Pusan National University
Computer Science and Engineering
Technical Report 2024-10

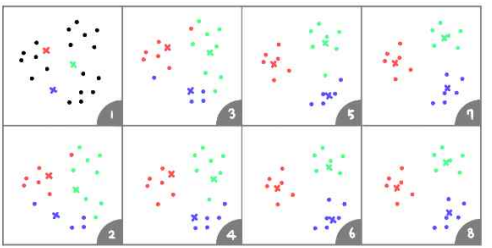
신약 개발 후보물질 발굴을 위한 거대 화합물 라이브러리 탐색 최적화

201924656 이정민
202155565 성가빈
201845928 최우영
지도교수 송길태

목 차

2.1.3. k-means 클러스터링

K-means 알고리즘은 주어진 데이터 포인트들을 K개의 클러스터로 나누는 비지도 학습 기법이다. 비지도 학습에서는 데이터 포인트들의 라벨을 지정하지 않고, 데이터의 패턴을 파악하게 한다.



[그림 1] k-means 클러스터링 과정

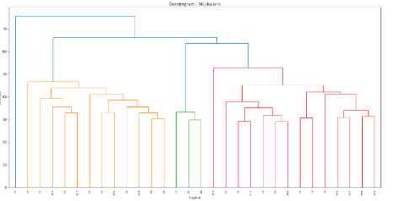
먼저 k개의 초기 클러스터 중심점(centroid)을 설정한다. 각 데이터 포인트는 가장 가까운 클러스터 중심점으로 할당되고, 나눠진 클러스터에서 중심점이 업데이트된다. 해당 과정을 중심점의 위치가 바뀌지 않을 때까지 반복한다. 이 방법은 계산 속도가 빠르지만, 초기 클러스터 중심점 설정에 따라 결과가 달라질 수 있다.

2.1.4. 병합 군집 클러스터링

병합 군집 알고리즘은 데이터 포인트들을 개별 클러스터로 시작하여, 가장 가까운 두 클러스터를 반복적으로 병합해 나가는 계층적 군집 기법이다.

클러스터 간의 거리 계산 방식에는 single, complete, average, centroid linkage 등이 존재한다. single linkage는 두 클러스터 간의 가장 가까운 거리를 사용하며, complete linkage는 가장 먼 거리를 사용한다. 그리고 average linkage는 각 클러스터의 모든 포인트 간의 거리의 평균을 사용한다. centroid linkage는 클러스터의 중심점 간의 거리를 사용하는 방식이다.

이처럼 다양한 거리 계산 방식을 적용할 수 있어 병합 군집 알고리즘은 유연성이 높다. 병합 군집 알고리즘의 결과로는 계층적인 트리 형태의 군집 구조를 만들어진다.



[그림 2] 병합 군집 클러스터링으로 만들어진 군집 구조

2.1.5. 베이지안 최적화

베이지안 최적화(bayesian optimization)는 알려지지 않은 목적 함수의 값을 최대 또는 최소로 만드는 입력 값을 찾는 것을 목표로 한다. 이 방법은 하이퍼파라미터 최적화에 주로 사용되며, 이 경우 하이퍼파라미터가 입력 값이 된다.

베이지안 최적화는 surrogate model과 acquisition function으로 구성되고, 이 두 과정을 반복하면서 목적 함수가 최대 또는 최소가 되는 지점을 찾게 된다. Surrogate model은 이전의 입력 값과 그에 따른 목적 함수의 값들을 이용해 목적 함수를 확률적으로 추정하며, GPR(Gaussian Process Regression)이 주로 사용된다. Acquisition function은 Surrogate model의 확률적 추정을 바탕으로 다음에 사용할 입력 값들을 추천한다.[5][6]

Acquisition function으로는 EI(Expected Improvement), PI(Probability of Improvement), UCB(Upper Confidence Bound) 등이 사용될 수 있다.[5][7] EI는 기대할 수 있는 목적 함수 값의 개선 정도가 가장 큰 다음 입력 값을 찾는다. f^*가 현재 가장 좋은 값일 때, $EI(x) = \exp[\max(0, f(x) - f^*)] = (\mu(x) - f^* - \zeta)CDF(\frac{\mu(x) - f^* - \zeta}{\sigma(x)}) + \sigma(x)PDF(\frac{\mu(x) - f^* - \zeta}{\sigma(x)})$ 로 계산된다.[5][7][8][9][10] PI는 선택했을 때 목적 함수 값이 개선될 가능성이 가장 큰 다음 입력 값을 찾는다. $PI(x) = P(\max(0, f(x) - f^*) > 0) = CDF(-\frac{f^* - \mu(x)}{\sigma(x)})$ 로 나타낼 수 있다.[5][7][8][9][10] UCB는 $UCB(x) = \beta\sigma(x) + \mu(x)$으로 계산되며, β값이 클수록 이전의 좋은 값 근처보다 아직 잘 모르는 값에 비중을 두고 입력 값을 선택한다.[5][7]

2.1.6. MEMES 알고리즘

MEMES(Machine learning framework for Enhanced MolEcular Screening)[11]는 Sarvesh Mehta등이 제시한 머신러닝 기반의 가상 스크리닝 프레임워크로, 베이지안 최적화를 통해 대규모 약물 라이브러리에서 잠재적인 신약 후보 물질을 효율적으로 식별하는 알고리즘이다.[11]

탐색 과정은 탐색 공간을 균형있게 커버할 수 있도록, 클러스터링된 리간드들을 이용해 각 클러스터에서 일정 수의 리간드를 무작위 선택하는 것으로 시작한다. 선택된 초기 리간드는 Surrogate model의 학습 데이터로 쓰이며, 이 모델을 이용해 리간드의 도킹 점수를 예측한다. 이후, 학습된 모델을 기반으로 Acquisition Function을 사용해 다음 탐색할 리간드를 선택한다. 여기서는 Surrogate Model로 GPR을 사용하였고, Acquisition Function으로 EI를 사용하였다. 이 과정은 최적의 리간드를 반복적으로 탐색하며 모델을 업데이트하는 방식으로 진행된다. 각 반복에서 새로운 리간드를 선택하고 도킹 점수를 예측해 탐색이 점진적으로 개선된다. 최종적으로, 모델에서 예측 점수가 높은 리간드를 선택하여 최종 결과를 도출한다.[11]

2.1.7. Diffdock

Diffdock[12]은 프로틴-리간드 도킹을 위한 딥러닝 기반 방법론으로 확산 모델(diffusion model)을 사용하여 단백질 결합 부위에 리간드의 3D 구조와 위치를 생성할 수 있다. Autodock Vina와 같이 분자 도킹 예측 모델이지만 둘은 접근 방식에 차이가 존재한다. Vina는 전통적인 분자 도킹 알고리즘을 사용하며, 주로 물리 기반 스코어링 함수와 최적화 알고리즘을 활용하지만 Diffdock은 확산 모델을 사용해 결합 예측을 더 정확하게 할 수 있다. [12]

2.2. 연구 일정 및 구성원 역할

2.2.1. 연구 일정

5월		6월				7월					8월					9월				10월		
4	5	1	2	3	4	1	2	3	4	5	1	2	3	4	5	1	2	3	4	1	2	3
착수 보고서 작성																						
	개발 환경 구축																					
		무작위 탐색 구현																				
				클러스터링 구현																		
							MEMES 알고리즘 구현															
									중간 보고서 작성													
													탐색 공간 확장									
									웹 애플리케이션 개발 & 시각화													
																				최종 보고서 작성		

2.2.2. 구성원 별 역할

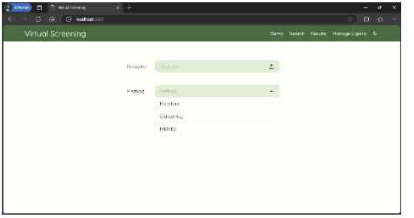
이름	진척도
이정민	보고서 작성 개발 환경 구축 및 관리 AutoDockTools, ChimeraX 등을 활용한 프로틴, 리간드 라이브러리 구축 spicy.cluster.hierarchy를 활용한 계층적 클러스터링 구현 MEMES에 Acquisition Function 추가 구현
성가빈	보고서 작성 rdkit을 이용한 clustering 구현 MO-MEMES 소스코드 수정 및 MEMES 알고리즘 테스트 웹 애플리케이션 UI 디자인 및 구현 MEMES에 Acquisition Function 추가 구현
최우영	보고서 작성 Docker와 Django를 이용해 환경 구축 리간드 데이터베이스 구축 및 리간드 추가 기능 구현 웹에서 탐색을 실행할 수 있도록 구현

3.3.5. 웹 애플리케이션 탑재

화합물 라이브러리 탐색의 접근성을 높이고 사용자 친화적인 환경을 제공하기 위해, Docker와 Django를 활용하여 편리하게 설치하고 이용할 수 있는 웹 애플리케이션을 개발하였다.

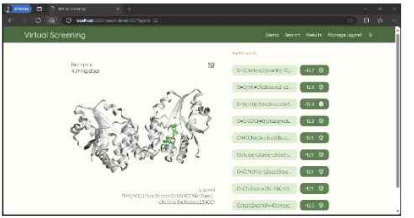
웹 애플리케이션에서 제공하는 주요 기능은 다음과 같다:

- Manage Ligand 페이지: 사용자는 이 페이지를 통해 팀이 수집한 리간드 데이터를 데이터베이스에 쉽게 추가할 수 있다. 또한, 직접 추가하고자 하는 리간드의 SMILES 문자열을 입력하여 데이터베이스에 저장할 수 있다.
- Search 페이지: 사용자는 자신의 컴퓨터에서 PDBQT 파일 형식의 리셉터를 선택하고, Random, Clustering, MEMES 세 가지 메소드 중 하나를 선택하여 탐색을 시작할 수 있다. MEMES 메소드를 선택할 경우 사용할 Acquisition Function도 지정할 수 있다.



[그림 10] 웹 애플리케이션 Search 페이지

- Demo 페이지: 미리 준비된 리셉터와 리간드들의 도킹 스코어를 활용하여, 실제 도킹 과정을 생략하고 각 메소드의 동작을 빠르게 비교할 수 있다. 이를 통해 사용자는 알고리즘의 성능과 특성을 직관적으로 파악할 수 있다.
- Results 페이지: 이전에 수행한 탐색 결과들의 목록을 확인하고, 각 결과의 세부 정보를 열람할 수 있다. 결과의 세부 정보 페이지에서는 추천된 리간드의 도킹 점수를 볼 수 있다. 그리고 Autodock Vina를 통해 얻은 도킹 포즈를 3Dmol.js[28]로 시각화하여 3D 모델을 볼 수 있게 하였다.



[그림 11] 웹 애플리케이션 결과 상세 페이지

웹 애플리케이션은 복잡한 설치 과정 없이도 사용자들이 쉽게 화합물 라이브러리 탐색을 수행할 수 있도록 설계되어 신약 후보물질 발굴 과정에서의 효율성 향상이 기대된다.

5. 결론 및 향후 연구 방향

이번 과제에서 신약 개발 후보물질 발굴을 위해 거대 화합물 라이브러리 탐색을 최적화하는 방법을 탐구하였다. 클러스터링 방식을 계층적으로 적용해 보았고, 베이지안 최적화 기법의 내부 목적함수에 변화를 주는 등의 방식으로 결합 예측도가 높을 것으로 추정되는 리간드를 선별해 보았다. 이외에도 과제에 도움이 될 만한 다양한 모델을 적용해 보았다. 결과적으로 전체 리간드의 1%를 탐색한 결과, 75%의 확률로 가장 결합도가 높은 리간드를 성공적으로 찾아내는 성과를 얻을 수 있었고 가장 좋은 성적을 내는 방식의 경우 샘플의 크기를 모집단의 0.07%까지 낮춰도 90%의 확률로 가장 결합도가 높은 리간드를 성공적으로 찾아낼 수 있었다.

하나 특이한 점은 200만개 라이브러리를 사용하였을 때, 결합 예측도와 유사도가 원하는 방향으로 같이 갔다는 것이다. 교수님께선 높은 결합 예측도와 낮은 유사도를 얻는 방향으로 가면 좋겠다고 말씀해주셨고, 4천개 라이브러리에선 예측도가 올라갈수록 유사도도 올라가는 트레이드 오프 관계였다. 의도치 않게 유사도 문제가 해결되어 아쉬움이 남는다. 이는 화합물 다양성을 확보하면서도 최적의 결합성을 갖춘 후보물질을 발굴하는 데 중요한 요소이므로, 향후 연구에서는 이러한 부분을 보완할 필요가 있다.

향후 연구 방향으로는 현재 급속히 발전하고 있는 신기술을 적극적으로 도입하고 대응해야 할 필요가 있다. 예를 들어, Google DeepMind의 AlphaProteo는 단백질 구조 예측을 최적화하는 AI 모델이며, NVIDIA의 BioNemo는 AI 기반의 신약 개발 플랫폼으로, 이러한 빅테크 기업들의 최신 인공지능 기반 기술은 신약 후보물질 발굴 과정의 효율성을 극대화할 수 있는 잠재력을 가지고 있다. 이들 외에도 다양한 기술들이 빠른 주기로 공개되고 있으며, 이는 신약 개발 분야의 패러다임을 바꿀 것이다. 따라서 향후 연구에서는 본 과제와 같이 최신 기술을 바탕으로 탐색 방법을 지속적으로 개선하는 접근이 필요하다. 신기술과의 결합이 단순히 새로운 도구를 사용하는 것에 그치지 않고, 각 기술의 강점을 최대한 활용하여 보다 다양하고 고도화된 신약 후보물질을 발굴하는 전략이 중요할 것이다. 이러한 방향은 향후 신약 개발의 성공 가능성을 높이는 데 큰 기여를 할 것으로 기대된다.

6. 참고 문헌

[1] 남궁석, *알파폴드: AI 신약개발 혁신*, biospectator, 2024

[2] Eberhardt, J., Santos-Martins, D., Tillack, A.F., Forli, S., "AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings", *Journal of Chemical Information and Modeling*, Vol. 61, No. 8, pp. 3891-3898, 2021

[3] Trott, O., & Olson, A. J., "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.", *Journal of computational chemistry*, Vol. 31, No. 2, pp. 455-461, 2010

[4] Center of Computational Structural Biology (CCSB) - Scripps Research(2021), AutoDock Vina [Online] Available: https://autodock-vina.readthedocs.io/en/latest/index.html

[5] Graff, David E. and Shakhnovich, Eugene I. and Coley, Connor W., "Accelerating high-throughput virtual screening through molecular pool-based active learning", *Chem. Sci.*, Vol. 12, No.22, pp. 7866-7881, 2021

[6] Brain_93 (2021), 베이지안 최적화(Bayesian Optimization) [Online]. Available: https://data-scientist-brian-kim.tistory.com/88

[7] Stathis Kamperis (2021), Acquisition functions in Bayesian Optimization [Online]. Available: https://ekamperi.github.io/machine%20learning/2021/06/11/acquisition-functions.html

[8] Meta Platforms (2024), BoTorch [Online]. Available: https://botorch.org/api

[9] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, "BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimizatio", *Advances in Neural Information Processing Systems 33*, 2020

[10] Jay Han (2022), Bayesiaan Optimization [Online]. Available: https://otzslayer.github.io/ml/2022/12/03/bayesian-optimization.html

[11] Sarvesh Mehta, Siddhartha Laghuvarapu, Yashaswi Pathak, Aaftaab Sethi, Mallika Alvala, U. Deva Priyakumar, "MEMES: Machine learning framework for Enhanced MolEcular Screening", *Chem. Sci.*, Vol.12, No.35, pp. 11710-11721, 2021

[12] Corso, Gabriele and Stärk, Hannes and Jing, Bowen and Barzilay, Regina and Jaakkola, Tommi, "DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking", International Conference on Learning Representations (ICLR), 2023

[그림 3] https://github.com/gcorso/DiffDock

[13] yunhuijang (2022), Diffusion model 설명 (Diffusion model 이란? Diffusion model 증명) [Online]. Available: https://process-mining.tistory.com/182

[14] Weng, Lilian. (2021). What are diffusion models? [Online]. Available: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

[15] 알파고라니 (2024), [만들면서 배우는 생성 AI] 8 장 - 확산 모델(diffusion model) [Online]. Available: https://velog.io/@running_learning/%EB%A7%8C%EB%93%A4%EB%A9%B4%EC%84%9C-%EB%B0%B0%EC%9A%B0%EB%8A%94-%EC%83%9D%EC%84%B1-AI-8%EC%9E%A5-%ED%99%95%EC%82%B0-%EB%AA%A8%EB%8D%B8diffusion-model

[16] RCSB PDB [Online]. Available: https://www.rcsb.org/

[17] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, "The Protein Data Bank", Nucleic Acids Research 28, pp. 235-242, 2000

[18] BioinformaticsCopilot (2022), Molecular Docking for Beginners | Autodock Full Tutorial [Online]. Available: https://www.youtube.com/watch?v=ZVKKsK5DsCY

[19] Open Babel Team (2023), Open Babel [Online], Available: http://openbabel.org

[20] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, Geoffrey R. Hutchison, "Open Babel: An open chemical toolbox.", *J. Cheminf.*, Vol.3, No. 33, 2011

[21] RDKit: Open-source cheminformatics. [Online], Available: https://www.rdkit.org

[22] CHML (2023), 파이썬 RDKit을 이용한 분자 유사도 (Tanimoto Similarity) 계산 [Online]. Available: https://untitledtblog.tistory.com/189

[23] Kang, Y.N., Stuckey, J.A. (2013), Crystal structure of apo CDK2 [Online]. Available: https://doi.org/10.2210/pdb4EK3/pdb

[24] Mehta S, Goel M and Priyakumar UD, "MO-MEMES: A method for accelerating virtual screening using multi-objective Bayesian optimization", *Front. Med*, 9:916481, 2022

[25] Mehta S, Goel M and Priyakumar UD (2022), MO-MEMES [Online]. Available: https://github.com/devalab/MO-MEMES

[26] Naik, M., Raichurkar, A., Bandodkar, B.S., Varun, B.V., Bhat, S., Kalkhambkar, R., Murugan, K., Menon, R., Bhat, J., Paul, B., Iyer, H., Hussein, S., Tucker, J.A., Vogtherr, M., Embrey, K.J., Mcmiken, H., Prasad, S., Gill, A., Ugarkar, B.G., Venkatraman, J., Read, J., Panda, M. (2015), Structure Guided Lead Generation for M. Tuberculosis Thymidylate Kinase (Mtb Tmk): Discovery of 3-Cyanopyridone and 1,6-Naphthyridin-2-One as Potent Inhibitors. [Online].

Available: https://doi.org/10.2210/pdb4UNN/pdb

[27] Gregory Gundersen (2020), Entropy of the Gaussian [Online]. Available: https://gregorygundersen.com/blog/2020/09/01/gaussian-entropy/

[28]Nicholas Rego and David Koes, 3Dmol.js: molecular visualization with WebGL, Bioinformatics, 31, 8, pp.1322-1324, 2015