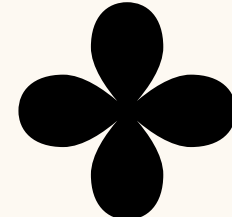


신약 개발 후보물질 발굴을 위한 거대 화합물 라이브러리 탐색 최적화



8조

201924656 이정민

202155565 성가빈

201845928 최우영

지도교수 송길태



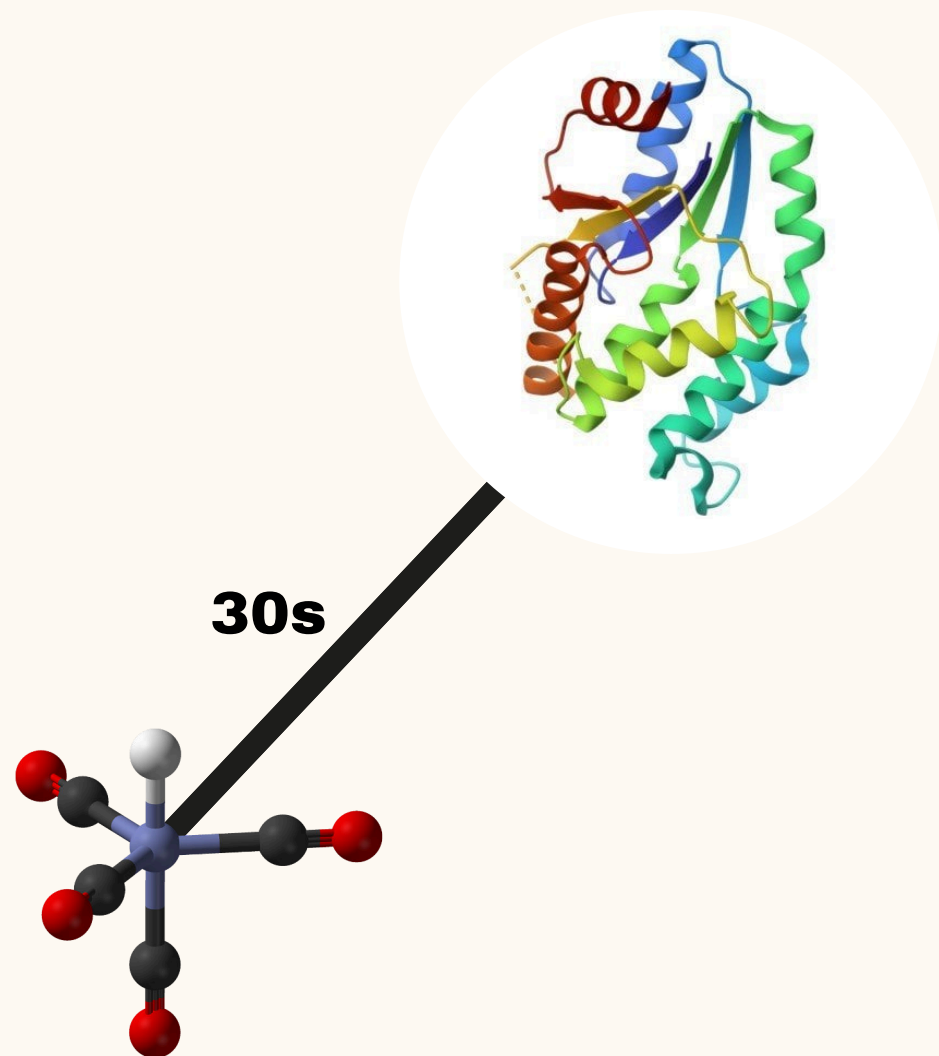
과제 배경

신약 개발은 현대 의학에서 중요한 과정이다.
이 과정에서 **거대 화합물 라이브러리**의 탐색 전략
은 성공적인 후보 물질을 찾기 위해 필수적이다.



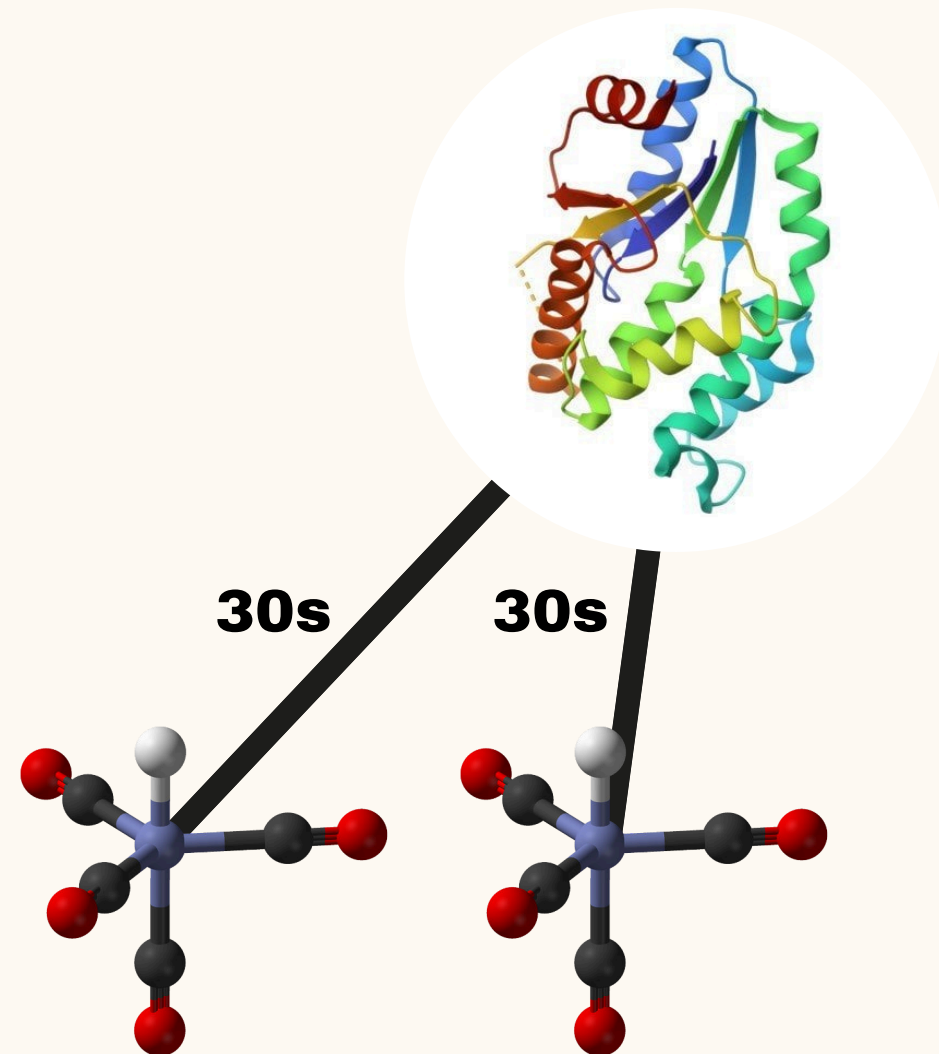
과제 배경

신약 개발은 현대 의학에서 중요한 과정이다.
이 과정에서 **거대 화합물 라이브러리**의 탐색 전략
은 성공적인 후보 물질을 찾기 위해 필수적이다.



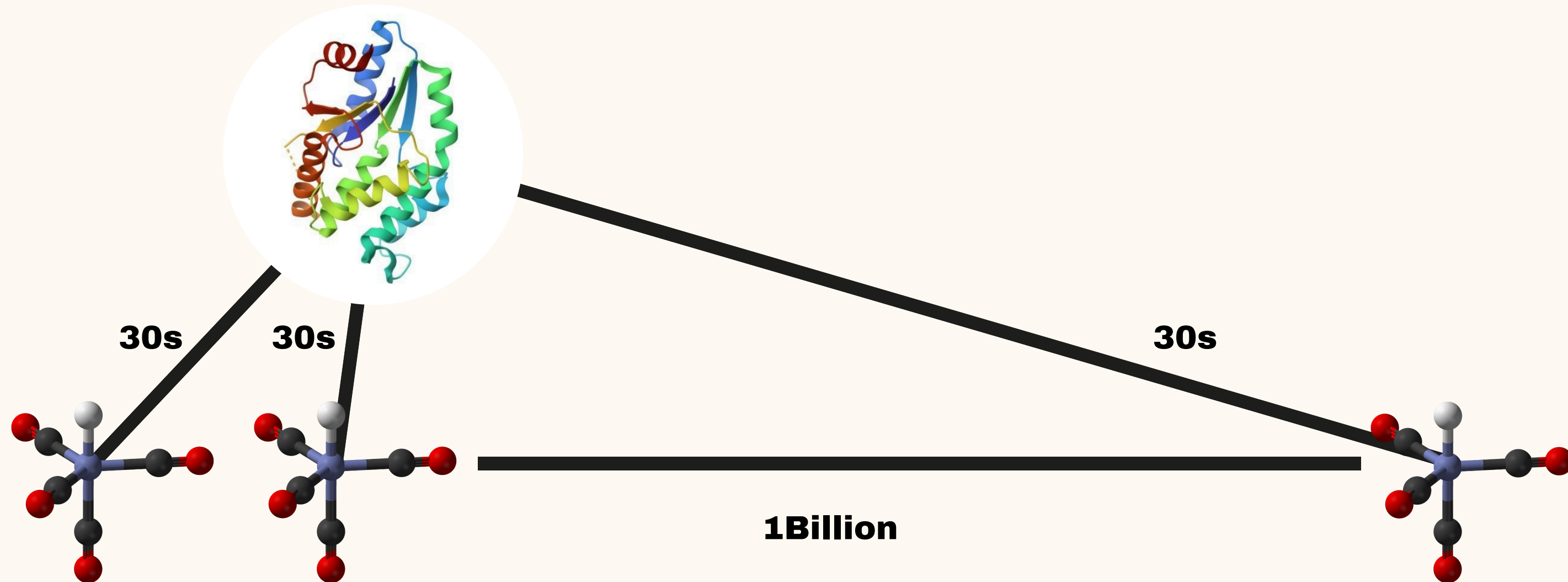
과제 배경

신약 개발은 현대 의학에서 중요한 과정이다.
이 과정에서 **거대 화합물 라이브러리**의 탐색 전략
은 성공적인 후보 물질을 찾기 위해 필수적이다.



과제 배경

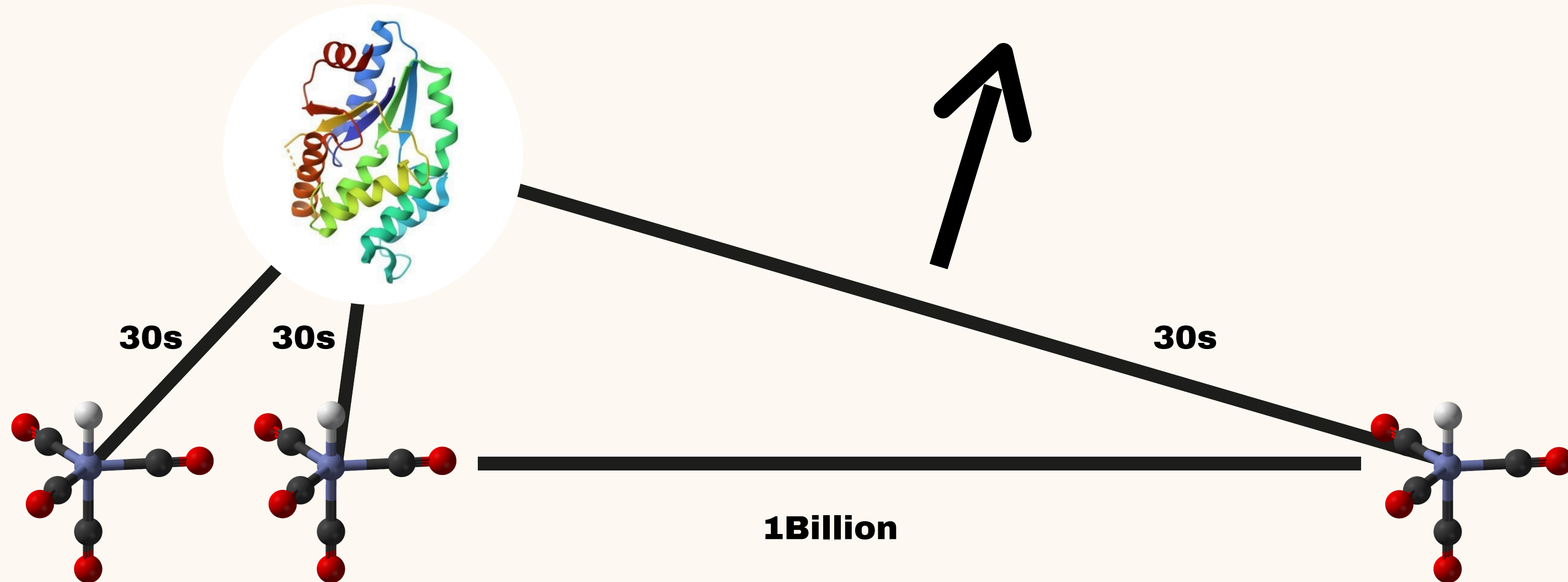
신약 개발은 현대 의학에서 중요한 과정이다.
이 과정에서 거대 **화합물 라이브러리**의 탐색 전략
은 성공적인 후보 물질을 찾기 위해 필수적이다.



과제 배경

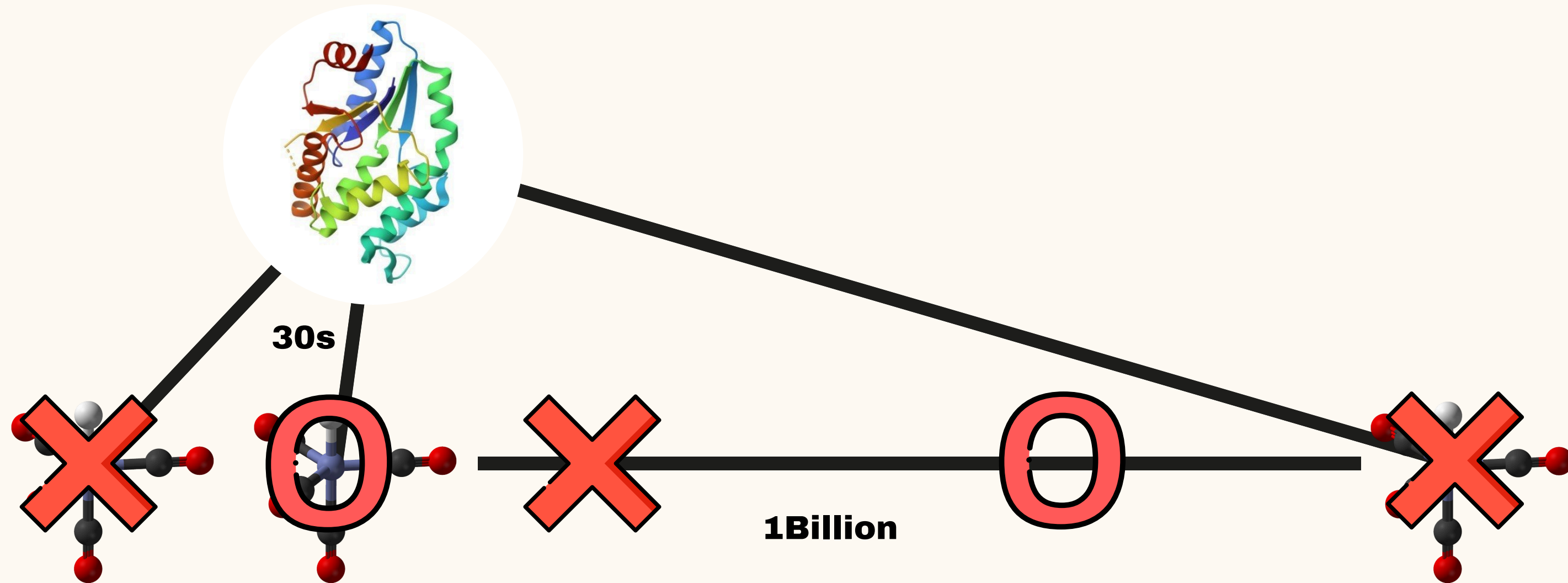
신약 개발은 현대 의학에서 중요한 과정이다.
이 과정에서 거대 **화합물 라이브러리**의 탐색 전략
은 성공적인 후보 물질을 찾기 위해 필수적이다.

10억개의 리간드 중 최적의 리간드를 찾기 위해
전수조사를 한다고 가정할 때, 하나의 리간드 당
30초의 시간이 소요된다면, **300억초**가 소요된다.
이는 약 **951년**이며 사실상 불가능하다.



과제 목표

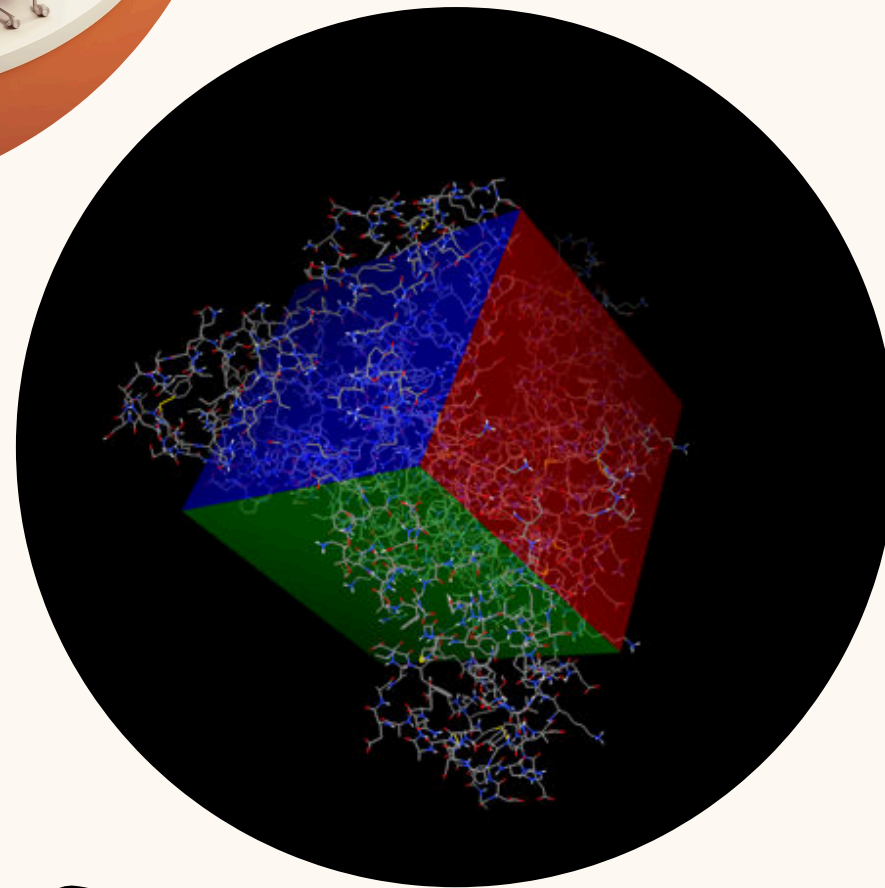
거대 화합물 라이브러리에서 주어진 타겟 단백질
과 **결합 확률이 높을 것으로 예측되는 화합물**을 효
율적으로 탐색하고 **선별**하는 것



가상 스크리닝

가상 스크리닝은 화합물과 타겟 단백질 간의 상호작용을 컴퓨터 **시뮬레이션**을 통해 **예측**하는 기술이다. 이를 통해 화합물이 타겟 단백질과 결합하는 방법을 알아볼 수 있고, 역장(forcefield) 기반이나 경험(empirical) 기반의 **스코어**를 얻을 수 있다.

과제에는 분자 도킹을 위한 오픈 소스 프로그램 **Autodock Vina**가 사용된다.



라이브러리 구축

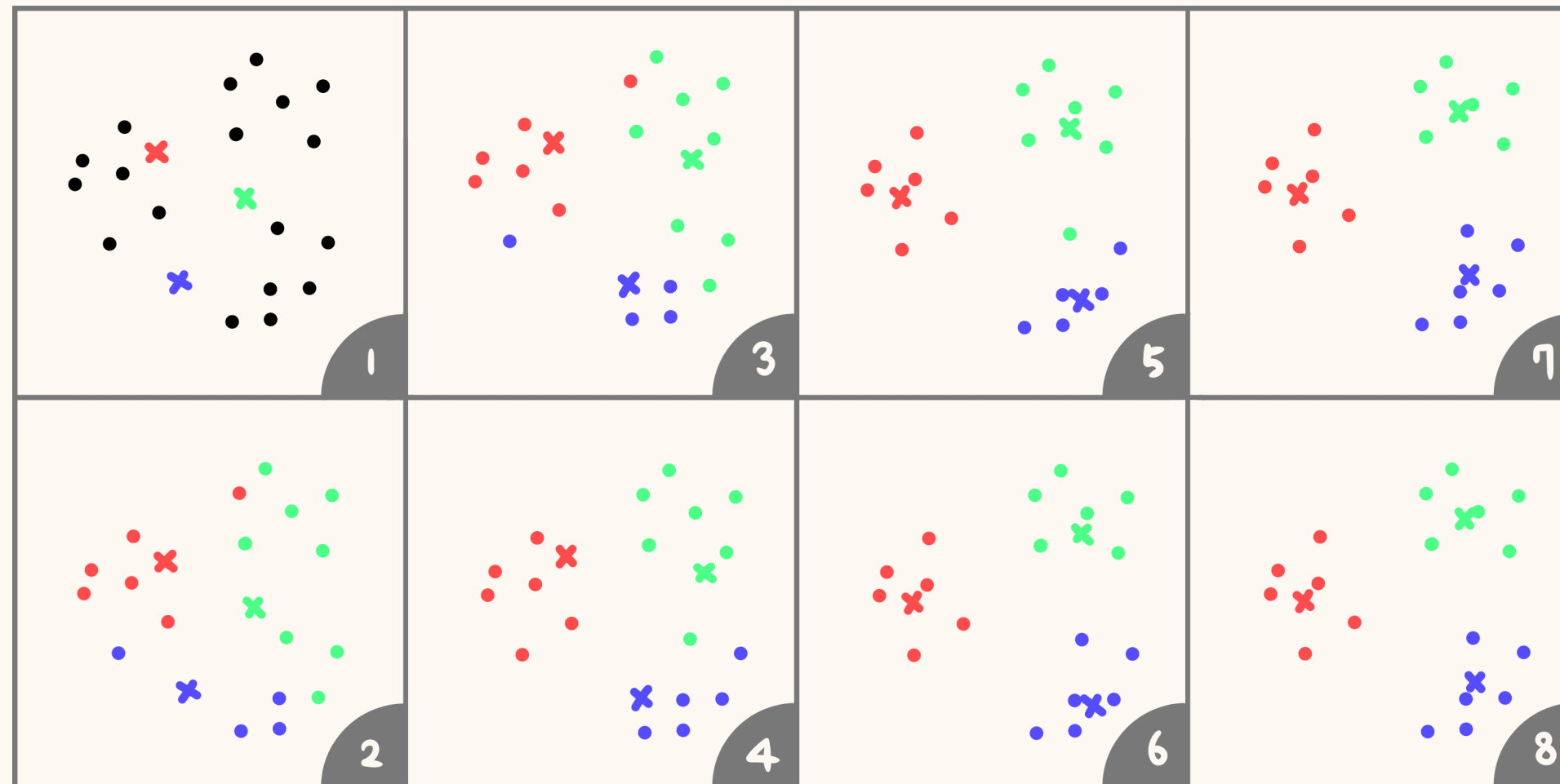
Vina로 단백질과 리간드의 결합을 예측하기 위해서는 타겟 단백질과 리간드의 **PDBQT 파일**이 필요하다.

단백질 파일은 PDB 파일을 다운로드 후 ChimeraX 툴과 AutoDockTools를 이용해 정리하는 작업이 필요하다.
ChimeraX를 활용하여 비표준 잔류물과 추가 체인을 제거, 이후 AutoDockTools를 통해 물 분자 제거, 수소 추가, 무극성 수소 병합, 콜먼 전하 추가, AD4 타입 원자 배치 등의 과정을 통해 PDBQT 파일로 변환하였다.

리간드 파일은 mol2 파일을 다운로드 후 OpenBabel GUI 툴을 통해 PDBQT파일로 변환하였다.

클러스터링

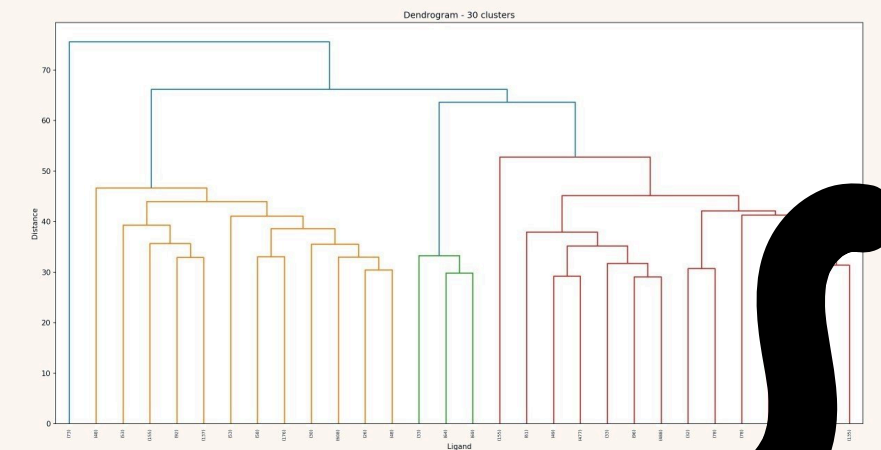
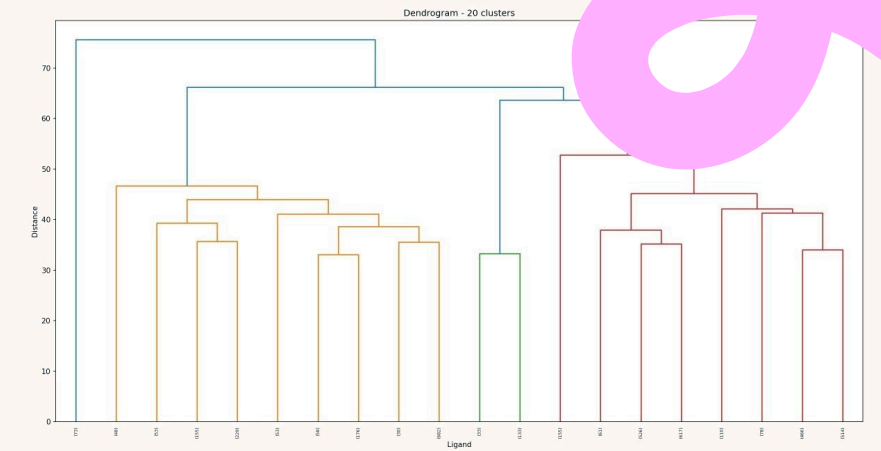
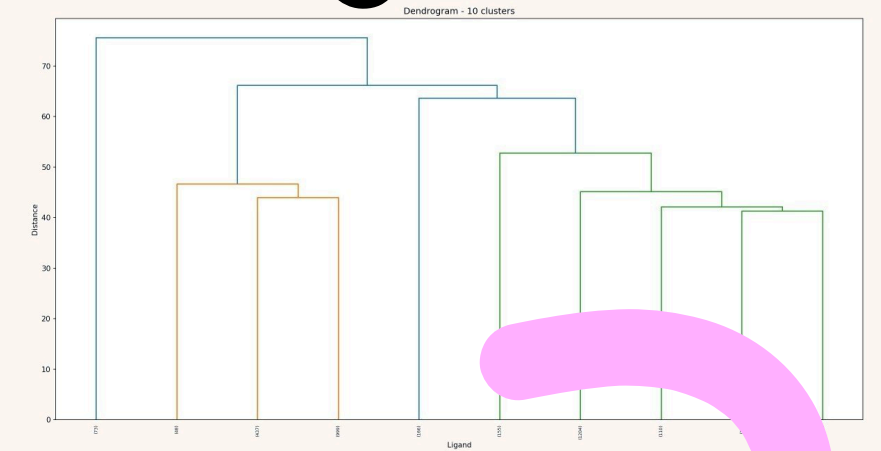
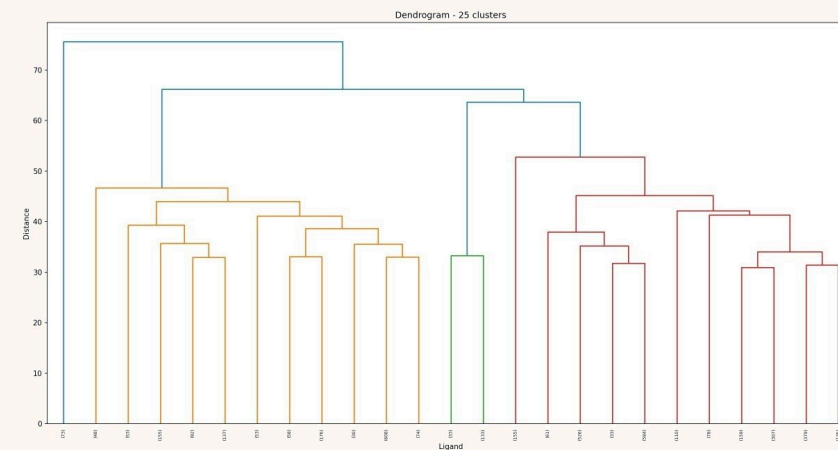
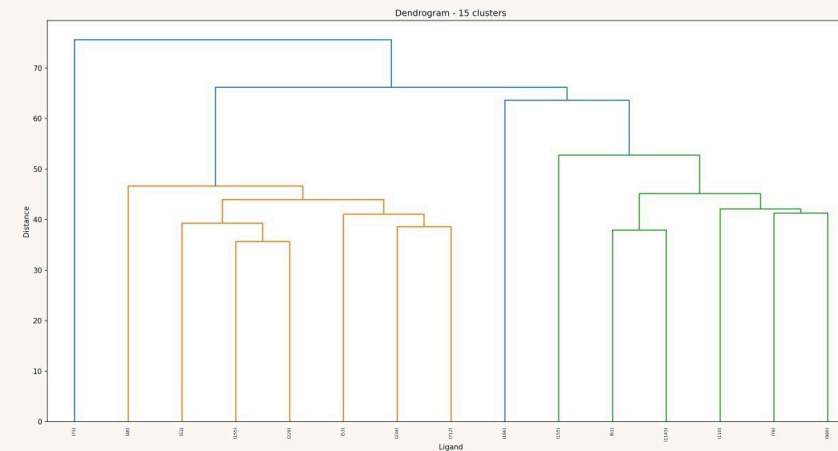
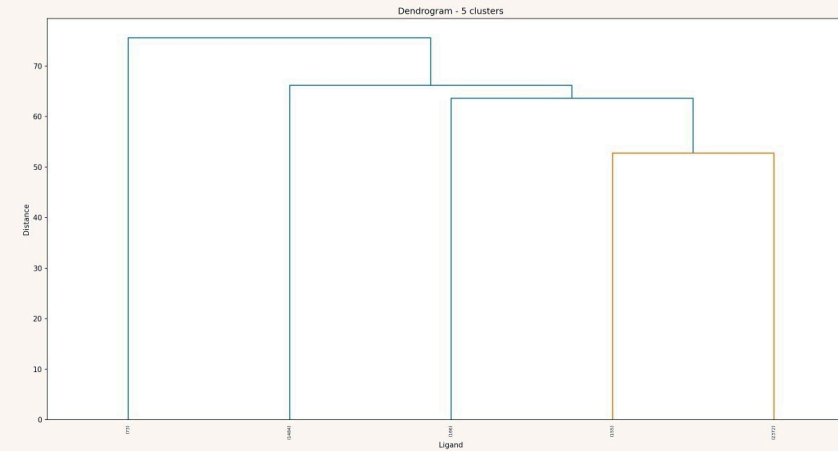
K-means 알고리즘은 주어진 데이터 포인트들을 k 개의 클러스터로 나누는 비지도 학습 기법이다. 비지도 학습에서는 데이터 포인트들의 라벨을 지정하지 않고, 데이터의 패턴을 파악하게 한다.



병합 군집 클러스터링

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

이를 통해 **점증적**으로 클러스터를 탐색해 결합 예측도가 높은 리간드를 선별한다.



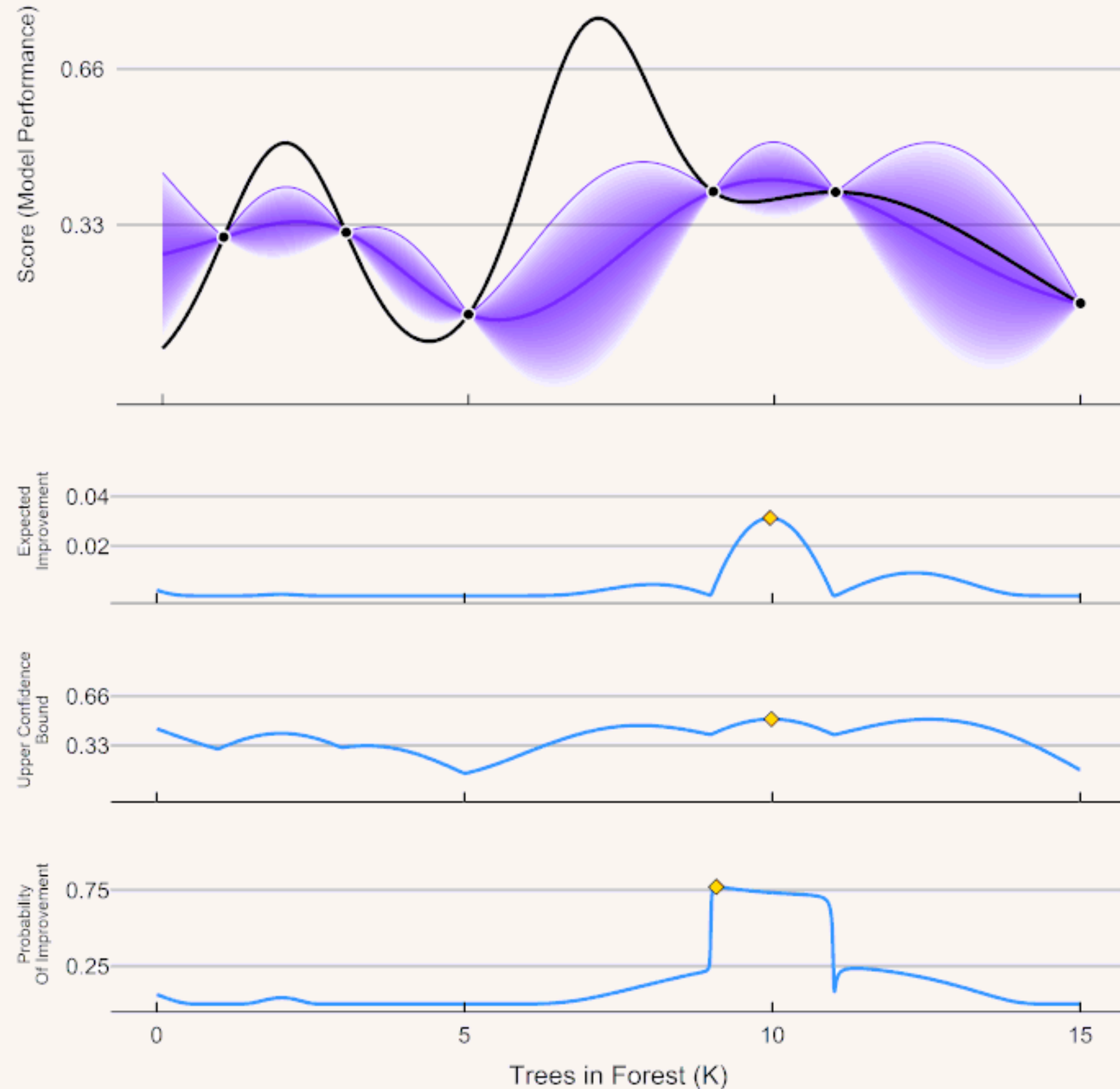
베이지안 최적화

베이지안 최적화는 알려지지 않은 목적 함수의 값을 최대 또는 최소로 만드는 입력 값을 찾는 것을 목표로 한다.

Surrogate model과 Acquisition function으로 구성되고, 이 두 과정을 반복하면서 목적 함수가 최대 또는 최소가 되는 지점을 탐색한다.

Surrogate model은 이전의 입력 값과 그에 따른 목적 함수의 값들을 이용해 목적 함수를 확률적으로 추정하며, 가우시안 프로세스 회귀가 주로 사용한다.

ParBayesianOptimization in Action (Round 1)



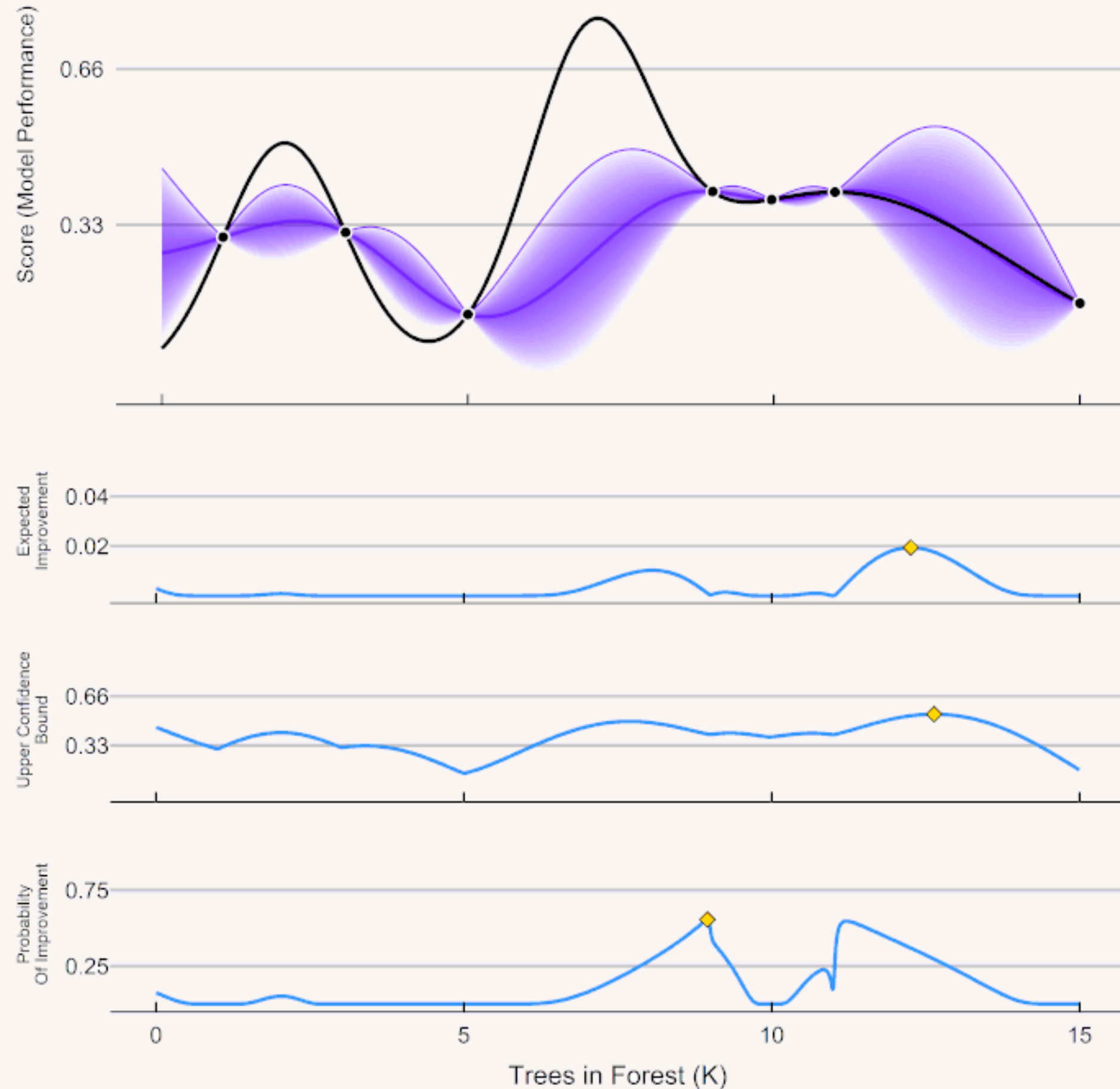
베이지안 최적화

베이지안 최적화는 알려지지 않은 목적 함수의 값을 최대 또는 최소로 만드는 입력 값을 찾는 것을 목표로 한다.

Surrogate model과 Acquisition function으로 구성되고, 이 두 과정을 반복하면서 목적 함수가 최대 또는 최소가 되는 지점을 탐색한다.

Surrogate model은 이전의 입력 값과 그에 따른 목적 함수의 값들을 이용해 목적 함수를 확률적으로 추정하며, 가우시안 프로세스 회귀가 주로 사용한다.

ParBayesianOptimization in Action (Round 2)



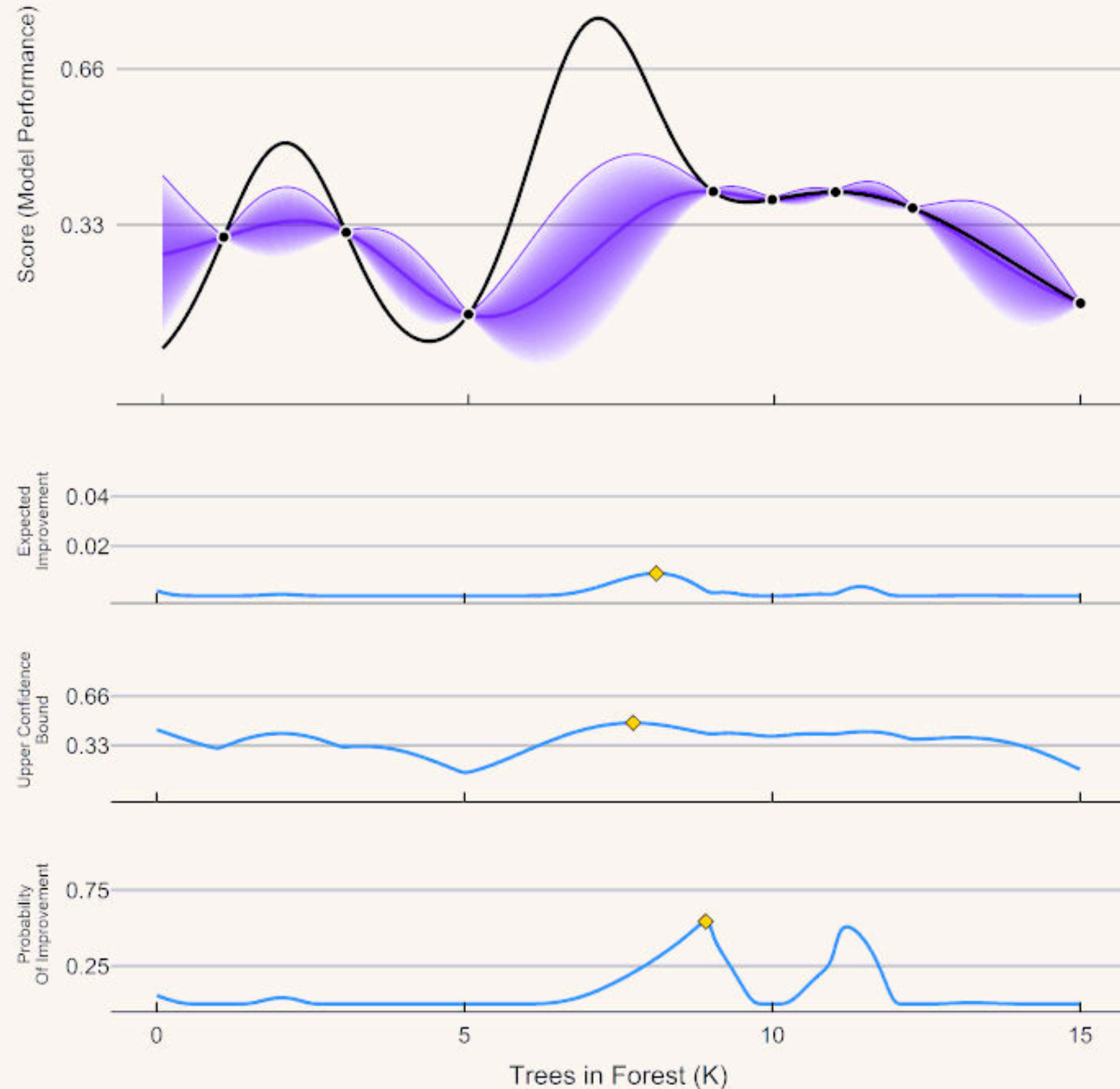
베이지안 최적화

베이지안 최적화는 알려지지 않은 목적 함수의 값을 최대 또는 최소로 만드는 입력 값을 찾는 것을 목표로 한다.

Surrogate model과 Acquisition function으로 구성되고, 이 두 과정을 반복하면서 목적 함수가 최대 또는 최소가 되는 지점을 탐색한다.

Surrogate model은 이전의 입력 값과 그에 따른 목적 함수의 값들을 이용해 목적 함수를 확률적으로 추정하며, 가우시안 프로세스 회귀가 주로 사용한다.

ParBayesianOptimization in Action (Round 3)



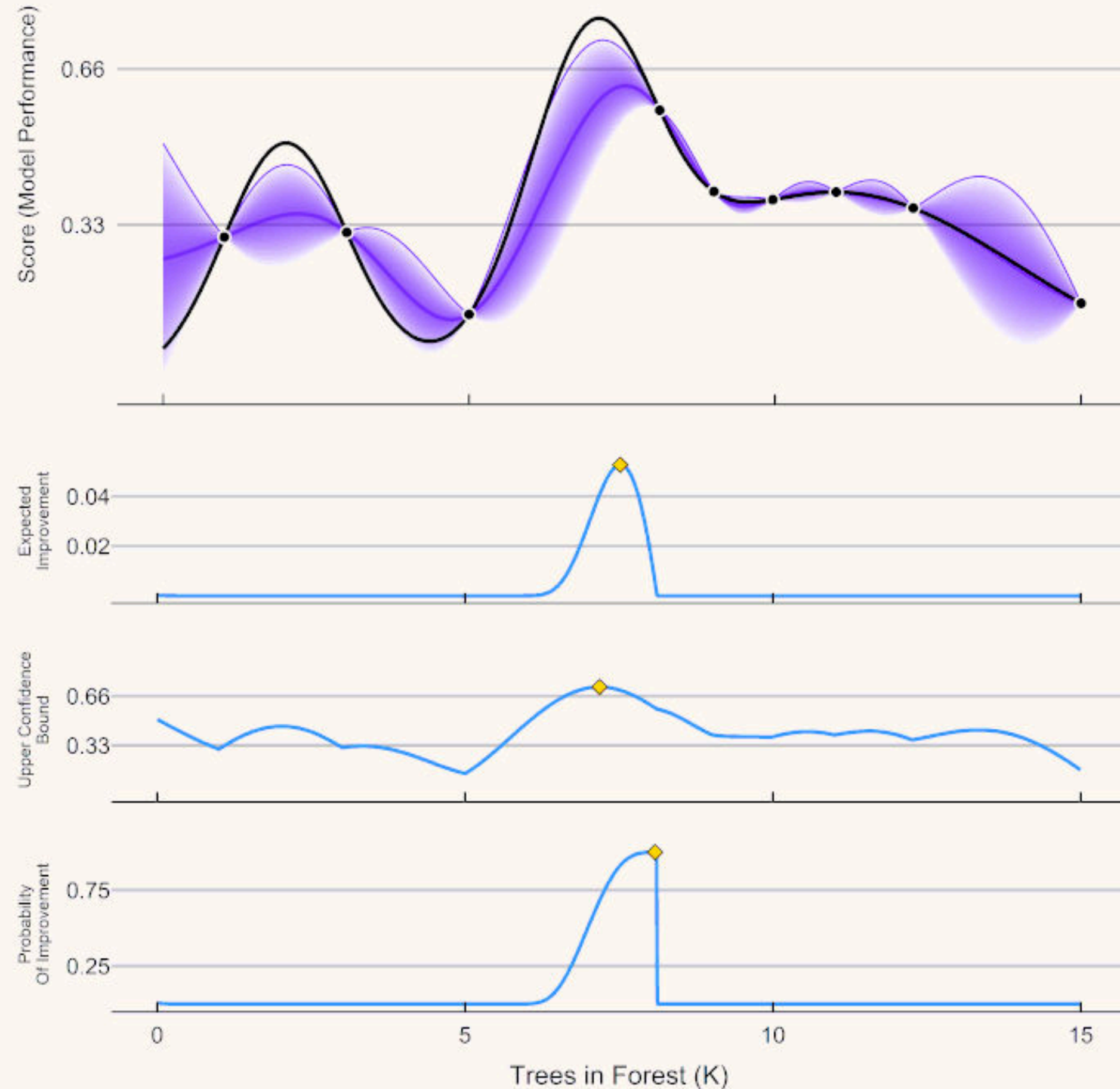
베이지안 최적화

베이지안 최적화는 알려지지 않은 목적 함수의 값을 최대 또는 최소로 만드는 입력 값을 찾는 것을 목표로 한다.

Surrogate model과 Acquisition function으로 구성되고, 이 두 과정을 반복하면서 목적 함수가 최대 또는 최소가 되는 지점을 탐색한다.

Surrogate model은 이전의 입력 값과 그에 따른 목적 함수의 값들을 이용해 목적 함수를 확률적으로 추정하며, 가우시안 프로세스 회귀가 주로 사용한다.

ParBayesianOptimization in Action (Round 4)



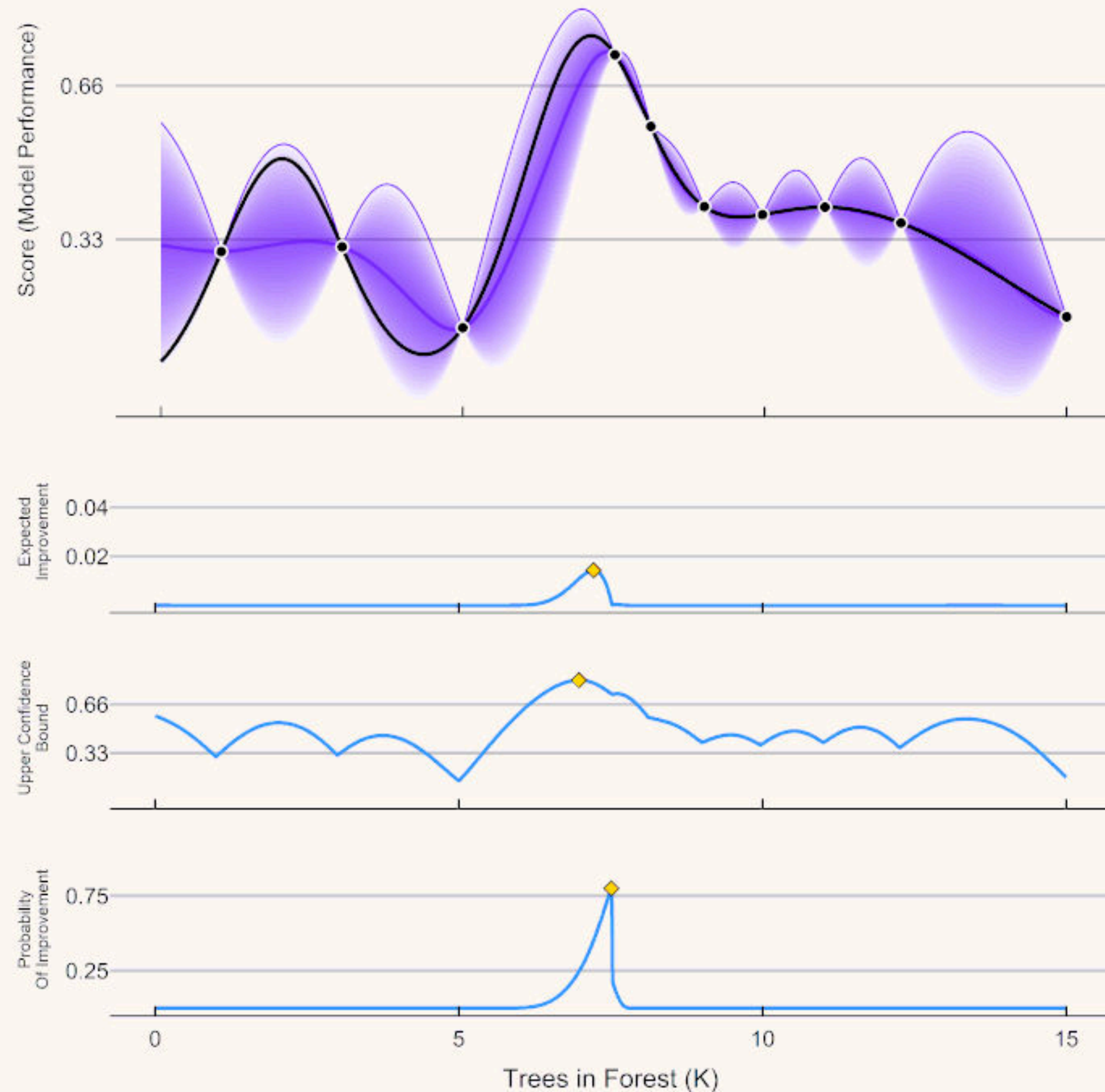
베이지안 최적화

베이지안 최적화는 알려지지 않은 목적 함수의 값을 최대 또는 최소로 만드는 입력 값을 찾는 것을 목표로 한다.

Surrogate model과 Acquisition function으로 구성되고, 이 두 과정을 반복하면서 목적 함수가 최대 또는 최소가 되는 지점을 탐색한다.

Surrogate model은 이전의 입력 값과 그에 따른 목적 함수의 값들을 이용해 목적 함수를 확률적으로 추정하며, 가우시안 프로세스 회귀가 주로 사용한다.

ParBayesianOptimization in Action (Round 5)



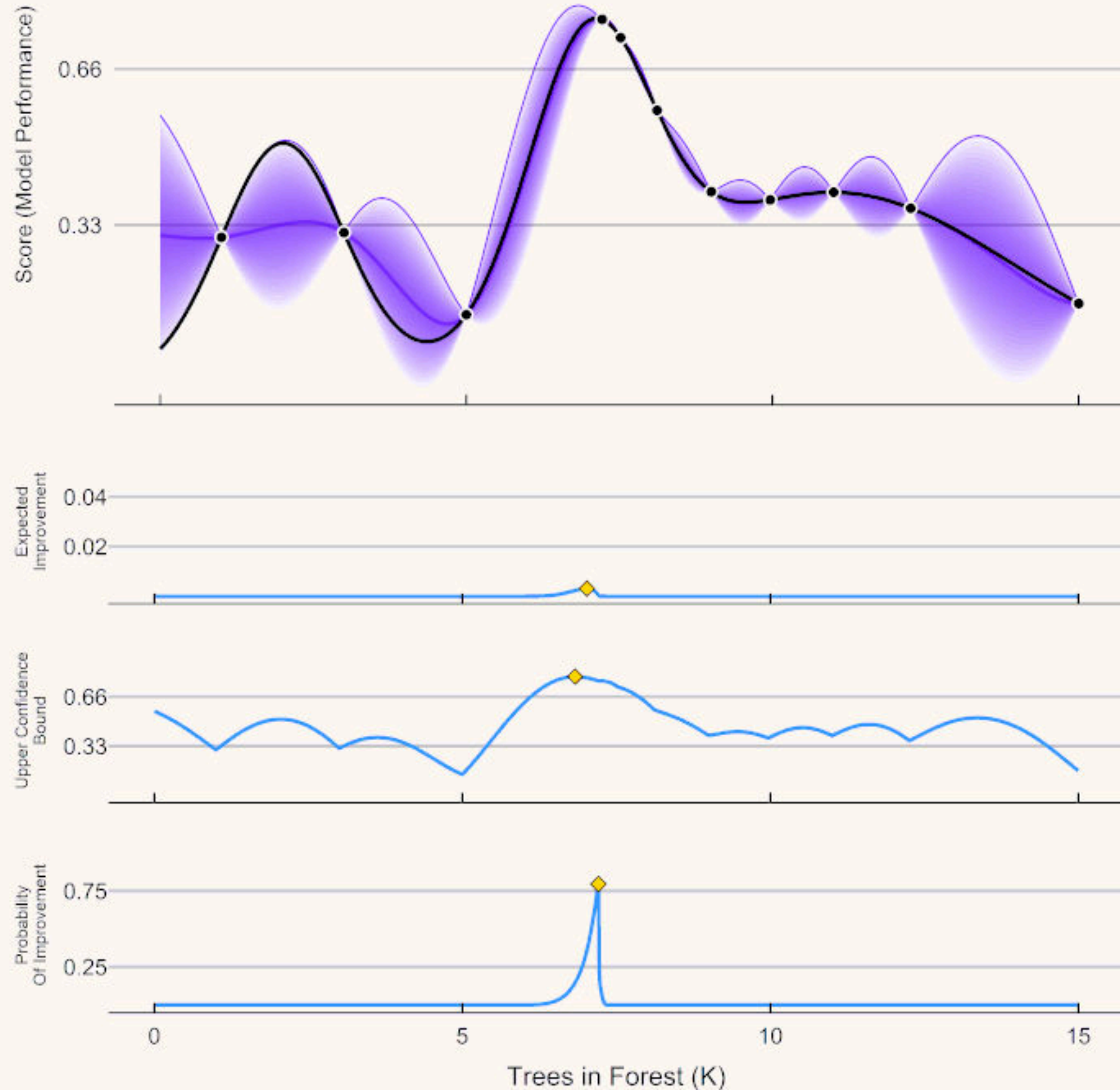
베이지안 최적화

베이지안 최적화는 알려지지 않은 목적 함수의 값을 최대 또는 최소로 만드는 입력 값을 찾는 것을 목표로 한다.

Surrogate model과 Acquisition function으로 구성되고, 이 두 과정을 반복하면서 목적 함수가 최대 또는 최소가 되는 지점을 탐색한다.

Surrogate model은 이전의 입력 값과 그에 따른 목적 함수의 값들을 이용해 목적 함수를 확률적으로 추정하며, 가우시안 프로세스 회귀가 주로 사용한다.

ParBayesianOptimization in Action (Round 6)



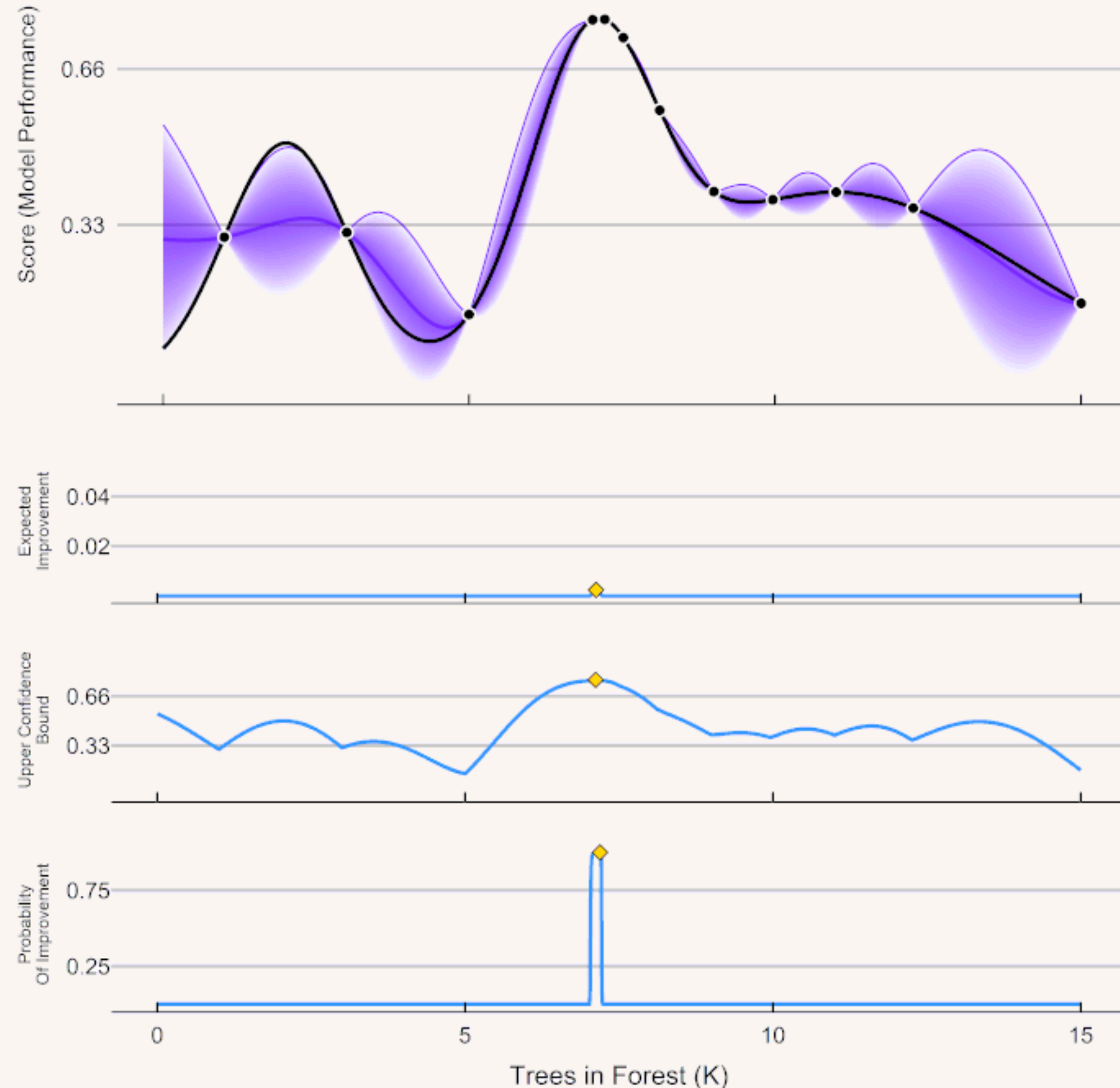
베이지안 최적화

베이지안 최적화는 알려지지 않은 목적 함수의 값을 최대 또는 최소로 만드는 입력 값을 찾는 것을 목표로 한다.

Surrogate model과 Acquisition function으로 구성되고, 이 두 과정을 반복하면서 목적 함수가 최대 또는 최소가 되는 지점을 탐색한다.

Surrogate model은 이전의 입력 값과 그에 따른 목적 함수의 값들을 이용해 목적 함수를 확률적으로 추정하며, 가우시안 프로세스 회귀가 주로 사용한다.

ParBayesianOptimization in Action (Round 7)

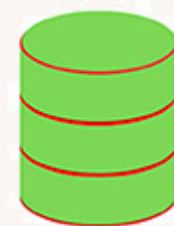


MEMES

MEMES(Machine learning framework for Enhanced MolEcular Screening)

는 Sarvesh Mehta등이 제시한 머신러닝 기반의 가상 스크리닝 프레임워크로, 베이지안 최적화를 통해 대규모 약물 라이브러리에서 잠재적인 신약 후보 물질을 효율적으로 식별하는 알고리즘이다.

클러스터링을 통해 다양한 초기 리간드를 선택해 Surrogate model의 학습 데이터로 쓰이며, 이 모델을 이용해 리간드의 도킹 점수를 예측한다. 이후, 학습된 모델을 기반으로 Acquisition Function을 사용해 다음 탐색할 리간드를 선택한다.



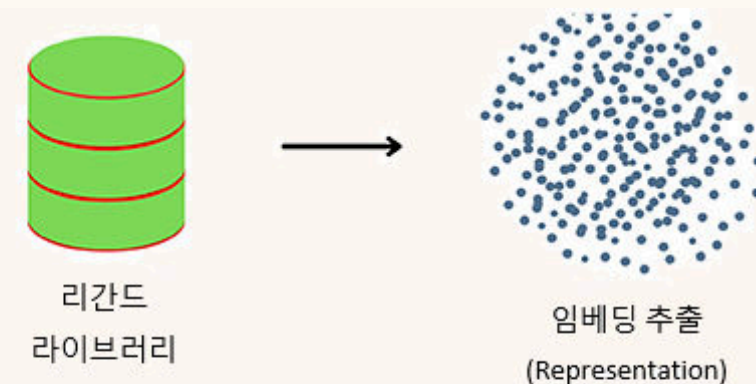
리간드
라이브러리

MEMES

MEMES(Machine learning framework for Enhanced MolEcular Screening)

는 Sarvesh Mehta등이 제시한 머신러닝 기반의 가상 스크리닝 프레임워크로, 베이지안 최적화를 통해 대규모 약물 라이브러리에서 잠재적인 신약 후보 물질을 효율적으로 식별하는 알고리즘이다.

클러스터링을 통해 다양한 초기 리간드를 선택해 Surrogate model의 학습 데이터로 쓰이며, 이 모델을 이용해 리간드의 도킹 점수를 예측한다. 이후, 학습된 모델을 기반으로 Acquisition Function을 사용해 다음 탐색할 리간드를 선택한다.

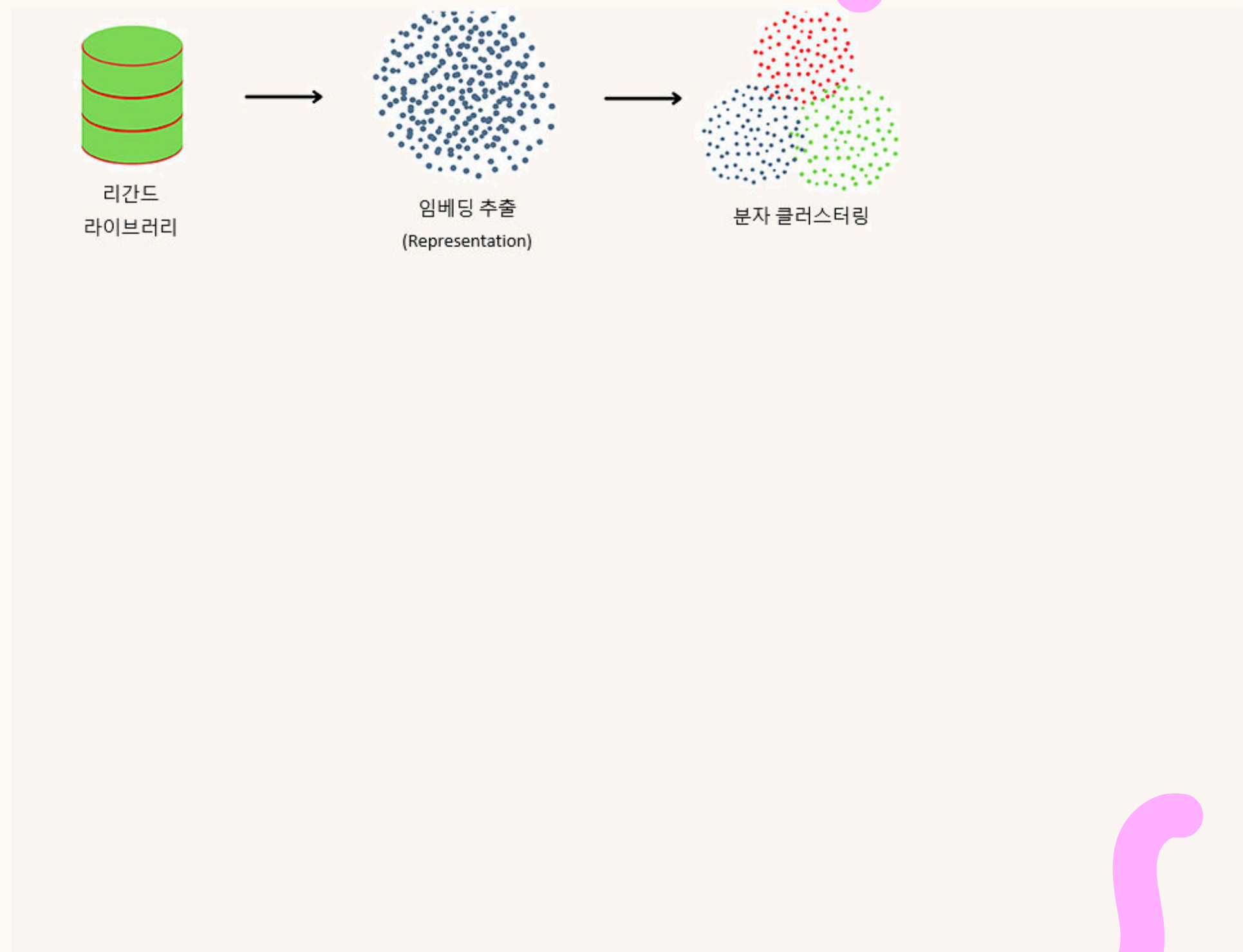


MEMES

MEMES(Machine learning framework for Enhanced MolEcular Screening)

는 Sarvesh Mehta등이 제시한 머신러닝 기반의 가상 스크리닝 프레임워크로, 베이지안 최적화를 통해 대규모 약물 라이브러리에서 잠재적인 신약 후보 물질을 효율적으로 식별하는 알고리즘이다.

클러스터링을 통해 다양한 초기 리간드를 선택해 Surrogate model의 학습 데이터로 쓰이며, 이 모델을 이용해 리간드의 도킹 점수를 예측한다. 이후, 학습된 모델을 기반으로 Acquisition Function을 사용해 다음 탐색할 리간드를 선택한다.

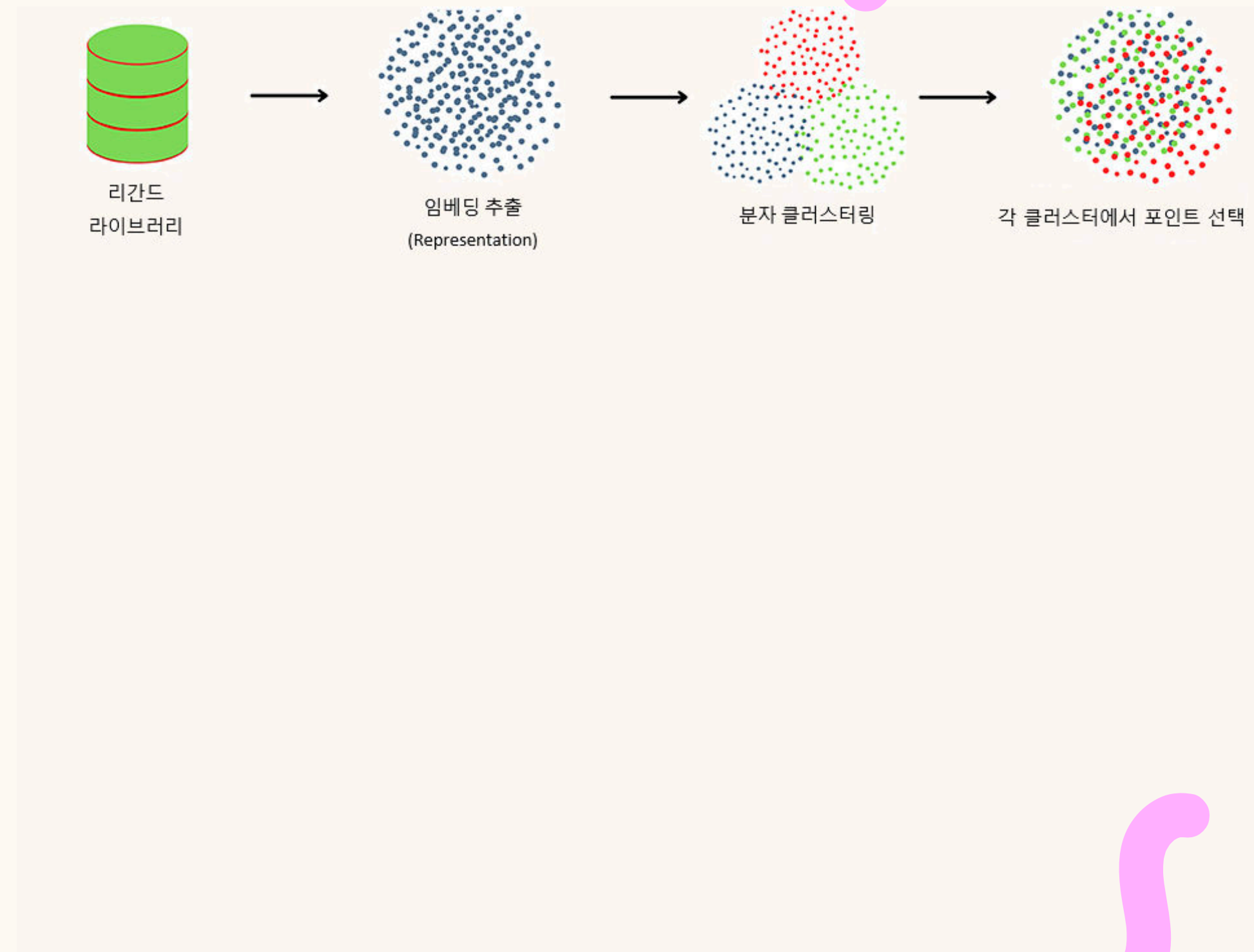


MEMES

MEMES(Machine learning framework for Enhanced MolEcular Screening)

는 Sarvesh Mehta등이 제시한 머신러닝 기반의 가상 스크리닝 프레임워크로, 베이지안 최적화를 통해 대규모 약물 라이브러리에서 잠재적인 신약 후보 물질을 효율적으로 식별하는 알고리즘이다.

클러스터링을 통해 다양한 초기 리간드를 선택해 Surrogate model의 학습 데이터로 쓰이며, 이 모델을 이용해 리간드의 도킹 점수를 예측한다. 이후, 학습된 모델을 기반으로 Acquisition Function을 사용해 다음 탐색할 리간드를 선택한다.

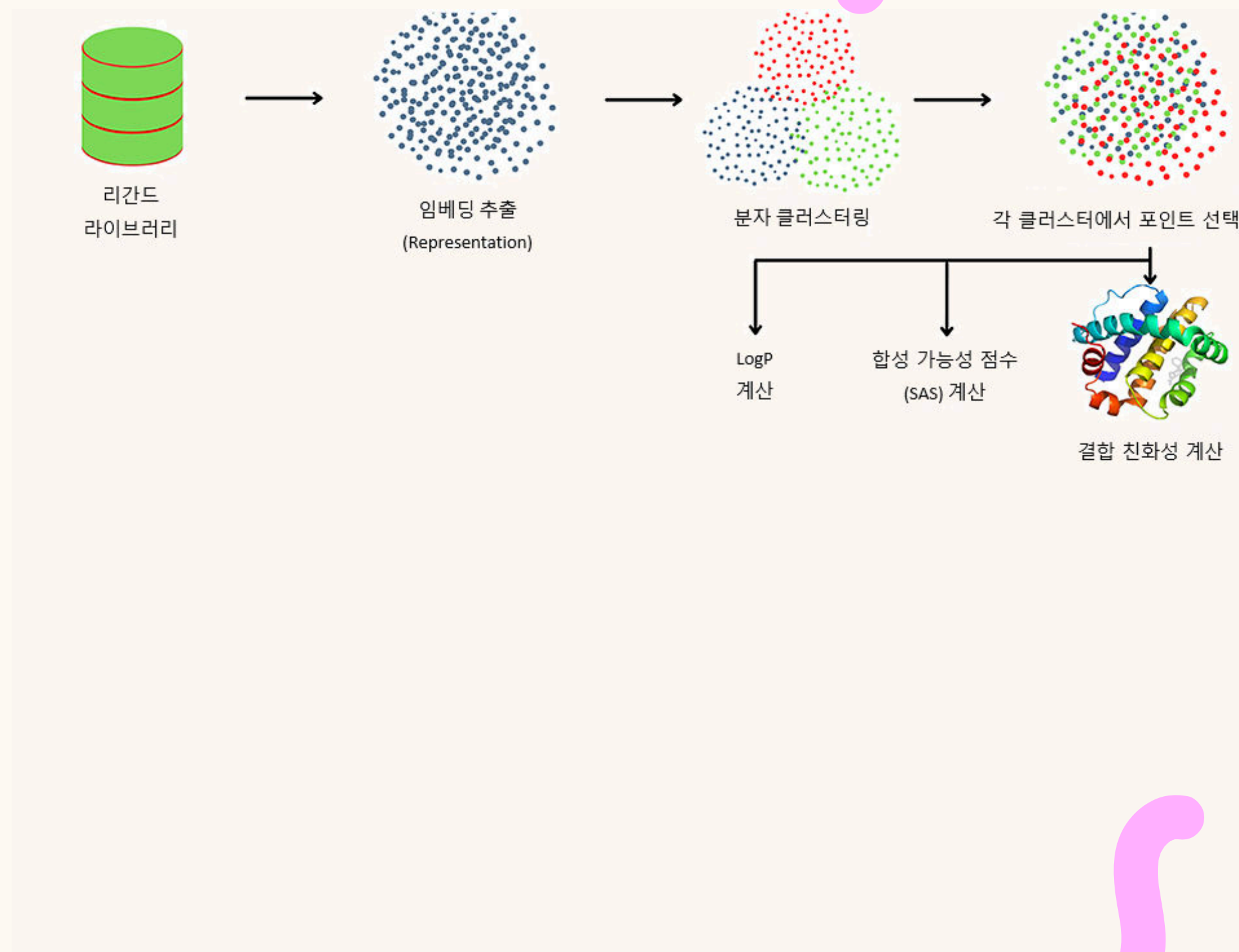


MEMES

MEMES(Machine learning framework for Enhanced MolEcular Screening)

는 Sarvesh Mehta등이 제시한 머신러닝 기반의 가상 스크리닝 프레임워크로, 베이지안 최적화를 통해 대규모 약물 라이브러리에서 잠재적인 신약 후보 물질을 효율적으로 식별하는 알고리즘이다.

클러스터링을 통해 다양한 초기 리간드를 선택해 Surrogate model의 학습 데이터로 쓰이며, 이 모델을 이용해 리간드의 도킹 점수를 예측한다. 이후, 학습된 모델을 기반으로 Acquisition Function을 사용해 다음 탐색할 리간드를 선택한다.

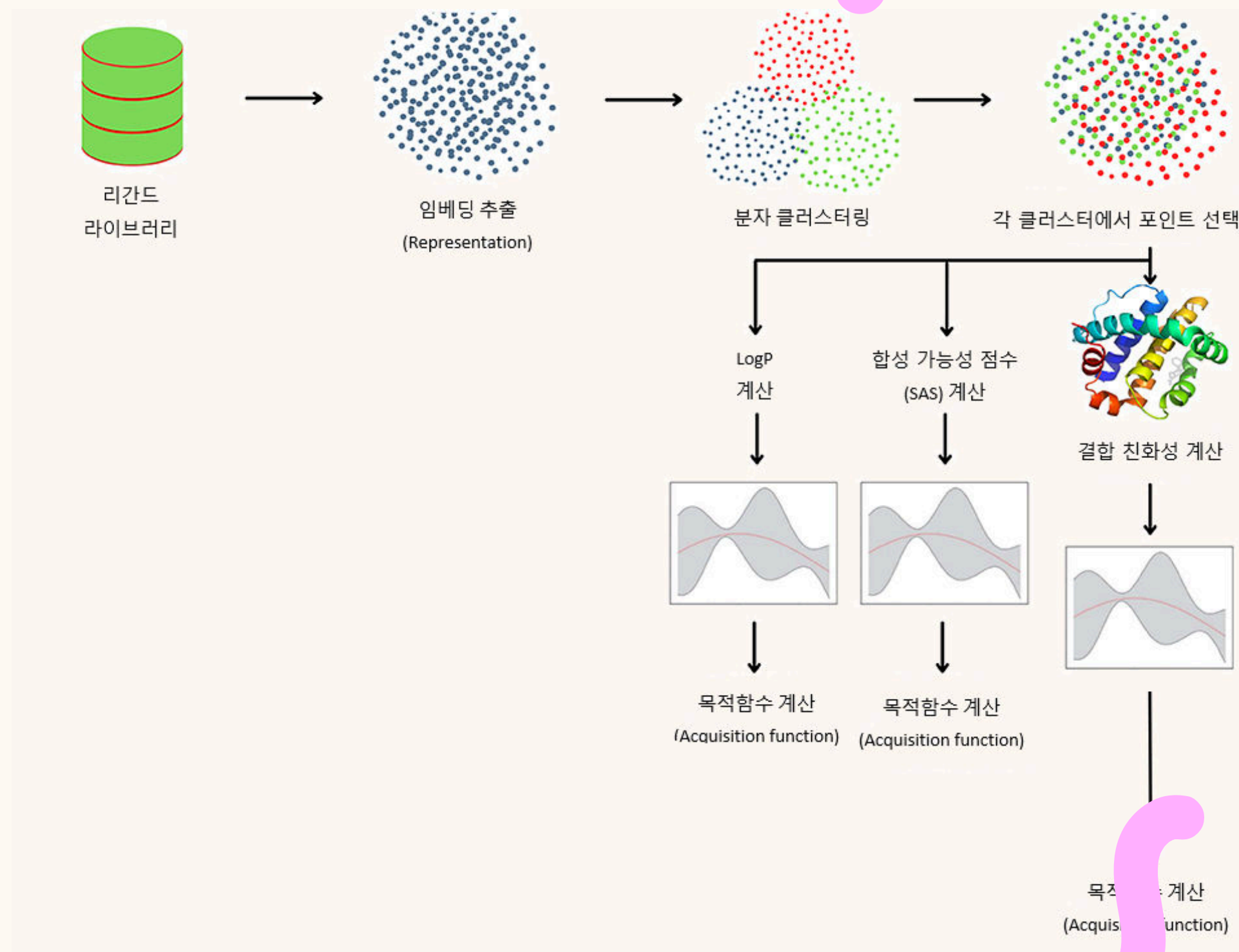


MEMES

MEMES(Machine learning framework for Enhanced MolEcular Screening)

는 Sarvesh Mehta등이 제시한 머신러닝 기반의 가상 스크리닝 프레임워크로, 베이지안 최적화를 통해 대규모 약물 라이브러리에서 잠재적인 신약 후보 물질을 효율적으로 식별하는 알고리즘이다.

클러스터링을 통해 다양한 초기 리간드를 선택해 Surrogate model의 학습 데이터로 쓰이며, 이 모델을 이용해 리간드의 도킹 점수를 예측한다. 이후, 학습된 모델을 기반으로 Acquisition Function을 사용해 다음 탐색할 리간드를 선택한다.

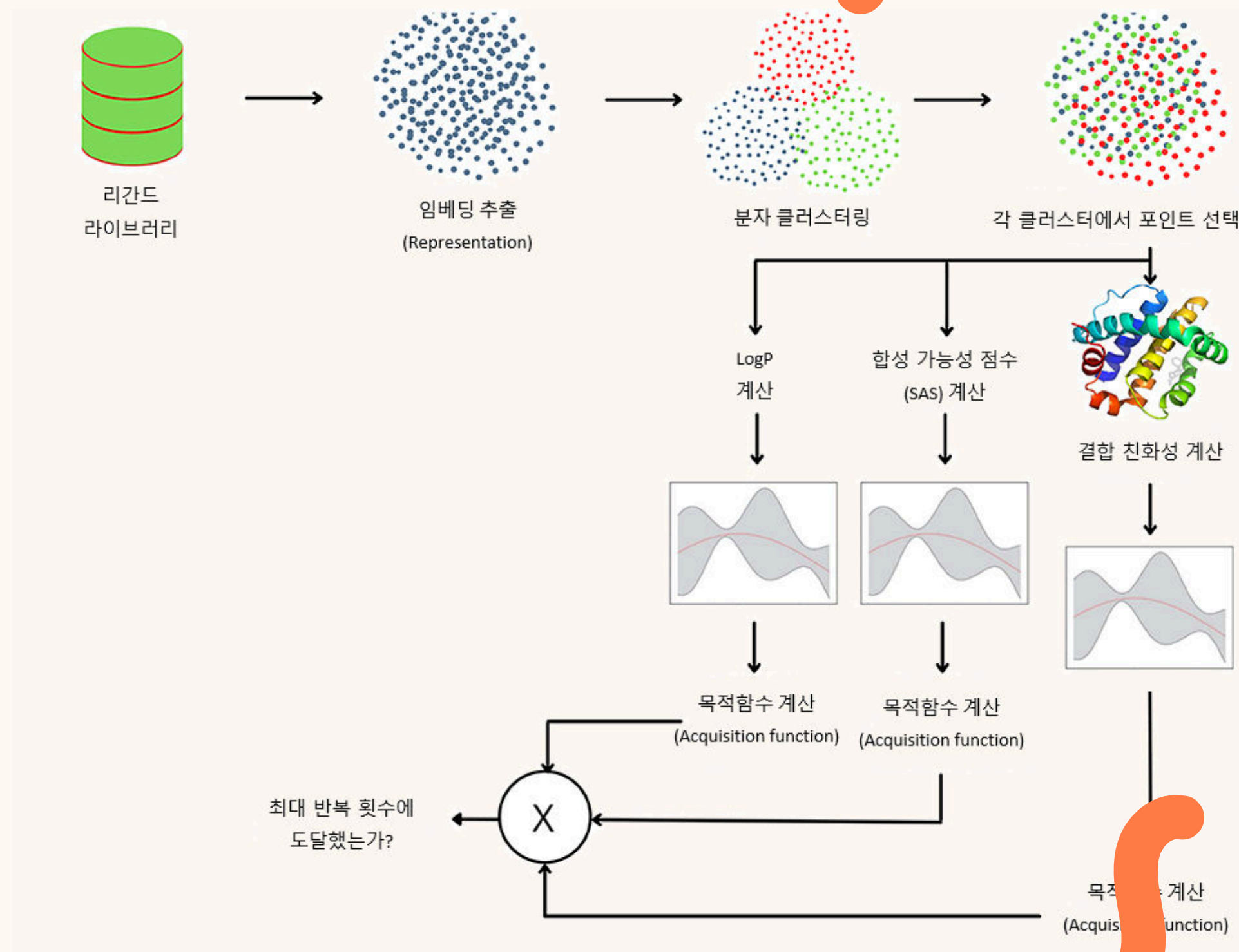


MEMES

MEMES(Machine learning framework for Enhanced MolEcular Screening)

는 Sarvesh Mehta등이 제시한 머신러닝 기반의 가상 스크리닝 프레임워크로, 베이지안 최적화를 통해 대규모 약물 라이브러리에서 잠재적인 신약 후보 물질을 효율적으로 식별하는 알고리즘이다.

클러스터링을 통해 다양한 초기 리간드를 선택해 Surrogate model의 학습 데이터로 쓰이며, 이 모델을 이용해 리간드의 도킹 점수를 예측한다. 이후, 학습된 모델을 기반으로 Acquisition Function을 사용해 다음 탐색할 리간드를 선택한다.

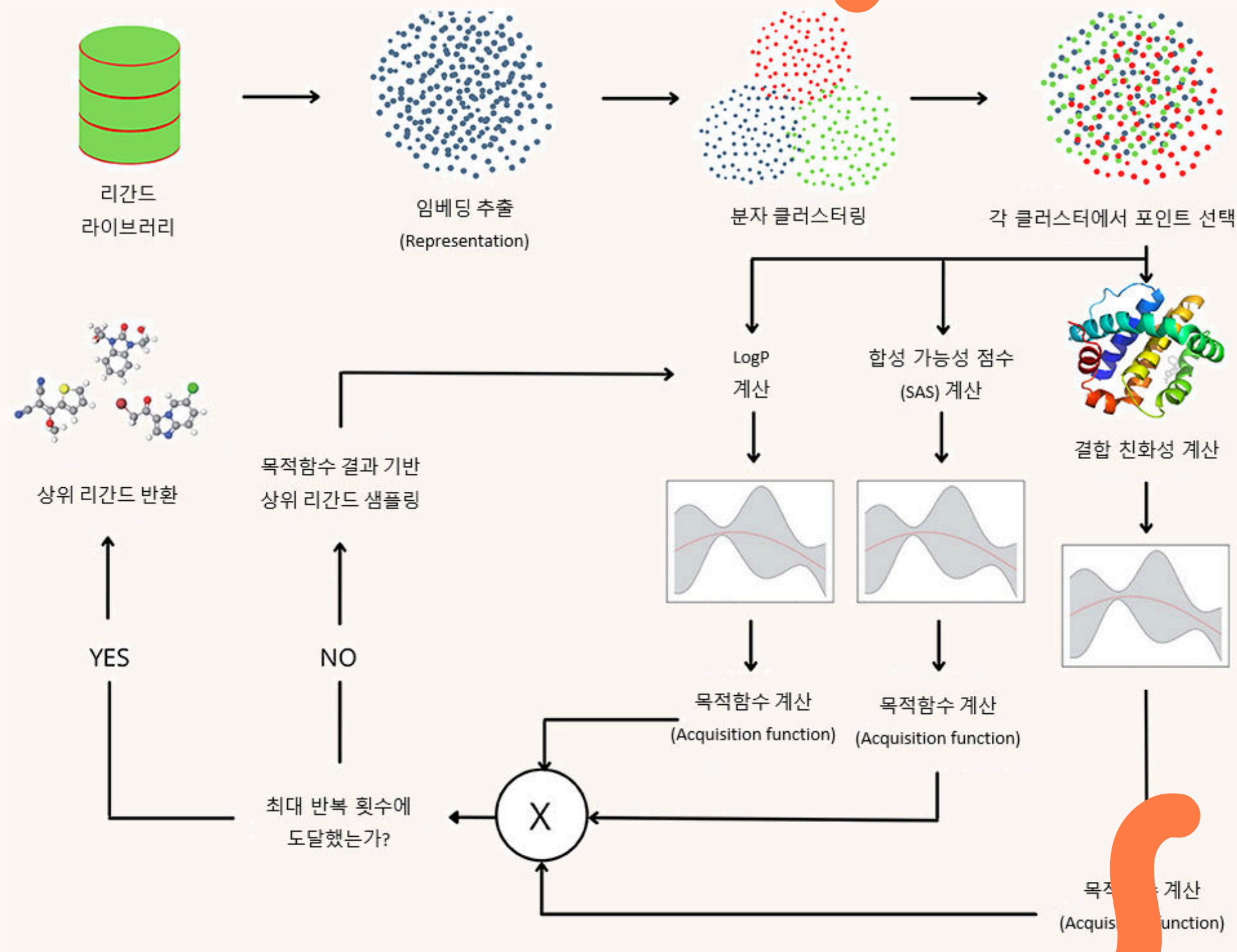


MEMES

MEMES(Machine learning framework for Enhanced MolEcular Screening)

는 Sarvesh Mehta등이 제시한 머신러닝 기반의 가상 스크리닝 프레임워크로, 베이지안 최적화를 통해 대규모 약물 라이브러리에서 잠재적인 신약 후보 물질을 효율적으로 식별하는 알고리즘이다.

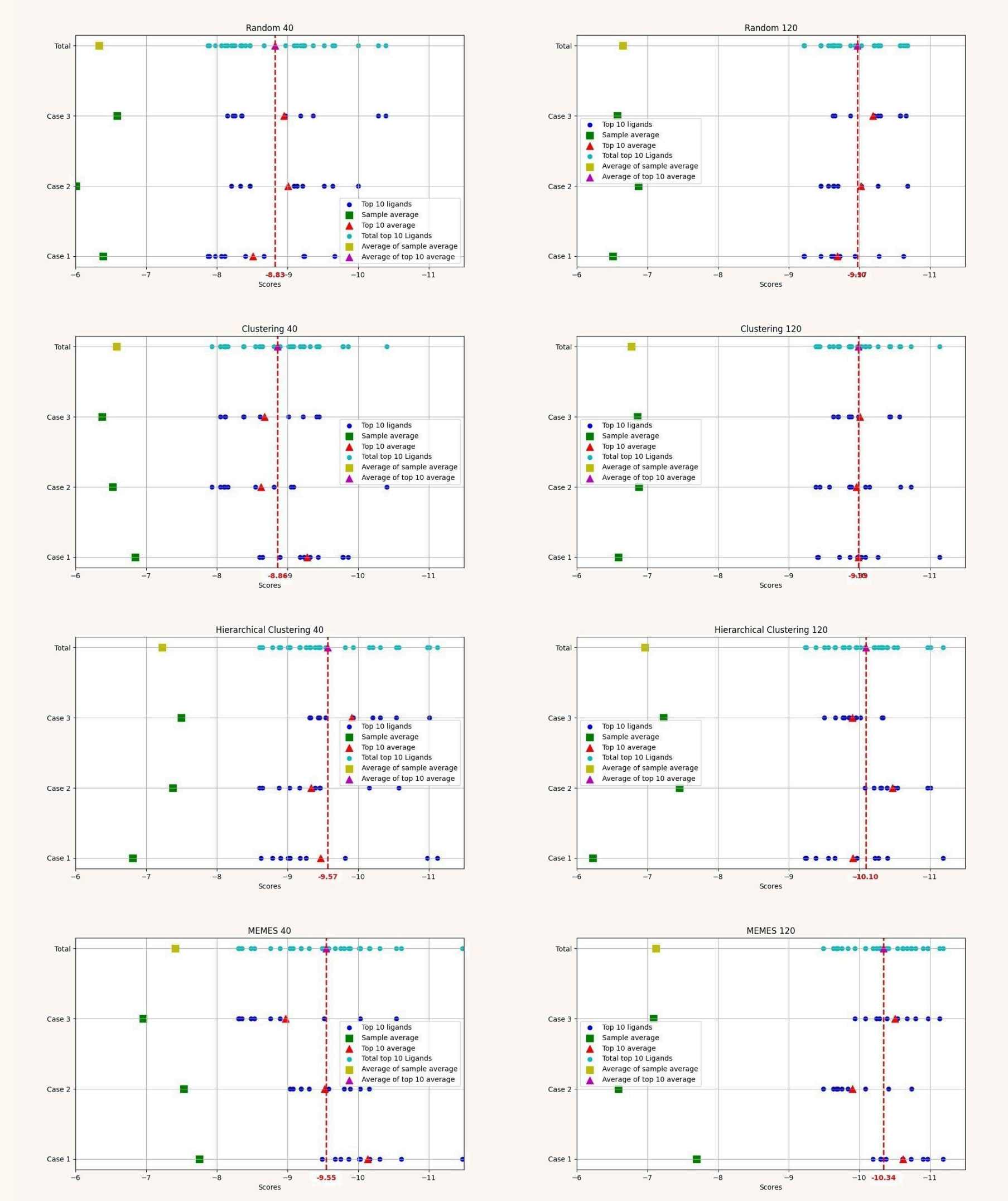
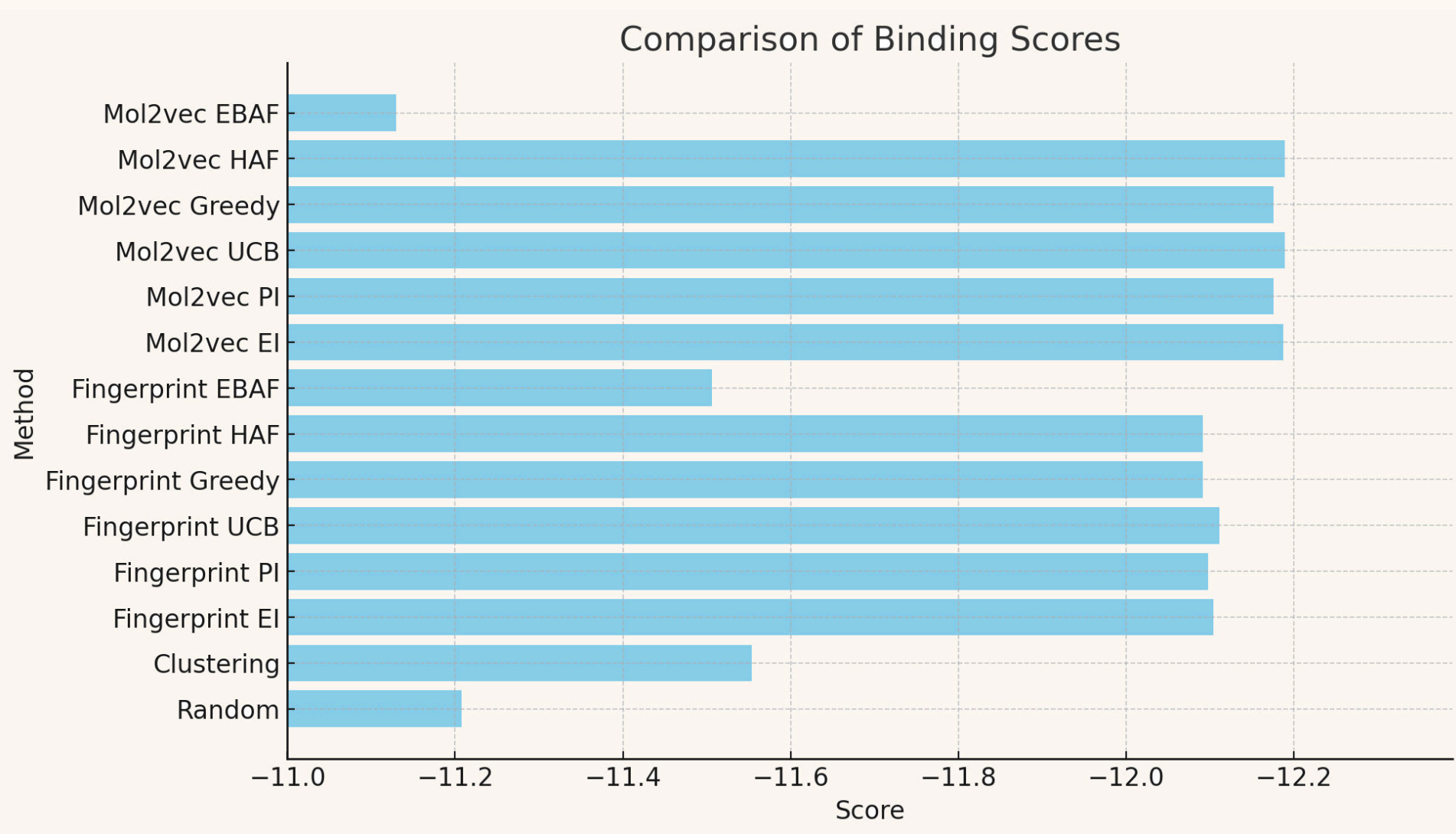
클러스터링을 통해 다양한 초기 리간드를 선택해 Surrogate model의 학습 데이터로 쓰이며, 이 모델을 이용해 리간드의 도킹 점수를 예측한다. 이후, 학습된 모델을 기반으로 Acquisition Function을 사용해 다음 탐색할 리간드를 선택한다.



결과

MEMES의 경우 전체 리간드의 1%를 탐색한 결과, 75%의 확률로 가장 결합도가 높은 리간드를 성공적으로 찾아낼 수 있었다.

가장 좋은 성적을 내는 방식의 경우 샘플의 크기를 모집단의 **0.07%**까지 낮춰도 **90%**의 확률로 가장 결합도가 높은 리간드를 성공적으로 찾아낼 수 있었다.



결론

거대 화합물 라이브러리 탐색은 신약 후보물질 발굴 과정에서 효율성을 높이는 핵심 요소이다.

향후에는 Google DeepMind의 AlphaProteo와 NVIDIA의 BioNemo와 같은 최신 AI 기술을 적극 도입해 탐색 방법을 개선할 필요가 있다. 이러한 기술들의 강점을 최대한 활용함으로써 보다 다양하고 고도화된 신약 후보물질을 발굴할 수 있을 것이다.



감사합니다!