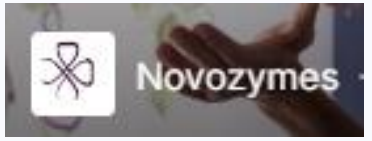
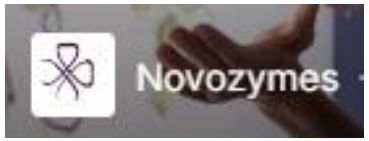
Kaggle Competition

Novozymes Enzyme Stability Prediction →

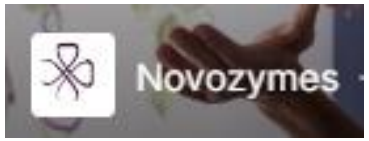
Bioinformatics # 정형 데이터 # NLP # GNN



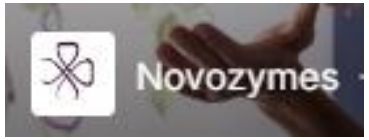
목차



목차



1. 대회 소개



대회 목적

Bioinformatics 분야에서 꾸준히 연구되어 온 Protein에 대한
Single-Point Mutation의 안정성 변화 예측 task 입니다.

Wildtype protein sequence(원형 단백질)에서 한 지점의 변이가 일어났을 때,
Mutant protein sequence(변이 단백질)의 안정성을 예측하는 것입니다.

Spearman 상관관계 공식

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where

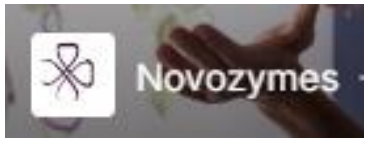
$$d_i = \mathrm{R}(X_i) - \mathrm{R}(Y_i)$$

평가 방법

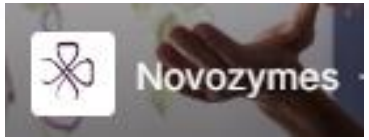
특정 wildtype에 대해 변이된 mutant들의 안정성(Tm)을 예측하여
실제 값과의 **Spearman 상관관계**을 측정하는 것으로 진행됩니다.
(학습 및 검증시에는 주로 ddG를 target으로 설정합니다.)

✓ 두 변수를 rank로 변환한 후의 Pearson 상관관계

목차



2. Background 정리



ddG (delta delta G)

DDG는 single-point mutation이 단백질 안정성에 영향을 미치는 정도를 측정한 metric 입니다. WT, MT 각각의 gibbs free energy 의 변화량에 대한 차이 입니다. 차이를 구하는 순서에 따라 다르긴 하나 **stability와 positive 방향으로 설정 후 학습시킵니다.**

Gibbs free energy (G) = Enthalpy (H) - Temperature (T) x Entropy (S).

PDB (Protein Data Bank)

단백질 정보 파일 형식은 텍스트로 된 파일 형식으로, 단백질 정보 은행에 실려 있는 분자들의 3차원 구조를 설명하는 데 사용됩니다. 해당 파일의 정보를 통해 **단백질 3차원 feature와 여러 biology feature를 추출할 수 있습니다.**

PDB 파일 추출 Public 딥러닝 모델

1. AlphaFold (Google)

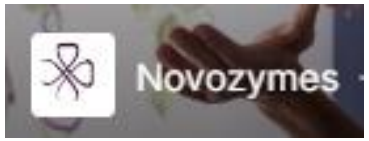
✓ 0~100 사이의 pLDDT 값이 산출, API가 없음
직접 모델을 다운로드 후 이용

2. ESMFold (Meta)

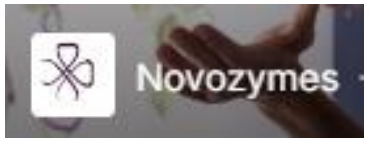
✓ 0~1 사이의 pLDDT 값이 산출, 무료 API 이용 가능

```python
# 파이썬에서 API 활용
import os
os.system(f'curl -X POST --data "{단백질시퀀스}" https://api.esmatlas.com/foldSequence/v1/p
time.sleep(1)
```

목차



3. 연구 결과 정리



1. Study For Deletion Mutant Stability

mutation이 아닌 deletion 된 sequence의 경우 모델 추론이 아닌 imputation 방식으로 예측을 하였는데, 이에 대해 어떤 값으로 imputation 해야 하는지에 대한 연구입니다.

Test environment

Model : pLDDT of wildtype (Rule-Based)

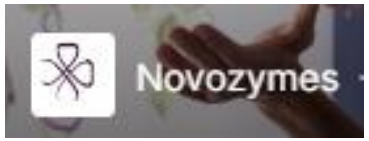
Result

deletion이 아닌 나머지 mutant에 대한 안정성의 Q1 지점으로 할당 한 후 mutation된 거리에 따라 scaling 한 값이 가장 좋았습니다.

Method	LB Score
Q1(25%) 지점으로 할당	0.293
가장 하위로 할당	0.29
Q1(25%) 지점으로 할당 + scaling by distance (*Std.)	0.291
가장 하위로 할당+ scaling by distance (*Std.)	0.291
Q1(25%) 지점으로 할당+ scaling by distance (*MAD)	**0.294**

Std. : 1 standard deviation, MAD : min absolute deviation

3. 연구 결과 정리



2. Study For Amino Acid Group Feature

변이된 아미노산이 본인과 다른 그룹의 아미노산으로 변이했을 때 안정성이 저하될 확률이 높습니다.
이에 대한 아이디어가 실제 유효한지에 대한 연구입니다.

Test environment

Model : Linear Regression (5-CV)

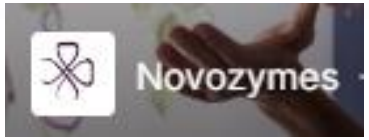
Target transformation : log1p (only ddG < 0)

Input feature 구성

변이된 위치의 WT & MT 아미노산을 오른쪽 코드와
같이 그룹핑한 후 onehot encoding을 합니다.

```
# 4가지 그룹외 나머지 그룹은 AA4라는 그룹으로 할당
aa_groups ={
  # Electrically Charged Side Chains - positive
  "AAG0": ["R", "H", "K"],
  # Electrically Charged Side Chains - negative
  "AAG1": ["D", "E"],
  # Polar Uncharged Side Chains'
  "AAG2": ["S", "T", "N", "Q"],
  # Hydrophobic Side Chains
  "AAG3": ["A", "V", "I", "L", "M", "F", "Y", "W"],
}
tmp = []
for i in list(aa_groups.values()):
    tmp.extend(i)
aa_groups["AAG4"] = diff(list(aa_map.values()), tmp) + "X"
```

3. 연구 결과 정리



2. Study For Amino Acid Group Feature

변이된 아미노산이 본인과 다른 그룹의 아미노산으로 변이했을 때 안정성이 저하될 확률이 높습니다.
이에 대한 아이디어가 실제 유효한지에 대한 연구입니다.

Test environment

Model : Linear Regression (5-CV)

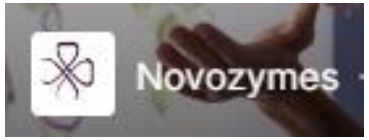
Target transformation : log1p (only ddG < 0)

Result

아미노산 그룹 feature만을 사용했을 때가 가장 성능이 좋았습니다. 이는 아미노산 개별 feature 를 사용하는 것보다 약 17.8% 더 높은 Spearman 상관관계를 보였습니다.

Method	LB Score
WT, MT 아미노산을 개별적으로 onehot encoding	0.118
WT, MT 아미노산 그룹을 개별적으로 onehot encoding	**0.139**
WT & MT 아미노산 + WT 그룹 & MT 그룹	0.134

3. 연구 결과 정리



3. Study For Target Transformation

참가자들이 주로 ddG가 0보다 작은 샘플만 추출하여 학습을 시키는 것을 보았는데,
이는 변이가 되면 안정성이 떨어지는 경우가 많기 때문입니다.
하지만 이러한 sample만을 가지고 학습 시키게 되면 **target의 왜도가 높아져서 예측 분포가 적절히 나오지 않는 문제가 발생하는데** 이를 해결하기 위한 방법에 대한 연구입니다.

Test environment

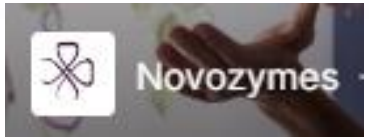
Model : Linear Regression (5-CV)
Feature : WT group, MT group feature

Result

Log1p transformation을 한 경우가 하지 않거나
Group ranking을 한 경우보다 약간 더 좋았습니다.

Method	LB Score
No transformation	0.119
Log1p Transformation	**0.126**
Group ranking (Ranking by WT)	0.114

3. 연구 결과 정리



4. Study For Scaling Factor on Wildtype Biology Feature

BLOSUM Matrix, pLDDT, SASA, Residue Depth, CA Depth와 같은 Biology feature가 안정성을 예측하는 데 큰 기여를 하는 것으로 파악했습니다. 그러나 이러한 feature는 추출하는 데 시간이 많이 소요됩니다. 예로, Residue Depth의 경우 PDB파일을 추출한 뒤 따로 특정 라이브러리 알고리즘을 통해 추출해야 합니다. 이러한 과다한 시간적 소요를 해결하고자 연구한 내용입니다.

Ideation

우연히 우측 그림과 같이 두 아키텍처의 성능이
유사하게 나오는 것을 발견하고,
mutant의 PDB 파일을 추출하지 않고도 wildtype의 PDB 파일 정보와 BLOSUM100 수치를 이용한다면 추정하여 비슷한 수준의 성능을 낼 수 있다고 생각하였습니다.

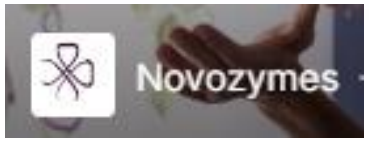
<그림1: WT pLDDT와 MT pLDDT의 차이를 이용>

NESP - Difference of pLDDT
Notebook copied with edits from Chris Deotte · Updated 2d ago
Score: 0.343 · Private · 0 comments · Novozymes Enzyme Stability Prediction +3

<그림2: WT pLDDT에 BLOSUM100 수치 만큼 패널티를 부여하여 stability 추정>

NESP - pLDDT with BLOSUM100 Score Scaling
Notebook copied with edits from Cafelatte1 · Updated 1d ago
Score: 0.343 · 0 comments · Novozymes Enzyme Stability Prediction +3

3. 연구 결과 정리



4. Study For Scaling Factor on Wildtype Biology Feature

BLOSUM Matrix, pLDDT, SASA, Residue Depth, CA Depth와 같은 Biology feature가 안정성을 예측하는 데 큰 기여를 하는 것으로 파악했습니다. 그러나 이러한 feature는 추출하는 데 시간이 많이 소요됩니다. 예로, Residue Depth의 경우 PDB파일을 추출한 뒤 따로 특정 라이브러리 알고리즘을 통해 추출해야 합니다. 이러한 과다한 시간적 소요를 해결하고자 연구한 내용입니다.

Test environment

PDB Extractor : ESM Fold

Operator : Plus / As a Rate

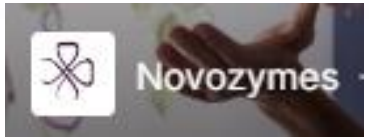
Operation value : BLOSUM100, Demask

Result

pLDDT 뿐만 아닌 다른 Biology feature에 적용했을 때 가장 좋은 성능을 보인 operation을 정리하였습니다.

Base Feature	Best Operation	Formula (ddG와 positive 방향)	LB Score
pLDDT	Demask / As a Rate	(-1) * (pLDDT * (1 - demask)	0.38841
SASA	Demask / As a Rate	SASA * (1 + demask)	0.41555
Residue Depth	**Demask / As a Rate**	**(-1) * (RD * (1 - demask)**	**0.42005**
CA Depth	Demask / As a Rate	(-1) * (CA * (1 - demask)	0.35859
RMSD	Demask / As a Rate	RMSD * (1 + demask)	0.38238

3. 연구 결과 정리



5. Study For HTMD 3D Input Feature

HTMD feature input 구성에 대한 연구입니다. 3D Conv 기반 모델에 적용할 다양한 input 구성을 테스트 하였습니다.

Test environment

Optimizer : AdamW

Learning Rate : 1e-3

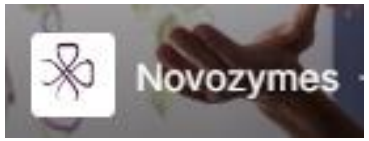
Weight Decay : 1e-4

Result

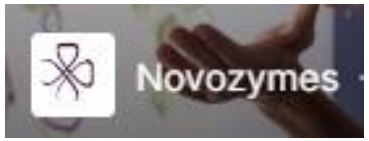
WT와 MT의 차이 및 비율 feature를 구하고 channel에 대해 concatenate 한 feature를 때 사용했을 때 가장 성능일 보여줬습니다.

Method	Formula (|| 은 concatenate를 의미)	LB Score
WT & MT 그대로 concat	WT raw || MT raw	0.397
WT & MT 차감	WT – MT	0.43
WT & MT 차감 및 WT와 MT의 절대 거리	WT – MT || ABS(WT – MT)	0.468
WT & MT 차감 및 WT와 MT의 비율	**WT – MT || (WT / (MT + 1e+1))**	**0.486**
WT & MT 차감 및 WT와 MT의 변화율	WT – MT || ((WT – MT) / (MT + 1e+1))	0.437

목차

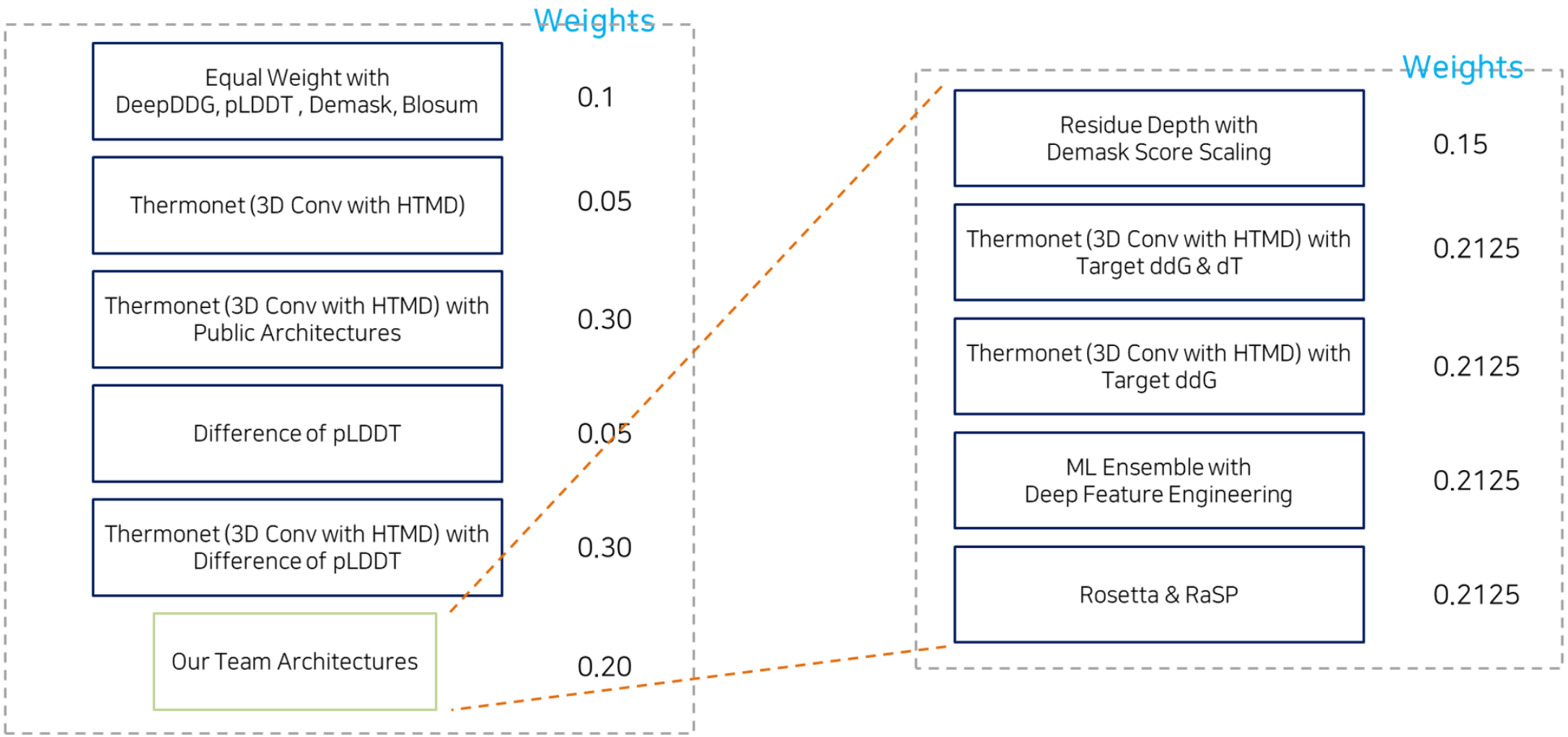


4. 최종 솔루션 소개

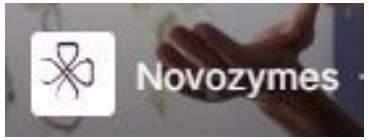


Our Team Architectures	Public Architectures
Residue Depth with Demask Score Scaling	Equal Weight with DeepDDG, pLDDT , Demask, Blosum
Thermonet (3D Conv with HTMD) with Target ddG & dT	Thermonet (3D Conv with HTMD)
Thermonet (3D Conv with HTMD) with Target ddG	Thermonet (3D Conv with HTMD) with Public Architectures
ML Ensemble with Deep Feature Engineering	Difference of pLDDT
Rosetta & RaSP	Thermonet (3D Conv with HTMD) with Difference of pLDDT

4. 최종 솔루션 소개


Weights
Equal Weight with DeepDDG, pLDDT , Demask, Blosum
0.1
Thermonet (3D Conv with HTMD)
0.05
Thermonet (3D Conv with HTMD) with Public Architectures
0.30
Difference of pLDDT
0.05
Thermonet (3D Conv with HTMD) with Difference of pLDDT
0.30
Our Team Architectures
0.20
Weights
Residue Depth with Demask Score Scaling
0.15
Thermonet (3D Conv with HTMD) with Target ddG & dT
0.2125
Thermonet (3D Conv with HTMD) with Target ddG
0.2125
ML Ensemble with Deep Feature Engineering
0.2125
Rosetta & RaSP
0.2125

4. 최종 솔루션 소개

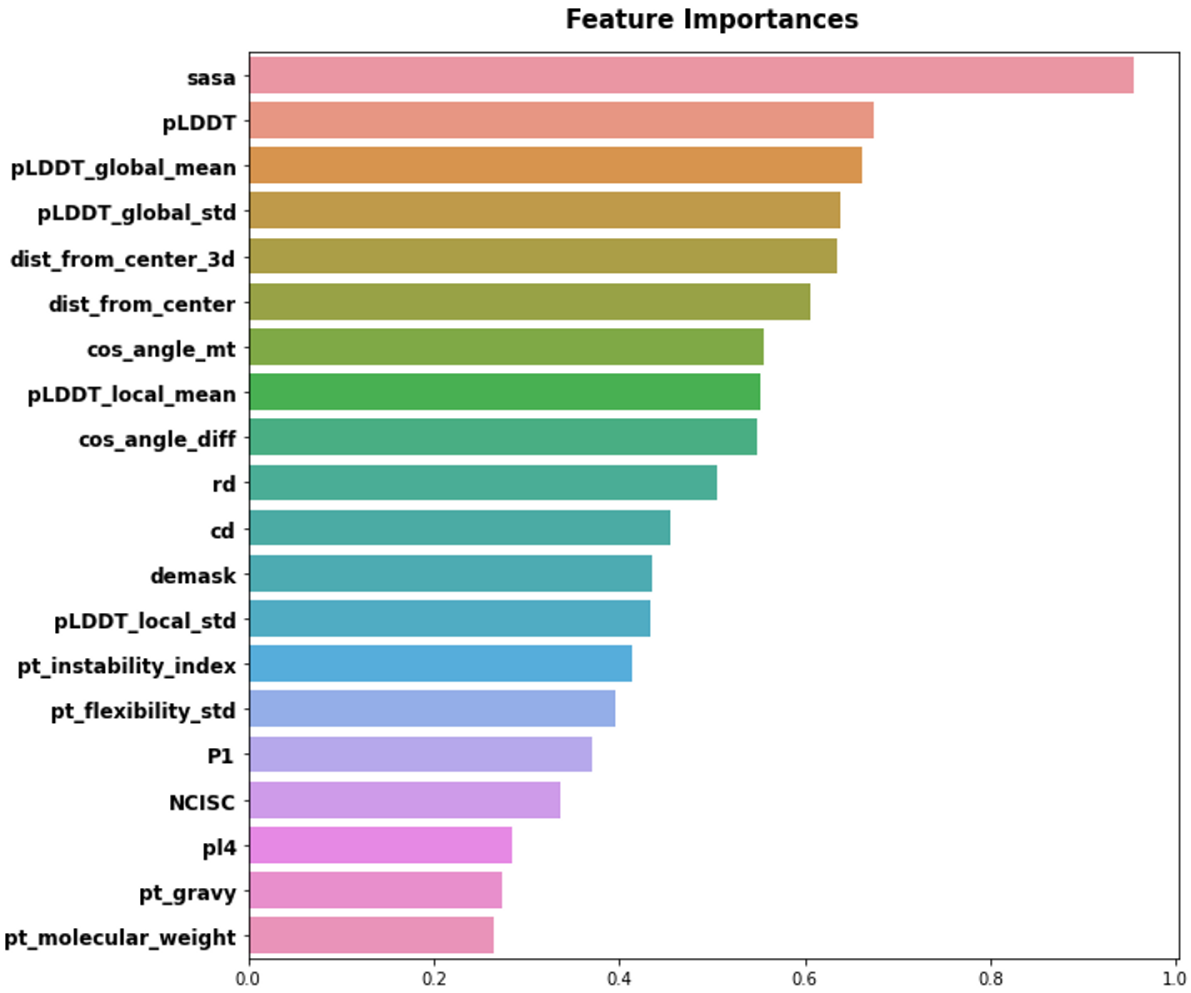


Key Points

1. Bioinformatics 관련 feature(SASA, Residue Depth 등)들이 비교적 큰 영향력이 있어, 이 feature 들을 활용한 Rule-Based 모델들이 좋은 성능을 보여주었습니다.

2. deep feature engineering를 통한 ML모델 앙상블 아키텍처는 **절대적인 성능은 높지 않았으나, Public-Private 성능 변동이 적었습니다.**
(Public : 0.454, Private : 0.462)

✓ 우측 그림과 같이 SASA, pLDDT feature와 같은 변이가 발생한 잔기의 biology feature와 더불어 변이가 일어난 위치도 높은 중요도를 보이고 있습니다



<그림: Feature importances>

4. 최종 솔루션 소개

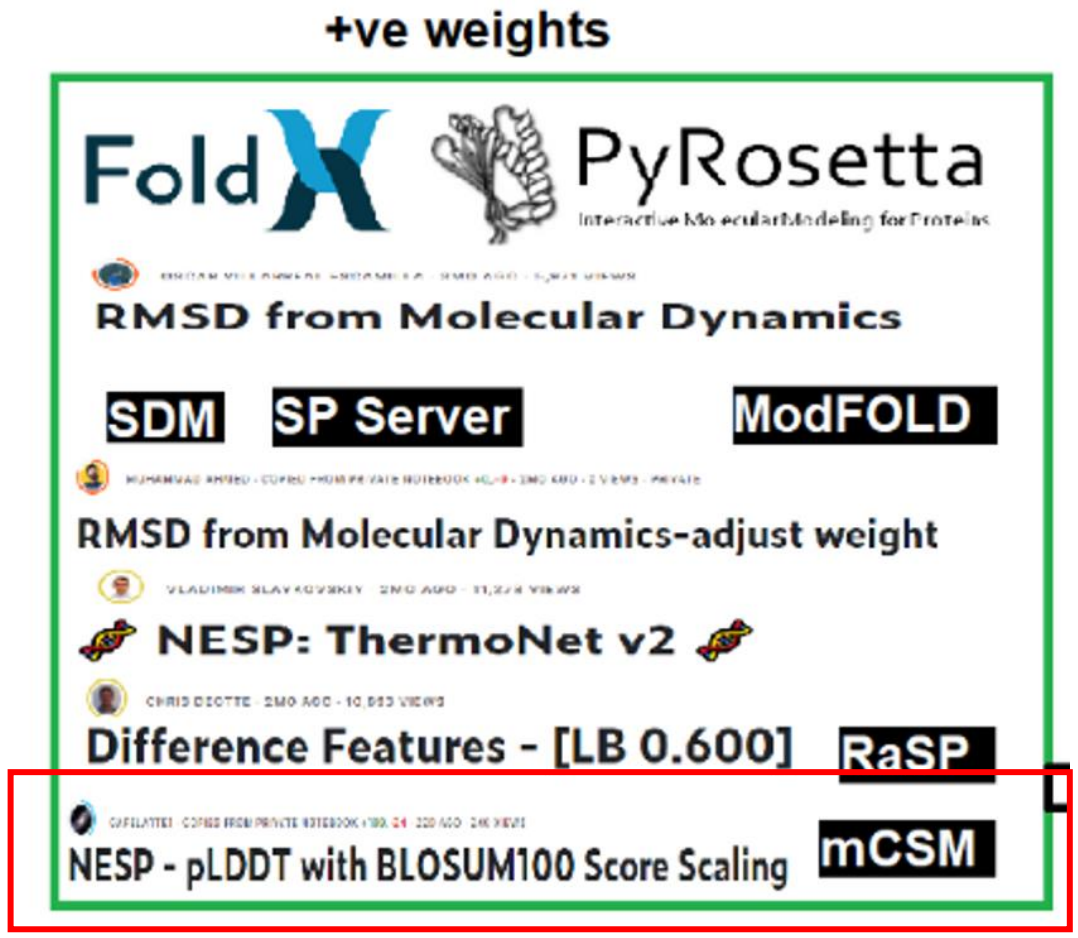
Key Points

3. HTMD 3D feature를 활용한 3D Conv 기반 DL 모델도 높은 성능을 보여주었습니다. (ThermoNet)

4. Rosetta, RaSP, DeepDDG 등 여러 ML 기반 Public 모델들을 사용했으나 압도적인 퍼포먼스를 보인 것은 없었습니다.

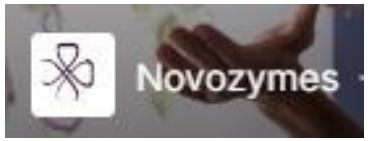
5. wildtype의 bioinformatics feature 들을 특정 factor로 scaling하여 mutant의 값을 추정하는 방법은 **mutant의 PDB 추출 과정 없이 높은 예측력을 보여주었습니다.**

✓ pLDDT에 BLOSUM100을 이용한 저희 팀 알고리즘이 **대회 2위를 차지한 팀 아키텍쳐에 사용되었습니다.**

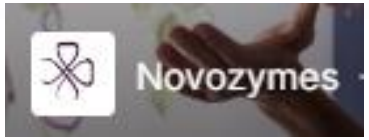


<그림: 2위 팀 Solution Architecture>

목차



5. 정리 및 한계점



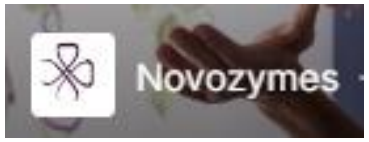
- 상위권만 고려하더라도 현업에서 사용될 만큼의 성능을 보여주지는 못했습니다. 아직 더 많은 연구가 필요해 보입니다. **그러나 biology feature가 퍼포먼스에 크게 영향을 미친다는 사실을 알 수 있었습니다.** 이를 더 정교하게 활용할 방법을 찾아야 할 것으로 보입니다.

- 본 대회는 하나의 wildtype protein sequence에 대한 여러 mutant의 안정성을 예측하는 것이어서, 예측 성능의 신뢰성이 크게 높지 않다고 보입니다.

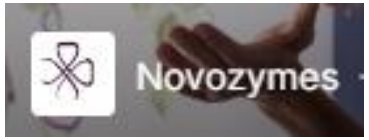
- PDB를 통해 추출한 **HTMD 3D voxel feature**를 활용한 3D Conv 기반 딥러닝 모델도 높은 성능을 보여주었습니다. 그러나 Private 성능이 모두 크게 떨어지는 현상이 있었습니다. **3차원의 정보인 만큼 많은 데이터를 확보하거나 augmentation 기법을 잘 활용해야 좋을 것 같다고 보았습니다.**

- 단백질 관련되어서는 **GNN 모델**도 많이 활용되나, 활용해보지 못한 부분이 아쉽습니다. 1위 Solution에 GNN 모델이 포함되어 있었던 만큼 추후 활용 가치가 있어 보입니다.

Thank You



Appendix



✓ Our Teams

김영준 (작성자 - Data & ML/DL Engineering)

은현종 (Domain knowledge Researching)

✓ Contact

E-Mail : flash659@gmail.com

✓ Kaggle Profile & More Details

Cafelatte1 | Expert | Kaggle

Notion Page - Kaggle NESP

✓ References

DDG-v2.pdf (cyrusbio.com)

AlphaFold Protein Structure Database (ebi.ac.uk)

facebookresearch/esm: Evolutionary Scale Modeling (esm): Pretrained language models for proteins (github.com)