

東京工業大学

修士論文

フラグメント分割に基づく高速な
化合物プレドッキング手法の開発

指導教員 秋山 泰 教授

平成28年1月

提出者

研究科 情報理工学研究科

専攻 計算工学専攻

学籍番号 14M38400

氏名 柳澤 溪甫

フラグメント分割に基づく 高速な化合物プレドッキング手法の開発

指導教員 秋山 泰 教授

計算工学専攻

14M38400 柳澤 深甫

創薬の上流工程では、大量の候補化合物から薬になりうる化合物を選別する作業が日常的に行われている。このうちコンピュータによる予測を利用してコスト削減および開発期間の短縮を目指す「バーチャルスクリーニング」と呼ばれる手法が近年盛んに用いられている。中でも、タンパク質や化合物の立体構造情報を用いたドッキングシミュレーション（以下ドッキング）を行う SBVS（Structure-Based Virtual Screening）は標的タンパク質に対する既知の薬剤情報がなくても適用できる手法であり、創薬の現場で大きく期待されている。

一方で、ドッキングは標的タンパク質と化合物との最適な複合体構造の探索を必要とするが、相対的な3次元位置、回転および、化合物内部の結合の回転という多くの探索パラメータが存在するため計算コストが高く、大量の化合物すべてを評価することは困難な場合が多い。この問題を解決するため、化合物を何らかの方法でフィルタリングし、数を減らした化合物サブセットについてドッキングを行う、という段階的な手法が多く実践されている。しかし従来のフィルタリング手法は既知の薬剤情報を用いてしまうため、SBVS と組み合わせた時に SBVS の利点の一部が失われてしまうという問題が存在していた。

そこで本研究では、ドッキングに基づいた上で、化合物の部分構造を用いることでより高速にフィルタリングを行う手法（プレドッキング）を提案する。提案手法では、まず全ての候補化合物をフラグメントと呼ばれる内部に回転可能な結合を持たない部分構造に分割し、標的タンパク質とフラグメントとの間でドッキングを行う。このとき、フラグメントへの分割により回転自由度がなくなるため、ドッキングの最適化問題の探索パラメータが減少し、計算の高速化が可能となる。さらに、フラグメントのスコアから化合物のスコアを高速に計算する方法を検討することで、高速だが粗い探索を行うドッキングの一つである Glide HTVS モードに比べ、提案手法は一般的なベンチマークデータセット（DUD-E）を

用いたときに精度は ROC 曲線下面積 (ROC-AUC) で 4% 程度劣るものの平均して約 9 倍の高速化を達成した。フラグメントごとのドッキング結果は化合物間で再利用が可能であるため、高速化率は評価対象の化合物数が多いほど高くなる。

また、提案法に基づくフィルタリングの後に通常のドッキングを行った場合の工程全体としての評価を行った。その結果、提案手法は Glide HTVS モードに比べて最大で精度の約 25% の向上、計算時間の約 40% の減少を達成するケースが存在することが示された。

目 次

第1章	序論	1
1.1	研究背景	1
1.2	研究目的	2
1.3	本論文の構成	2
第2章	ドッキングシミュレーションによる薬物候補化合物の選別	4
2.1	SBVS (Structure-Based Virtual Screening) とは	4
2.2	タンパク質-化合物ドッキングシミュレーション	4
2.2.1	ドッキングシミュレーションの要素	6
2.2.2	ドッキングシミュレーションの問題点	8
2.3	化合物のフィルタリング	9
2.3.1	既存のフィルタリング手法	10
2.3.2	既存手法の問題点	11
2.3.3	提案手法の目標	11
第3章	提案手法：化合物の部分構造を利用したフィルタリング（プレドッキング）手法の開発	13
3.1	提案手法の概説	13
3.1.1	フィルタリングの要件	13
3.1.2	提案手法へのアイデア	14
3.2	提案手法の詳細の説明	14
3.2.1	提案手法のフローチャート	15
3.2.2	化合物のフラグメントへの分割	15
3.2.3	フラグメント単位でのドッキングシミュレーション	18
3.2.4	化合物のフィルタリングスコアの算出	19
第4章	実験	22
4.1	データセット	22
4.2	予測精度の評価指標	22

4.3 計算環境	23
4.4 比較対象	25
4.5 評価実験	25
4.5.1 フラグメント分割	25
4.5.2 ドッキング速度の評価	26
4.5.3 予測精度の評価	26
4.5.4 フィルタリング手法としての性能評価実験	28
第5章 考察	31
5.1 総和法におけるフラグメント数に対するペナルティ	31
5.2 提案手法が得意とするケースの調査	34
5.3 提案手法の利用例	35
第6章 結論	39
6.1 本研究の結論	39
6.2 今後の課題	41
謝辞	42
参考文献	43
付録A DUD-E データセットの詳細	47
付録B 各手法を単体で用いた場合のROC曲線	53

図 目 次

2.1	ドッキングシミュレーションのイメージ	5
2.2	ドッキングシミュレーションの探索空間	7
2.3	Glide のワークフロー ¹⁾	8
2.4	eHiTS のクリーク探索	9
2.5	AutoDock の GA で用いる変数群	9
2.6	提案手法の速度の数値目標	12
3.1	フラグメント単位でのドッキング結果例	15
3.2	提案手法の手順	16
3.3	化合物のフラグメント分割アルゴリズム	17
3.4	ZINC “drugs now” 10,639,555 化合物を分割した例	18
3.5	複数のドッキング結果の出力例および最良構造の選択	19
3.6	score_sum, score_max の算出	21
3.7	maxsumBS の算出	21
4.1	ROC-AUC 計算例	24
4.2	EF 計算例	24
4.3	DUD-E ターゲットにおける化合物数とフラグメント種類数の関係	25
4.4	EF (1%), EF (2%) 算出までの流れ	28
5.1	ターゲット fnta の全ての化合物のうち重原子数 32 の化合物の単純加算スコア	31
5.2	薬剤化合物の例 : (a) ホルモテロール, (b) カンデサルタン	35
B.1	各手法の単体性能 ROC 曲線 (1)	54
B.2	各手法の単体性能 ROC 曲線 (2)	55
B.3	各手法の単体性能 ROC 曲線 (3)	56
B.4	各手法の単体性能 ROC 曲線 (4)	57
B.5	各手法の単体性能 ROC 曲線 (5)	58
B.6	各手法の単体性能 ROC 曲線 (6)	59
B.7	各手法の単体性能 ROC 曲線 (7)	60

B.8 各手法の単体性能 ROC 曲線 (8)	61
B.9 各手法の単体性能 ROC 曲線 (9)	62
B.10 各手法の単体性能 ROC 曲線 (10)	63
B.11 各手法の単体性能 ROC 曲線 (11)	64
B.12 各手法の単体性能 ROC 曲線 (12)	65
B.13 各手法の単体性能 ROC 曲線 (13)	66
B.14 各手法の単体性能 ROC 曲線 (14)	67
B.15 各手法の単体性能 ROC 曲線 (15)	68
B.16 各手法の単体性能 ROC 曲線 (16)	69
B.17 各手法の単体性能 ROC 曲線 (17)	70

表 目 次

1.1 従来のフィルタリング手法と本研究で提案する手法との比較	3
2.1 主なドッキングシミュレーションソフトウェア	5
2.2 ドッキングシミュレーションソフトウェアの計算速度	9
4.1 DUD-E のターゲットの化合物数	22
4.2 利用した計算環境	23
4.3 フラグメント分割を行った時のフラグメント 1 種類あたりの化合物数	25
4.4 ドッキング計算時間の比較（括弧内は Glide HTVS との速度比）	26
4.5 提案手法の予測精度	27
4.6 フィルタリング手法としての提案手法間の精度評価	29
4.7 フィルタリング手法としての提案手法と従来手法の比較	29
5.1 小さなフラグメントを無視することによる score_sum の精度の変化	32
5.2 フラグメント数に対する線形ペナルティによる score_sum の精度の変化	32
5.3 小さなフラグメントを無視することによる maxsumBS の精度の変化	33
5.4 フラグメント数に対する線形ペナルティによる maxsumBS の精度の変化	33
5.5 提案手法が上手く行ったケース	34
5.6 提案手法が得意なターゲットの性質	34
5.7 化合物全体を評価するのに要する時間の比較	36
5.8 総化合物数が 1 万以上存在する DUD-E のターゲットに対する評価実験	37
5.9 提案手法が従来手法に速度・精度ともに勝る例	38
6.1 提案手法のフィルタリング利用時の精度	39
6.2 提案手法の性能	40
6.3 通常ドッキング（Glide SP）と組み合わせた速度・精度評価	40
A.1 DUD-E の詳細 (1)	49
A.2 DUD-E の詳細 (2)	50
A.3 DUD-E の詳細 (3)	51
A.4 DUD-E の詳細 (4)	52

B.1 各手法单体性能 AUC-ROC 值 (1)	71
B.2 各手法单体性能 AUC-ROC 值 (2)	72
B.3 各手法单体性能 AUC-ROC 值 (3)	73
B.4 各手法单体性能 AUC-ROC 值 (4)	74

第1章

序論

1.1 研究背景

近年、創薬の上流工程において、コンピュータによる予測を通して大量の化合物から薬剤候補化合物を選別するバーチャルスクリーニング（Virtual Screening, VS）と呼ばれる手法が用いられ、創薬コストの削減および創薬にかかる時間の短縮が試みられている。このコンピュータを用いた化合物の選別手法は大きく3つに分けられる。

1. タンパク質や化合物の立体構造を用いた手法（Structure-Based Virtual Screening, SBVS）
 - タンパク質-化合物ドッキングシミュレーション¹⁾⁻³⁾
2. 既知の薬剤やタンパク質の活動を阻害する化合物（阻害剤）等の情報を用いた手法（Ligand-Based Virtual Screening, LBVS）
 - 構造活性相関（Quantitative Structure-Activity Relationship, QSAR）を用いた手法⁴⁾
 - 機械学習による分類手法⁵⁾
 - 化合物の官能基の性質を用いたファーマコフォアモデルに基づく化合物分類手法⁶⁾
3. 化合物だけではなくタンパク質も複数活用することでタンパク質と薬剤との2部グラフなどのネットワークを構築し、類似度から予測を行う創薬手法（Chemical Genomics-Based Virtual Screening, CGBVS）⁷⁾

このうち、タンパク質-化合物ドッキングシミュレーションによるSBVSは物理的なエネルギーを計算する演繹的な手法であり、既知の薬剤や阻害剤等が存在しない創薬標的であってもタンパク質の構造が得られれば薬物候補化合物を選別することができる有用な方法である。また、既知の薬剤や阻害剤から法則性を見つけ出すなど帰納的な手法であるLBVSやCGBVSに比べて既知の薬剤や

阻害剤等と大きく性質の異なる、「新規の構造を持った」薬剤候補化合物を発見する能力が高いこともドッキングシミュレーションによる SBVS のメリットである。

ドッキングシミュレーションは Glide,¹⁾ eHiTS,²⁾ AutoDock³⁾ を始めとして多様なツールが開発されており、その中でも Glide は予測精度が高く⁸⁾、広く利用されている⁹⁾。

一方、ドッキングシミュレーションはタンパク質と化合物との複合体構造の予測を行うためには化合物の回転や平行移動を行いながら探索を行う最適化問題が必要となり、計算コストが非常に高いという問題点が存在する。これを解決するためにドッキング手法の高速化の研究^{10)–12)} が行われているが、速度の点で未だ不十分である。例えば、購入可能な化合物の立体構造データベースを開いている ZINC¹³⁾ に存在する 22,724,825 件の化合物を一斉にドッキングシミュレーションをしようとすると、現状では 1 CPU コアでの計算では 5 年以上の時間を要する。また、Glide は計算に利用する CPU コア数に応じてライセンスを購入しなければならない形式の商用ソフトであるため、TSubame 2.5 などのスーパーコンピュータの大規模利用による高速化を行うことができない。

以上の理由から、SBVS を用いた創薬研究ではドッキングシミュレーションを行う前に化合物を選別するフィルタリングが行われることが多い^{14),15)}。しかし、これらのフィルタリング手法の多くは LBVS のように、既知の薬剤等の化合物情報を用いるものであり、前述した SBVS の長所である「既知の薬剤等と大きく性質の異なる薬剤候補化合物」をフィルタリングで落としてしまうことが多く、SBVS とは相性が悪いという問題がある。また、Glide の簡易ドッキングモードである HTVS モードを用いてフィルタリングを行うこともあり^{16),17)}、この手法を用いれば SBVS の長所を損なうことなくフィルタリングを行うことができるが、前述したような数千万単位の化合物数では Glide HTVS モードですら 1 CPU コアで半年程度の計算時間をする。

1.2 研究目的

1.1 節で示したように、SBVS におけるフィルタリングは未だ研究が不十分であり、新規の構造を持つ化合物を残す高速なフィルタリング手法を開発する必要がある。本論文では、ドッキングに基づいた、既知の薬剤に依存しない高速なフィルタリング手法を提案することを目的とする（表 1.1）。

1.3 本論文の構成

第 2 章では、ドッキングシミュレーションに基づいた SBVS についての説明を行い、同時に既存のフィルタリング手法について説明する。第 3 章では提案手法について述べ、第 4 章でこの提案手法と簡易ドッキングである Glide HTVS モードとの比較を行う。また、第 5 章では第 4 章で行った実験の結果についての考察を加え、第 6 章で結論および今後の展望を述べる。

表 1.1 従来のフィルタリング手法と本研究で提案する手法との比較

手法	既知薬剤への依存性	計算時間
既知薬剤情報を用いたフィルタリング	✗ 依存性があり、既知の薬剤等と大きく性質の異なる化合物が除外されてしまう	◎ 機械学習などに基づいており、高速にフィルタリングが可能
高速なドッキングによるフィルタリング	○ 既知薬剤を必要とせず、新規の構造を持った化合物も残すことができる	✗ 計算量が大きく、数千万化合物のフィルタリングには時間を要する
提案手法	○ 既知薬剤を必要とせず、新規の構造を持った化合物も残す	○ 従来の高速ドッキング手法に比べて高速にフィルタリングが可能

第2章

ドッキングシミュレーションによる薬物候補化合物の選別

この章ではドッキングシミュレーションに基づく化合物の選別手法を説明し、既存の化合物フィルタリング手法を紹介する。

2.1 SBVS (Structure-Based Virtual Screening) とは

バーチャルスクリーニング (Virtual Screening, VS) とは、コンピュータを用い、データベースに存在する化合物について、創薬標的となっているタンパク質の活性部位への結合のしやすさを仮想的 (Virtual) に評価、選別 (Screening) することを指す。化合物の評価・選別を創薬標的のタンパク質や化合物の立体構造に基づいて行う手法のことを SBVS (Structure-Based Virtual Screening) と呼ぶ。このSBVSは、化合物の評価・選別を既知の創薬標的タンパク質へ結合する化合物（リガンド, ligand）を用いて行う LBVS (Ligand-Based Virtual Screening) と比べて

- 既知のリガンド情報を必要とせず
- 既知のリガンドにとらわれない、多様な薬剤候補化合物を得ることができる

という長所を持っている。

2.2 タンパク質-化合物ドッキングシミュレーション

SBVSにおける化合物の評価にはタンパク質-化合物ドッキングシミュレーションが一般に用いられる。ドッキングシミュレーションは、1つのタンパク質の立体構造と1つの化合物の立体構造を入力として、化合物がタンパク質中でどのような構造をとるとエネルギー的に最も安定であるかとい

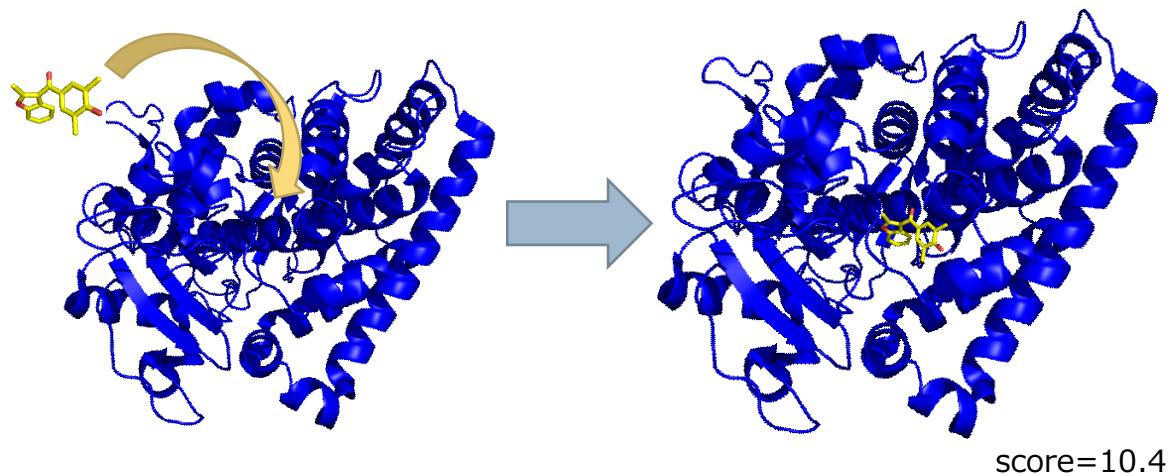


図 2.1 ドッキングシミュレーションのイメージ

う最適化問題を解き、最安定と考えられる化合物の構造とその時のスコアを近似的に求め出力する(図 2.1)。この得られたスコアを直接、あるいは何らかの形で変換を行った評価値を用いて複数の化合物の選別を行う。

このドッキングシミュレーションを行うツールは Glide,¹⁾ eHiTS²⁾などの有償ソフトウェア、AutoDock³⁾などのオープンソースソフトウェアを始めとして、有償無償問わず様々開発されている。表 2.1 に主なドッキングシミュレーションソフトウェアを示す。

表 2.1 主なドッキングシミュレーションソフトウェア

ソフトウェア名	論文	無償/有償	ホームページ
AutoDock	(Morris <i>et al.</i> , 2009) ³⁾	無償	http://autodock.scripps.edu/
AutoDock Vina	(Trott <i>et al.</i> , 2010) ¹²⁾	無償	http://autodock.scripps.edu/
DOCK	(Ewing <i>et al.</i> , 2001) ¹⁸⁾	無償	http://dock.compbio.ucsf.edu/
eHiTS	(Zsoldos <i>et al.</i> , 2007) ²⁾	有償	http://www.simbiosys.ca/ehits/
FlexX	(Rarey <i>et al.</i> , 1996) ¹⁹⁾	有償	http://www.biosolveit.de/flexx
Glide	(Friesner <i>et al.</i> , 2004) ¹⁾	有償	http://www.schrodinger.com/Glide/
GOLD	(Jones <i>et al.</i> , 1997) ²⁰⁾	有償	http://www.ccdc.cam.ac.uk/
ICM	(Abagyan <i>et al.</i> , 1993) ²¹⁾	有償	http://www.molsoft.com/docking.html

2.2.1 ドッキングシミュレーションの要素

SBVS の薬物候補化合物の選別はドッキングシミュレーションによって得られたスコアを基に行われるため算出されるスコアは重要となるが、後述するように探索空間が非常に広く、さらに最適化を行うべきスコア値も一般的に探索空間内で単調ではないため、厳密な最適スコアを求めることは事実上不可能である。そのため、ドッキングシミュレーションにおいては

- 非常に広い探索空間からなる最適化問題で良い準最適解を効率良く見つける探索アルゴリズム
- 適度に高速に計算でき、タンパク質-化合物の結合構造の良し悪しを適切に見積もるスコア関数

の 2 つが重要であり、これらは 1982 年に最初のドッキングシミュレーションツールである DOCK²²⁾ が開発されて以来、様々なグループによって研究が進められている。

探索空間

ドッキングシミュレーションでは、タンパク質の位置を固定として、化合物がタンパク質とどのような構造をとるとスコアが良いかを探索する。この際、探索しなければならない空間は化合物の並進運動および回転運動の 6 次元に加え、化合物の内部に回転可能な結合を持つため化合物の内部自由度を考慮しなければならない（図 2.2）。すでに上市され利用されている薬剤が収録された ZINC Drug Database に登録されている 2,924 個の薬剤化合物の内部自由度の平均は 4.61 であり、これが計算量に大きな影響を及ぼす。

探索アルゴリズム

前述のように探索空間の広さのために大域的最適解を求めるることは困難であるため、より良い局所的最適解を求めるための工夫がツール毎になされている。

- Glide¹⁾

段階的な全探索を行うことで局所最適解を得る。具体的には、最初の段階では化合物を球体に近似して化合物とタンパク質が衝突しないかどうかの見積もりから始め、徐々に化合物の近似を厳密なものにしていく。それぞれの段階で上位の位置・構造のみを残し次の段階へ進めることで、全探索の空間を現実的な量に制限し、探索を完了させる（図 2.3）。

- eHiTS²⁾

化合物を部分構造に分割し、部分構造にとって良い構造をそれぞれ多数記録し、ノードにする。その後、2 つの部分構造が構造を構成するのに適度な距離、適度な向きになっているノード

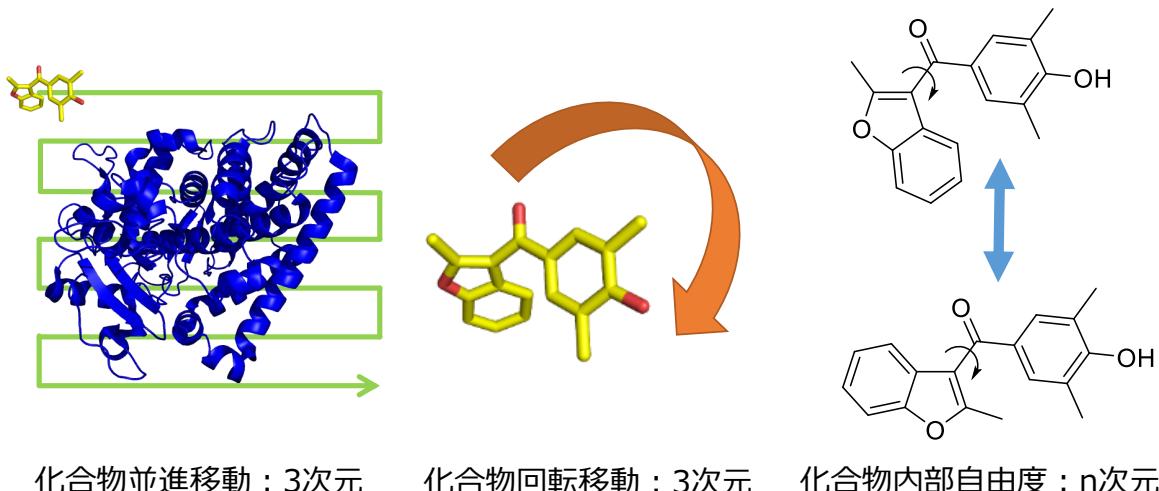


図 2.2 ドッキングシミュレーションの探索空間

ド間にエッジを張り、作成されたグラフに関して最大クリーク問題を解くことで適切な構造を得る（図 2.4）。

- AutoDock³⁾

並進運動位置、回転運動位置、化合物の内部回転角を用いた遺伝的アルゴリズム（Genetic Algorithm, GA）でより良い局所最適解を得る（図 2.5）。

スコア関数

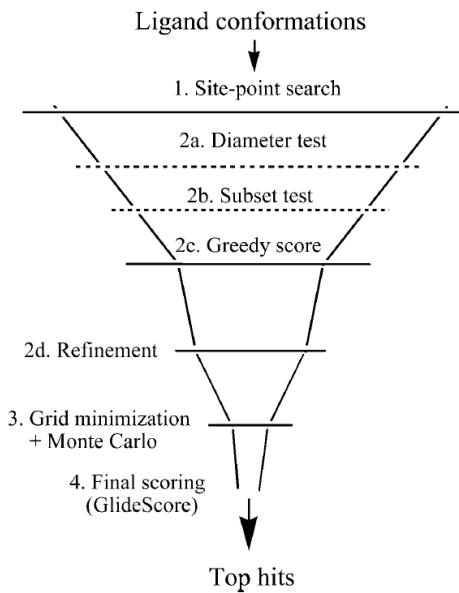
探索アルゴリズムがどれほど良く、大域最適なスコアを得たとしても、そのスコアがタンパク質と化合物との物理的な結合エネルギーとの相関がなければ意味がない。しかし、結合エネルギーを厳密に計算するには量子化学計算が必要となり、実用的な時間では計算が完了しないので、近似計算が必要となる。したがって、スコア関数に関しても様々な提案がなされている。

1. 物理化学的スコア関数 (Force field)

静電相互作用力やファンデルワールス力など、原子と原子との間に働く物理化学的な力を基としたスコア関数であり、考慮する物理的な項や、原子に対するパラメータが異なるなどによって多数のスコア関数が提案されている^{18), 20), 23)}。

2. 経験的スコア関数 (Empirical)

水素結合など物理化学的な要素に基づきつつも、実験から得られたタンパク質と化合物との

図 2.3 Glide のワークフロー¹⁾

結合エネルギーを再現できるように関数の係数ではなく関数そのものをパラメータ化することによって作成されたスコア関数である²⁴⁾⁻²⁶⁾.

3. 統計的スコア関数 (Knowledge-based)

タンパク質と化合物との複合体構造は Protein Data Bank (PDB)²⁷⁾ に多数登録されており、この複合体構造におけるそれぞれの原子種間の距離や角度の確率を計算し、それをエネルギーに変換することで作成されたスコア関数である²⁸⁾⁻³¹⁾. この手法では、どの複合体構造のセットを用いるかによってスコア関数が変化するため、一部のターゲットに特化したスコア関数なども作成されている³²⁾.

2.2.2 ドッキングシミュレーションの問題点

2.2.1 節に述べたように、ドッキングシミュレーションツールはそれぞれ高速化のための工夫を凝らしているが、それでも現時点では計算速度が不十分である。例えば、CPU 1 コアを用いて 1 つの化合物を評価するのに Glide で 0.2–2.4 分程度¹⁾、eHiTS は最速で数秒²⁾を要すると述べられているが、この速度で 1,000 万化合物を選別しようとすると 10 秒で 1 つの化合物を評価できたとしても 1,200 CPU days もの時間を必要とする。このような場合に一般的に用いられる手段である大規模計算化に関しても、Glide や eHiTS はライセンス式の有償ソフトウェアであるために、大量のライセンスを購入する必要があり現実的ではない。一方、AutoDock はライセンスが必要なく大規模並列計算が可能であるが、Glide と比べて 250 倍程度も遅いという報告がなされている³³⁾。AutoDock は

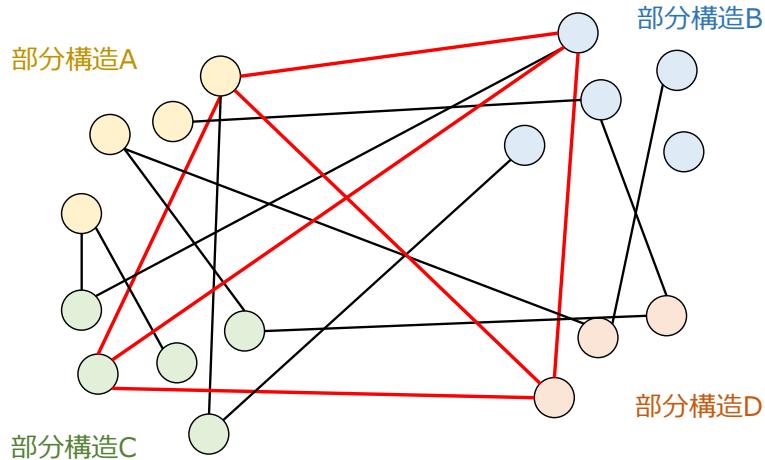


図 2.4 eHiTS のクリーク探索

Tx	Ty	Tz	Qx	Qy	Qz	Qw	R1	…	Rn
----	----	----	----	----	----	----	----	---	----

平行移動変数 4 元数による回転移動変数 化合物内の結合回転変数

図 2.5 AutoDock の GA で用いる変数群

オープンソースソフトウェアであるため、GPU 実装による高速化も提案されているが、遺伝的アルゴリズムやスコア関数の計算が最大 50 倍程度高速になる程度であり¹⁰⁾、Glide に及ばない。

表 2.2 ドッキングシミュレーションソフトウェアの計算速度

ソフトウェア名	無償/有償	CPU 1 コアでの計算速度
AutoDock	無償	Glide の約 250 倍の計算時間 ³³⁾
eHiTS	有償	<10 sec/compound (最速) ²⁾
Glide	有償	10–150 sec/compound ¹⁾

2.3 化合物のフィルタリング

ドッキングシミュレーションは大きな計算量を必要とするために、1,000 万個もの化合物から薬物候補化合物を選別しようとすることが非常に難しいことを 2.2.2 節で述べた。このため、ドッキングシミュレーションを高速化するのではなく、ドッキングシミュレーションの入力とする化合物の数をあらかじめ減することで総計算時間を削減するという戦略が創薬研究では良く用いられる。

2.3.1 既存のフィルタリング手法

既存のフィルタリング手法は大きく分けて、1. 化合物の物理的特徴に基づくフィルタリング、2. 化合物の構造に基づくフィルタリング、3. ドッキングベースのフィルタリングの3種類が存在している。

1. 化合物の物理的特徴に基づくフィルタリング

分子量や水溶性か油溶性かを示す分配係数 (LogP) など、化合物の物理的な特徴を示す値は体内での吸収などの上で重要な値であることから、これらの値を用いたフィルタリングが提案されている。

- **Lipinski の法則³⁴⁾**

経口薬として優れた薬物の物理的特徴を4つの法則にまとめたもの

- **Quantitative Estimate of Druglikeness (QED)³⁵⁾**

既知薬剤の物理的な値からヒストグラムを作成し、化合物の薬物らしさ (Druglikeness) のスコアを付ける手法

2. 化合物の構造に基づくフィルタリング

タンパク質の活動の阻害はタンパク質や化合物1分子単位の非常に微視的なメカニズムによって発生しており、したがって化合物の分子構造はタンパク質との複合体を形成する上で非常に重要な情報である。特に、構造が似ている化合物同士は同じタンパク質との複合体を形成することが多く、この性質を利用したフィルタリング手法が複数提案されている。

- **フィンガープリント (fingerprint) を用いたフィルタリング¹⁴⁾**

化合物の分子構造式を数百～数千のあらかじめ定めた局所構造が存在するか否かのバイナリであるフィンガープリントに変換し、これを用いて既知の薬剤やタンパク質の阻害剤にどれほど近いかを判定する手法

- **ファーマコフォア (pharmacophore) を用いたフィルタリング¹⁵⁾**

分子の構造式のみではなく、化合物の立体構造も用いて化合物の類似性を評価する手法

3. ドッキングシミュレーションベースのフィルタリング

2.2.2節で述べた通り、ドッキングシミュレーションは一般的に計算コストが高くフィルタリングには適していないが、Glideには化合物の構造について強い仮定を置くことで計算を簡易化し、通常ドッキングモード (SPモード) の10倍程度の速度³⁶⁾で計算を完了させる高速ドッキングモード (HTVSモード) が存在する。このモードをフィルタリングとして利用し、フィルタリング後の化合物群に対してSPモードによるドッキングシミュレーションを行うという手法が用いられることがある^{16),17)}。

2.3.2 既存手法の問題点

2.3.1 節で述べたように、既存のフィルタリング手法は多く存在するものの、以下の 2 点からこれらの手法は改善する余地が残されている。

- 化合物の物理的特徴や化合物の構造に基づくフィルタリング手法は帰納的な手法であり、標的タンパク質を狙った既知の薬剤や阻害剤が必須となる。さらに既知の薬剤や阻害剤を利用してきたとしても、化合物の類似性を利用する手法であるためにフィルタリング結果の化合物が既知の化合物に似てしまうという問題がある。
- Glide の高速ドッキングモード (HTVS モード) はドッキングシミュレーションとしては高速であるが、それでも 1 化合物 1 秒程度を要する。1,000 万件の化合物のフィルタリングを行う場合 1 CPU コアの利用で 4 か月程度の期間を要してしまうため、十分な速度とは言えない。

2.3.3 提案手法の目標

前節までの既存手法の現状を踏まえ、提案手法の目標は以下のように定める。

- 既知の薬剤や阻害剤の情報に依存しない、ドッキングシミュレーションに基づいたフィルタリング手法を提案する。
- 1 CPU コアの利用で、1,000 万化合物のフィルタリングを 2 週間以内に終了できるようにする。これは、1 化合物につき約 0.1 秒で評価を行うことに相当する。

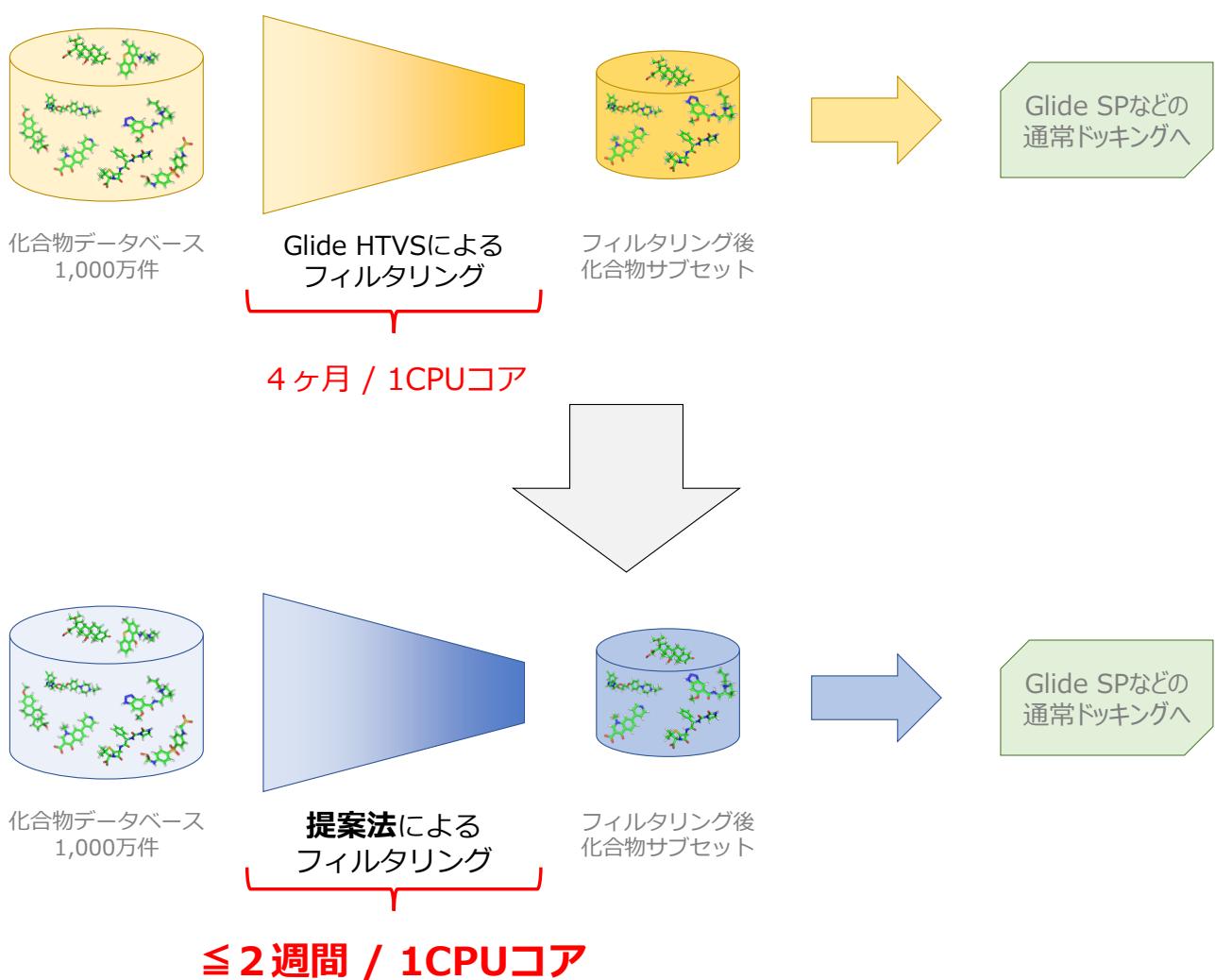


図 2.6 提案手法の速度の数値目標

第3章

提案手法：化合物の部分構造を利用したフィルタリング（プレドッキング）手法の開発

ここでは本研究で新たに提案する、化合物を部分構造に分割することで高速にドッキングを完了させるフィルタリング（プレドッキング）手法の内容を記述する。

3.1 提案手法の概説

前章で述べた通りドッキングシミュレーションは多くの計算時間を要するが、その理由は探索空間の広さとドッキング計算を行うべき化合物の数の多さにある。この節では、この2つの問題を解決するアイデア、および高速にフィルタリングを行うために追加する仮定を説明する。

3.1.1 フィルタリングの要件

フィルタリングに求められる要件は2つ存在する。

1. 高速に化合物を評価すること

フィルタリングを実用的に行うためには、フィルタリング後に行うドッキングシミュレーションよりも十分に高速である必要がある。

2. 予測の精度がある程度保持されていること

一般に計算速度と予測精度はトレードオフの関係にあるが、どれほど高速であってもある程度予測精度が保持されていること、特に偽陰性（False Negative）を出さないように弁別することがフィルタリングには求められる。

一方、フィルタリングはその後に複合体構造を予測する通常のドッキングシミュレーションを行うことを前提とするため、必ずしもタンパク質と化合物との複合体構造を出力する必要はなく、偽陽性（False Positive）を発生させることもある程度は許容される。

3.1.2 提案手法へのアイデア

前節で示したフィルタリングの要件を満たすために、2つのアイデアを考案した。

1. 化合物を部分構造に分割し、部分構造のドッキングシミュレーションを行う

2.2.1節で述べたように、化合物の内部自由度が及ぼす計算量への影響は大きい。そのため、eHiTS²⁾ や FlexX¹⁹⁾ など一部のドッキングシミュレーションツールでは化合物を内部自由度より少ない部分構造に分割し、タンパク質と部分構造との結合能力を評価しつつ最終的な複合体構造を構成する、という手法を用いている。本提案手法では、小峰ら³⁷⁾ による化合物の分割方法を用いて化合物を内部自由度を考慮しなくて良い「フラグメント」に分割、これらをドッキングすることで必要最低限の探索空間でのドッキングシミュレーションを実現する。

2. フラグメントから化合物の構造を再構成せず、フラグメントの結合スコアから化合物のフィルタリングスコアを算出する

構造分割に基づくドッキングシミュレーションツールでは、対象の化合物の内部構造に衝突などの問題が発生しないように部分構造を相互配置してタンパク質と化合物との複合体構造を形成する。しかし部分構造同士の衝突などの考慮を行うと、単純なアルゴリズムでは計算量が化合物を構成する部分構造数の指数オーダーとなってしまい、近似アルゴリズムを用いたとしても時間を要してしまう。

一方、本研究で提案するプレドッキング手法は、その後に複合体構造を予測するドッキングシミュレーションを行うことを前提とするためタンパク質と化合物との正しい複合体構造を必ずしも出力する必要はない。そこで提案手法ではフラグメントに分割する前の化合物の構造に関する考慮を行わないアプローチを選択した。こうすることで、図3.1のようにフラグメント同士の衝突が許容されてしまうが、化合物のスコアをフラグメント数の線形オーダー $O(n)$ で算出することが可能になり、高速な化合物の評価ができるようになる。

3.2 提案手法の詳細の説明

前節で本研究で用いる2つのアイデアを示したが、それを用いてどのようにフィルタリングを実現しているのかをこの節で詳説する。

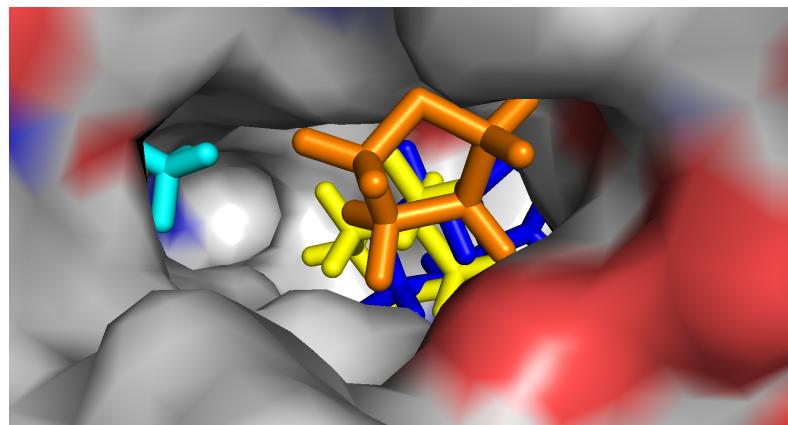


図 3.1 フラグメント単位でのドッキング結果例

3.2.1 提案手法のフローチャート

提案手法は以下の手順で構成される。

1. 入力化合物群をフラグメントに分割する
2. ドッキングシミュレーションツールを用いて各フラグメントの標的タンパク質への結合スコアを算出する
3. 各フラグメントの結合スコアから化合物全体としてのフィルタリングスコアを算出する
4. フィルタリングスコアの上位 N% をフィルタを通過した化合物として出力する

ワークフローを図 3.2 に示す。

3.2.2 化合物のフラグメントへの分割

化合物の分割は小峰らによる手法³⁷⁾を用い、内部自由度を持たない部分構造であるフラグメントを生成する。実装には C++ を用い、ケモインフォマティクスツールである OpenBabel³⁸⁾ および OpenMP, Boost を利用している。フラグメント分割のアルゴリズムを以下に示し、このアルゴリズムによるフラグメント分割の進行を図 3.3 に示す。

1. 元の分子のうち、重原子（水素以外の原子）のみに着目し、原子一つひとつを初期フラグメントとする。(図 3.3 (b))
2. 回転可能な单結合以外の結合の両端の 2 原子を同一フラグメントとする。
3. 環構造を構成している原子を同一フラグメントとする。(図 3.3 (c))

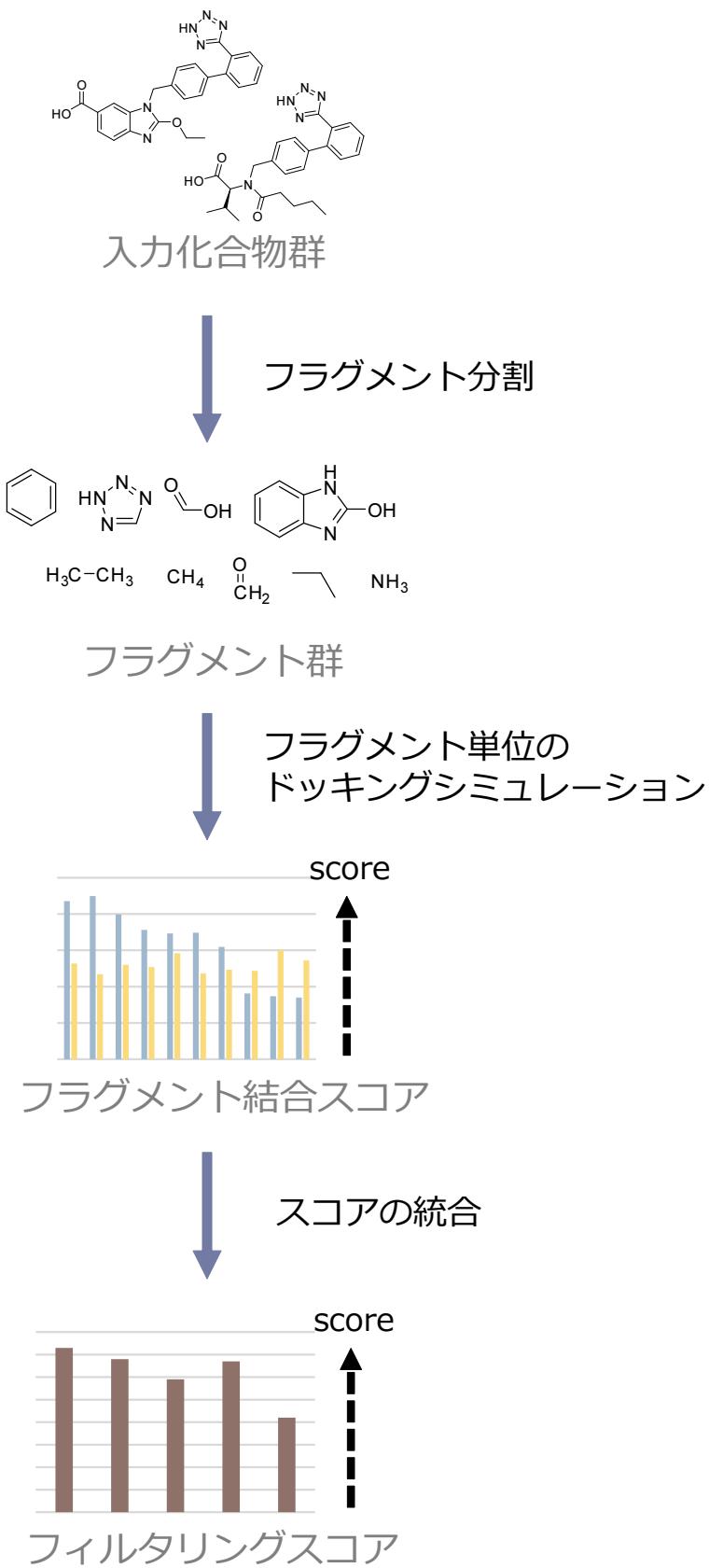


図 3.2 提案手法の手順

4. 回転可能な単結合を構成する原子ペアのうち、片方にそれ以上原子がつながっていない場合には同一フラグメントとする。これは、片方にそれ以上の原子がつながっていない場合、回転可能な単結合を回転させてもその原子がその場で回転するだけとなり、化合物の原子の位置関係には影響を与えないためである。
5. 2つの単結合の切断により孤立してしまう原子は、切断された先に存在する2つのフラグメントのどちらかに併合する。なお、3つ以上の単結合の切断により孤立してしまう原子に関してはこの操作を行わない。(図3.3 (d))
6. 全ての水素原子について、その原子が結合している重原子の属するフラグメントに含める。
7. 切断面に水素原子を付加する。

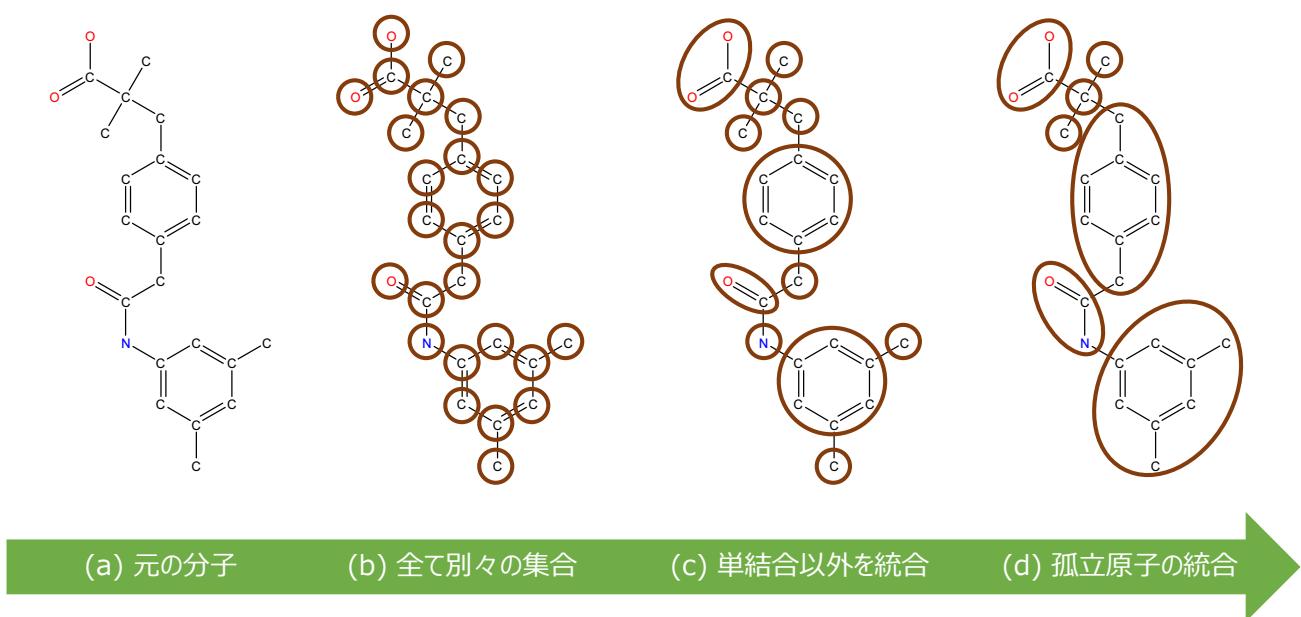


図 3.3 化合物のフラグメント分割アルゴリズム

この化合物のフラグメントへの分割により、内部自由度を考慮することなくドッキングシミュレーションを行うことができる。

また、複数の化合物間で部分構造に共通性が見られることが多く、本研究で用いている分割手法によって得られるものの中にも多数の共通フラグメントが発生する。例えば、ZINCの“drugs now”データセットに含まれている10,639,555化合物をフラグメント分割するとフラグメントの種類数は約20万となり(図3.4)、通常の化合物ドッキングを行う場合と比べ、計算を必要とする構造の数は約50分の1に低減される。このフラグメント種類数の増加は化合物数の増加に比べて緩やかであり、化合物数が多いほどフラグメント分割による速度への貢献が大きいと考えられる。

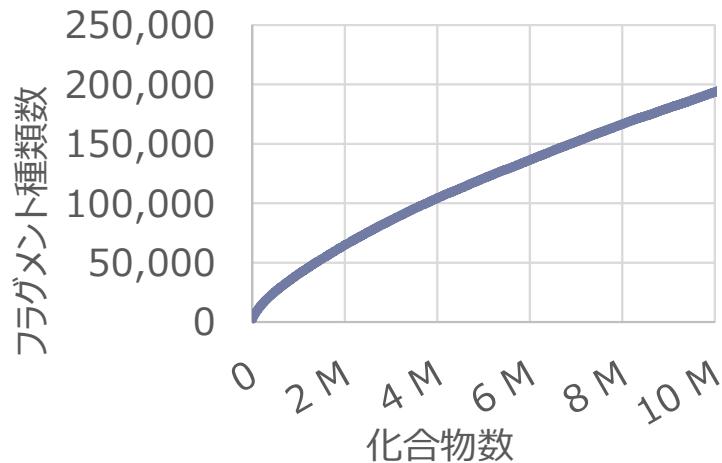


図 3.4 ZINC “drugs now” 10,639,555 化合物を分割した例

3.2.3 フラグメント単位でのドッキングシミュレーション

次に、分割されたフラグメントについて、標的タンパク質との結合スコアを求めるためにドッキングシミュレーションを行う。本研究では、有償ソフトである Glide¹⁾ を用いる。Glide には

- HTVS (High-Throughput Virtual Screening, 高速) モード
- SP (Standard Precision, 通常) モード
- XP (eXtra Precision, 精密) モード

の 3 種類のモードが存在するが、本研究では HTVS モードと SP モードを利用した場合の評価を行う。また、一般的に 1 つのタンパク質と 1 つの化合物とのドッキング結果では、複数のポーズを持つタンパク質-フラグメント結合予測構造および結合スコアが output されるが、この後の化合物のフィルタリングスコアの算出ではこのうち最良の結合スコアを利用する（図 3.5）。

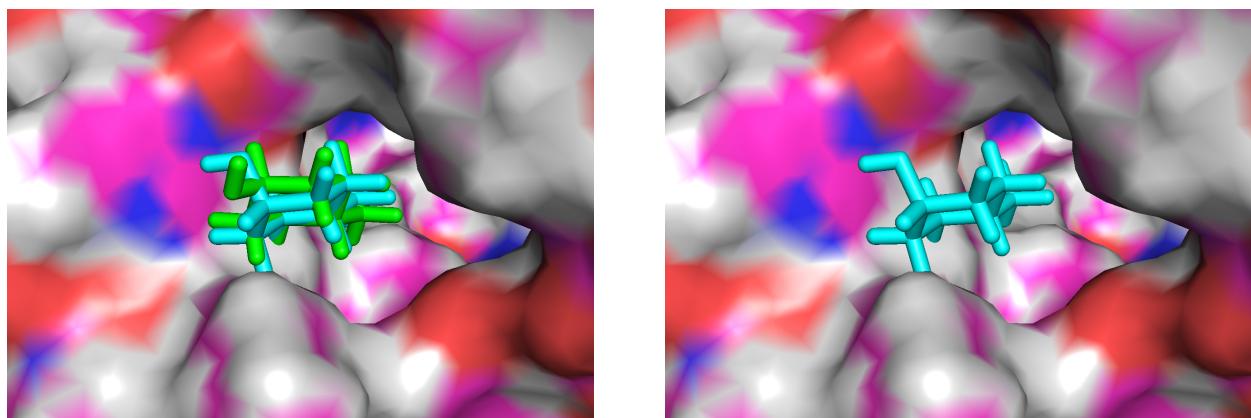


図 3.5 複数のドッキング結果の出力例および最良構造の選択

3.2.4 化合物のフィルタリングスコアの算出

フラグメント単位でのドッキングシミュレーションによって、フラグメントの結合構造およびその結合スコアを得た。続いて、このフラグメント結合スコアから化合物のフィルタリングに用いるスコアを算出する。本研究では、3種類のスコアの算出方法の実験を行った。

1. 総和法 (score_sum)

各フラグメント結合スコアの総和をとり、それを化合物全体のフィルタリングスコアとする。全てのフラグメントが高いスコアでタンパク質と結合できる化合物の評価を高くする手法である。フラグメントは化合物内に存在する結合という束縛条件を一部緩和したものであるため、一般に総和法によって得られた化合物フィルタリングスコアは化合物そのものの結合スコアよりも高くなり、特に分割数が多くなるほどその傾向は顕著になる。このため、重原子（水素以外の原子）の数が2個以下の小さなフラグメントの結合スコアはフィルタリングスコア算出から除外することでフィルタリングスコアの無意味な向上を抑えている（図3.6）。

$$SCORE_{\text{化合物}} = \sum_{\substack{\text{重原子数} > 2 \text{ の} \\ \text{フラグメント}}} SCORE_{\text{フラグメント}} \quad (3.1)$$

2. 最良値法 (score_max)

各フラグメント結合スコアのうちの最良値をとり、それを化合物全体のフィルタリングスコアとする（図3.6）。タンパク質との結合スコアが特に高いフラグメントを1つでも持っている化合物の評価が高くなる手法である。総和法とは異なり最良値法によって得られた化合物フィルタリングスコアは化合物そのものの結合スコアの下界となる。

$$SCORE_{\text{化合物}} = \max_{\substack{\text{すべての} \\ \text{フラグメント}}} SCORE_{\text{フラグメント}} \quad (3.2)$$

3. 総和法と最良値法の値の線形和 (maxsumBS)

これまでに示した総和法と最良値法はフラグメント結合スコアの全て，もしくはただ1つを見る手法であり両極端であるため，これらを統合して用いることで，より良い指標となるのではないかと考えた。しかし，総和法の値域が最良値法の値域よりも大きいために単純和では総和法の影響を大きく受けてしまう。そこで，2つの手法を適當なバランスで組み合わせるために，フィルタリングを行いたい化合物群の総和法によるスコア，最良値法によるスコアをそれぞれ平均0，分散1にし（すなわち z スコア化し），変換後のスコアを足し合わせることでバランスよくスコアを統合する（図3.7）。この手法は総和法によるスコアと最良値法によるスコアのバランスをとったスコアであるので，maxsumBS (max-sum Balanced Score) として以下では記述する。

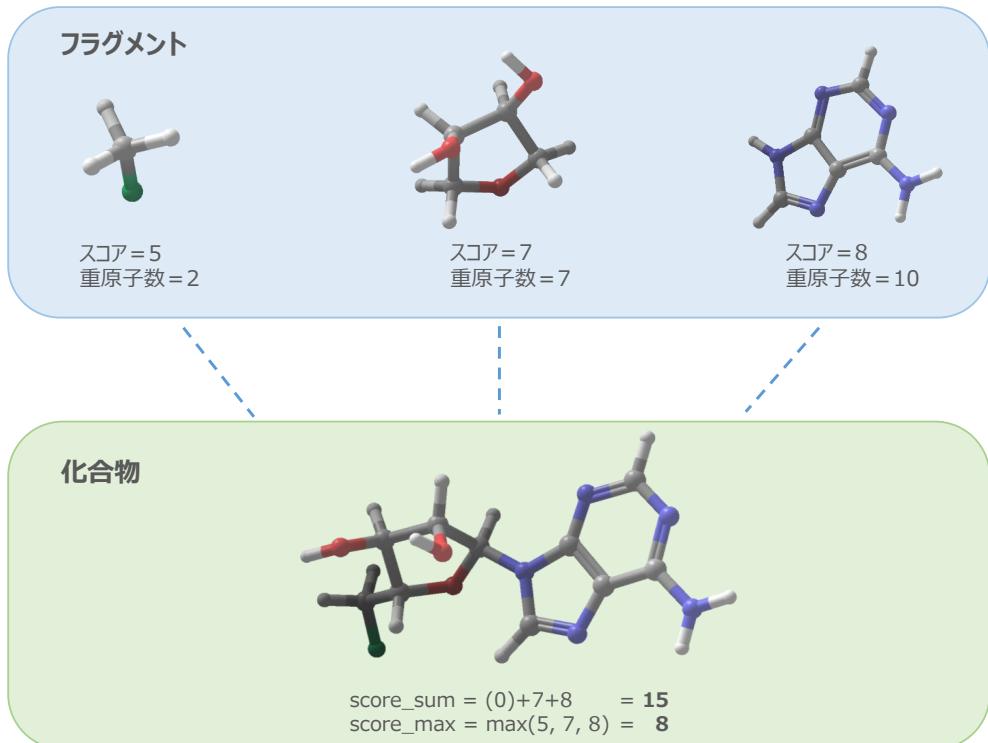


図 3.6 score_sum, score_max の算出

	sum	max	和
	15	8	$z\text{-score} = \frac{x - \mu}{\sigma}$
	18	6	
	13	9	

図 3.7 maxsumBS の算出

第4章

実験

ここでは、提案手法と既存手法との比較実験を行い、提案手法の長所を示す。

4.1 データセット

本実験では、データセットとして Directory of Useful Decoys, Enhanced (DUD-E)⁴⁰⁾ を用いた。DUD-E はカリフォルニア大学サンフランシスコ校の Mysinger らによって作成されたドッキングミュレーション手法を評価するためのデータセットである。DUD-E には 102 種類の標的タンパク質(ターゲット)が登録されており、それぞれに対してタンパク質構造・正例化合物・負例化合物を用意している。表 4.1 にターゲットごとの化合物数、正例と負例の比率の最小値、最大値、平均値を示す。各ターゲットの詳細については付録 A に記載する。なお、DUD-E のターゲットのうち fgfr1 および fa10 は記載されている負例数とデータセットに実際に含まれている負例数が大きく異なっているが、そのまま扱うこととする。

表 4.1 DUD-E のターゲットの化合物数

	総化合物数	正/負例の比率
最大値	52,022 (fnta)	1:104 (fnta)
平均値	13,881	1:60
中央値	9,297	1:59
最小値	472 (fgfr1)	1:2.4 (fgfr1)

4.2 予測精度の評価指標

バーチャルスクリーニングでは、計算機による選別を通過して実際に活性実験が行われる化合物数が元の化合物と比べてきわめて少ない状況を想定することが多い。また、化合物データベースの

中で実際に標的タンパク質を阻害する化合物は約1,000個に1個であると言われており、したがって正例と負例の比率が大きく偏っていると言える。そのためこの分野における予測精度の評価指標は以下の2種類が多く用いられている。

• ROC-AUC

Receiver Operating Characteristic (ROC) 曲線は、正例/負例の予測の閾値を変化させながら、縦軸に True Positive (TP) 率、横軸に False Positive (FP) 率をとった曲線である。TP 率とはデータセット中の正例の中で正しく正例と判別されたものの割合であり、FP 率とはデータセット中の負例の中で誤って正例と判別されたものの割合である。TP 率、FP 率はそれぞれ以下の式で求められる。

$$\text{TP 率} = \frac{\# \text{TP}}{\# \text{TP} + \# \text{FN}} \quad (4.1)$$

$$\text{FP 率} = \frac{\# \text{FP}}{\# \text{FP} + \# \text{TN}} \quad (4.2)$$

この方法によって描かれた ROC 曲線の曲線下面積 (Area Under the Curve, AUC) を用いた評価指標が ROC-AUC である。具体例を図 4.1 に示す。

• Enrichment Factor

Enrichment Factor (EF) とは、予測結果の上位のみを取り出したときに、元々のデータセットからどれだけ正例が「濃縮されたか」を表す指標である。具体例を図 4.2 に示す。上位どのくらいを取り出すかによって値が異なり、上位 x% 取り出したときの集合の正例率 (x%)、EF を EF (x%) と表記することにすると、これらは以下の式で求められる。

$$\text{正例率 (x\%)} = \frac{\text{正例数 (x\%)}}{\text{正例数 (x\%) + 負例数 (x\%)}} \quad (4.3)$$

$$\text{EF (x\%)} = \frac{\text{正例率 (x\%)}}{\text{正例率 (100\%)}} \quad (4.4)$$

本研究においては、ROC-AUC、EF (1%)、EF (2%)、EF (5%)、EF (10%) の5つの指標を用いて手法の評価を行う。

4.3 計算環境

本研究では、東京工業大学のスーパーコンピュータである TSUBAME 2.5 の Thin ノードを利用した。利用した計算環境を表 4.2 に示す。

表 4.2 利用した計算環境

CPU	Intel Xeon X5670, 2.93 GHz (6 cores) ×2
Memory	54 GB RAM

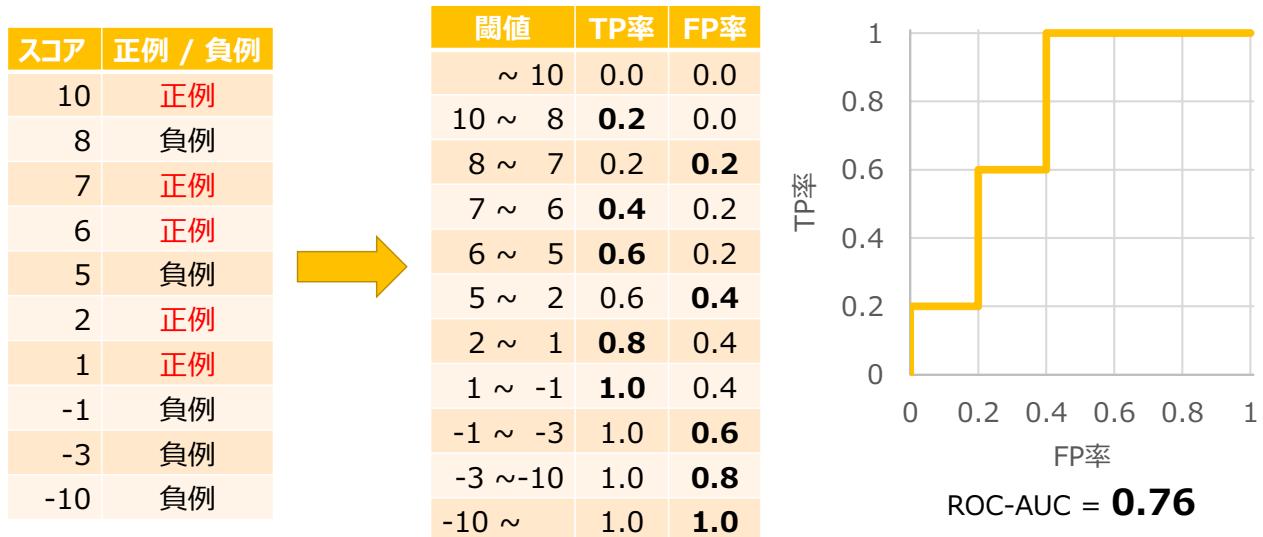


図 4.1 ROC-AUC 計算例

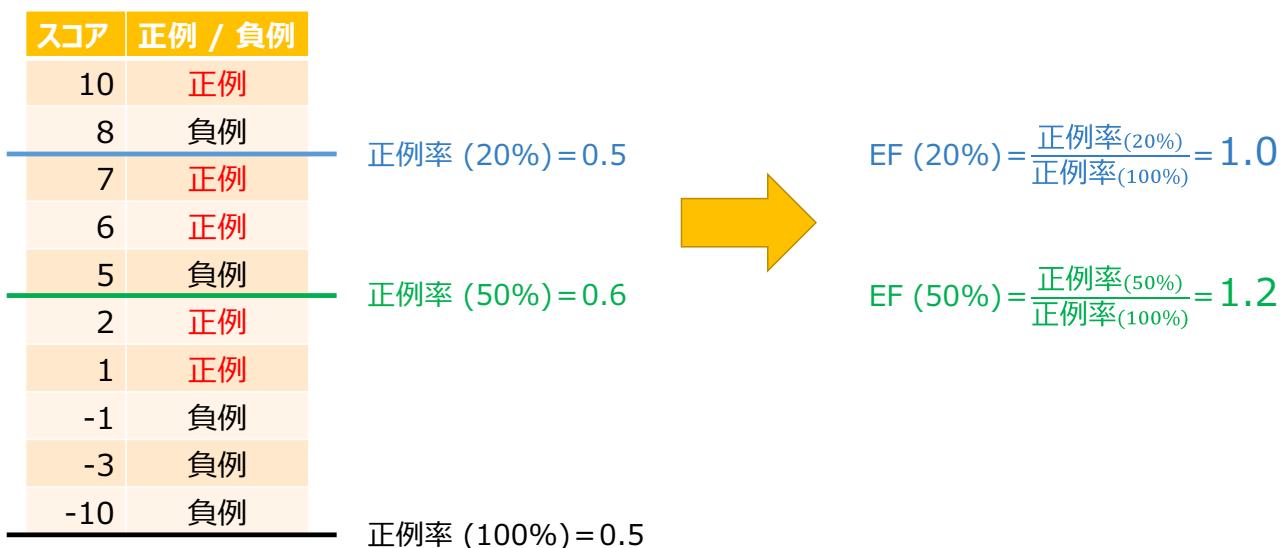


図 4.2 EF 計算例

4.4 比較対象

本提案手法はドッキングに基づくフィルタリング手法であるため、同様の用途に用いられることがある Glide HTVS（高速）モードを比較対象として用いる。また、フィルタリングとしての性能を評価するために、Glide SP（通常）モードによる化合物ドッキングシミュレーションと組み合わせた評価も行うため、計算時間などの評価に関しては Glide SP モードも比較対象とする。

4.5 評価実験

4.5.1 フラグメント分割

まず、今回用いる複数のターゲットについて、フラグメント分割を行うことでドッキングの必要数をどの程度減らせるのかを示す。それぞれターゲットにフラグメント分割を適用した場合における化合物数とフラグメント種類数の推移は図 4.3 の通りとなり、DUD-E ターゲット全体で平均するとフラグメント種類数は化合物数の約 4 分の 1 に抑えられている（表 4.3）。化合物数が多いほど、化合物数に対するフラグメント種類数が抑えられる傾向にあることも確認された。

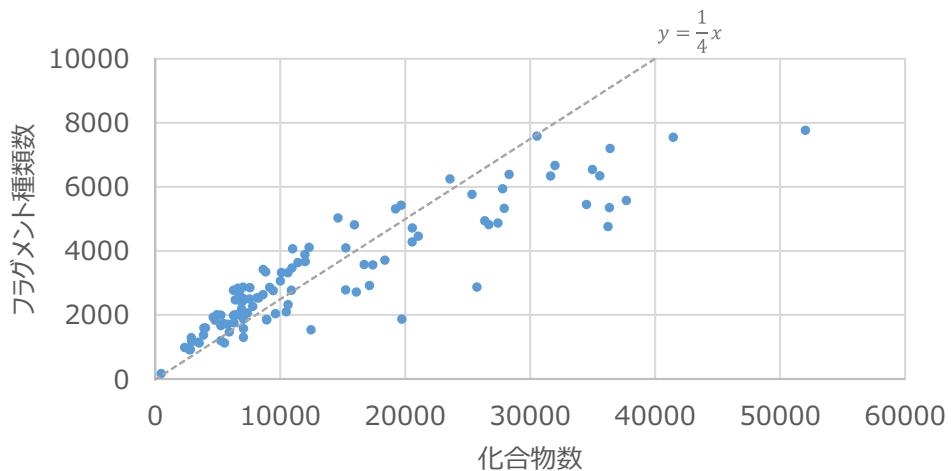


図 4.3 DUD-E ターゲットにおける化合物数とフラグメント種類数の関係

表 4.3 フラグメント分割を行った時のフラグメント 1 種類あたりの化合物数

ターゲット数	フラグメント 1 種類あたりの平均化合物数
化合物数 1 万未満の DUD-E ターゲット	53
化合物数 1 万以上の DUD-E ターゲット	49
全 DUD-E ターゲット	102

4.5.2 ドッキング速度の評価

つづいてフィルタリング手法の計算速度を評価する。ここでは以下に示す4種類の手法の計算時間を比較する。

1. Glide SP モードを利用した通常のドッキングシミュレーション 「Glide SP」
2. Glide HTVS モードを利用した簡易なドッキングシミュレーション 「Glide HTVS」
3. フラグメントのドッキングシミュレーションに Glide SP モードを用いた提案手法 「提案手法 (SP)」
4. フラグメントのドッキングシミュレーションに Glide HTVS モードを用いた提案手法 「提案手法 (HTVS)」

4.5.1節で述べたように、1つのターゲットに含まれる化合物数が多ければ多いほどフラグメント数は相対的に少なくなり提案手法の計算コストの削減が増幅される。そこでDUD-E 102 ターゲット全てでの所要計算時間の平均以外に、総化合物数が最小であるターゲット fgfr1、総化合物数が平均値に近いターゲット adrb2、総化合物数が最大であるターゲット fnta の3種類について独立して結果を示す。

結果は表4.4の通りであり、Glide HTVS モードと比較すると提案手法 (SP) は平均約9倍、提案手法 (HTVS) は平均約15倍の速度向上を達成している。

表 4.4 ドッキング計算時間の比較（括弧内は Glide HTVS との速度比）

問題 サイズ	ターゲット名	化合物数	フラグ メント 種類数	計算時間 [CPU sec.]			
				Glide SP	Glide HTVS	提案手法 (SP)	提案手法 (HTVS)
小	fgfr1	472	166	3,523	566 (x1.0)	164 (x3.5)	140 (x4.0)
中	adrb2	15,224	2,779	338,511	17,043 (x1.0)	1,481 (x11.5)	899 (x19.0)
大	fnta	52,022	7,767	1,770,967	98,665 (x1.0)	4,149 (x24.0)	2,549 (x38.7)
全ての平均		13,881	3,231	236,156	14,813 (x1.0)	1,673 (x8.9)	987 (x15.0)

4.5.3 予測精度の評価

次に、提案手法の予測精度の評価を行う。提案手法は2つのドッキングモード (SP モードおよび HTVS モード)、3つのフィルタリングスコア算出方法が存在するため合計6通りを示す。

表 4.5 提案手法の予測精度

ドッキング計算	化合物スコア 算出方法	ROC-AUC	Enrichment Factor			
			EF(1%)	EF(2%)	EF(5%)	EF(10%)
提案手法 (SP)	総和 (score_sum)	0.624	5.08	4.14	3.02	2.34
	最良値 (score_max)	0.637	6.78	5.65	3.81	2.60
	線形和 (maxsumBS)	0.679	6.03	5.03	3.96	3.00
提案手法 (HTVS)	総和 (score_sum)	0.618	4.84	3.97	2.99	2.29
	最良値 (score_max)	0.627	6.94	5.55	3.32	2.55
	線形和 (maxsumBS)	0.665	5.98	4.84	3.58	2.82
簡易ドッキングシミュレーション (Glide HTVS モード)		0.705	16.67	11.18	6.38	4.11

結果は表 4.5 の通りである。なお、各手法を用いた場合のターゲットごとの ROC 曲線は付録 B に記載している。この結果から、単体での予測精度に関しては、どの評価指標においても既存手法である Glide HTVS モードが高速性を重視した本研究の提案手法よりも勝っていることが分かる。

また、提案手法間の比較を行うことで以下のことが言える。

- ドッキング計算について、化合物フィルタリングスコアの算出方法に関わらず、提案手法 (SP) は提案手法 (HTVS) と比べてほぼすべての評価指標で予測精度が良くなる。
- ROC-AUC は maxsumBS が他の 2 つの提案手法に比べて良い結果が出ているが、EF (1%) や EF (2%) に関しては score_max が maxsumBS を上回っている。

4.5.2 で述べたように提案手法 (SP) の速度は 9 倍程度、従来手法に比べて高速である。提案手法 (HTVS) は従来手法に比べて 15 倍程度高速であり、計算時間の短縮が特に重要な場合には考慮に値するが、本研究では高速化を達成した中での精度を重視し、フラグメントのドッキングシミュレーションには Glide SP モードを用いることとする。

4.5.4 フィルタリング手法としての性能評価実験

4.5.2節および4.5.3節では、フィルタリング手法を単体で用いた場合の性能を評価し、速度では提案手法が勝っているものの、精度では Glide HTVS モードに後塵を拝する結果となった。しかし、本研究で提案した手法はフィルタリングを想定したものであり、その次に行われる通常のドッキングシミュレーション手法と組み合わせた場合の速度や精度の評価はより重要となる。

この節では通常のドッキングシミュレーションである Glide SP モードとの組み合わせを通じた評価を行う。組み合わせを通じた評価は

1. フィルタリング手法で 2%, 5%, 10%まで化合物を削減（以下、この割合を「通過率」と示す）
2. 残った化合物を通常のドッキングシミュレーション（Glide SP モード）で再計算
3. 再計算の結果の上位 1%および上位 2%の濃縮率 (EF (1%), EF (2%)) を評価

という手順を用いる。なおフィルタリング手法を用いて 2%まで削減した場合、EF (2%) はフィルタリング手法単体の性能と変わらなくなるため、「-」と表記する。

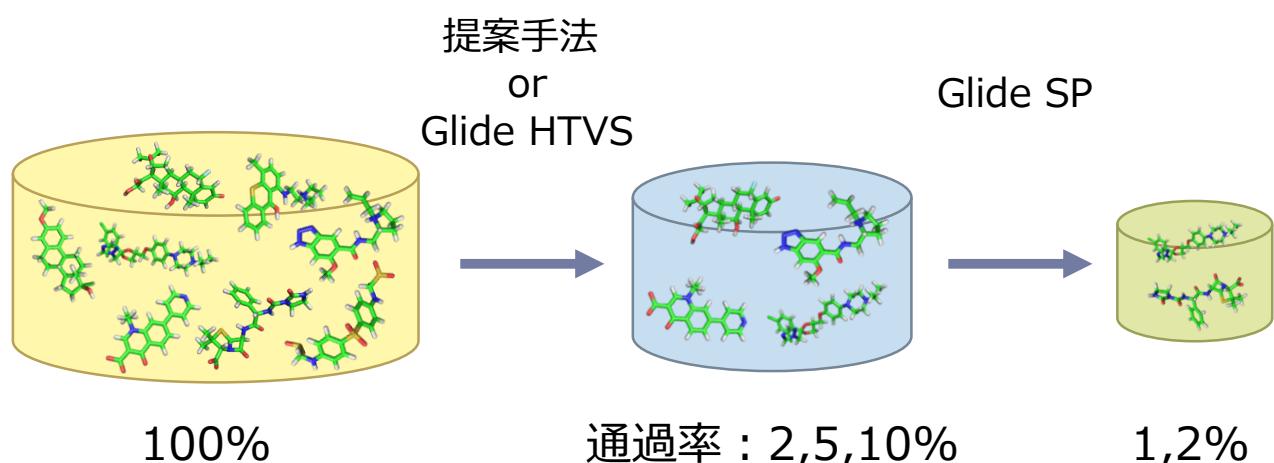


図 4.4 EF (1%) , EF (2%) 算出までの流れ

提案手法間の精度比較

まず提案手法間の精度比較を行い、化合物フィルタリングスコアの算出方法を検討した。結果は表4.6のようになり、多くの場合フィルタリングスコア算出方法はmaxsumBSを用いるのが最適であることが分かった。以降ではmaxsumBSを用いることとする。

表 4.6 フィルタリング手法としての提案手法間の精度評価

フィルタリング 手法	通過率	EF (1%)	EF (2%)	合計計算時間 [CPU sec.]
総和 (score_sum)		6.84	—	
最良値 (score_max)	2%	9.09	—	6,396
線形和 (maxsumBS)		8.75	—	
総和 (score_sum)		9.61	5.92	
最良値 (score_max)	5%	10.93	7.49	13,481
線形和 (maxsumBS)		12.92	7.99	
総和 (score_sum)		12.41	7.67	
最良値 (score_max)	10%	11.85	8.24	25,289
線形和 (maxsumBS)		15.45	10.00	

予測精度の従来手法との比較

続いて、提案手法と従来手法との比較を行った。

表 4.7 フィルタリング手法としての提案手法と従来手法の比較

フィルタリング 手法	通過率	EF (1%)	EF (2%)	合計計算時間 [CPU sec.]	提案手法の 高速化率
提案手法 (maxsumBS)		8.75	—	6,396	
従来手法 (Glide HTVS モード)	2%	17.85	—	19,536	x 3.05
提案手法 (maxsumBS)		12.92	7.99	13,481	
従来手法 (Glide HTVS モード)	5%	18.97	12.50	26,621	x 1.97
提案手法 (maxsumBS)		15.46	10.00	25,289	
従来手法 (Glide HTVS モード)	10%	19.60	12.92	38,429	x 1.52
通常ドッキング (Glide SP モード)		21.54	14.68	236,156	—

表4.7の結果より、以下のことが言える。

- 4.5.3 節で示した単体での性能評価と同様に, Glide HTVS モードが提案手法よりも精度が良くなっている.
- 一方, 計算速度について, フィルタリングで元の化合物群の 2% を通過させる場合, 提案手法と通常ドッキング計算の合計必要時間が従来用いられていた Glide HTVS モードよりも少なくなっている, これまででは達成できなかった速度での化合物の選別が可能になっていることがこの結果からわかる. この利点はフィルタを通過させる化合物の割合を高めるほど薄れて行く. これは, 通常のドッキングシミュレーションの計算時間が支配的となり, 提案しているフィルタリング手法の計算時間の面での利点が失われてしまうためである.

第5章

考察

5.1 総和法におけるフラグメント数に対するペナルティ

もし、フラグメントの結合スコアを単純に全て加算し、それを化合物のフィルタリングスコアとすると、図5.1のように化合物の総原子数が同じであっても分割数が多いほどフィルタリングスコアが向上してしまう。これは化合物のフラグメント化によって発生してしまった誤った傾向である。この分割数と総和法のスコアとの相関は最適化問題の条件緩和と考えることで説明できる。すなわち、本来化合物には原子間の結合距離という拘束条件が存在している。フラグメント分割によって切断された原子間の結合は距離を考えずにスコア付けして良いので、分割は原子間の結合という拘束条件を1つずつ緩和することに対応する。このため、フラグメント分割がされればされるほどスコアが良くなってしまうのである。

このような現象を改善するための手法として、以下2つの実験を行った。

1. 小さなフラグメントの無視

重原子（水素以外の原子）の個数に閾値を設け、その閾値を超えているフラグメントの結合ス

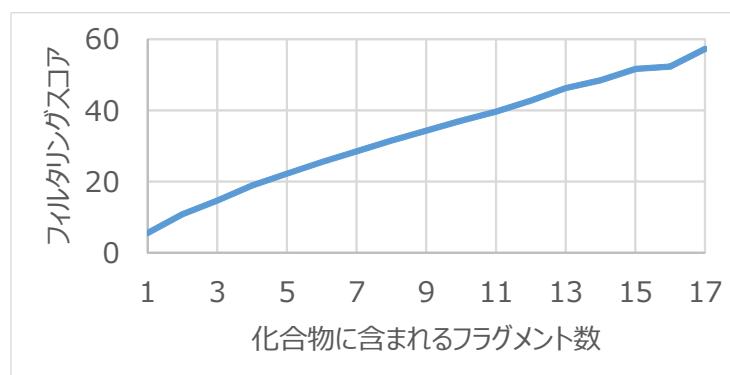


図 5.1 ターゲット fnta の全ての化合物のうち重原子数 32 の化合物の単純加算スコア

コアのみを総和に用いる。分割が多ければ多いほど小さなフラグメントが発生するため、小さなフラグメントの結合スコアを無視することで事実上のフラグメント数に対するペナルティとなる。

2. フラグメント数に対する線形ペナルティ

全てのフラグメントの結合スコアを加算した後、化合物が持つフラグメントの個数に応じたペナルティを付与する。図5.1を見ると、フィルタリングスコアの平均とフラグメント数との関係は線形に近く、フラグメント数に対して線形なペナルティを課すことでフラグメント数に依存しないフィルタリングスコアとなることが想定される。

この2つの手法を個別に利用した場合の総和法(score_sum)の精度は表5.1および表5.2のようになり、重原子数3以下のフラグメントの結合スコアを無視することが最も精度を高めている。

表 5.1 小さなフラグメントを無視することによる score_sum の精度の変化

無視するフラグメントのサイズ	ROC-AUC	Enrichment Factor			
		EF(1%)	EF(2%)	EF(5%)	EF(10%)
全てのフラグメントを利用	0.545	3.46	2.76	2.01	1.63
重原子数1	0.557	2.38	2.16	1.85	1.66
重原子数2以下	0.624	5.08	4.14	3.02	2.34
重原子数3以下	0.634	5.75	4.34	3.03	2.49
重原子数4以下	0.620	4.27	3.43	2.79	2.32
重原子数5以下	0.614	4.43	3.68	2.75	2.13
重原子数6以下	0.537	2.20	1.86	1.53	1.43

表 5.2 フラグメント数に対する線形ペナルティによる score_sum の精度の変化

フラグメント1つあたりの ペナルティ c	ROC-AUC	Enrichment Factor			
		EF(1%)	EF(2%)	EF(5%)	EF(10%)
ペナルティなし	0.545	3.46	2.76	2.01	1.63
$c = 1$	0.559	3.81	2.98	2.18	1.77
$c = 2$	0.586	4.70	3.65	2.66	2.08
$c = 3$	0.622	5.03	4.03	2.86	2.32
$c = 4$	0.588	3.80	3.19	2.51	2.14
$c = 5$	0.549	3.57	2.96	2.20	1.78
$c = 6$	0.530	3.30	2.53	1.91	1.57
$c = 7$	0.520	3.06	2.29	1.72	1.46

一方、同様に総和法のペナルティを変化させながら総和法と最良値法の線形和（maxsumBS）の精度について実験を行うと、重原子数2以下のフラグメントの結合スコアを無視した総和法を用いた場合に最良のROC-AUCとなった（表5.3、表5.4）。maxsumBSの精度はscore_sumよりも良いことから、本研究の提案手法では重原子数2以下のフラグメントの結合スコアを無視した総和法を利用することとしている。

表 5.3 小さなフラグメントを無視することによる maxsumBS の精度の変化

無視するフラグメントのサイズ	ROC-AUC	Enrichment Factor			
		EF(1%)	EF(2%)	EF(5%)	EF(10%)
全てのフラグメントを利用	0.652	5.35	4.56	3.32	2.60
重原子数1	0.652	4.67	4.18	3.25	2.56
重原子数2以下	0.679	6.03	5.03	3.96	3.00
重原子数3以下	0.672	5.57	4.79	3.78	2.85
重原子数4以下	0.653	4.89	4.32	3.46	2.67
重原子数5以下	0.643	4.95	4.28	3.29	2.55
重原子数6以下	0.566	2.76	2.46	2.03	1.79

表 5.4 フラグメント数に対する線形ペナルティによる maxsumBS の精度の変化

フラグメント1つあたりの ペナルティ c	ROC-AUC	Enrichment Factor			
		EF(1%)	EF(2%)	EF(5%)	EF(10%)
ペナルティなし	0.652	5.35	4.56	3.32	2.60
$c = 1$	0.657	5.67	4.78	3.40	2.67
$c = 2$	0.665	6.40	5.02	3.59	2.80
$c = 3$	0.665	5.97	4.84	3.78	2.88
$c = 4$	0.630	6.16	4.69	3.41	2.66
$c = 5$	0.609	6.49	4.71	3.17	2.45
$c = 6$	0.600	6.49	4.69	3.11	2.36
$c = 7$	0.591	6.43	4.62	3.06	2.32

5.2 提案手法が得意とするケースの調査

4.5.3 節の実験結果より、提案手法は簡易なドッキングシミュレーションである Glide HTVS モードと比べて精度が低調に終わることが判明している。しかし、本研究で用いた 102 ターゲット中 46 ターゲットに関しては提案手法が従来手法である Glide HTVS モードよりも精度が良く、ROC-AUC で 0.2 以上上回っているケースも表 5.5 に示す通り 3 例存在している。

どのような場合において提案手法が有用であるかを調べるために、この 3 つのターゲットについて化合物の持つフラグメント数の平均、小さなフラグメントを削減した後のフラグメント数の平均、sitemap⁴¹⁾ によって計算された各タンパク質の結合部位のサイズを求めた。その結果、結合部位のサイズやデータセット全体を通してのフラグメント数などに傾向は見受けられなかったが、

- 化合物の持つフラグメント数の平均
- 重原子数が 2 以下のフラグメント数の平均

どちらも正例より負例が上回っているということが判明した（表 5.6）。

表 5.5 提案手法が上手く行ったケース

提案手法（maxsumBS）が従来手法（Glide HTVS モード）よりも ROC-AUC で 0.2 以上上回ったケースについて、ROC-AUC の差の降順で示している。

ターゲット名	ROC-AUC 差	ROC-AUC	
		従来手法	提案手法
mcr	0.319	0.466	0.785
akt1	0.285	0.539	0.824
gcr	0.252	0.528	0.780

表 5.6 提案手法が得意なターゲットの性質

ターゲット	フラグメント数の平均			重原子数 2 以下の フラグメント数の平均			結合部位のサイズ [Å ³]
	全体	正例	負例	全体	正例	負例	
akt1	7.98	6.94	8.00	4.64	3.08	4.66	637
gcr	5.93	5.33	5.94	2.68	2.00	2.69	471
mcr	5.90	5.43	5.91	2.67	2.06	2.68	179
全 102 ターゲット 平均	7.22	7.43	7.21	3.83	3.78	3.83	437

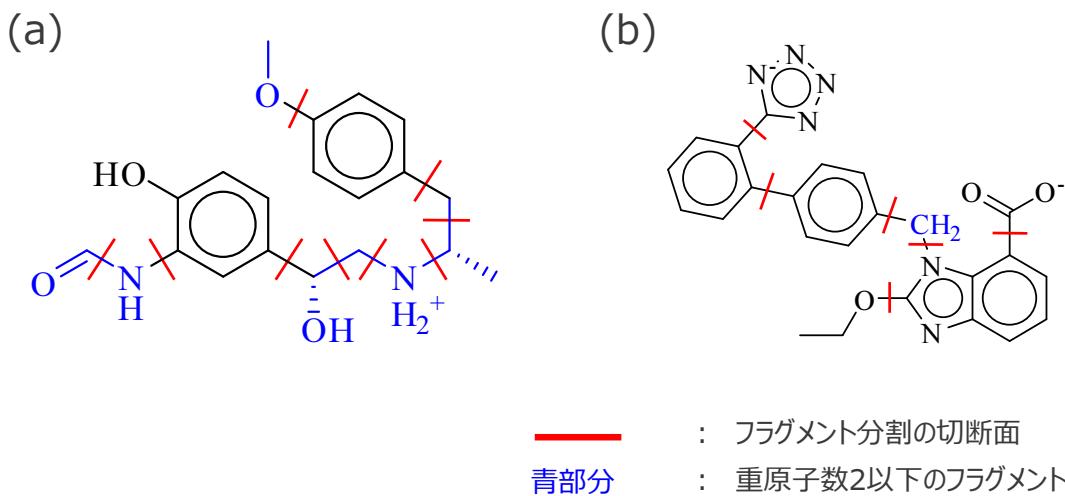


図 5.2 薬剤化合物の例：(a) ホルモテロール, (b) カンデサルタン

例えば、ホルモテロール（図 5.2-(a)）とカンデサルタン（図 5.2-(b)）はそれぞれ薬剤として認められている化合物だが、重原子数が 2 以下になるフラグメントの量がかなり異なる。このような場合、前者 (a) のような重原子数 2 以下のフラグメントが多く発生してしまう化合物が結合するタンパク質を対象としたフィルタリングには従来通り簡易ドッキングシミュレーションを用い、後者 (b) のような重原子数 2 以下のフラグメントがあまり発生しない化合物が結合するタンパク質を対象としたフィルタリングには提案手法を用いることでより高い精度でフィルタリングを行うことができると推定される。

一方この性質は 5.1 節で示したペナルティの影響を受けていると考えられるため、化合物の構造の有利不利が発生しないようなスコア計算手法およびペナルティの考案を引き続き行う必要がある。

5.3 提案手法の利用例

バーチャルスクリーニングでは数百万化合物から数百化合物程度を選別することが多く、上位 0.01% など、非常に小さな比率における Enrichment Factor の計算などが本来必要となる。また、計算時間に関しても数百万化合物を用いた場合に何日を要するのか、という評価が必要である。

しかし本研究で評価に利用したデータセットである DUD-E は表 4.1 で示したように 472 化合物しか存在しないターゲットも存在しており、このようなターゲットは上位 0.01% を計算することは不可能である。そこで、ここでは DUD-E のターゲットのうち総化合物数が 10,000 以上である 49 ターゲットを用いることで、フィルタリングにおける化合物の通過率がより少ない場合や、EF (0.1%) などの小さな割合における EF の評価（表 5.8）を用い、実際のバーチャルスクリーニングでの提案

手法の利用例を示す。計算速度に関しては線形に計算量が増大すると仮定することで数百万化合物を評価した場合の計算時間の概算値を算出する。

なお、総化合物数が 10,000 以上である 49 ターゲットの平均化合物数は 22,259、平均フラグメント種類数は 4,588 である。

1. 大規模な化合物データベース全体の超高速な評価

表 5.8 によると、提案手法で 0.5% の化合物をフィルタリングし、それらを通常のドッキングミュレーションで再評価することで Glide HTVS モードを用いる場合の約 8 分の 1 の計算時間で評価を完了させることができる。例えば 1,000 万化合物を評価する場合、今回のケースの 450 倍程度の化合物数となるので、Glide HTVS モードは 1 CPU 換算で 4 か月程度を要してしまう。一方、提案手法と通常ドッキングである Glide SP モードの組み合わせでは 1 CPU でも半月程度で済む計算となる。この差は大きく、提案手法は有用であると言える。

表 5.7 化合物全体を評価するのに要する時間の比較

	合計計算時間 [CPU sec.]	1,000 万化合物評価の 推定時間 [CPU days]
提案手法で 0.5% フィルタリング	3,280	17.1
Glide HTVS モード単独性能	23,552	122.5

2. 従来手法以下の所要時間の中での予測精度の向上

表 5.8 に示されている通り、提案手法と従来手法とで単純に比較を行うと精度は従来手法に分がある。しかしいくつかのケースについては、化合物ライブラリのサイズを変えることで同程度の所要時間の中で精度を高めることができる。例えば、100 万化合物を Glide HTVS モードを用いて 10% にフィルタリングし、Glide SP でリランギングした場合、上位 1 万化合物の濃縮率 (EF 1% に相当する) は 16.89、この時の推定必要計算時間は 32.0 CPU days となる。一方、1,000 万化合物を提案手法で 1% にフィルタリングし、Glide SP でリランギングした場合、上位 1 万化合物の濃縮率 (EF 0.1% に相当する) は 20.26、この時の推定必要計算時間は 26.9 CPU days となり、速度を向上させつつ、予測精度を高めることができる。このようなケースは複数存在しており（表 5.9）、これらの場合においては提案手法を利用すべきであると言える。

表 5.8 総化合物数が 1 万以上存在する DUD-E のターゲットに対する評価実験

フィルタリング 手法	通過率	EF (0.1%)	EF (0.2%)	EF (0.5%)	EF (1%)	EF (2%)	合計計算時間 [CPU sec.]
提案手法 (maxsumBS)	0.5%	14.33	9.39	—	—	—	3,280
従来手法 (Glide HTVS モード)	35.16	29.97	—	—	—	—	25,452
提案手法 (maxsumBS)	20.26	13.91	7.14	—	—	—	5,180
従来手法 (Glide HTVS モード)	35.36	31.17	22.19	—	—	—	27,352
提案手法 (maxsumBS)	24.87	19.14	10.57	6.32	—	—	8,979
従来手法 (Glide HTVS モード)	35.54	31.69	23.10	15.50	—	—	31,151
提案手法 (maxsumBS)	29.24	25.04	16.80	10.51	6.29	—	20,378
従来手法 (Glide HTVS モード)	35.54	31.40	23.56	16.28	10.59	—	42,550
提案手法 (maxsumBS)	31.94	27.48	19.68	13.58	8.36	—	39,377
従来手法 (Glide HTVS モード)	35.70	31.78	23.80	16.89	11.07	—	61,549
通常ドッキング (Glide SP モード)	35.98	32.82	25.57	18.96	12.82	379,965	

表 5.9 提案手法が従来手法に速度・精度ともに勝る例

	化合物数	上位 1 万化合物の	推定計算時間
		濃縮率 (EF)	[CPU days]
提案手法で 1% フィルタリング	1,000 万	20.26 (EF 0.1%)	26.9
Glide HTVS モードで 10% フィルタリング	100 万	16.89 (EF 1%)	32.0
提案手法で 1% フィルタリング	500 万	13.91 (EF 0.2%)	13.5
Glide HTVS モードで 10% フィルタリング	50 万	11.07 (EF 2%)	16.0
提案手法で 2% フィルタリング	1,000 万	24.87 (EF 0.1%)	46.7
Glide HTVS モードで 10% フィルタリング	200 万	23.80 (EF 0.5%)	64.0
提案手法で 5% フィルタリング	200 万	16.80 (EF 0.5%)	21.2
Glide HTVS モードで 5% フィルタリング	100 万	16.28 (EF 1%)	22.1

第6章

結論

6.1 本研究の結論

本研究ではドッキングに基づいた高速なフィルタリング手法を提案した。この手法は以下の3つの要素を用いて高速化を実現している。

- 化合物の内部自由度を無くすような分割によるドッキングの高速化
- 共通フラグメントの発生によるドッキング回数の削減
- 高速な化合物のスコア付け

化合物のフィルタリングスコアは3通りの計算式を実験し、フラグメントの結合スコアの総和と最良値の線形和を行う maxsumBS が最良であることを示した（表6.1）。

表 6.1 提案手法のフィルタリング利用時の精度

手法	通過率	EF (1%)	EF (2%)	合計計算時間 [CPU sec.]
総和 (score_sum)	2%	6.84	—	6,396
最良値 (score_max)		9.09	—	
線形和 (maxsumBS)		8.75	—	
総和 (score_sum)	5%	9.61	5.92	13,481
最良値 (score_max)		10.93	7.49	
線形和 (maxsumBS)		12.92	7.99	
総和 (score_sum)	10%	12.41	7.67	25,289
最良値 (score_max)		11.85	8.24	
線形和 (maxsumBS)		15.45	10.00	

また、提案手法を DUD-E の全ターゲットである 102 種のデータセットを用いて評価すると、予測精度は簡易ドッキングシミュレーションである Glide HTVS モードに比べ劣っているものの、計算速度は既存手法の約 9 倍の高速化を達成した（表 6.2）。

表 6.2 提案手法の性能

手法	ROC-AUC	Enrichment Factor				平均計算時間 [CPU sec.]
		EF(1%)	EF(2%)	EF(5%)	EF(10%)	
提案手法（maxsumBS, フラグメントドッキング： Glide SP モード）	0.679	6.03	5.03	3.96	3.00	1,673
従来手法（簡易ドッキング シミュレーション Glide HTVS モード 通常利用）	0.705	16.67	11.18	6.38	4.11	14,813

最後に、フィルタリング後に行う通常のドッキングシミュレーションと組み合わせた場合の速度・精度の評価を行い、提案手法をフィルタリング手法として用いるべきユースケースを示した（表 6.3）。

表 6.3 通常ドッキング（Glide SP）と組み合わせた速度・精度評価

	化合物数	上位 1 万化合物の 濃縮率 (EF)		推定計算時間 [CPU days]
		1,000 万	20.26 (EF 0.1%)	
提案手法で 1% フィルタリング	1,000 万	20.26 (EF 0.1%)		26.9
従来手法で 10% フィルタリング	100 万	16.89 (EF 1%)		32.0

6.2 今後の課題

本研究の今後の課題として、以下の事項が考えられる。

1. 速度をなるべく維持しつつの精度の向上

提案手法は高速な計算を可能にしている一方、精度は簡易ドッキングシミュレーションに劣っており、改善の余地がある。改善の方策として以下が考えられる。

- フラグメント切断面に付与する仮想原子などの考案
- フラグメントをドッキングする際のスコア関数の改善
- 通常ドッキングシミュレーションの化合物の結合スコアへの、フィルタリングスコアのフィッティング
- 総和法(score_sum)のペナルティのターゲットごとの調整
- 化合物のスコア算出時の非現実的なフラグメント配置に対するペナルティの付与

2. 数百万～数千万化合物程度の、より現実のバーチャルスクリーニングに即した化合物データセットを用いた速度評価

本研究では一般的なベンチマークデータセットである DUD-E を用いた評価を行ったが、これは1つのターゲットに対して最大でも約 50,000 個しか化合物が登録されていない。現実のバーチャルスクリーニングで一般的に行われている数百万個の化合物を用いた評価を行うことで、提案手法の有用性をより明示的に示すことができると考えられる。

3. 提案手法と従来手法の2段階フィルタリングを行った場合の性能・速度評価

本研究の結果、提案手法は簡易ドッキングシミュレーションの Glide HTVS モードよりも高速であるが精度は劣っていた。通常のドッキングシミュレーションに対するフィルタリングのように、簡易ドッキングシミュレーションの前に提案手法を用いることで予測精度を保つつつ計算量を削減することができると考えられる。

4. Glide 以外のツールを利用した場合の提案手法の評価

本研究ではフラグメントのドッキングシミュレーションには Glide SP モードないしは Glide HTVS モードを利用したが、ここで用いるソフトウェアは自由に選択することができる。他のソフトウェアを使った場合の精度や速度の評価を行うことで、手法の汎用性を確かめることは重要である。

謝辞

本研究を進めるにあたり、貴重な時間を割いてご指導を賜り、本論文をまとめる際ににおいても細やかなご助言をいただきました秋山 泰教授に深く感謝申し上げます。

また、研究内容の方向性のディスカッションや発表資料のとりまとめなど、多くの事柄に対して丁寧なご指導をいただきました石田 貴士准教授、ならびに大上 雅史助教に感謝の意を表します。

さらに、本研究を進めるにあたり秋山研究室・関嶋研究室・石田研究室合同ゼミを通して物理化学的な背景を含めた細かく的確なアドバイスを頂いた関嶋 政和准教授に御礼申し上げます。

最後に、本研究を行うにあたり秋山研究室・関嶋研究室・石田研究室の皆様には多大なるご協力を賜りました。暖かく、時には厳しいご指摘を通してご支援いただきましたことを心より感謝いたします。

参考文献

- [1] R. A. Friesner *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, 2004.
- [2] Z. Zsoldos *et al.* eHiTS: A new fast, exhaustive flexible ligand docking system. *Journal of Molecular Graphics and Modelling*, 26(1):198–212, 2007.
- [3] G. M. Morris *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, 2009.
- [4] C. Hansch and T. Fujita. ρ - σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of the American Chemical Society*, 86(8):1616–1626, 1964.
- [5] O. Ivanciu. Applications of support vector machines in chemistry. *Reviews in computational chemistry*, 23:291–400, 2007.
- [6] G. Wolber *et al.* Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today*, 13(1-2):23–29, 2008.
- [7] J. B. Brown and Y. Okuno. Systems Biology and Systems Chemistry : New Directions for Drug Discovery. *Chemistry & Biology*, 19(1):23–28, 2012.
- [8] D. M. Krüger and A. Evers. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *ChemMedChem*, 5(1):148–158, 2010.
- [9] E. Yuriev and P. A. Ramsland. Latest developments in molecular docking: 2010-2011 in review. *Journal of Molecular Recognition*, 26(5):215–239, 2013.
- [10] S. Kannan and R. Ganji. Porting Autodock to CUDA. *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pages 1–8, 2010.
- [11] S. McIntosh-Smith *et al.* High performance in silico virtual drug screening on many-core processors. *International Journal of High Performance Computing Applications*, 29(2):119–134, 2015.

- [12] O. Trott and A. J. Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010.
- [13] J. J. Irwin and B. K. Shoichet. ZINC — A Free Database of Commercially Available Compounds for Virtual Screening ZINC. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005.
- [14] R. Nilakantan *et al.* New method for rapid characterization of molecular shapes: applications in drug design. *Journal of chemical information and computer sciences*, 33(1):79–85, 1993.
- [15] M. D. Parenti *et al.* Docking and Database Screening Reveal New Classes of Plasmodium falciparum Dihydrofolate Reductase Inhibitors. *Journal of Medicinal Chemistry*, 46(14):2834–2845, 2003.
- [16] T. Fujimoto *et al.* In silico multi-filter screening approaches for developing novel beta-secretase inhibitors. *Bioorganic & medicinal chemistry letters*, 18(9):2771–2775, 2008.
- [17] A. Grover *et al.* A leishmaniasis study: Structure-based screening and molecular dynamics mechanistic analysis for discovering potent inhibitors of spermidine synthase. *Biochimica et Biophysica Acta*, 1824(12):1476–1483, 2012.
- [18] T. J. A. Ewing *et al.* DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design*, 15(5):411–428, 2001.
- [19] M. Rarey *et al.* A fast flexible docking method using an incremental construction algorithm. *Journal of molecular biology*, 261(3):470–489, 1996.
- [20] G. Jones *et al.* Development and validation of a genetic algorithm for flexible docking. *Journal of molecular biology*, 267(3):727–748, 1997.
- [21] R. Abagyan *et al.* ICM - A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *Journal of Computational Chemistry*, 15(5):488–506, 1993.
- [22] I. D. Kuntz *et al.* A geometric approach to macromolecule-ligand interactions. *Journal of molecular biology*, 161(2):269–288, 1982.
- [23] G. M. Morris *et al.* Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *Journal of Computational Chemistry*, 19(14):1639–1662, 1998.

- [24] R. Wang *et al.* Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design*, 16(1):11–26, 2002.
- [25] D. K. Gehlhaar *et al.* Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming. *Chemistry and Biology*, 2(5):317–324, 1995.
- [26] M. D. Eldridge *et al.* Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of computer-aided molecular design*, 11(5):425–445, 1997.
- [27] H. M. Berman *et al.* The Protein Data Bank. *Nucleic acids research*, 28(1):235–242, 2000.
- [28] I. Muegge. PMF scoring revisited. *Journal of Medicinal Chemistry*, 49(20):5895–5902, 2006.
- [29] M. Xue *et al.* Knowledge-based scoring functions in drug design. 1. Developing a target-specific method for kinase-ligand interactions. *Journal of Chemical Information and Modeling*, 50(8):1378–1386, 2010.
- [30] H. Gohlke *et al.* Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology*, 295(2):337–356, 2000.
- [31] S. Y. Huang and X Zou. Inclusion of solvation and entropy in the knowledge-based scoring function for protein – ligand interactions. *Journal of chemical information and modeling*, 50(2):262–273, 2010.
- [32] M. H. J. Seifert. Targeted scoring functions for virtual screening. *Drug Discovery Today*, 14(11-12):562–569, 2009.
- [33] T. Tuccinardi *et al.* Protein kinases: Docking and homology modeling reliability. *Journal of Chemical Information and Modeling*, 50(8):1432–1441, 2010.
- [34] C. A. Lipinski *et al.* Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.
- [35] G. R. Bickerton *et al.* Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.
- [36] How long does it take to screen 10,000 compounds with glide? <http://www.schrodinger.com/kb/1012>. (2016年1月24日閲覧)

- [37] 小峰 駿汰 他. フラグメント伸長型タンパク質-化合物ドッキングのビームサーチによる高速化. 情報処理学会研究報告, 2015-BIO-4(62):1–8, 2015.
- [38] N. M. O’Boyle *et al.* Open Babel: An open chemical toolbox. *Journal of cheminformatics*, 3(33):1–14, 2011.
- [39] M. L. Verdonk *et al.* Virtual screening using protein-ligand docking: Avoiding artificial enrichment. *Journal of Chemical Information and Computer Sciences*, 44(3):793–806, 2004.
- [40] M. M. Mysinger *et al.* Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 2012.
- [41] T. A. Halgren. Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling*, 49(2):377–389, 2009.
- [42] A. Gaulton *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 2012.
- [43] G. W. Bemis and M. A. Murcko. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996.

付録 A

DUD-E データセットの詳細

DUD-E (A Database of Useful Decoys: Enhanced) はカリフォルニア大学サンフランシスコ校の Mysinger らによって作成されたドッキングシミュレーションツールを評価するためのデータセットであり⁴⁰⁾、無料で公開されている (<http://dude.docking.org/>)。ターゲットは薬剤開発の標的となるものを多様性を考慮しながら 102 種類選ばれており、それぞれに対して 1. 正例化合物群、2. 負例化合物群、3. 代表タンパク質構造の 3 つが用意されている。

正例化合物群、負例化合物群、代表タンパク質構造はそれぞれ以下の手順で選択されている。

1. 正例化合物群

- (a) ターゲットタンパク質への実験データがある化合物を ChEMBL⁴²⁾ から取得
- (b) (a) の化合物のうち、タンパク質の機能の阻害性能を示す IC₅₀ (タンパク質の機能を 50% 阻害する時の化合物濃度) が 1μM 以下である化合物を既知の阻害剤とする
- (c) 既知阻害剤に対し Bemis-Murcko atomic frameworks (化合物の構造に基づいたクラスタリング手法⁴³⁾) を用いたクラスタリングを行い、化合物数が 100 以上になるように N 個の化合物を IC₅₀ の良い順番に各クラスタから選択する

2. 負例化合物群

負例化合物群は以下 2 種類の化合物で構成されている。

- デコイ化合物（実験は行われていないが、標的タンパク質の阻害を行わないと考えられる化合物）
- 既知の負例化合物（実験によって 標的タンパク質の阻害をしないことが確認されている化合物）

デコイ化合物は 1. で作成した正例化合物群のそれぞれに対して以下の操作を行うことで作成される。

- (a) 注目している化合物の分子量、分配係数など物理化学的な値を計算し、これと近い物性を持つ化合物を ZINC¹³⁾ から数千個取得する
- (b) ZINC から得た数千個の化合物のうち、正例化合物群のどれかに構造が似ている化合物を除去する
- (c) 残った化合物群の中から、50 個をランダムサンプリングする

また、実験の結果 IC₅₀ が 30μM よりも悪い化合物を「既知の負例化合物」として負例化合物群に追加する。

3. 代表タンパク質構造

- (a) ターゲットタンパク質の立体構造を PDB より多数取得する
- (b) (a) で得た全ての立体構造に対し、上記 1. 2. の正例/負例化合物群を用いてドッキングを実施
- (c) 立体構造の解像度が良く、ドッキングの結果正例/負例の弁別精度が高い、人間のタンパク質を (a) の立体構造から選択し、代表タンパク質構造とする

表 A.1 DUD-E の詳細 (1)

ターゲット	PDBID	タンパク質詳細	化合物数		平均フラグメント数	
			正例	負例	正例	負例
aa2ar	3EML	Adenosine A2a receptor	482	31,498	6.26	7.04
abl1	2HZI	Tyrosine-protein kinase ABL	182	10,746	7.08	7.33
ace	3BKL	Angiotensin-converting enzyme	282	16,860	9.90	7.52
aces	1E66	Acetylcholinesterase	453	26,233	9.48	8.07
ada	2E1W	Adenosine deaminase	93	5,449	8.16	8.01
ada17	2OI0	ADAM17	532	35,809	8.81	8.32
adrb1	2VT4	Beta-1 adrenergic receptor	247	15,842	10.05	9.62
adrb2	3NY8	Beta-2 adrenergic receptor	231	14,993	10.35	9.72
akt1	3CQW	Serine / threonine-protein kinase AKT	293	16,426	6.94	8.00
akt2	3D0E	Serine / threonine-protein kinase AKT2	117	6,893	6.61	7.39
aldr	2HV5	Aldose reductase	159	8,995	4.81	5.34
ampc	1L2S	Beta-lactamase	48	2,832	4.94	4.95
andr	2AM9	Androgen Receptor	269	14,343	3.91	4.67
aofb	1S3B	Monoamine oxidase B	122	6,900	4.37	4.39
bace1	3L5D	Beta-secretase 1	283	18,080	10.72	8.79
braf	3D4Q	Serine / threonine-protein kinase Braf	152	9,942	7.27	7.24
cah2	1BCD	Carbonic anhydrase II	492	31,133	7.14	7.10
casp3	2CNK	Caspase-3	199	10,692	10.47	8.72
cdk2	1H00	Cyclin-dependent kinase 2	474	27,830	6.25	6.70
comt	3BWM	Catechol O-methyltransferase	41	3,848	4.85	5.15
cp2c9	1R9O	Cytochrome P450 2C9	120	7,446	7.18	6.78
cp3a4	3NXU	Cytochrome P450 3A4	170	11,796	7.71	7.42
csf1r	3KRJ	Macrophage colony stimulating factor receptor	166	12,144	6.86	6.92
cxcr4	3ODU	C-X-C chemokine receptor type 4	40	3,406	6.75	7.30
def	1LRU	Peptide deformylase	102	5,696	9.78	7.77
dhi1	3FRJ	11-beta-hydroxysteroid dehydrogenase 1	330	19,340	5.89	5.54
dpp4	2I78	Dipeptidyl peptidase IV	533	40,916	6.52	6.44

表 A.2 DUD-E の詳細 (2)

ターゲット	PDBID	タンパク質詳細	化合物数		平均フラグメント数	
			正例	負例	正例	負例
drd3	3PBL	Dopamine D3 receptor	480	34,022	7.58	7.30
dyr	3NXO	Dihydrofolate reductase	231	17,170	6.68	7.10
egfr	2RGP	Epidermal growth factor receptor erbB1	542	35,020	7.27	7.74
esr1	1SJ0	Estrogen receptor alpha	383	20,663	5.55	6.72
esr2	2FSZ	Estrogen receptor beta	367	20,182	5.21	6.56
fa10	3KL6	Coagulation factor X	537	20,023	9.01	8.18
fa7	1W7X	Coagulation factor VII	114	6,245	10.29	8.26
fabp4	2NNQ	Fatty acid binding protein adipocyte	47	2,749	6.68	6.53
fak1	3BZ3	Focal adhesion kinase 1	100	5,350	8.44	8.12
fgfr1	3C4F	Fibroblast growth factor receptor 1	139	333	7.33	7.55
flkb1a	1J4H	FK506-binding protein 1A	111	5,800	9.75	8.42
fnta	3E37	Protein farnesyltransferase / geranylgeranyltransferase type I alpha subunit	592	51,430	8.20	7.65
fpps	1ZW5	Farnesyl diphosphate synthase	85	8,822	7.08	7.01
gcr	3BQD	Glucocorticoid receptor	258	14,987	5.33	5.94
glcm	2V3F	Beta-glucocerebrosidase	54	3,799	8.57	7.93
gria2	3KGC	Glutamate receptor ionotropic, AMPA 2	158	11,832	6.47	6.52
grik1	1VSO	Glutamate receptor ionotropic kainate 1	101	6,547	5.83	6.32
hdac2	3MAX	Histone deacetylase 2	185	10,299	10.02	8.29
hdac8	3F07	Histone deacetylase 8	170	10,448	9.52	7.95
hivint	3NF7	Human immunodeficiency virus type 1 integrase	100	6,644	6.42	6.35
hivpr	1XL2	Human immunodeficiency virus type 1 protease	536	35,688	11.16	8.75
hivrt	3LAN	Human immunodeficiency virus type 1 reverse transcriptase	338	18,879	4.98	5.65
hmdh	3CCW	HMG-CoA reductase	170	8,743	9.79	8.56

表 A.3 DUD-E の詳細 (3)

ターゲット	PDBID	タンパク質詳細	化合物数		平均フラグメント数	
			正例	負例	正例	負例
hs90a	1UYG	Heat shock protein HSP 90-alpha	88	4,848	5.56	7.20
hxk4	3F9M	Hexokinase type IV	92	4,696	6.78	6.97
igf1r	2OJ9	Insulin-like growth factor I receptor	148	9,291	7.95	8.27
inha	2H7L	Enoyl-[acyl-carrier-protein] reductase	43	2,300	7.14	6.17
ital	2ICA	Leukocyte adhesion glycoprotein LFA-1 alpha	138	8,487	7.72	7.62
jak2	3LPB	Tyrosine-protein kinase JAK2	107	6,495	6.23	6.63
kif11	3CJO	Kinesin-like protein 1	116	6,848	6.78	6.18
kit	3G0E	Stem cell growth factor receptor	166	10,447	7.80	7.57
kith	2B8T	Thymidine kinase	57	2,850	6.70	7.60
kpcb	2I0E	Protein kinase C beta	135	8,692	5.67	6.85
lck	2OF2	Tyrosine-protein kinase LCK	420	27,374	7.20	7.32
lkha4	3CHP	Leukotriene A4 hydrolase	171	9,448	7.70	7.59
mapk2	3M2W	MAP kinase-activated protein kinase 2	101	6,147	4.67	5.58
mcr	2AA2	Mineralocorticoid receptor	94	5,146	5.43	5.91
met	3LQ8	Hepatocyte growth factor receptor	166	11,240	7.45	7.48
mk01	2OJG	MAP kinase ERK2	79	4,548	7.25	6.84
mk10	2ZDT	c-Jun N-terminal kinase 3	104	6,599	6.63	6.71
mk14	2QD9	MAP kinase p38 alpha	578	35,810	7.31	7.03
mmp13	830C	Matrix metalloproteinase 13	572	37,126	8.98	8.29
mp2k1	3EQH	Dual specificity mitogen-activated protein kinase kinase 1	121	8,147	6.92	8.02
nos1	1QW6	Nitric-oxide synthase, brain	100	8,050	4.86	5.50
nram	1B9V	Neuraminidase	98	6,199	8.39	7.01
pa2ga	1KVO	Phospholipase A2 group IIA	99	5,146	10.12	9.63
parp1	3L3M	Poly [ADP-ribose] polymerase-1	508	30,035	5.07	5.57
pde5a	1UDT	Phosphodiesterase 5A	398	27,521	6.95	7.63
pgh1	2OYU	Cyclooxygenase-1	195	10,797	4.66	4.72
pgh2	3LN1	Cyclooxygenase-2	435	23,135	5.14	5.28

表 A.4 DUD-E の詳細(4)

ターゲット	PDBID	タンパク質詳細	化合物数		平均フラグメント数	
			正例	負例	正例	負例
plk1	2OWB	Serine / threonine-protein kinase PLK1	107	6,797	7.16	7.95
pnph	3BGS	Purine nucleoside phosphorylase	103	6,950	3.93	6.08
ppara	2P54	Peroxisome proliferator-activated receptor alpha	373	19,356	9.75	8.98
ppard	2ZNP	Peroxisome proliferator-activated receptor delta	240	12,223	8.95	8.69
pparg	2GTK	Peroxisome proliferator-activated receptor gamma	484	25,256	9.19	8.61
prgr	3KBA	Progesterone receptor	293	15,642	3.95	4.65
ptn1	2AZR	Protein-tyrosine phosphatase 1B	130	7,243	8.98	7.62
pur2	1NJS	GAR transformylase	50	2,694	12.18	8.54
pygm	1C8K	Muscle glycogen phosphorylase	77	3,940	6.74	6.19
pyrd	1D3G	Dihydroorotate dehydrogenase	111	6,446	5.66	5.64
reni	3G6Z	Renin	104	6,955	14.59	11.68
rock1	2ETR	Rho-associated protein kinase 1	100	6,297	5.54	6.30
rxra	1MV9	Retinoid X receptor alpha	131	6,935	5.95	5.79
sahh	1LI4	Adenosylhomocysteinase	63	3,450	3.46	5.67
src	3EL8	Tyrosine-protein kinase SRC	524	34,454	7.18	7.61
tgfr1	3HMM	TGF-beta receptor type I	133	8,498	5.02	5.66
thb	1Q4X	Thyroid hormone receptor beta-1	103	7,441	6.52	7.19
thrb	1YPE	Thrombin	461	26,948	11.09	8.75
try1	2AYW	Trypsin I	449	25,914	10.42	8.34
tryb1	2ZEC	Tryptase beta-1	148	7,643	11.55	9.19
tysy	1SYN	Thymidylate synthase	109	6,738	8.57	7.14
urok	1SQT	Urokinase-type plasminogen activator	162	9,841	7.29	6.88
vgfr2	2P2I	Vascular endothelial growth factor receptor 2	409	24,927	7.23	7.37
wee1	3BIZ	Serine / threonine-protein kinase WEE1	102	6,148	5.87	7.57
xiap	3HL5	Inhibitor of apoptosis protein 3	100	5,145	11.89	8.87

付録B

各手法を単体で用いた場合のROC曲線

4.5.3節にて示した総和法(score_sum), 最良値法(score_max), 総和法と最良値法の値の線形和(maxsumBS)の3つの提案手法およびGlide HTVSモードをそれぞれ単体で用いた場合のROC曲線を図B.1から図B.17に示す。3つの提案手法についてはフラグメントの結合スコア算出をGlide SPモード/Glide HTVSモードで行った場合がそれぞれ示されている。例えば、「score_max_SP」とはフラグメントの結合スコアをGlide SPモードで求め、そのフラグメントスコアの最良値をとった場合の精度がROC曲線で示されている。また、ROC-AUCの値については表B.1から表B.4にも記載する。

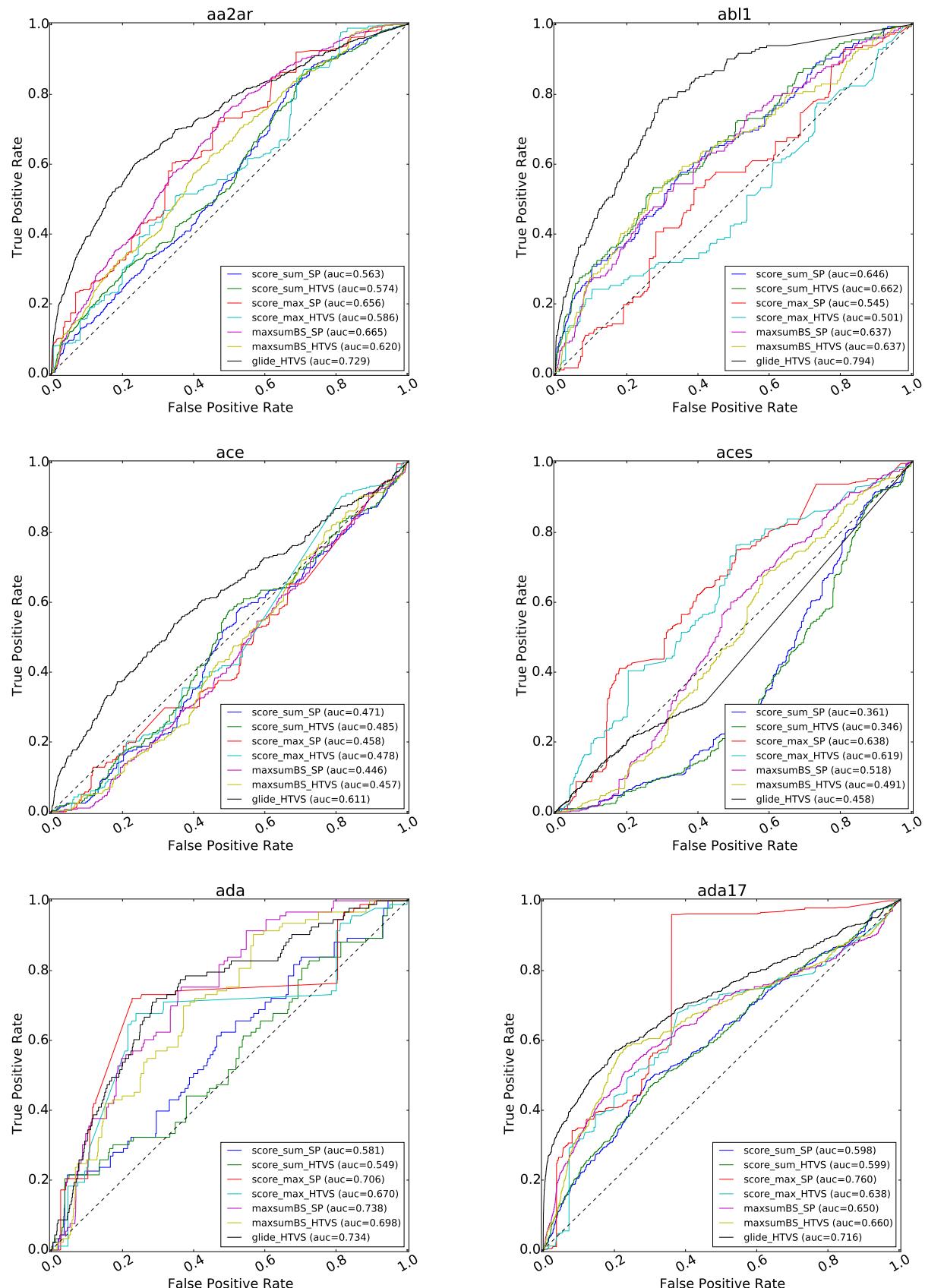


図 B.1 各手法の単体性能 ROC 曲線 (1)

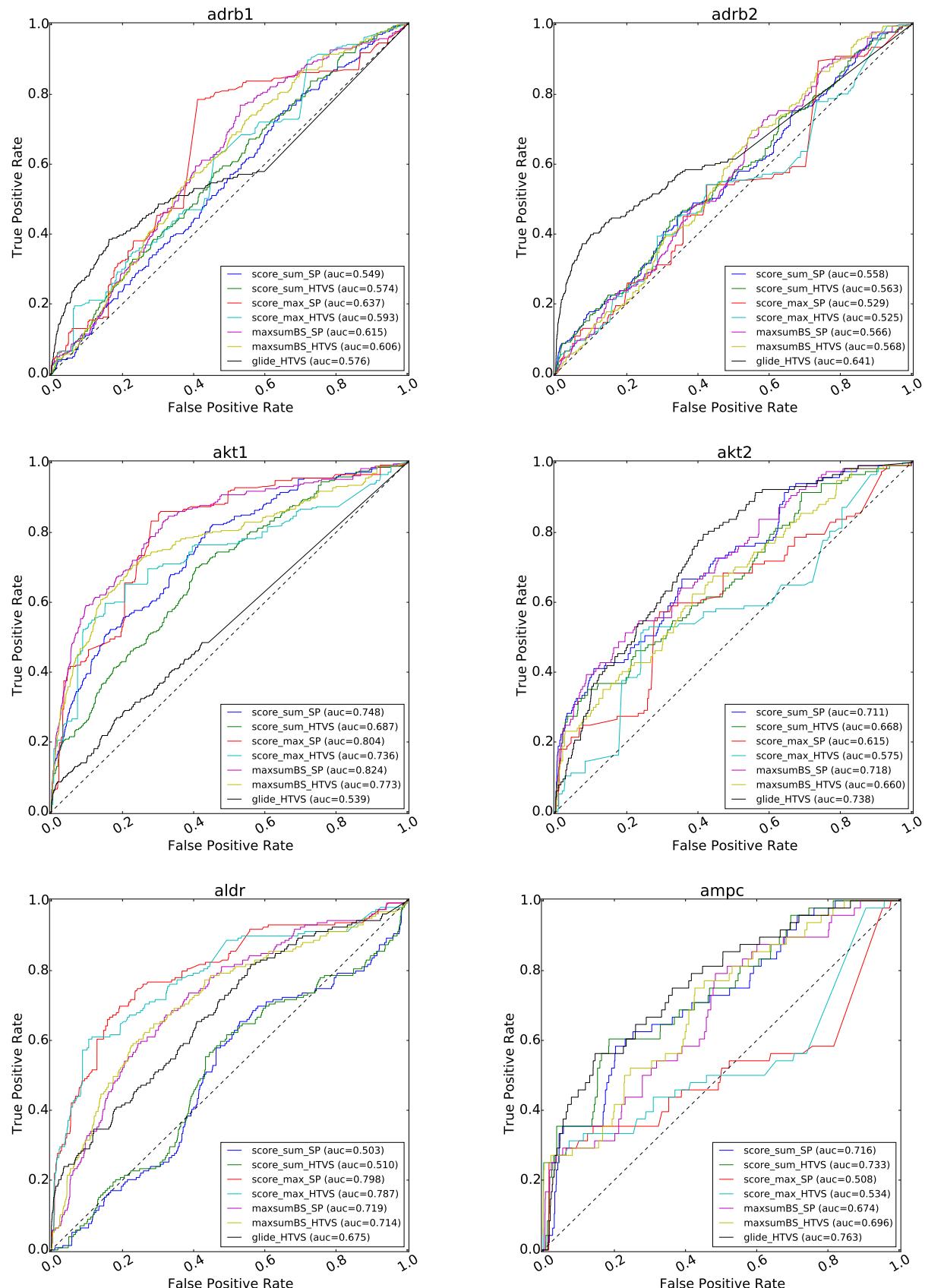


図 B.2 各手法の単体性能 ROC 曲線 (2)

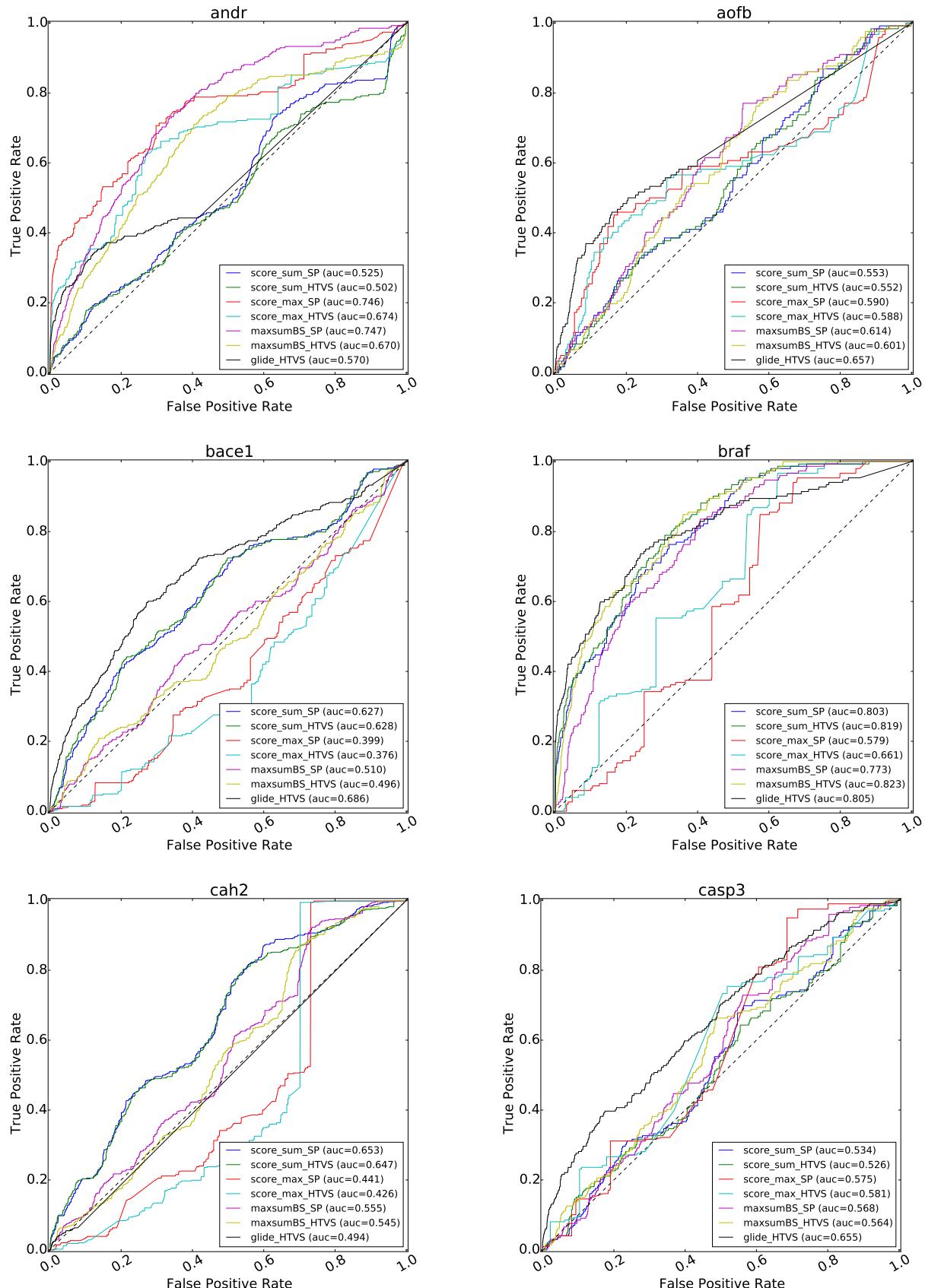


図 B.3 各手法の単体性能 ROC 曲線 (3)

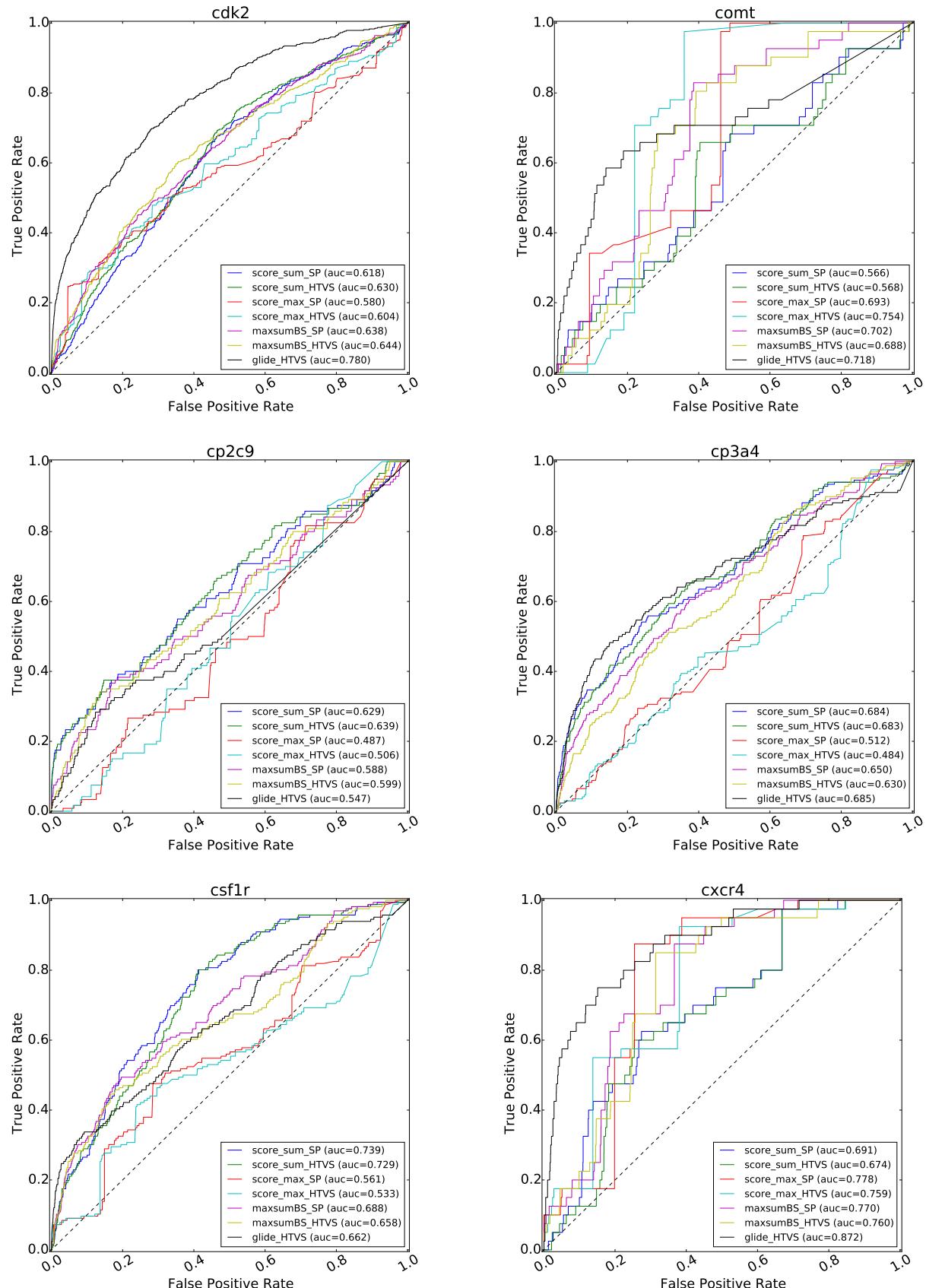


図 B.4 各手法の単体性能 ROC 曲線 (4)

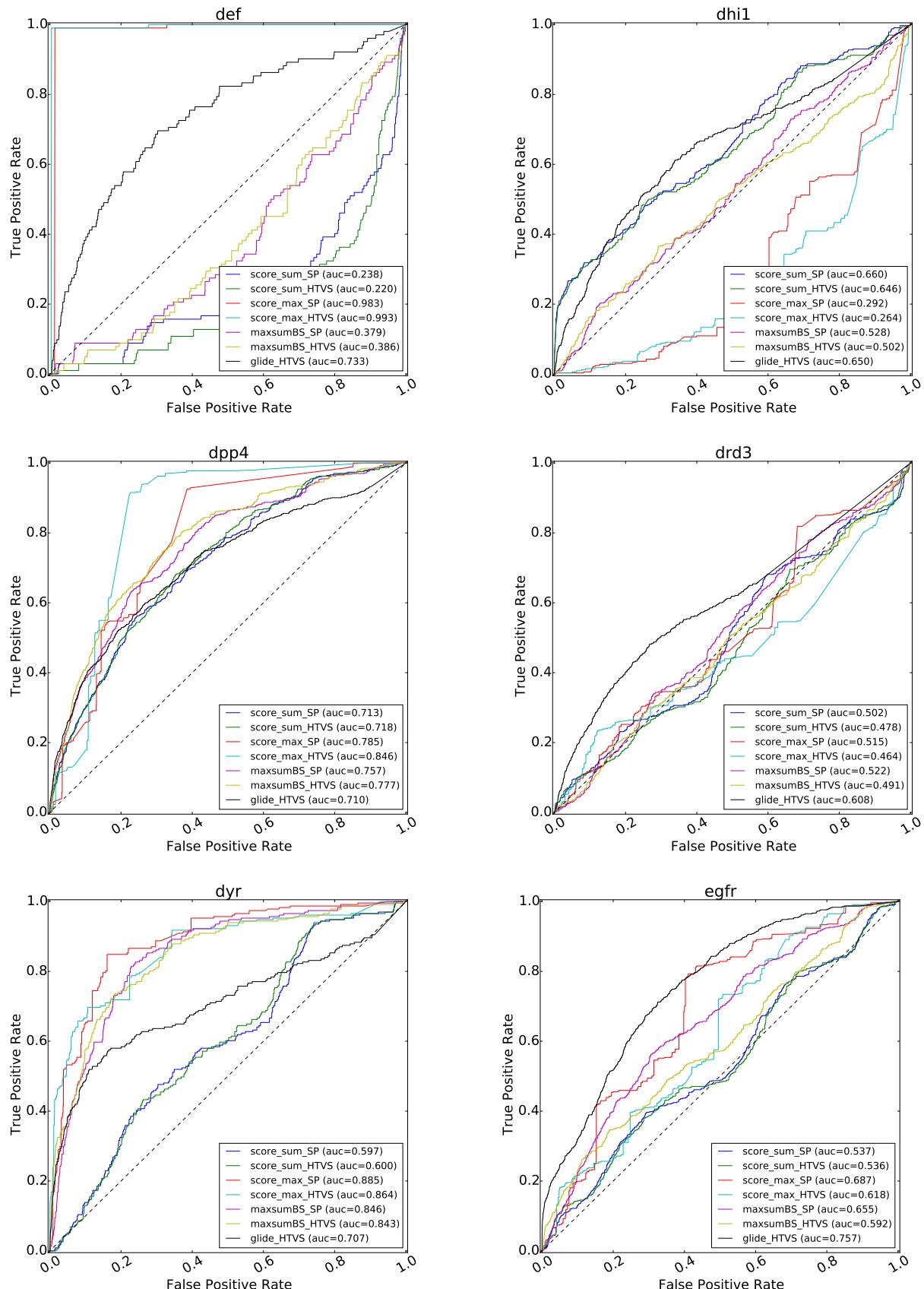


図 B.5 各手法の単体性能 ROC 曲線 (5)

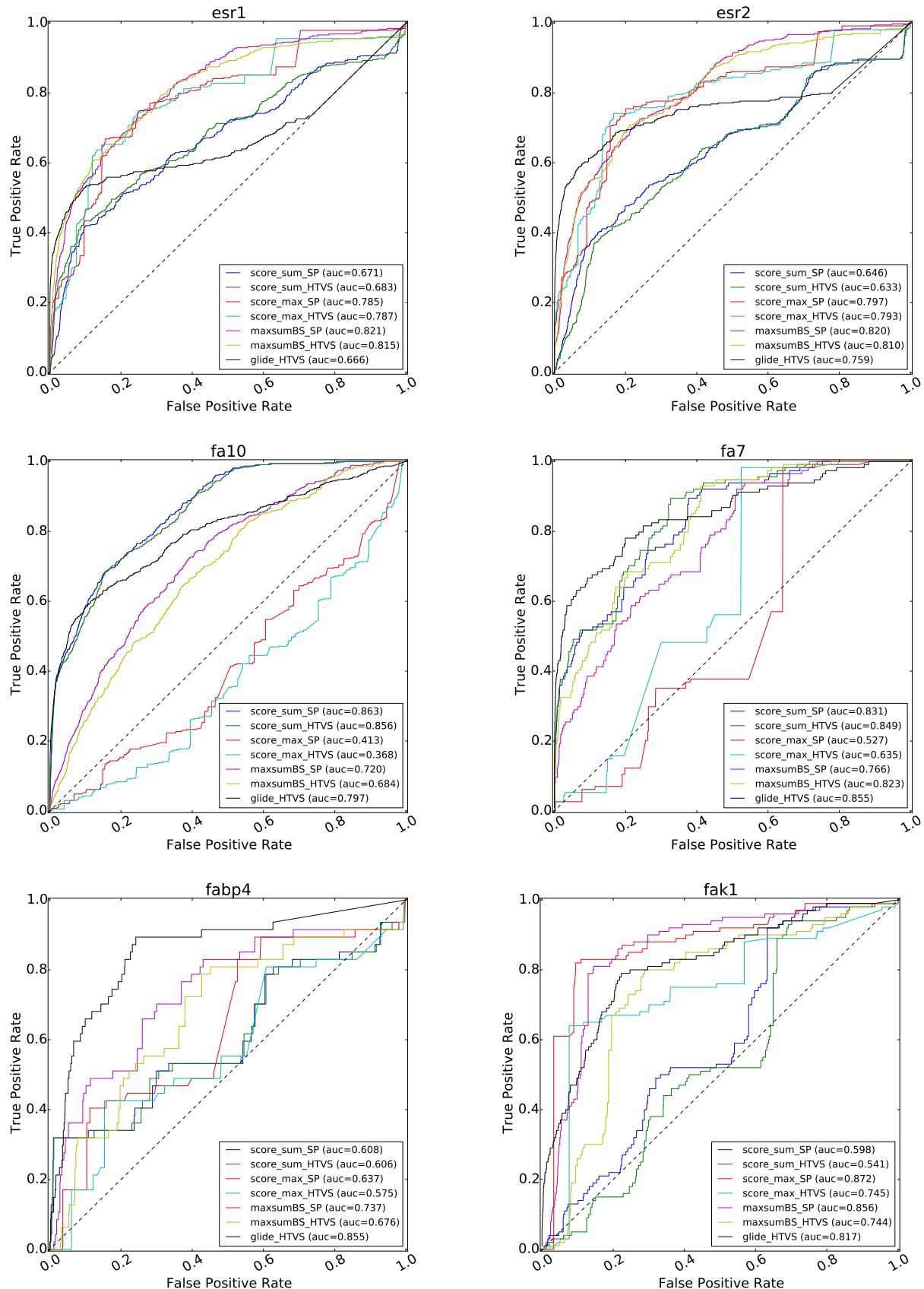


図 B.6 各手法の単体性能 ROC 曲線 (6)

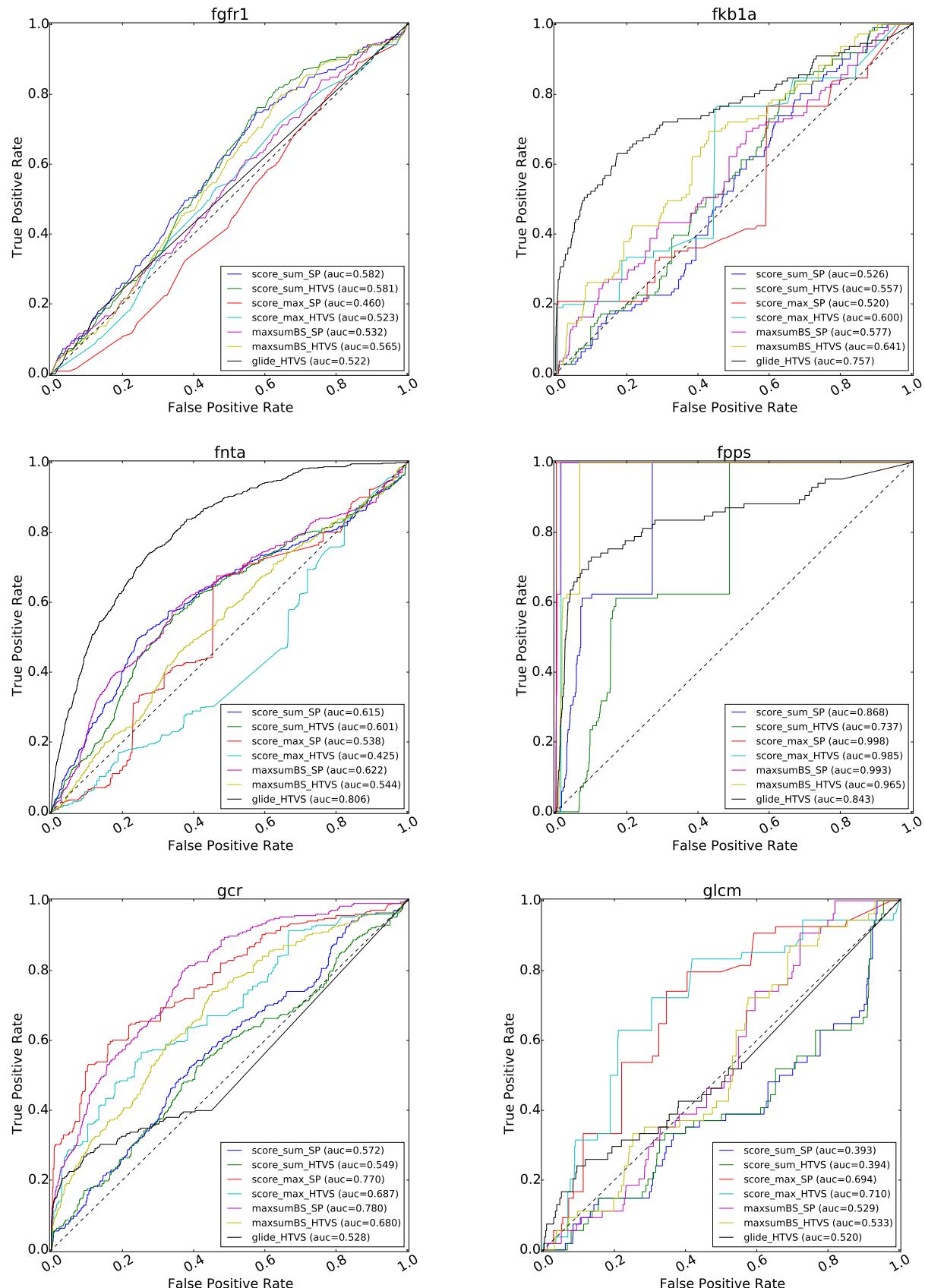


図 B.7 各手法の単体性能 ROC 曲線 (7)

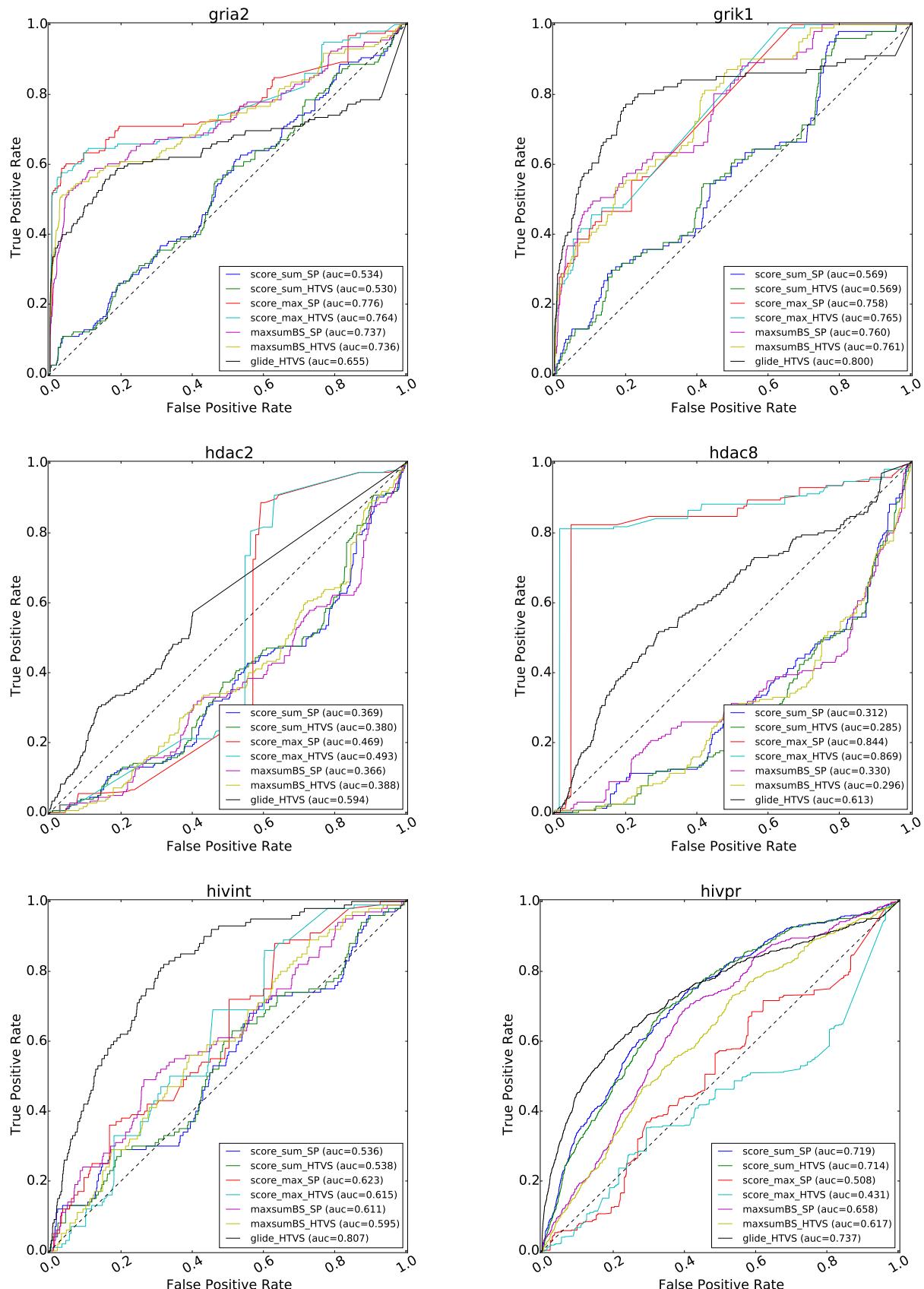


図 B.8 各手法の単体性能 ROC 曲線 (8)

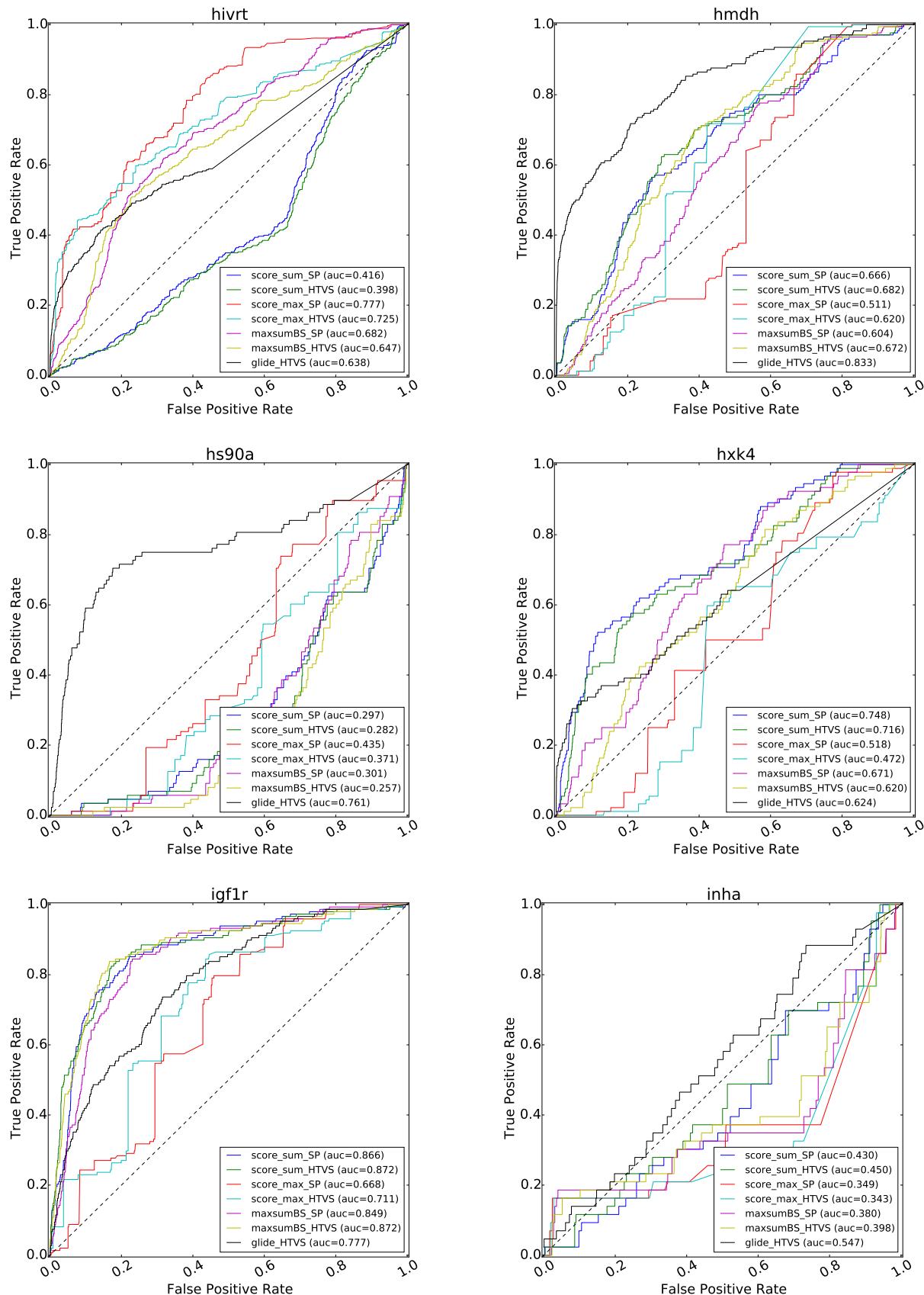


図 B.9 各手法の単体性能 ROC 曲線 (9)

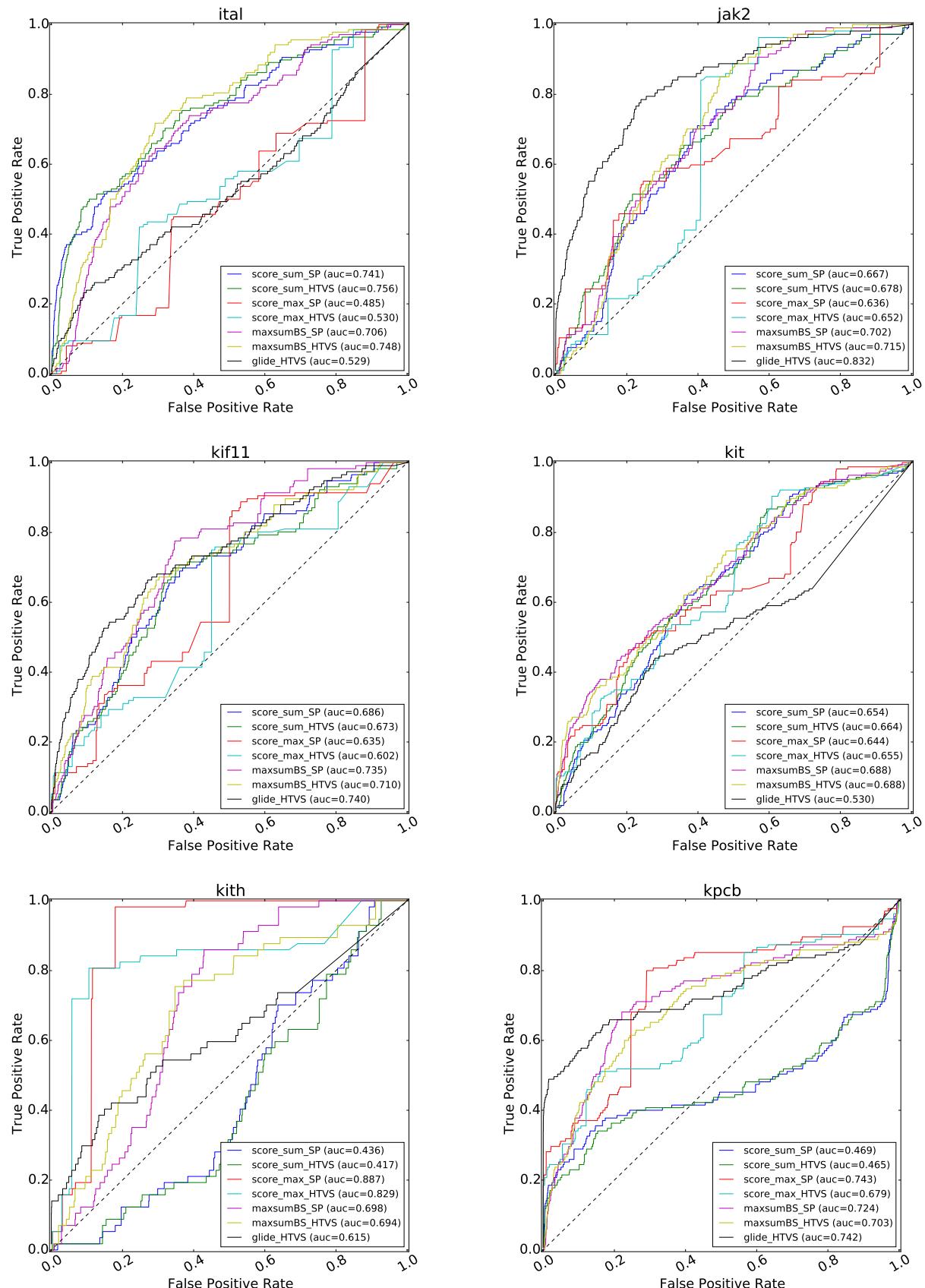


図 B.10 各手法の単体性能 ROC 曲線 (10)

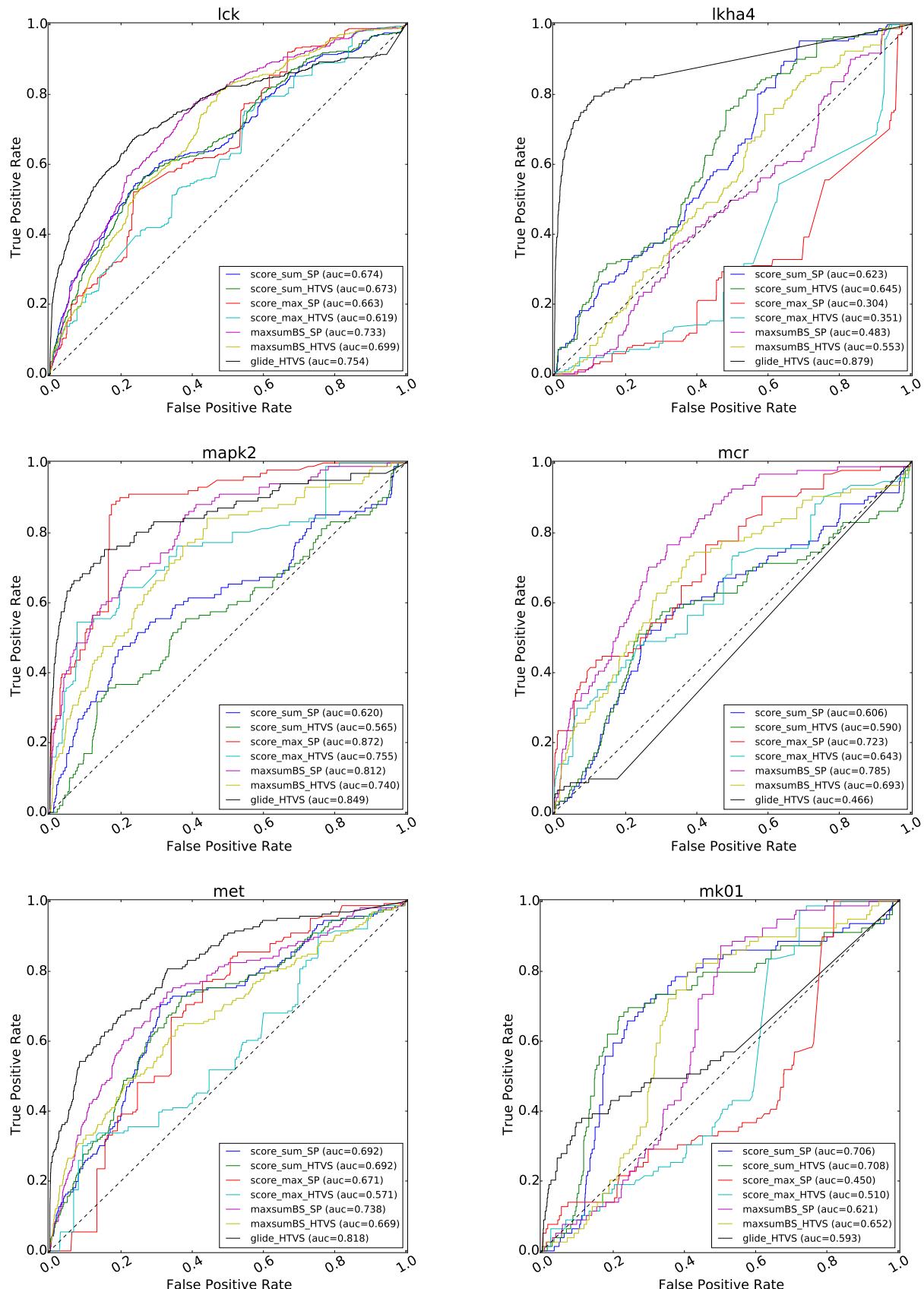


図 B.11 各手法の単体性能 ROC 曲線 (11)

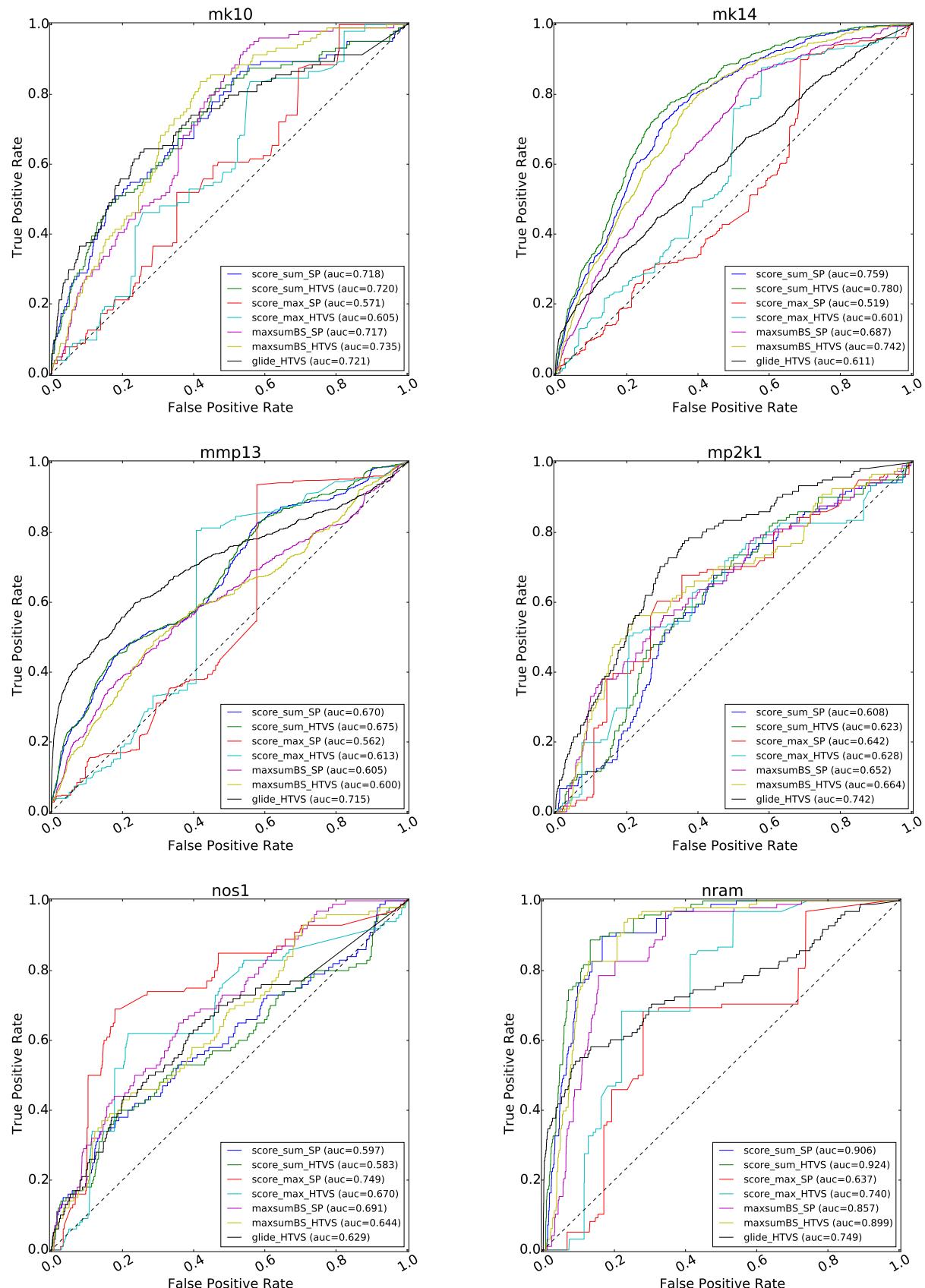


図 B.12 各手法の単体性能 ROC 曲線 (12)

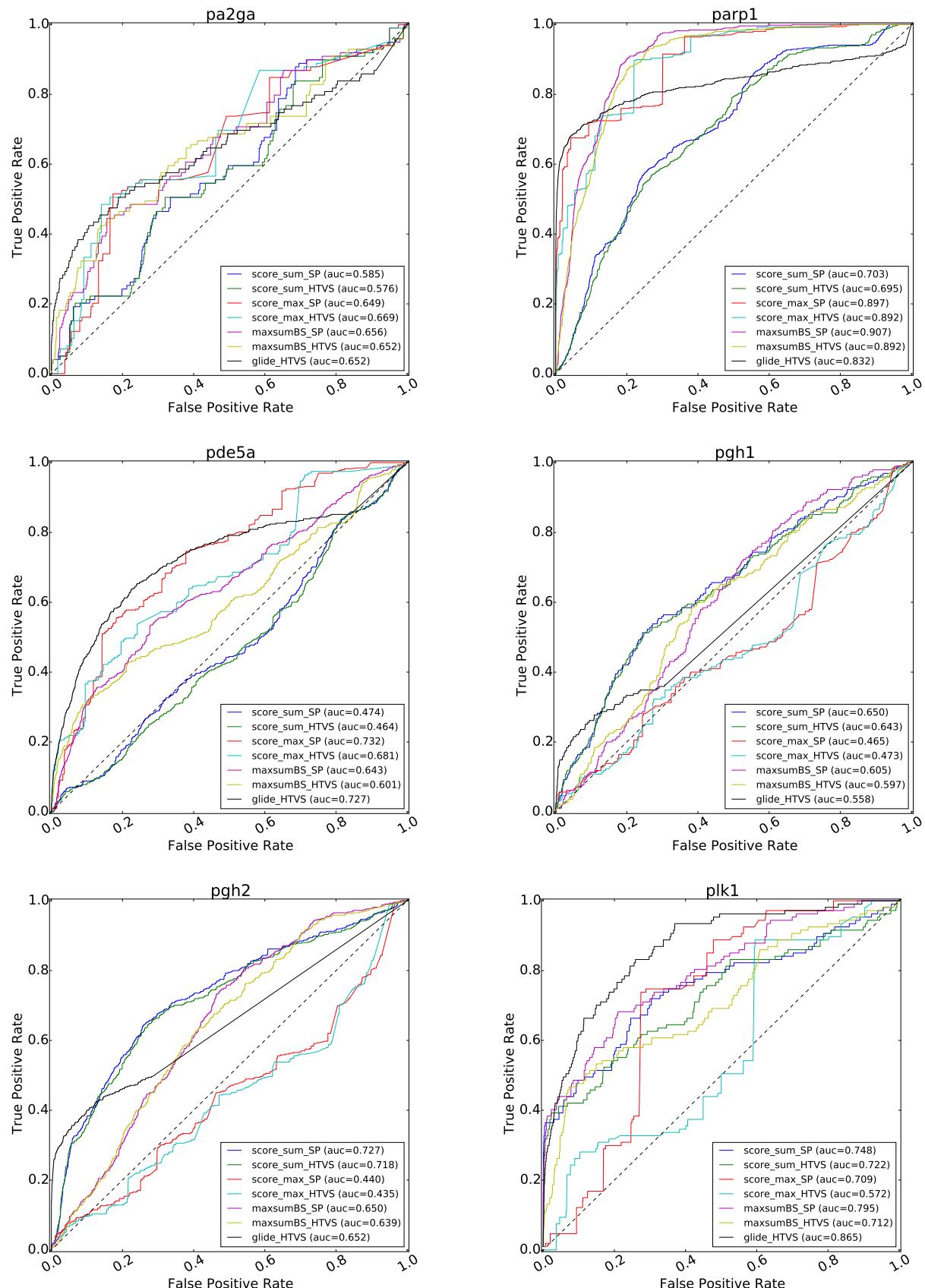


図 B.13 各手法の単体性能 ROC 曲線 (13)

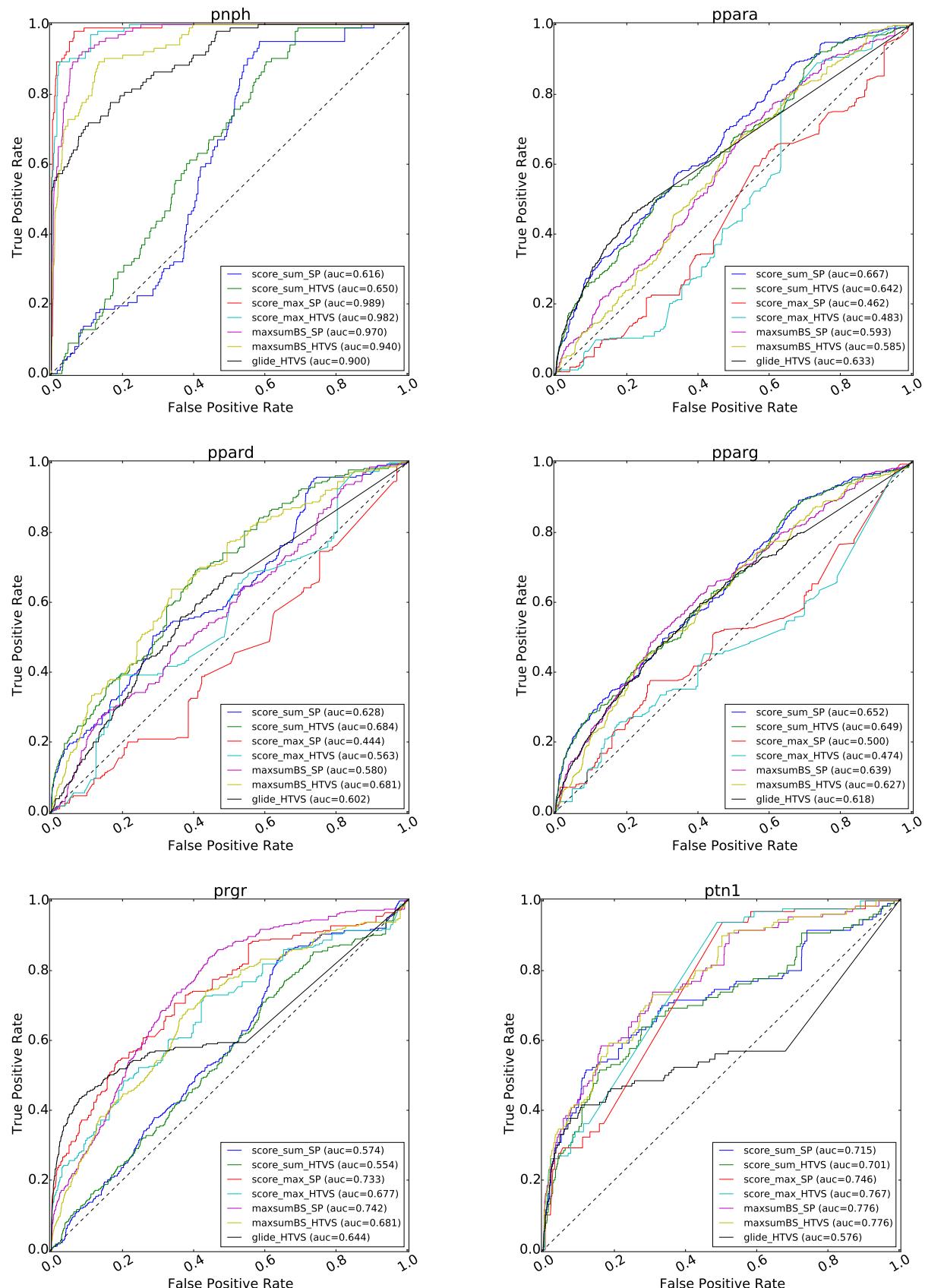


図 B.14 各手法の単体性能 ROC 曲線 (14)

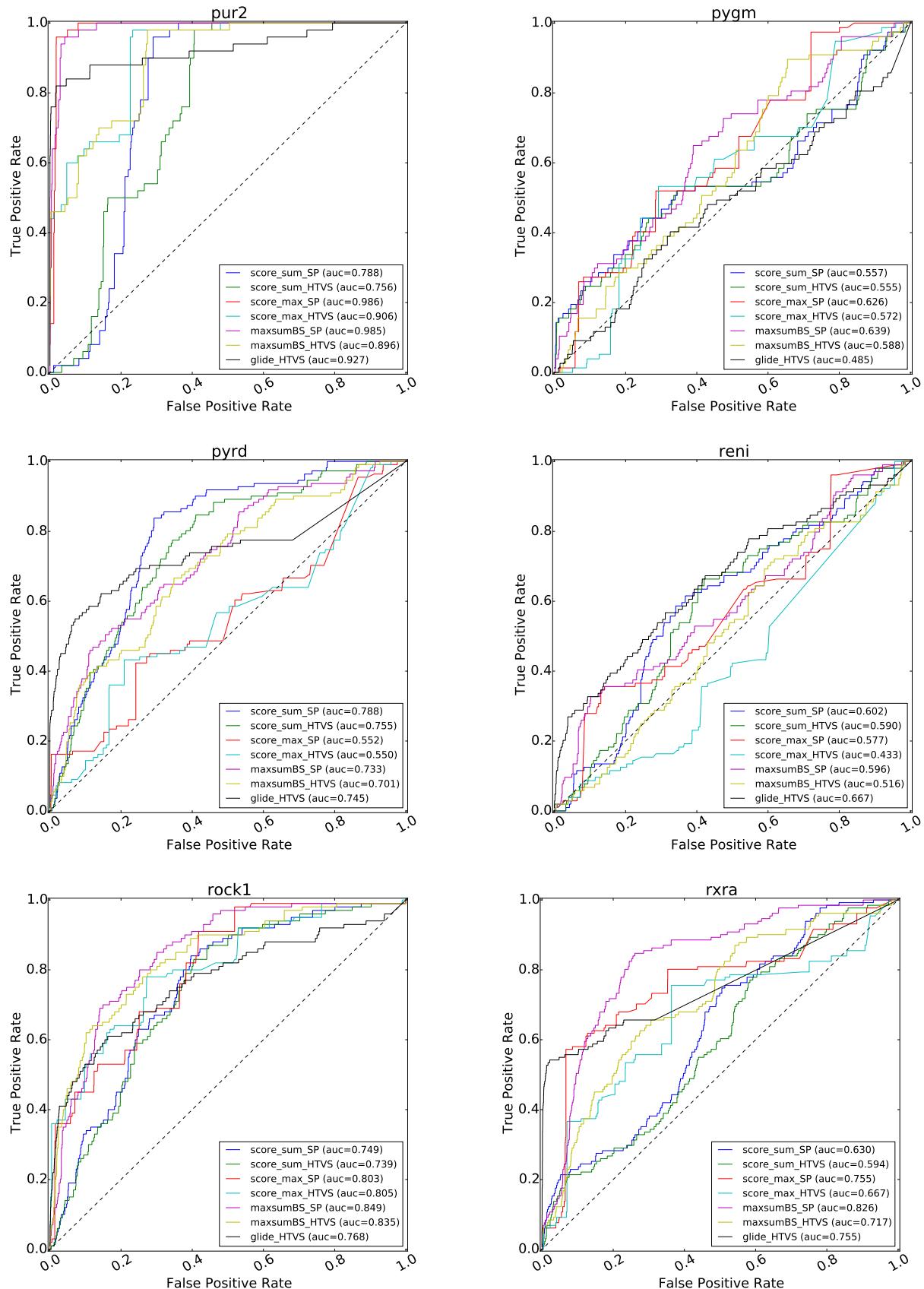


図 B.15 各手法の単体性能 ROC 曲線 (15)

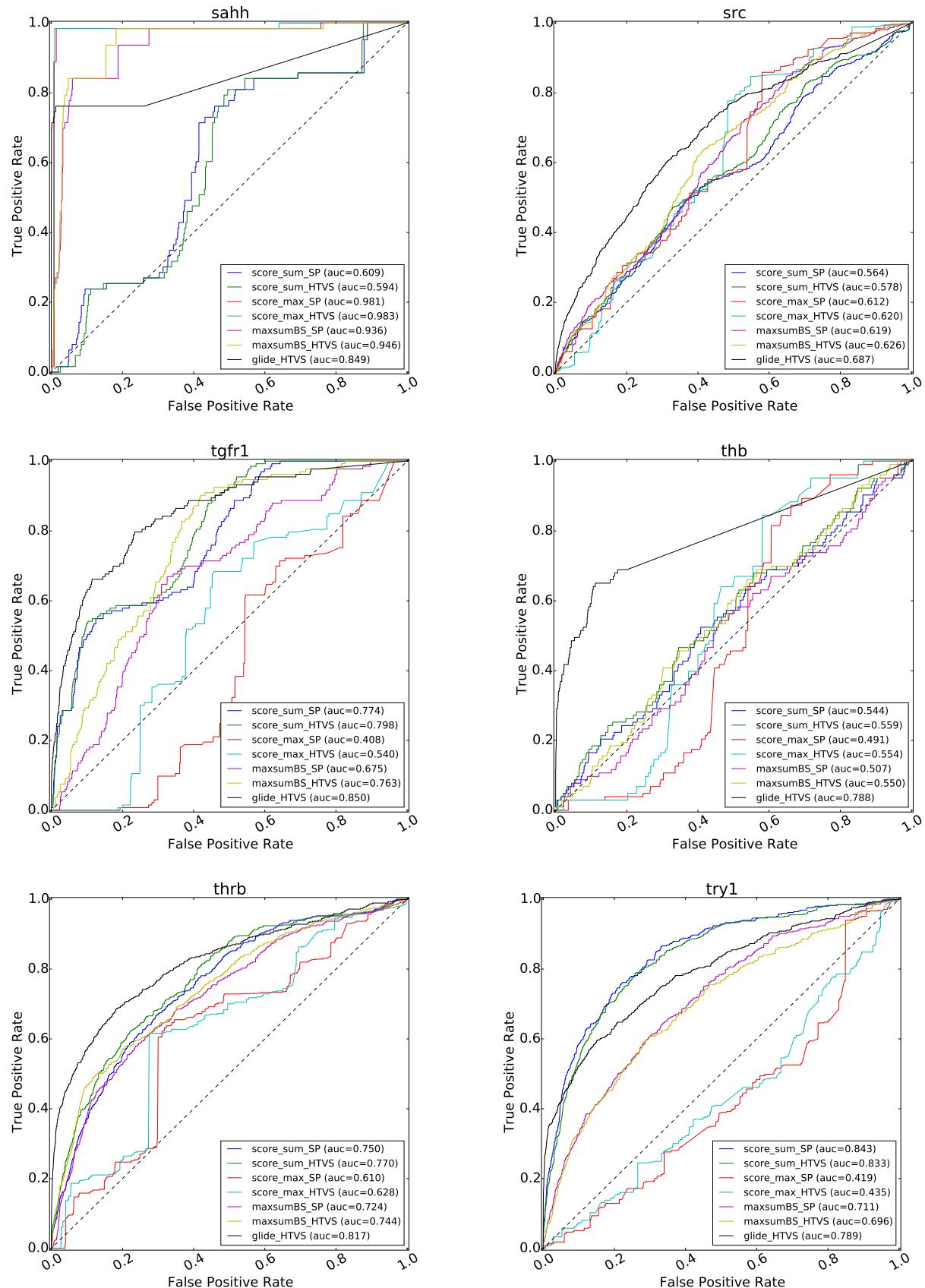


図 B.16 各手法の単体性能 ROC 曲線 (16)

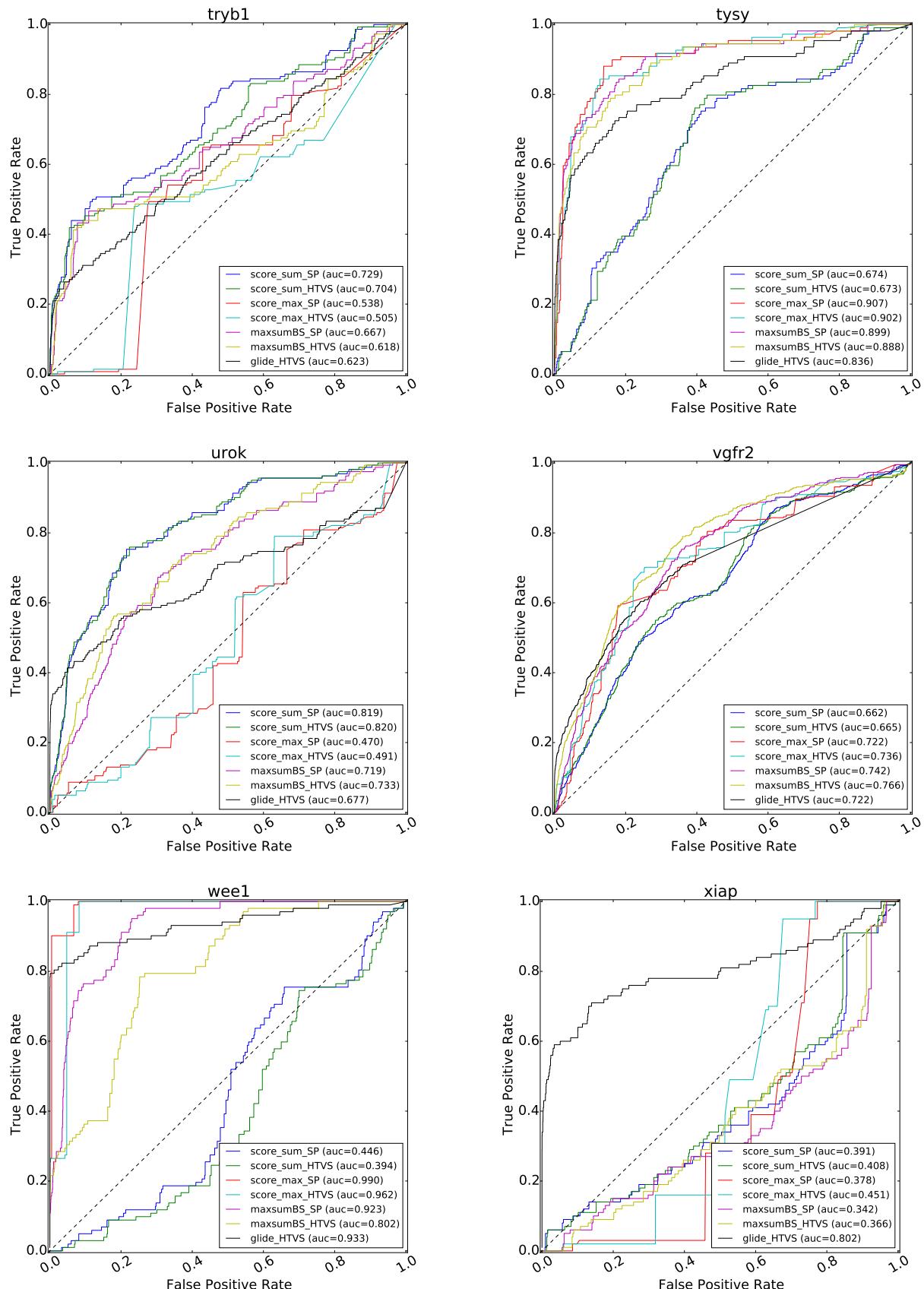


図 B.17 各手法の単体性能 ROC 曲線 (17)

表 B.1 各手法単体性能 AUC-ROC 値(1)

ターゲット	提案手法						簡易ドッキング シミュレーション (Glide HTVS)	
	score_sum		score_max		maxsumBS			
	SP	HTVS	SP	HTVS	SP	HTVS		
aa2ar	0.563	0.574	0.656	0.586	0.665	0.620	0.729	
abl1	0.646	0.662	0.545	0.501	0.637	0.637	0.794	
ace	0.471	0.485	0.458	0.478	0.446	0.457	0.611	
aces	0.361	0.346	0.638	0.619	0.518	0.491	0.458	
ada	0.581	0.549	0.706	0.670	0.738	0.698	0.734	
ada17	0.598	0.599	0.760	0.638	0.650	0.660	0.716	
adrb1	0.549	0.574	0.637	0.593	0.615	0.606	0.576	
adrb2	0.558	0.563	0.529	0.525	0.566	0.568	0.641	
akt1	0.748	0.687	0.804	0.736	0.824	0.773	0.539	
akt2	0.711	0.668	0.615	0.575	0.718	0.660	0.738	
aldr	0.503	0.510	0.798	0.787	0.719	0.714	0.675	
ampc	0.716	0.733	0.508	0.534	0.674	0.696	0.763	
andr	0.525	0.502	0.746	0.674	0.747	0.670	0.570	
aofb	0.553	0.552	0.590	0.588	0.614	0.601	0.657	
bace1	0.627	0.628	0.399	0.376	0.510	0.496	0.686	
braf	0.803	0.819	0.579	0.661	0.773	0.823	0.805	
cah2	0.653	0.647	0.441	0.426	0.555	0.545	0.494	
casp3	0.534	0.526	0.575	0.581	0.568	0.564	0.655	
cdk2	0.618	0.630	0.580	0.604	0.638	0.644	0.780	
comt	0.566	0.568	0.693	0.754	0.702	0.688	0.718	
cp2c9	0.629	0.639	0.487	0.506	0.588	0.599	0.547	
cp3a4	0.684	0.683	0.512	0.484	0.650	0.630	0.685	
csf1r	0.739	0.729	0.561	0.533	0.688	0.658	0.662	
cxcr4	0.691	0.674	0.778	0.759	0.770	0.760	0.872	
def	0.238	0.220	0.983	0.993	0.379	0.386	0.733	
dhi1	0.660	0.646	0.292	0.264	0.528	0.502	0.650	

表 B.2 各手法単体性能 AUC-ROC 値 (2)

ターゲット	提案手法						簡易ドッキング シミュレーション (Glide HTVS)	
	score_sum		score_max		maxsumBS			
	SP	HTVS	SP	HTVS	SP	HTVS		
dpp4	0.713	0.718	0.785	0.846	0.757	0.777	0.710	
drd3	0.502	0.478	0.515	0.464	0.522	0.491	0.608	
dyr	0.597	0.600	0.885	0.864	0.846	0.843	0.707	
egfr	0.537	0.536	0.687	0.618	0.655	0.592	0.757	
esr1	0.671	0.683	0.785	0.787	0.821	0.815	0.666	
esr2	0.646	0.633	0.797	0.793	0.820	0.810	0.759	
fa10	0.863	0.856	0.413	0.368	0.720	0.684	0.797	
fa7	0.831	0.849	0.527	0.635	0.766	0.823	0.855	
fabp4	0.608	0.606	0.637	0.575	0.737	0.676	0.855	
fak1	0.598	0.541	0.872	0.745	0.856	0.744	0.817	
fgfr1	0.582	0.581	0.460	0.523	0.532	0.565	0.522	
flkb1a	0.526	0.557	0.520	0.600	0.577	0.641	0.757	
fnta	0.615	0.601	0.538	0.425	0.622	0.544	0.806	
fpps	0.868	0.737	0.998	0.985	0.993	0.965	0.843	
gcr	0.572	0.549	0.770	0.687	0.780	0.680	0.528	
glcm	0.393	0.394	0.694	0.710	0.529	0.533	0.520	
gria2	0.534	0.530	0.776	0.764	0.737	0.736	0.655	
grik1	0.569	0.569	0.758	0.765	0.760	0.761	0.800	
hdac2	0.369	0.380	0.469	0.493	0.366	0.388	0.594	
hdac8	0.312	0.285	0.844	0.869	0.330	0.296	0.613	
hivint	0.536	0.538	0.623	0.615	0.611	0.595	0.807	
hivpr	0.719	0.714	0.508	0.431	0.658	0.617	0.737	
hivr	0.416	0.398	0.777	0.725	0.682	0.647	0.638	
hmdh	0.666	0.682	0.511	0.620	0.604	0.672	0.833	
hs90a	0.297	0.282	0.435	0.371	0.301	0.257	0.761	
hxk4	0.748	0.716	0.518	0.472	0.671	0.620	0.624	

表 B.3 各手法単体性能 AUC-ROC 値(3)

ターゲット	提案手法						簡易ドッキング シミュレーション (Glide HTVS)	
	score_sum		score_max		maxsumBS			
	SP	HTVS	SP	HTVS	SP	HTVS		
igf1r	0.866	0.872	0.668	0.711	0.849	0.872	0.777	
inha	0.430	0.450	0.349	0.343	0.380	0.398	0.547	
ital	0.741	0.756	0.485	0.530	0.706	0.748	0.529	
jak2	0.667	0.678	0.636	0.652	0.702	0.715	0.832	
kif11	0.686	0.673	0.635	0.602	0.735	0.710	0.740	
kit	0.654	0.664	0.644	0.655	0.688	0.688	0.530	
kith	0.436	0.417	0.887	0.829	0.698	0.694	0.615	
kpcb	0.469	0.465	0.743	0.679	0.724	0.703	0.742	
lck	0.674	0.673	0.663	0.619	0.733	0.699	0.754	
lkha4	0.623	0.645	0.304	0.351	0.483	0.553	0.879	
mapk2	0.620	0.565	0.872	0.755	0.812	0.740	0.849	
mcr	0.606	0.590	0.723	0.643	0.785	0.693	0.466	
met	0.692	0.692	0.671	0.571	0.738	0.669	0.818	
mk01	0.706	0.708	0.450	0.510	0.621	0.652	0.593	
mk10	0.718	0.720	0.571	0.605	0.717	0.735	0.721	
mk14	0.759	0.780	0.519	0.601	0.687	0.742	0.611	
mmp13	0.670	0.675	0.562	0.613	0.605	0.600	0.715	
mp2k1	0.608	0.623	0.642	0.628	0.652	0.664	0.742	
nos1	0.597	0.583	0.749	0.670	0.691	0.644	0.629	
nram	0.906	0.924	0.637	0.740	0.857	0.899	0.749	
pa2ga	0.585	0.576	0.649	0.669	0.656	0.652	0.652	
parp1	0.703	0.695	0.897	0.892	0.907	0.892	0.832	
pde5a	0.474	0.464	0.732	0.681	0.643	0.601	0.727	
pgh1	0.650	0.643	0.465	0.473	0.605	0.597	0.558	
pgh2	0.727	0.718	0.440	0.435	0.650	0.639	0.652	
plk1	0.748	0.722	0.709	0.572	0.795	0.712	0.865	

表 B.4 各手法単体性能 AUC-ROC 値(4)

ターゲット	提案手法						簡易ドッキング シミュレーション (Glide HTVS)	
	score_sum		score_max		maxsumBS			
	SP	HTVS	SP	HTVS	SP	HTVS		
pnph	0.616	0.650	0.989	0.982	0.970	0.940	0.900	
ppara	0.667	0.642	0.462	0.483	0.593	0.585	0.633	
ppard	0.628	0.684	0.444	0.563	0.580	0.681	0.602	
pparg	0.652	0.649	0.500	0.474	0.639	0.627	0.618	
prgr	0.574	0.554	0.733	0.677	0.742	0.681	0.644	
ptn1	0.715	0.701	0.746	0.767	0.776	0.776	0.576	
pur2	0.788	0.756	0.986	0.906	0.985	0.896	0.927	
pygm	0.557	0.555	0.626	0.572	0.639	0.588	0.485	
pyrd	0.788	0.755	0.552	0.550	0.733	0.701	0.745	
reni	0.602	0.590	0.577	0.433	0.596	0.516	0.667	
rock1	0.749	0.739	0.803	0.805	0.849	0.835	0.768	
rxra	0.630	0.594	0.755	0.667	0.826	0.717	0.755	
sahh	0.609	0.594	0.981	0.983	0.936	0.946	0.849	
src	0.564	0.578	0.612	0.620	0.619	0.626	0.687	
tgfr1	0.774	0.798	0.408	0.540	0.675	0.763	0.850	
thb	0.544	0.559	0.491	0.554	0.507	0.550	0.788	
thrb	0.750	0.770	0.610	0.628	0.724	0.744	0.817	
try1	0.843	0.833	0.419	0.435	0.711	0.696	0.789	
tryb1	0.729	0.704	0.538	0.505	0.667	0.618	0.623	
tysy	0.674	0.673	0.907	0.902	0.899	0.888	0.836	
urok	0.819	0.820	0.470	0.491	0.719	0.733	0.677	
vgfr2	0.662	0.665	0.722	0.736	0.742	0.766	0.722	
wee1	0.446	0.394	0.990	0.962	0.923	0.802	0.933	
xiap	0.391	0.408	0.378	0.451	0.342	0.366	0.802	