Vol.2018-BIO-53 No.9
2018/3/10

多様な活性化合物発見のための化合物 fingerprint アンサンブル手法の開発

松山 祐輔[1,3,a)]　石田 貴士[1,2,3]

概要：機械学習による薬剤活性予測は十分な既知データが存在すれば高い予測精度を達成可能であり，大きな注目を集めている．しかし，そこで活性があると予測される化合物は既知活性化合物に構造などの特徴が類似したものとなる傾向があり，予測結果の上位は同じような構造の化合物ばかりとなってしまうことがある．そのため，構造多様な候補化合物を出力することが現在の機械学習による薬剤活性予測の課題の一つとなっている．そこで本研究では多数の化合物 fingerprint による予測モデルの予測結果を，構造の多様性が出るように工夫して統合することで，単一の fingerprint に基づく従来の予測に比べて，予測精度を低下させずに構造多様性の高い活性候補化合物群を予測する手法を提案する．評価実験の結果，予測精度をあまり落とさずに提案化合物の構造多様性を確保できることを確認し，既存の分子骨格によるクラスタリング手法に比べより高い構造多様性が得られることを確認した．

キーワード：薬剤活性予測, 機械学習, 集団学習, 化合物 fingerprint

Development of a molecular fingerprint ensemble method for discovery of active compounds with diverse structures

Abstract: Machine learning-based drug activity prediction methods often output candidate compounds with similar structures. Thus, discoverying active compounds with diverse structures is one of the important problem in the field. In this research, we propose a method for predicting active candidate compound group with high structural diversity by integration of multiple molecular fingerprint prediction models.

Keywords: drug activity prediction, machine learning, emsemble learning, molecular fingerprint

1. はじめに

近年，創薬プロセスの効率化のため，計算機を用いた化合物ヴァーチャルスクリーニングが注目されている．その中でもリガンドベースのヴァーチャルスクリーニング（Ligand Based Virtual Screening, LBVS）の手法, 特に機械学習を用いた化合物ヴァーチャルスクリーニングはその予測精度の高さから注目を集めている．

このような化合物ヴァーチャルスクリーニングの創薬プロセスにおける適用先として，リード化合物探索におけるヒット化合物探索と言うものがある．これは，すでに標的タンパク質との薬剤活性がわかっているデータを用いて予測モデルを構築し，数百万から数千万もあるような膨大な数の化合物ライブラリから活性がありそうな化合物を選択するというものである．

この過程で選び出された化合物はその後，実際に薬剤活性があるか確かめられた上で体内安定性や毒性，そして副作用と言った観点から厳しい選別および最適化を経て薬として上市される． この際に似たような構造をもつ化合物では毒性や体内安定性などについても同様な傾向を示すことが多く，ヒット化合物が似たような構造ばかりの場

[1] 東京工業大学 大学院情報理工学研究科 計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Meguro, Tokyo, 1552-8660 Japan
[2] 東京工業大学 情報理工学院 情報理工学系
Department of Computer Science, School of Computing, Tokyo Institute of Technology, Meguro, Tokyo, 1552-8660 Japan
[3] 東京工業大学 情報生命博士教育院
Education Academy of Computational Life sciences, Tokyo Institute of Technology
[a)] matsuyama@cb.cs.titech.ac.jp

合，全てのヒット化合物が同じような毒性をもっていて創薬プロセスが終了してしまうということが起こりうる．このため，ヒット化合物探索においてはできるだけ活性がありそうであると予測される化合物群（以下，候補化合物群という）はある程度多様な構造を持っていることが望ましい [1]．しかし残念ながら，一般に機械学習による LBVS では一般に，予測最上位にくる化合物は似たような構造を持ったものばかりに偏りやすいといった問題が存在する．理由としては，機械学習による予測では学習データ中の既知化合物と類似した構造の化合物を予測する傾向があり，特に学習データが偏ってしまっていた場合，そのデータで学習されたモデルでは数千万の化合物ライブラリから上位数パーセントを取り出して提案化合物群とするような操作をしても，必然的にそれらは似たような構造を持つ化合物ばかりになってしまうということになる．

なお，「予測化合物群の偏り」という問題は大きく分けて二つの観点がある．一つ目は提案化合物群そのものが似たような化合物によって占められているという問題であり，もう一つは，既知の活性あり化合物と同じような化合物ばかりが提案化合物群に含まれているという問題である．本研究では前者の観点に注目して研究対象とする．

2. 既存手法

構造に多様性のある予測化合物群を得たいという状況に対し，既存手法では提案化合物を多めにとり，クラスタリングによって似たような化合物を除去するといった手法を適用することにより解決することがある．特によく使われる手法としては，次の二つがあげられる．

2.1 分子骨格によるクラスタリング

ここに属する手法は，化合物の構造をより抽象的な分子骨格 (molecular scaffolds) で表現し，同じ分子骨格を持つ化合物を類似するものとして除去していく手法である．このような手法として代表的なものには Bemis murco scaffolds が挙げられる [2]．この手法においては，環構造とそれらをつなぐリンカによって化合物が構成されていると定義して scaffold 表現をする．

2.2 ベクトルの類似度基づく構造クラスタリング

これらに属する手法は，化合物をベクトル表現した上で，そのベクトル間の距離にもとづいてクラスタリングをする手法である．ケモインフォマティクスでよく使われるこのような手法の一つとして，Butina clustering[3] がある．この手法では，一定以上に類似した化合物によるグループを作成し，一番大きいグループを一つのクラスタとして出力する．そしてそのグループ内の化合物をデータから削除した上でこれらの操作を再帰的に繰り返す．ベクトル間の距離としては Tanimoto 係数が用いられることが多い．

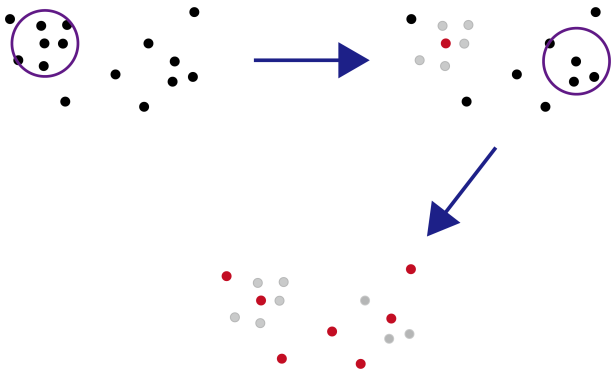

図 1 Butina クラスタリングの概要．黒い点が化合物，紫の丸が中心の化合物からのベクトルの距離を示す． 紫の範囲内に存在する化合物が最も多い化合物がクラスタの中心となり，残りの円内の化合物を消去するという操作を繰り返すと最終的に赤い点で示す化合物が残る．

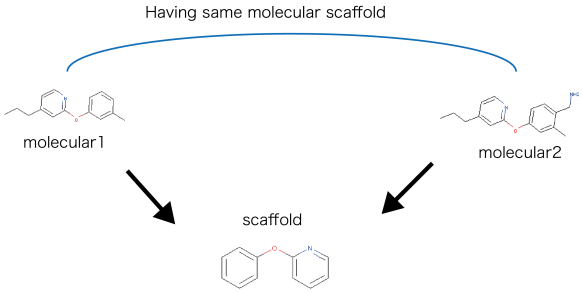

図 2 Bemis Murco Scaffold の例．左と右の化合物は同じ種類の環二つとそれをつなぐリンカ鎖の 3 つから成り立つと表現でき，同一の分子骨格を持つといえる．

これらは簡便であり利用しやすい手法であるが，かなり大雑把に候補化合物を削ってしまうことから，予測性能の低下を引き起こし，構造多様性と予測性能のトレードオフが生じてしまうものと考えられる．

3. Multiple Fingerprint Ensemble と Topic Diversification

3.1 Multiple Fingerprint Ensemble

機械学習における入力として，一般には固定長のベクトルが必要である．一方，化合物は複雑なグラフ構造をしているため，何らかの手法で固定長ベクトルに変換する手法が必要となる．これらの手法は総称して molecular fingerprint と呼ばれ，過去数十年に渡って多数提案されてきた．従来，molecular fingerprint 手法は標的タンパク質やタスクごとに適切な手法を選択するというのが主流であった．それに対して我々のグループでは，主要な molecular fingerprint は異なった特徴を捉えたベクトルを生成していること，さらにそれぞれの fingerprint で学習された予測モデルの予測上位化合物はある程度異なることを示した [4]．その上で，メタ学習の枠組みを用いて多数の異なる fingerprint による

予測結果のアンサンブル（二段階の予測モデルの構築）を行うことにより，単一の fingerprint を用いた場合に比べて高い精度で薬剤活性予測を行う手法を提案した [5].

そこで本研究では，その予測手法を改良することで，機械学習による薬剤活性予測タスクにおける問題の一つである予測化合物群の低い構造多様性を改善する手法を提案する．具体的には，

(1) 多数の化合物 fingerprint を用いることにより，より多くの有望な化合物を拾い出す
(2) 二段階目の予測モデルに変わり，予測化合物群の構造多様性を確保できるような最適化手法を提案する

ことにより，予測精度をあまり落とさずに予測化合物群の構造多様性を確保するような手法を提案する．

3.2 Topic Diversification

Topic diversification は，情報検索の分野における一つの概念である．例えばショッピングサイトにはユーザの過去の検索や購入履歴に基づいた商品の推薦が行われることが多いが，全く同じような商品を並べるよりは，ユーザが興味がありそうでありかつ多岐にわたる商品を提示された方が望ましい場合がある．この問題に取り組んだ研究として，Ziegler らによる手法 [6] が提案されている．これは，ユーザが購入しそうな順に並べたスコアと，すでに表示したアイテム群との距離のスコアを組み合わせてリランキングするという手法である．

本研究においては，この Topic Diversification の手法に範をとり，予測化合物群の構造多様化を図る手法を考案する．

4. 提案手法

提案手法では，多数の fingerprint それぞれによって別々に薬剤活性予測モデルを作成し，その予測結果を提案化合物群の構造多様性に注意しながら統合する．

4.1 薬剤活性予測モデルの構築

まず，それぞれの molecular fingerprint を用いて薬剤活性予測モデルを作成する．この予測モデルの学習アルゴリズムには教師あり機械学習手法であれば利用可能であるが，本研究では安定した予測性能と実行速度の観点から Random Forest を用いた．学習データ内において 3-fold Cross validation を行い，パラメータチューニングを行った．チューニングには GridSearch を用いており，探索したパラメータを以下に示す．

- $num\_estimators$: $50, 100, 150, 200, 250, 300, 350, 400$
- $max\_features$: $50, 100, 200, 300, 400$

4.2 予測結果の統合

本研究では，次のようにして貪欲的な方法により前節で作成した薬剤活性予測モデルを統合する．

(1) 各 Fingerprint ごとの薬剤活性モデルに対してテストデータを予測する
(2) $pool = \emptyset$ を提案化合物群とする
(3) 各予測のうち最上位の化合物を $pool$ に入れ， 予測結果から削除
(4) 現在の予測結果それぞれのうち上位 10 件に対して，現在の $pool$ に入っている化合物群との平均 Tanimoto 係数がもっとも小さいものを $pool$ に入れ，すべての予測結果から削除
(5) 同じくその中で現在の $pool$ に入ってる化合物群の中に，Tanimoto 係数が一定以上となる化合物が存在する場合，該当化合物を予測結果から削除
(6) 3.，4. を $pool$ が必要な大きさになるまで繰り返す

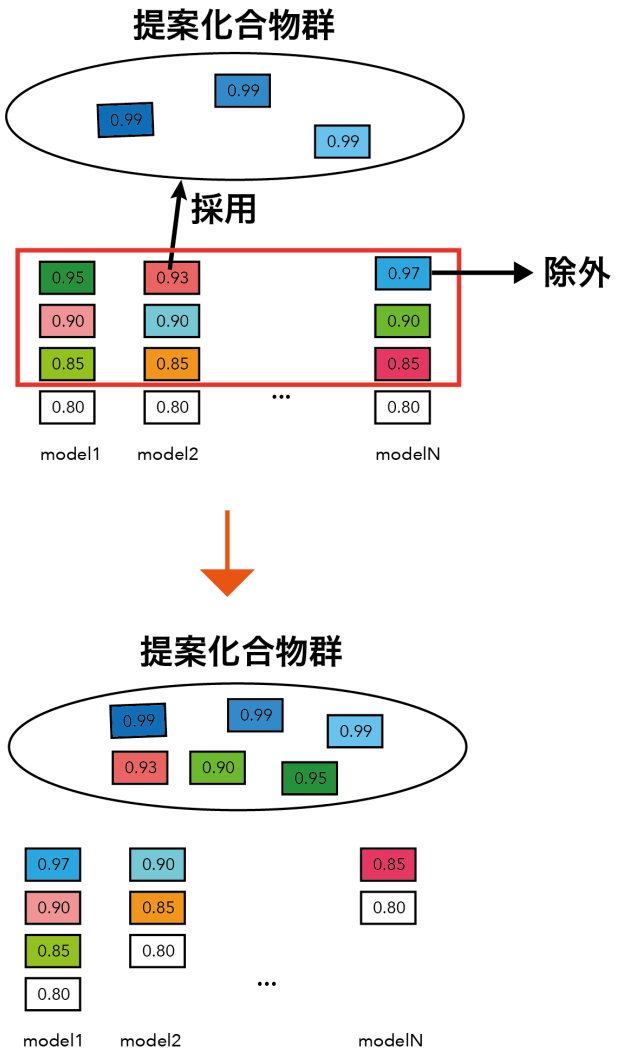

図 3 提案手法の概要図．小さな四角は化合物を表しており，それぞれの色は化合物の構造を表現していて，色が異なるほど構造が異なるという意味である．

表 1　本研究で用いたアッセイデータおよびデータ個数の一覧

Pubchem AID	Target or goal	#Active	#Inactive
1915	Group A Streptokinase expression inhibition, streptkinase	5266	7751
463213	Identify small molecule inhibitors of tim10-1 yeast	4141	3235
463215	Identify small molecule inhibitors of tim10 yeast	2941	1695
492992	two pole domain potassium channel	2094	2820
504607	Mdm2/MdmX interaction	4830	1412
624504	mitochndrial permeablity transition pore	3944	1090
651739	Inhibition of Trypanosoma cruzi	4051	1324
651744	NIH, 3T3 toxicity	3102	2306
652065	bind r(CAG) RNA repeats	2966	1287

表 2　本研究で用いた化合物特徴ベクトルの一覧

Feature	length	Ref
ECFP4/6	2048	David rogers(2010)
2DPF	3348	A.Gobbi(1998)
Maccs key	166	Symyx Technologies
daylight	2048	Daylight Chemical Information Systems Inc.
Topological Torsion	2048	R.Nilakantan(1987)
atompair	2048	R.E.Carhart(1985)
Chemical/Structure descriptors	193	Using RDKIT

5. 実験と考察

5.1 データセット

まずアッセイデータとして，創薬系公開データベースである PubChem より，**表 1** に示すアッセイデータを取得した．

本研究では，実際のヒット化合物予測における問題設定をある程度再現するため，次のような手順を用いて評価用データセットを構築した．

(1) アッセイデータより，活性化合物をランダムに 100 個選択する． さらに非活性化合物からランダムに 1000 個を選択し，これらを学習用データとする．学習データ中の正例と負例の比率は 1 対 10 となる．

(2) 残りのデータをテストデータとする． しかしこのままではテストデータ中に正例の比率が高いため，ZINC データベースよりランダムに化合物を選び出し，比率が学習データと同じく 1 対 10 となるように調整した．

これにより，実際に LBVS を適用する際に起こりがちな，「探索対象のライブラリに対して学習データ内の活性化合物のケミカルスペースがとても小さい」状況と，「テストデータにおいて実際の活性化合物は少ない」という状況を同時に再現した．

また使用する molecular fingerprint について**表 2** に示す． いずれも非常によく使われる主要な fingerprint であり，生成にはいずれも RDKit を用いた．

また，物理化学的 descriptor に関しては，RDKit がサポートしている 193 次元の特徴ベクトルを用い，0 から 1 の範囲で正規化した

5.2 評価方法と評価対象

本研究においては，各手法の評価として予測性能と提案化合物群の構造多様性の二つの観点について同時に評価を行う．ここではそれぞれの評価手法について述べる．

5.2.1 予測性能に対する評価

予測性能の指標については，タスクとして提案化合物「群」についての精度評価を行うことから，今回は提案化合物群の予測性能の評価指標として，Enrichment Factor (EF) を用いる．Enrichment Factor は，元のテストデータから，スクリーニングした事によりどれだけの正解化合物を「濃縮」できたかを表す指標である．テストデータにおける活性化合物の数を $\#pos$， 非活性化合物の数を $\#neg$ とする．ここでテストデータ全体の $\chi\%$ の化合物を選び出す事になったとする．この時提案化合物群における活性化合物の数を $\#pos_x$，非活性化合物の数を $\#neg_x$ とすると，Enrichment Facot($\chi\%$) は以下のように定義される

$$EF(\chi\%) = \frac{\#pos_x}{\#pos_x + \#neg_x} / \frac{\#pos}{\#neg}$$

この数値が高いほど，薬剤活性予測モデルは活性化合物をよくスクリーニングしているといえる．なお，ランダムな予測モデルであれば，この数値は 1.0 に近くなる．また，今回のテストデータは活性化合物と非活性化合物の割合が 1 対 10 固定であるため，理想的な予測が行われた場合の $EF1\%$ は 11.0 となる．

5.2.2 構造多様性に対する評価

現在のところ構造多様性の評価には一般的な評価指標は提案されていない．そこで本研究では，提案化合物群の構造多様性の評価について次のような指標を提案する．まず，N 個の化合物が含まれる化合物群 P を考える．その中に含まれる各化合物について，化合物間の類似性が最も高い化合物を選び出す操作を行い，N 個のペアを作る．本研究では類似性の計算には ECFP4 による Tanimoto 係数を用いた．ここで，P に含まれる i 番目の化合物を ECFP4 で fingerprint 化したものを v_i とするとき，

(1) 単純平均スコア

$$Score_1 = \frac{1}{N} \sum_{i \in P} maxT(v_i, v_j) \| i, j \in P$$

$T(x, y)$ は x,y の間の Tanimoto 係数を示す．

(2) Tanimoto 係数の特性 (0.3 を下回るような低い値域ではその数値の高低にあまり意味があるとは言えない) を考慮したスコア

$$Score_2 = \frac{1}{N}\sum_{i \in P} Clip(max_j T(v_i, v_j), 0.3, 1.0)$$

$$\| i, j \in P$$

5.2.3 比較対象

ここでは評価実験において比較対象となる手法について述べる．比較対象の 1 つ目は幅広く用いられる fingerprint の一つである ECFP4 を用いて random forest により予測を行った結果そのままである．これは特に構造多様性を考慮しない場合の予測精度，構造多様性を示している．それに加え，既存研究の項目でも述べた Butina クラスタリングおよび Murco Scaffold による類似化合物の除去を予測上位化合物に対して行った場合についても評価を行い予測結果の比較を行った．

5.3 実験結果

表 3 に各アッセイデータごとの実験結果を示す．まず，ベースラインとなる ECFP4 単独で学習させた予測モデルと比較すると，提案手法および既存手法はともに予測精度の低下と引き換えに構造多様性のスコアは改善していることがわかる．

まず確認すべき点として，AID651734 および AID651744 の結果を見ると，なんらかの構造多様化手法を適用することにより予測性能が大幅に大きく低下してしまっているということがある．これらのアッセイデータに対する ECFP4 のみの結果を見ると，予測結果はほとんど同じような化合物で固まっておりかつ，ほぼ完璧な予測が行われていることがわかる．故にテストデータのうちの正例があまりにも偏っているための結果だと思われる．

これらの結果を除いた上で各手法における評価指標の平均値をとったものを**表 4** に示す．これを見ると，データセットごとに見れば適した手法はことなるものの，平均的に見ると提案手法は既存手法と比べ，提案化合物群の構造多様性を確保しつつ，予測性能の低下をある程度緩和できていることがわかる．

表 5 は各 fingerprint による予測モデルの予測上位 1% に含まれていた活性化合物の数と，それらの数の合計のうちいくつが unique な化合物であったかを示したものである．前述した AID651744 以外について，複数の化合物 fingerprint の予測モデルを統合することにより，より多くの化合物を候補として拾い出すことができていることがわかる．今回予測性能という点では提案手法を適用することにより低下してしまっているが，このことを踏まえると複数の化合物 fingerprint による予測結果を統合することにより構造多様な提案化合物を出力するという考え方自体は妥当であると考えられる．

表 3 各アッセイデータに対する実験結果の表

AID	methods	Score1	Score2	EF1
1915	ECFP4	0.540	0.563	11.0
	Bemis Murco	0.343	0.395	9.11
	Butina	0.287	0.366	5.22
	提案手法	0.285	0.344	9.68
463213	ECFP4	0.507	0.534	6.14
	Bemis Murco	0.338	0.387	5.91
	Butina	0.304	0.370	2.63
	提案手法	0.277	0.349	5.33
463215	ECFP4	0.446	0.482	7.65
	Bemis Murco	0.272	0.339	4.57
	Butina	0.323	0.388	4.92
	提案手法	0.260	0.337	5.57
492992	ECFP4	0.525	0.550	10.40
	Bemis Murco	0.260	0.346	7.60
	Butina	0.331	0.394	7.40
	提案手法	0.255	0.328	8.18
504607	ECFP4	0.482	0.513	6.77
	Bemis Murco	0.315	0.365	4.50
	Butina	0.312	0.377	3.65
	提案手法	0.296	0.358	6.33
624504	ECFP4	0.545	0.558	10.40
	Bemis Murco	0.360	0.388	7.72
	Butina	0.295	0.340	7.12
	提案手法	0.398	0.414	9.07
651739	ECFP4	0.993	0.993	10.95
	Bemis Murco	0.380	0.414	2.28
	Butina	0.554	0.591	8.80
	提案手法	0.342	0.381	1.774
651744	ECFP4	0.990	0.990	10.87
	Bemis Murco	0.417	0.459	4.07
	Butina	0.475	0.495	8.80
	提案手法	0.305	0.360	1.833
652065	ECFP4	0.641	0.652	10.97
	Bemis Murco	0.340	0.380	8.39
	Butina	0.322	0.388	6.63
	提案手法	0.350	0.387	7.82

表 4 各手法の平均評価指標（AID651739 および AID651744 を除く）

method	Score1	Score2	EF1
ECFP4	0.527	0.550	9.05
Bemis Murco	0.311	0.374	5.36
Butina	0.318	0.371	6.83
提案手法	0.303	0.360	7.43

表 5　各 fingerprint による予測結果上位 1% に含まれていた活性化合物の数．

fingerprint /AID	463213	492992	504607	651744
2dpf	193	201	202	323
ecfp4	248	207	320	326
ecfp6	232	201	399	321
descriptor	209	137	267	322
daylight	304	200	264	319
atompair	221	203	396	319
maccskey	294	187	238	311
topo_tor	336	207	247	320
上位 1 パーセントに当たる化合物数	444	219	520	330
Unique な化合物数	683	416	1011	181

6. まとめ

本研究では機械学習を用いたリガンドベースのヴァーチャルスクリーニング手法における予測化合物群の構造多様性という問題に対し，多数の異なる fingerprint による予測モデルの予測結果を構造多様性が得られるように統合することで，予測精度の低下を避けながら活性があると予測された化合物群の構造多様性を確保する手法を提案した．評価実験の結果，提案手法は予測精度をあまり落とさずに提案化合物の構造多様性を確保できることを示した．しかし，分子骨格によるクラスタリング手法に比べて大きな改善は得られなかったため，予測精度の低下を抑える更なる改良が必要であると考えられる．

参考文献

[1] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Igor, I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Consonni, V., Kuz, V. E. and Cramer, R.: QSAR Modeling: Where have you been? Where are you going to?, Vol. 57, No. 12, pp. 4977–5010 (online), DOI: 10.1021/jm4004285.QSAR (2015).
[2] Bemis, G. W. and Murcko, M. A.: The Properties of Known Drugs . 1 . Molecular Frameworks, Vol. 2623, No. 96, pp. 2887–2893 (1996).
[3] Butina, D.: Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets, *Journal of Chemical Information and Computer Sciences*, Vol. 39, No. 4, pp. 747–750 (online), DOI: 10.1021/ci9803381 (1999).
[4] 石田貴士松山祐輔：薬剤活性予測の改良のための化合物フィンガープリントの比較解析，情報処理学会研究報告，2017-BIO-49 (2017).
[5] Matsuyama, Y. and Ishida, T.: Using multiple molecular fingerprints improves ligand-based virtual screening, *submitted* (2017).
[6] Ziegler, C.-N., McNee, S. M., Konstan, J. A. and Lausen, G.: Improving recommendation lists through Topic Diversification, *Proceedings - 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, ICIS 2009*, Vol. 3, pp. 222–225 (online), DOI: 10.1109/ICICISYS.2009.5358201 (2009).