第26回オープンバイオ研究会
2022年3月12日(土) @Zoom

ランク学習を用いた化合物スクリーニングにおける多様なアッセイデータの統合戦略

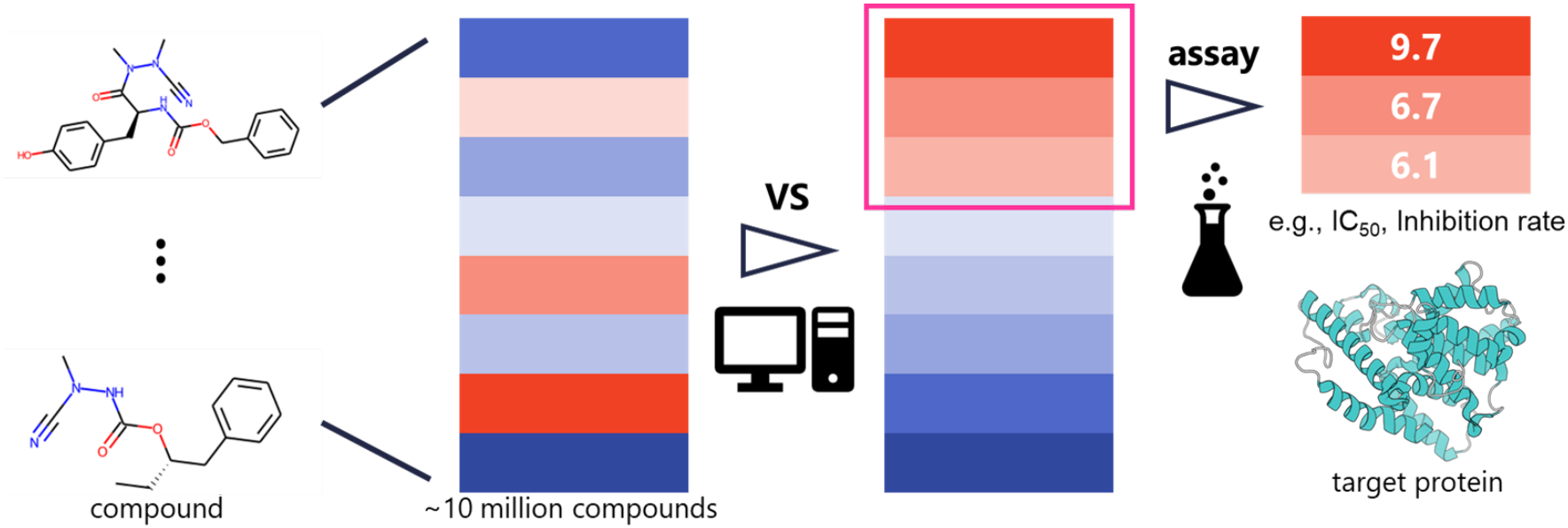
古井 海里　　**大上 雅史**

東京工業大学 情報理工学院 情報工学系

東京工業大学

バーチャルスクリーニング (Virtual Screening, VS)

標的タンパク質と化合物間の活性を計算機によって予測



化合物をより適切に並び替えるのがVSの目的 ▶ **ランク学習**

背景　2.ランク学習によるバーチャルスクリーニング

ランク学習によるVS（Zhang, et al., 2015[1]）

1. **ランキング予測の精度が高い**
   - ただし値の直接の予測はできない
2. **異なる実験データを混ぜて訓練データに用いることができる**

   IC_{50}など生化学アッセイは環境によって分布が異なる
   - 単純に回帰学習を行うのは難しい
   - 順序関係を考慮するランク学習は異なる環境データの統合に適している

[1] Zhang, Wei, et al. Journal of Cheminformatics, 7.1 2015. 1-13.

研究目的

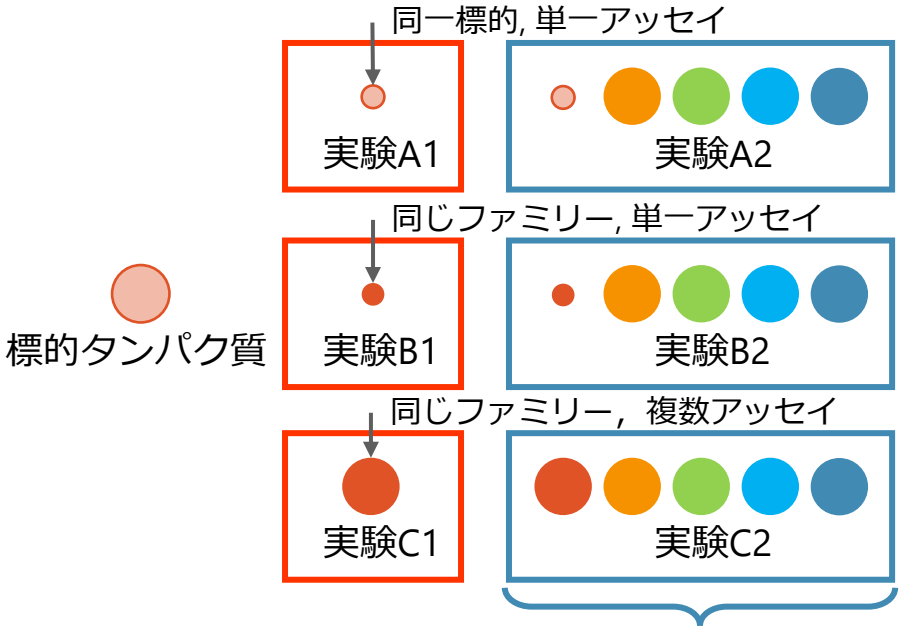
- **多様な学習データの状況*でランク学習手法のVS性能を検証**
- **既存の評価指標 (NDCG) よりも予測の良さを直接評価できる指標の考案**

*同じファミリーのタンパク質とのアッセイ情報が全く無い・少しある等

結果

1. ランク学習は回帰学習を**上回るか同等**の予測精度
2. 標的と**関連するタンパク質情報のみ**を用いるのが効果的
3. 新規に提案した**NEDCG**という評価指標がVSに適していた

実験設定



標的タンパク質について単一アッセイのデータセットがある

標的と同一ファミリーのタンパク質について**単一アッセイ**のデータセットがある

標的と同一ファミリーのタンパク質について**複数アッセイ**のデータセットがある

標的と関連性の低いのタンパク質情報を訓練データに追加
▶幅広い情報は学習に役立つか？

標的に関連するアッセイデータが存在しない

実験A1と実験C1の組み合わせ

VS性能を評価するランキング評価指標 (NDCG, NEDCG)

Discounted Cumulative Gain (DCG)

$$\mathrm{DCG@}K = \sum_{i=1}^{K} \frac{g_i}{D_i}$$

$g_i = 2^{y_i} - 1$
$D_i = \log_2(1+i)$
i : 予測された順位
y : 正解ラベル

予測の上位K件に高い活性を持つ化合物を列挙できているかを評価

➡ **VSの性能評価に適している**

Normalized DCG (NDCG)

- 正しい予測のとき1になるようにDCGを正規化
  - 情報検索分野で提案され，既存のランク学習によるVSでも用いられていた

$$\mathrm{NDCG@}K = \frac{\mathrm{DCG@}K}{\mathrm{maxDCG@}K}$$

NDCGは他のモデルと比較して優れているかどうかしか分からない
▶ VSでは**実際にどれくらい価値のある予測をしているか**が知りたい

Normalized Enrichment DCG (NEDCG)　（本研究で新規に提案）

- ランダムな予測のとき0，正しい予測のとき1になるように補正したNDCG
  - VSにとって**予測がどれくらい価値があるのか**を評価するために導入

$$\mathrm{NEDCG@}K = \frac{\mathrm{DCG@}K - \mathbf{randomDCG@}K}{\mathrm{maxDCG@}K - \mathbf{randomDCG@}K}$$

データセット

- 24個のタンパク質のIC_{50}に関する生化学アッセイを**ChEMBLデータベース**から収集
- normalized Smith-Waterman Scoreが0.2以上のペアのあるタンパク質をテストデータとして扱う

タンパク質ごとのデータ数

タンパク質 z	総データ数	アッセイ数
ATP-binding cassette sub-family G member 2	1,193	41
Acetylcholinesterase	5,450	199
Arachidonate 5-lipoxygenase	2,840	107
Cannabinoid CB1 receptor	1,559	52
Cyclooxygenase-1	3,030	89
Cyclooxygenase-2	5,475	172
Cytochrome P450 19A1	1,876	75
Dipeptidyl peptidase IV	3,956	145
Epidermal growth factor receptor erbB1	10,598	384
Epoxide hydratase	2,283	72
Estrogen receptor alpha	3,633	93
Estrogen receptor beta	2,150	56
Glucocorticoid receptor	3,394	108
HERG	7,979	260
Hepatocyte growth factor receptor	3,556	120
Histone deacetylase 1	4,913	176
Monoamine oxidase A	3,263	104
Monoamine oxidase B	3,882	161
P-glycoprotein 1	1,201	50
Peroxisome proliferator-activated receptor gamma	1,734	55
Prostaglandin E synthase	1,837	57
Serotonin transporter	3,636	108
Sodium:glucose cotransporter 2	1,406	47
Vascular endothelial growth factor receptor 2	10,796	306

実験A,Bの訓練データとして用いるデータ

Target	標的名	データ数
MO-A	Monoamine oxidase A	40
MO-B	Monoamine oxidase B	53
CO-1	Cyclooxygenase-1	34
CO-2	Cyclooxygenase-2	36
ER-α	Estrogen receptor alpha	107
ER-β	Estrogen receptor beta	108

ex) 実験B1（同一ファミリーで学習）
- 訓練データ　：MO-A
- テストデータ：MO-B

本研究で用いる手法

本研究で用いた手法

lambdarank (GBDT) [2]
NDCGの変化を考慮した損失関数

比較する手法

regression (GBDT)
回帰学習による損失関数

GBDTモデルにはLightGBMの実装を利用

特徴量

- 化合物
  - mordredによる化合物記述子の特徴量(1,452次元)
- タンパク質
  - PyBioMedのCTD記述子(147次元)
  - (composition, transition and distribution)

評価指標

- NEDCG%20
  （上位20%に関するNEDCG）

学習方法

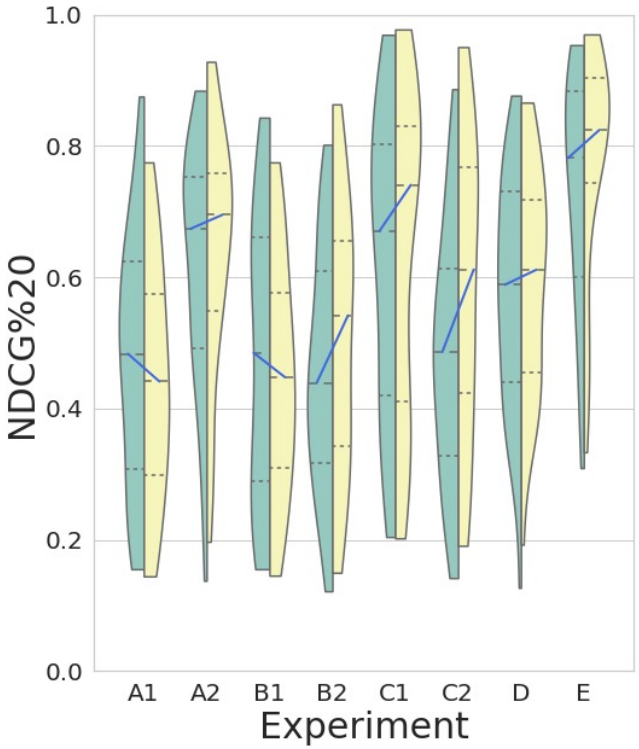
- 3-foldの交差検証でグリッドサーチによるパラメータ探索
- テストデータと同じアッセイは訓練データに含まれない

[2] Burges, Christopher, Robert Ragno, and Quoc Le. Advances in Neural Information Processing Systems, 19, 2006.

結果：NEDCGとNDCG

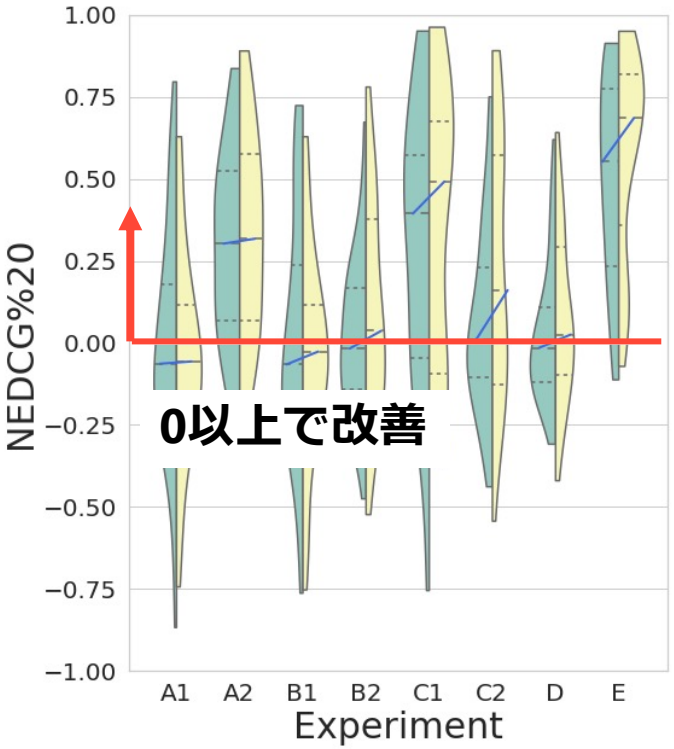
既存の評価指標

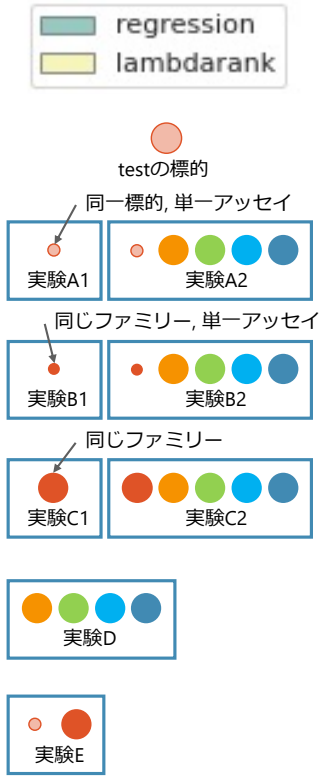
NDCG%20



提案した評価指標

NEDCG%20

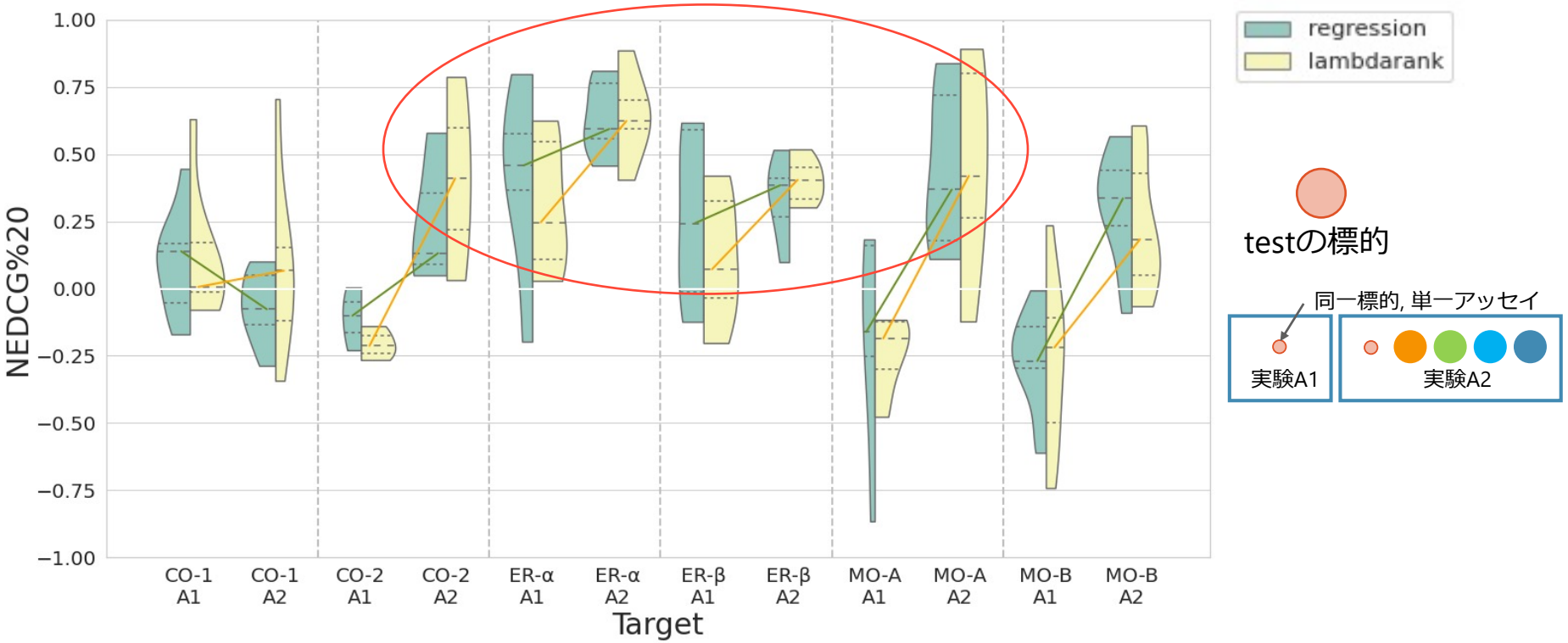




NEDCGは複数のテストデータの分布の評価に優れる

- NEDCGが0より大きいかで**簡便に予測の良し悪しが評価できる**
- NDCGでは複数データを統括して予測結果の良さを比較できない
  (テストデータごとにrandomDCGは異なるため)

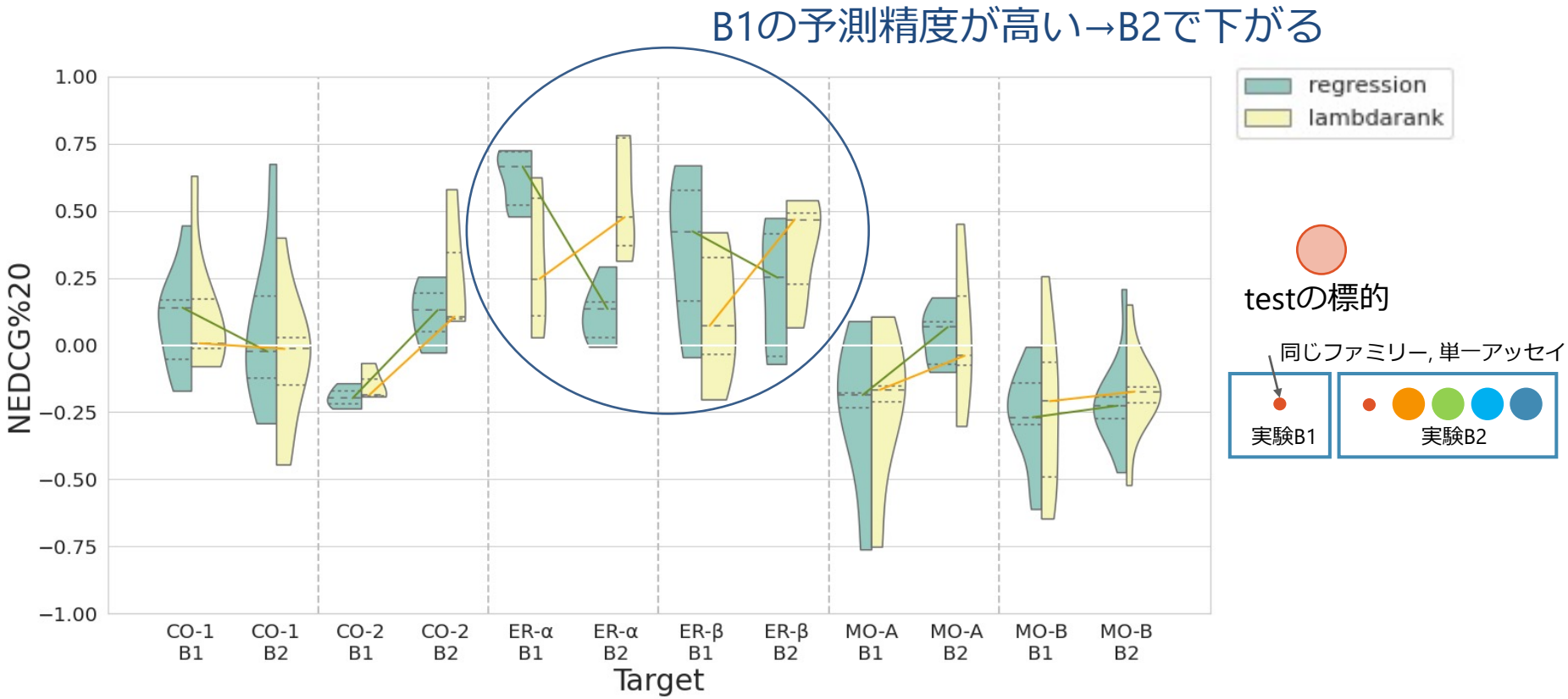
結果：実験A（標的タンパク質の単一アッセイ）



周辺情報を与えることで予測精度が向上する

- 概ねA2の方がA1より良い（中央値が右肩下がり）
- **新規標的でも数十件程度のアッセイ情報があれば学習可能**

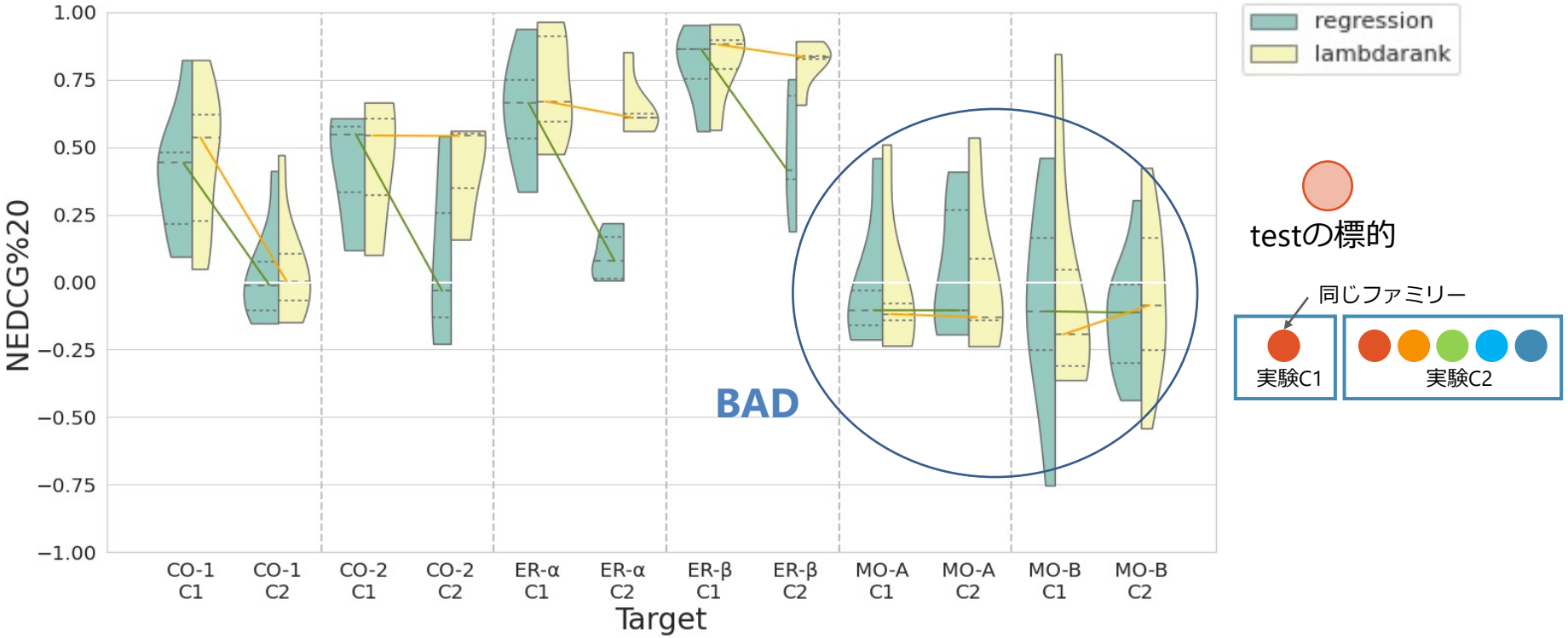
結果：実験B（同じファミリーの単一アッセイ）



周辺情報を集めると学習に悪影響を及ぼすことがある

- 実験B1の予測精度が高いとき，実験B2の予測精度が低下する

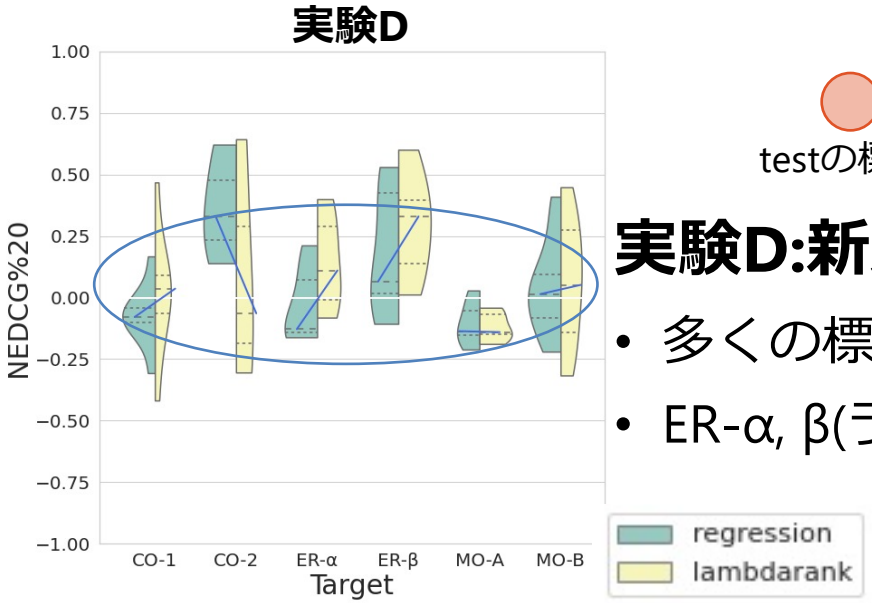
結果：実験C（同じファミリーの複数アッセイ）



関連しないタンパク質情報が学習のノイズになっている

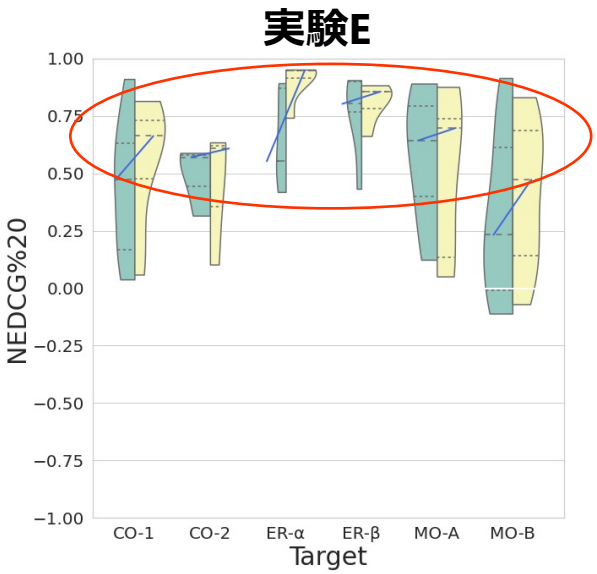
- 実験C1の方が良い（中央値は右肩下がり）
- MO-A, Bの予測精度は低い
  - 同じファミリーのタンパク質に関するアッセイ情報が十分あってもうまく学習できないことがある

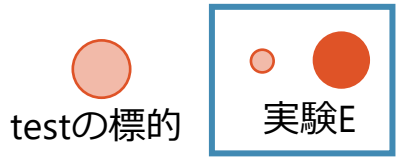
結果：実験D （関連なし）・実験E（多数データ）



実験D:新規標的に関する予測は困難

- 多くの標的でNEDCGの中央値が0を下回る
- ER-α, β(ランク)，CO-1(回帰)でわずかに改善あり

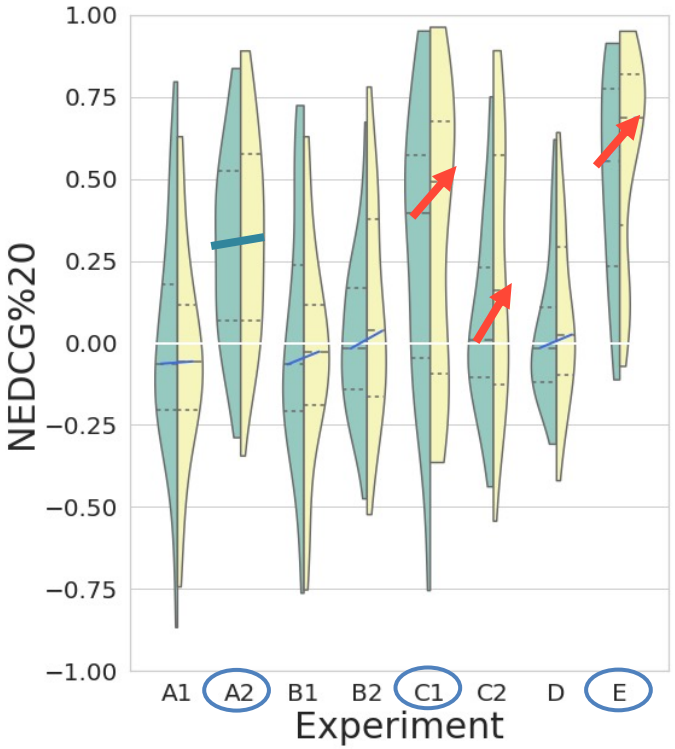


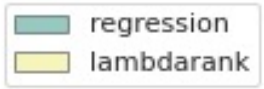


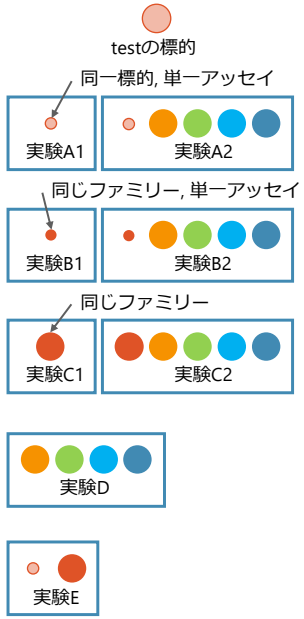
実験E:多くのデータで改善

- **ランク学習の予測精度が全体的に高い**
- 実験Cで精度の低かったMO-A, MO-Bが向上

全体の結果・考察







中央値が0を大きく上回る

標的と関連するアッセイ情報が学習に有効

– 実験Eが最も良く，次点で実験C1, A2が優れていた

ランク学習は同等か上回る予測精度

- 複数アッセイの統合（実験C1, C2, E）● ではランク学習が大きく上回る

結論

1. ランク学習によるVSの効果を様々な学習設定で検証

2. ランク学習は回帰学習を上回るか同等のランキング予測精度

- 特に複数アッセイデータの統合する設定で回帰手法を大きく上回る
- 学習には標的と関連するタンパク質のみを用いるのが効果的

3. NEDCGは予測がVSにとってどのくらい価値があるか評価できる

今後の展望

1. 深層ランク学習やGBDTの改良によって更に**高い予測精度**を目指す
2. **データ統合に優れたランク学習の適用範囲を広げる**ため，膜透過性予測や毒性予測などケモインフォマティクスの他のタスクでの有効性を検証する

謝辞 本研究は、JST ACT-X (No. JPMJAX20A3)、上原記念生命科学財団、JSPS 科研費基盤研究(B)(No. 20H04280)の支援を受けて行われました