึงทำให้เกิดหัวข้อ drug repurposing based on network และในการหัวข้อนี้ยังอาจใช้ความรู้ด้านอื่น ๆ เช่น สถิติ รวมถึงความรู้ทางการแพทย์เพื่อหาความเป็นไปได้ที่มากที่สุดที่จะเพิ่มประสิทธิภายของยาได้

Data preparation

เก็บข้อมูลยาตัวที่ส่งผลต่อ covid-19 จาก chembl database และนำมา filter โดยใช้ standard type เป็น IC50 เพื่อแบ่งกลุ่มยาเหล่านั้นออกเป็น 3 กลุ่มโดยใช้เกณฑ์ของค่า standard value ได้แก่ active, inactive, intermediate โดยค่า standard value(IC50) ที่มีค่ามากกว่า 10,000 จะอยู่ในกลุ่ม inactive กรณีมีค่าน้อยกว่า 10,000 แต่มากกว่า 1,000 จะอยู่ในกลุ่ม intermediate กรณีมีค่าน้อยกว่า 1,000 จะอยู่ในกลุ่ม active หลังจากนั้นจะคำนวณค่า LogP, QED และ Binding affinity ของแต่ละกลุ่ม โดยค่า Binding affinity ได้ทำนายโดยใช้โมเดล BertDTA

	molecule_chembl_id	canonical_smiles	standard_value	bioactivity_class
0	CHEMBL187579	Cc1noc(C)c1CN1C(=O)C(=O)c2cc(C#N)ccc21	7200.0	intermediate
1	CHEMBL188487	O=C1C(=O)N(Cc2ccc(F)cc2Cl)c2ccc(I)cc21	9400.0	intermediate
2	CHEMBL185698	O=C1C(=O)N(CC2COc3ccccc3O2)c2ccc(I)cc21	13500.0	inactive
3	CHEMBL426082	O=C1C(=O)N(Cc2cc3ccccc3s2)c2ccccc21	13110.0	inactive
4	CHEMBL187717	O=C1C(=O)N(Cc2cc3ccccc3s2)c2c1cccc2[N+](=O)[O-]	2000.0	intermediate
...	...	...	...	...
128	CHEMBL2146517	COC(=O)[C@@]1(C)CCCc2c1ccc1c2C(=O)C(=O)c2c(C)c...	10600.0	inactive
129	CHEMBL187460	C[C@H]1COC2=C1C(=O)C(=O)c1c2ccc2c1CCCC2(C)C	10100.0	inactive
130	CHEMBL363535	Cc1coc2c1C(=O)C(=O)c1c-2ccc2c(C)cccc12	11500.0	inactive
131	CHEMBL227075	Cc1cccc2c3c(ccc12)C1=C(C(=O)C3=O)[C@@H](C)CO1	10700.0	inactive
132	CHEMBL45830	CC(C)C1=Cc2ccc3c(c2C(=O)C1=O)CCCC3(C)C	78900.0	inactive

Active

Table1: ตารางสรุปค่า LogP, QED และ binding affinity ของ active group

	LogP	QED	Binding affinity
count	15.000	15.000	15.000
mean	3.778	0.628	5.548
standard deviation	1.056	0.157	0.195
min	2.411	0.207	5.208
Percentile 25th	2.816	0.613	5.418
Percentile 50th	3.700	0.675	5.559
Percentile 75th	4.314	0.732	5.624
max	6.101	0.766	6.031

link

Intermediate

Table2: ตารางสรุปค่า LogP, QED และ binding affinity ของ intermediate group

	LogP	QED	Binding affinity
count	14.000	14.000	14.000
mean	3.594	0.569	5.550
standard deviation	1.166	0.158	0.275
min	1.127	0.287	5.071
Percentile 25th	3.319	0.490	5.326
Percentile 50th	3.687	0.553	5.603
Percentile 75th	4.025	0.665	5.704
max	6.305	0.862	6.071

link

Inactive

Table3: ตารางสรุปค่า LogP, QED และ binding affinity ของ inactive group

	LogP	QED	Binding affinity
count	104.000	104.000	104.000
mean	3.969	0.466	5.553
standard deviation	1.455	0.197	0.255
min	-0.055	0.036	5.055
Percentile 25th	3.220	0.305	5.408
Percentile 50th	3.996	0.438	5.531
Percentile 75th	4.808	0.651	5.718
max	7.052	0.925	6.287

link

Dataset preparation

เราได้แบ่ง dataset ออกเป็น 3 กลุ่มได้แก่

1. zinc250k เป็น dataset ที่รวบรวมโมเลกุลของยามากกว่าสองแสนห้าหมื่นโมเลกุล
2. Chembl จาก Chembl database ที่มีการทดลองว่ามี target protein เป็น SARS-CoV 3C-like protease
3. zinc250k รวมกับ Chembl

zinc250k

Table4: ตารางแสดงค่า LogP และ QED ของ pre-training dataset ของ zinc250k

	LogP	QED
count	249455.000	249455.000
mean	2.457	0.728
standard deviation	1.434	0.728
min	-6.876	0.062
Percentile 25th	1.575	0.655
Percentile 50th	2.606	0.764
Percentile 75th	3.487	0.832
max	8.252	0.936

link

Chembl

Table5: ตารางแสดงค่า LogP และ QED ของ pre-training dataset ของ chembl

	LogP	QED
count	133.000	133.000
mean	3.908	0.495
standard deviation	1.386	0.197
min	-0.055	0.036
Percentile 25th	3.220	0.336
Percentile 50th	3.770	0.486
Percentile 75th	4.667	0.665
max	7.052	0.925

link

Chembl และ zinc250k

Table6: ตารางแสดงค่า LogP และ QED ของ pre-training dataset ของ zinc250k รวมกับ chembl

	LogP	QED
count	249588.000	249588.000
mean	2.458	0.728
standard deviation	1.434	0.145
min	-6.876	0.036
Percentile 25th	1.575	0.655
Percentile 50th	2.606	0.764
Percentile 75th	3.487	0.832
max	8.252	0.936

link

Generate new drug

Zinc250k

ผู้จัดทำได้รวบรวมข้อมูลของ zinc250k แล้วนำไปสร้างออบเจ็คของคลาส dataset ของ torchdrug เพื่อที่จะนำมาเทรนโมเดลต่อไป โดยโมเดลที่ใช้มีกราฟ RGCN และ กราฟ GCPN โดยมีการกำหนดโมเดลดังนี้

Colab link: link

1. RGCN Graph:
   - Input dimension: Dataset.node_feature_dim (class Dataset)
   - Hidden dimension: [256,256,256,256]
   - Batch normalization: ไม่ได้ทำ batch normalization
2. GCPN Graph:
   - Model: RGCN
   - GraphAtom type: Dataset.atom_type
   - Criterion: nll (ไม่ได้ใช้ reanforcement)

การเทรนใช้จำนวน epoch เท่ากับ 5 ซึ่งมีการตรวจสอบแล้วว่าค่า loss มีค่าคงที่

Table7: ตารางแสดงผลลัพธ์เมื่อใช้ zinc250k เป็น pretraining dataset

Smiles	LogP	QED
CC=CC(C)C	2.219	0.452
CCC=C(C)C	2.363	0.452
C#CC=C=C=C	1.116	0.292
CC=C(C)CC	2.363	0.452
C#CC(C)=CC	1.586	0.411

link

Table8: ตารางแสดงค่า LogP และ QED ของผลลัพธ์เมื่อใช้ zinc250k เป็น pretraining dataset

	LogP	QED
count	66.000	66.000
mean	1.657	0.406
standard deviation	0.556	0.060
min	0.359	0.292
Percentile 25th	1.294	0.356
Percentile 50th	1.664	0.410
Percentile 75th	2.136	0.449
max	2.612	0.526

link

Chembl

ผู้จัดทำได้รวบรวมข้อมูลของ chembl แล้วนำไปสร้างออบเจ็คของคลาส dataset ของ torchdrug เพื่อที่จะนำมาเทรนโมเดลต่อไป โดยโมเดลที่ใช้มีกราฟ RGCN และ กราฟ GCPN โดยมีการกำหนดโมเดลดังนี้

Colab link: link

1. RGCN Graph:
   - Input dimension: Dataset.node_feature_dim (class Dataset)
   - Hidden dimension: [256,256,256,256]
   - Batch normalization: ไม่ได้ทำ batch normalization
2. GCPN Graph:
   - Model: RGCN
   - GraphAtom type: Dataset.atom_type
   - Criterion: nll (ไม่ได้ใช้ reanforcement)

การเทรนใช้จำนวน epoch เท่ากับ 450 ซึ่งมีการตรวจสอบแล้วว่าค่า loss มีค่าคงที่ หลังจากการเทรนได้มีการทำ reinforcement โดยมีการกำหนด GCPN Graph ใหม่ดังนี้

GCPN Graph:

Model: RGCN

GraphAtom type: Dataset.atom_type

Task: qed, plogp

Criterion: ppo (ใช้ reinforcement โดยเรียนรู้ค่าจาก task นั่นก็คือค่า qed และ logp)

reward discount rate: 0.9

agent update interval: 3

โดยการทำ reinforcement training ได้ใช้จำนวน epoch เท่ากับ 10 ซึ่งมีการตรวจสอบแล้วว่าค่า loss มีค่าคงที่

Table9: ตารางแสดงผลลัพธ์เมื่อใช้ chembl เป็น pretraining dataset

Smiles	LogP	QED
C=C(C)C(C)C	2.2185	0.452
CCCC(C)C	2.4425	0.525
CC(C)C(C)C	2.2984	0.498
CCC(S)CC	2.1048	0.542
CC1=CC=C1C	1.8926	0.452

link

Table10: ตารางแสดงค่า LogP และ QED ของผลลัพธ์เมื่อใช้ chembl เป็น pretraining dataset

	LogP	QED
count	100.000	100.000
mean	4.178	0.559
standard deviation	1.711	0.103
min	1.752	0.191
Percentile 25th	2.866	0.507
Percentile 50th	3.745	0.567
Percentile 75th	5.385	0.626

max	9.271	0.795

link

Zinc250k และ Chembl

ผู้จัดทำได้รวบรวมข้อมูลของ zinc250k รวมกับ chembl แล้วนำไปสร้างออบเจ็คของคลาส dataset ของ torchdrug เพื่อที่จะนำมาเทรนโมเดลต่อไป โดยโมเดลที่ใช้มีกราฟ RGCN และ กราฟ GCPN โดยมีการกำหนดโมเดลดังนี้

Colab link: link

1. RGCN Graph:
   Input dimension: Dataset.node_feature_dim (class Dataset)
   Hidden dimension: [256,256,256,256]
   Batch normalization: ไม่ได้ทำ batch normalization
2. GCPN Graph:
   Model: RGCN
   GraphAtom type: Dataset.atom_type
   Criterion: nll (ไม่ได้ใช้ reinforcement)

การเทรนใช้จำนวน epoch เท่ากับ 5 ซึ่งมีการตรวจสอบแล้วว่าค่า loss มีค่าคงที่

Table11: ตารางแสดงผลลัพธ์เมื่อใช้ zinc250k และ chembl เป็น pretraining dataset

Smiles	LogP	QED
CC=C(C)CC	2.363	0.452
CC#CC=CC	1.586	0.411
C#CCC(C)C	1.666	0.447
C#CC(=C)CC	1.586	0.445
CC#CC#CP	0.846	0.302

link

Table12: ตารางแสดงค่า LogP และ QED ของผลลัพธ์เมื่อใช้ zinc250k และ chembl เป็น pretraining dataset

	LogP	QED
count	56.000	56.000
mean	1.738	0.401
standard deviation	0.505	0.069
min	0.256	0.213
Percentile 25th	1.414	0.329
Percentile 50th	1.811	0.412
Percentile 75th	2.128	0.450
max	2.609	0.524

link

Calculate binding affinity

ผู้จัดทำได้ทำนายค่า binding affinity โดยใช้โมเดล BertDTA และตั้งค่า target protein เป็น SARS-CoV 3C-like protease

Fasta sequence:

SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDTVYCPRHVICTAEDMLNPNYEDLLIRKSNHSFLVQAGNV
QLRVIGHSMQNCLLRLKVDTSNPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNHTIKGSFLNGSCGSV
GFNIDYDCVSFCYMHHMELPTGVHAGTDLEGKFYGPFVDRQTAQAAGTDTTITLNVLAWLYAAVINGDRWFLNR
FTTTLNDFNLVAMKYNYEPLTQDHVDILGPLSAQTGIAVLDMCAALKELLQNGMNGRTILGSTILEDEFTPFDVVR
QCSGVTFQ

Colab link: link

Zinc250k

Table13: ตารางแสดงผลลัพธ์และค่า binding affinity เมื่อใช้ zinc250k เป็น pretraining dataset

Smiles	LogP	QED	Binding affinity
CC=CC(C)C	2.219	0.452	5.748
CCC=C(C)C	2.363	0.452	5.771
C#CC=C=C=C	1.116	0.292	6.008
CC=C(C)CC	2.363	0.452	5.967
C#CC(C)=CC	1.586	0.411	6.026

link

Table14: ตารางสรุปค่า LogP, QED และ binding affinity ของผลลัพธ์เมื่อใช้ zinc250k เป็น pretraining dataset

	LogP	QED	Binding affinity
count	56.000	56.000	56.000
mean	1.738	0.401	5.968
standard deviation	0.505	0.069	0.183
min	0.256	0.213	5.4001
Percentile 25th	1.414	0.329	5.861
Percentile 50th	1.811	0.412	5.995
Percentile 75th	2.128	0.450	6.078
max	2.609	0.524	6.300

link

Chembl

Table15: ตารางแสดงผลลัพธ์และค่า binding affinity เมื่อใช้ chembl เป็น pretraining dataset

Smiles	LogP	QED	Binding affinity
C=C(C)C(C)C	2.2185	0.452	5.729
CCCC(C)C	2.4425	0.525	5.994
CC(C)C(C)C	2.2984	0.498	5.780
CCC(S)CC	2.1048	0.542	6.211
CC1=CC=C1C	1.8926	0.452	5.867

link

Table16: ตารางสรุปค่า LogP, QED และ binding affinity ของผลลัพธ์เมื่อใช้ chembl เป็น pretraining dataset

	LogP	QED	Binding affinity
count	100.000	100.000	100.000
mean	4.178	0.559	5.849
standard deviation	1.711	0.103	0.230
min	1.752	0.191	5.249
Percentile 25th	2.866	0.507	5.693
Percentile 50th	3.745	0.567	5.888
Percentile 75th	5.385	0.626	6.012
max	9.271	0.795	6.289

link

Zinc250k และ Chembl

Table17: ตารางแสดงผลลัพธ์และค่า binding affinity เมื่อใช้ zinc250k และ chembl เป็น pretraining dataset

Smiles	LogP	QED	Binding affinity
CC=CC(C)C	2.219	0.452	5.967
CCC=C(C)C	2.363	0.452	6.036
C#CC=C=C=C	1.116	0.292	5.590
CC=C(C)CC	2.363	0.452	6.096
C#CC(C)=CC	1.586	0.411	6.194

link

Table18: ตารางสรุปค่า LogP, QED และ binding affinity ของผลลัพธ์เมื่อใช้ zinc250k และ chembl เป็น pretraining dataset

	LogP	QED	Binding affinity
count	56.000	56.000	56.000
mean	1.738	0.401	6.025
standard deviation	0.505	0.069	0.177
min	0.256	0.213	5.590
Percentile 25th	1.414	0.329	5.943
Percentile 50th	1.811	0.412	6.047
Percentile 75th	2.128	0.450	6.158
max	2.609	0.524	6.274

link

Conclusion

ผลของการสร้างยาใหม่โดยใช้กราฟ RGCN และ GCPN นั้นค่าของ LogP และ QED นั้นมีค่าไม่แตกต่างจาก dataset ที่ใช้เป็น pretraining ของแต่ละ dataset แต่มี standard variation ดีกว่า pretraining dataset ในส่วนของค่า binding affinity นั้นในส่วนของยาที่สร้างมาใหม่ทั้ง 3 ชุดมีค่า binding affinity พอๆกับของ chembl dataset ที่ได้มีการทดลองว่าสามารถจับกับโปรตีน SARS-CoV 3C-like protease ได้ดี

Google Colab link

Collect data from Chembl database: link

Analyze dataset: link

Generate new drug using zinc250k as a pretraining dataset: link

Generate new drug using chembl as a pretraining dataset: link

Generate new drug using zinc250k and chembl as a pretraining dataset: link

Predict binging affinity: link

Summary result: link

Project folder: link