

# COURBES DE PREDICTIVITE APPLIQUEES AU CRIBLAGE VIRTUEL

Charly Empereur-mot<sup>1</sup>, H el ene Guillemain<sup>1</sup>, Aur elien Latouche<sup>2</sup>, Jean-Fran ois Zagury<sup>1</sup>, Vivian Viallon<sup>3</sup> & Matthieu Montes<sup>1</sup>

<sup>1</sup> *Laboratoire G enomique Bioinformatique et Applications, EA 4627, Conservatoire National des Arts et M etiers, 292 rue Saint Martin, 75003 Paris, France ([charly.empereur@gmail.com](mailto:charly.empereur@gmail.com), [heleneguillemain@gmail.com](mailto:heleneguillemain@gmail.com), [zagury@cnam.fr](mailto:zagury@cnam.fr), [matthieu.montes@cnam.fr](mailto:matthieu.montes@cnam.fr))*

<sup>2</sup> *Equipe MSDMA, Laboratoire CEDRIC, EA 4629, Conservatoire National des Arts et M etiers, 292 rue Saint Martin, 75003 Paris, France ([aurelien.latouche@cnam.fr](mailto:aurelien.latouche@cnam.fr))*

<sup>3</sup> *UMRESTTE, IFSTTAR, Universit  Lyon 1, 8 avenue Rockefeller, 69373 Lyon, France ([vivian.viallon@univ-lyon1.fr](mailto:vivian.viallon@univ-lyon1.fr))*

**R esum .** Cette  tude a pour objectif de transf erer au domaine du criblage virtuel (VS) les courbes de pr edictivit  (PC), une m etricque de r ef erence dans le domaine de l' pid emiologie. De mani ere similaire, nous avons utilis  des mod eles de r egression logistique pour calculer des probabilit s d'activit  en fonction des scores d'affinit  obtenus lors d'exp eriences de criblage virtuel. Les PC sont des outils graphiques et intuitifs permettant de comparer la capacit  pr edictive des m ethodes de criblage. Comme les courbes de ROC, les PC d ependent de la distribution des scores et fournissent une  chelle commune pour comparer les m ethodes. Contrairement aux courbes de ROC, les PC prennent en compte la dispersion des scores. Cette propri t  permet de quantifier la capacit  pr edictive des m ethodes de criblage sur un sous-ensemble des mol cules du jeu de donn es et fait des courbes de pr edictiv  un outil int ressant pour l' valuation de la reconnaissance pr coce. A cette fin, nous introduisons l'usage du gain total partiel (pTG) pour quantifier la reconnaissance pr coce des compos s actifs attribuable aux variations des scores. En plus de leur utilit  pour la comparaison des m ethodes de VS, les PC peuvent  tre utilis es pour d finir des seuils de score optimaux pour la s lection de compos s   tester exp erimentalement dans les processus de conception de m dicaments. Pour d montrer cet usage, nous avons r alis  des criblages sur les jeux de donn es de la DUD avec trois m ethodes diff erentes et compar  les analyses possibles par les courbes de ROC et PC. Les courbes de pr edictivit , compl mentaires aux m etricques de r ef erence du domaine, compl tent efficacement les outils d'analyse des r esultats de criblage virtuel.

**Mots-cl s.** Criblage virtuel, courbes de pr edictivit , reconnaissance pr coce, gain total partiel

**Abstract.** In the present work, we aimed to transfer to the field of virtual screening (VS) the predictiveness curve (PC), a metric that has been advocated in clinical epidemiology. Similarly, we used logistic regression models to calculate activity probabilities related to the scores compounds obtained in VS experiments. The PC then provides an intuitive and graphical tool to compare the predictive power of VS methods. As receiver operating characteristic (ROC) curves, PC are functions of the scores distribution and provide a common scale to benchmark the methods. Contrary to ROC curves, score dispersion is well described by PC. This property allows to quantify the predictive performance of VS methods on a fraction of a molecular data-set and makes PC an interesting tool to address the early recognition problem. To this last end, we introduced the use of the partial total gain (pTG) to quantify early recognition of active compounds attributed to VS scores variations. Additionally to its usefulness in benchmarking VS methods, PC can be used to define optimal score thresholds for the selection of compounds to be tested experimentally in a drug discovery program. To demonstrate this usage, we performed VS on the DUD data-set using three different methods and analyzed the insights brought by ROC and PC. We believe PC efficiently complete the set of tools available for the analysis of VS results.

**Keywords.** Virtual screening, predictiveness curves, early recognition, partial total gain

## 1 Contexte

Le criblage virtuel (VS) permet de calculer des scores d'affinité de liaison entre de vastes bibliothèques de molécules et une cible d'intérêt thérapeutique. Il est utilisé extensivement dans le processus de conception de médicaments afin de réduire le nombre de molécules à tester expérimentalement<sup>1</sup>. Son objectif est d'enrichir un sous-ensemble du jeu de données moléculaire en composés actifs et d'éliminer les composés supposés inactifs<sup>2</sup>. En criblage virtuel, le problème principal est de choisir un seuil de score optimal pour la sélection d'un petit sous-ensemble de molécules à tester expérimentalement. Ce score seuil est généralement estimé empiriquement par l'analyse de résultats de criblage rétrospectifs, sur des jeux de données comparatifs, incluant des composés actifs et inactifs à propriétés similaires.

## 2 Limites des métriques appliquées à l'analyse des résultats de criblage virtuel

Ainsi, plusieurs métriques sont appliquées à l'analyse des résultats de criblage : les facteurs d'enrichissements (EF), courbes de ROC<sup>2</sup> (Receiver Operating Characteristic), l'aire sous la courbe de ROC<sup>2</sup> (ROC AUC), l'aire partielle sous la courbe de ROC<sup>3</sup> (pAUC), les métriques BEDROC<sup>4</sup> (Boltzmann-Enhanced Discrimination of ROC), RIE<sup>5</sup> (Robust Initial Enhancement), et plusieurs autres, ayant chacune leurs avantages et inconvénients. Les courbes de ROC et EF font autorité dans le domaine.

Une propriété attractive de la courbe de ROC est de fournir une échelle commune pour comparer les résultats de criblage. L'AUC peut être calculée à partir de la courbe et quantifie élégamment les performances globales d'une expérience de VS. Cependant, les courbes de ROC et leur AUC comportent deux désavantages. Premièrement, les méthodes de criblage sont utilisées pour prioriser un petit nombre de molécules tandis que les courbes de ROC et l'AUC capturent la capacité d'une méthode de criblage à ordonner une base de données de molécules dans son entièreté<sup>4</sup>. Deuxièmement, ces deux métriques sont exclusivement basées sur le rang des composés et ne prennent pas en compte les scores d'affinité calculés<sup>2</sup>. Ainsi, les scores seuils de sélection calculés à partir des courbes de ROC ne prennent pas en compte correctement ni les différences de scores, ni les intervalles entre scores, potentiellement représentatifs d'importantes différences pour la détection de l'activité des molécules criblées.

Concernant la reconnaissance précoce, les EF sont plus fiables puisqu'ils se concentrent principalement sur une mesure de classification : le taux de vrais positifs. Cependant, les EF sont critiqués car ils dépendent d'un seuil de sélection, défini par l'utilisateur, permettant d'étudier une fraction du jeu de données, et ne prennent pas en compte la précision du classement sur la fraction étudiée<sup>4</sup>. De plus, la valeur maximale des EF est fortement dépendante du ratio de composés actifs dans le jeu de données<sup>4</sup>.

Il est important de définir notre position concernant l'utilisation des métriques appliquées à l'analyse des résultats de criblage virtuel. Les courbes de ROC sont utiles puisqu'elles répondent directement à la question : « Est-ce que cette méthode de criblage est supérieure à une sélection aléatoire des composés ? ». Cette question est importante, mais dans le domaine du VS, il est plus important d'évaluer si une méthode peut nous aider à sélectionner des composés actifs (« leads »), permettant de poursuivre des phases d'optimisation dans le processus de développement de médicaments.

## 3 Courbes de prédictivité appliquées à l'analyse des résultats de criblage virtuel

En épidémiologie, les bio-marqueurs sont utilisés pour formuler des diagnostics, pronostiques et estimer les risques d'occurrence d'une maladie. La littérature décrit l'utilisation des courbes de prédictivité pour évaluer la performance des bio-marqueurs, leur adéquation statistique et estimer l'utilité des modèles cliniques appliqués à une population. Notre étude démontre que les courbes de prédictivité peuvent être appliquées à l'analyse des résultats de criblage virtuel, pour améliorer l'évaluation des méthodes et la sélection des composés à tester expérimentalement.

L'approche choisie pour notre étude consiste à modéliser la capacité des scores d'affinité à expliquer l'activité des composés dans une expérience de criblage virtuel. Nous avons utilisé des modèles de régression logistique pour calculer une probabilité d'activité pour le score de chaque composé, séparément pour différentes cibles thérapeutiques et méthodes de criblage. Les PC sont ensuite construites comme une fonction de distribution cumulative des probabilités d'activité des composés.

Les courbes de prédictivité exposent donc les probabilités d'activité en fonction de la dispersion des scores et des composés actifs dans un jeu de données, et assurent une échelle commune pour la comparaison des méthodes de criblage. La particularité principale des PC est leur capacité à décrire la dispersion des scores, ce qui fournit des informations complémentaires à l'utilisation des courbes de ROC. Le gain total partiel<sup>6</sup> (pTG), récemment proposé comme mesure partielle d'une PC, permet de quantifier l'association entre scores d'affinité et activité des composés, sur une fraction d'un jeu de données, en conservant une échelle commune pour la comparaison des méthodes. Les PC et leurs pTG peuvent être utilisés pour 1. quantifier et comparer la capacité prédictive des méthodes de criblage au-delà d'un quantile de scores et 2. définir des seuils de score optimaux pour la sélection de composés à tester expérimentalement dans les processus de conception de médicaments. Notre étude démontre que les courbes de prédictivité sont particulièrement adaptées pour évaluer la reconnaissance précoce et permettent une appréciation intuitive des performances des méthodes de criblage virtuel.

#### **4 Données d'essai et méthodes de criblage utilisées**

Nous avons réalisé des criblages virtuels sur les données d'essai de la DUD (Directory of Useful Decoys), composée de 40 protéines cibles, un total de 2950 composés actifs, pour chacun desquels existent en moyenne 36 composés « leurres » (« decoys »), supposés inactifs, pour un total de 98266 composés dans la base de données. Nous avons utilisé trois outils de criblage virtuel : Surflex-dock, ICM et Autodock Vina, afin de comparer les analyses possibles par les courbes de ROC et PC.

#### **5 Analyse des graphes de prédictivité**

Les graphes de prédictivité permettent une visualisation intuitive de plusieurs métriques : le taux de vrais positifs (TPF), le taux de faux positifs (FPF), la valeur prédictive positive (PPV) et la valeur prédictive négative (NPV) (Fig. 1). La visualisation des PC permet de quantifier l'association entre scores de criblage et activité des composés, particulièrement efficace pour l'inclusion de critères de décision et guider la détermination de seuils de probabilité d'activité pertinents pour la sélection d'un sous-ensemble de composés. Un seuil de probabilité d'activité peut ensuite être converti en seuil de score à partir des données initiales.

Sur les graphes de prédictivité, la PPV associée au quantile sélectionné par le critère : « probabilité d'activité  $\geq$  seuil d'activité » est la proportion du cadre bleu situé sous la courbe de prédictivité. Pour un modèle nul, qui détecte l'absence d'association entre scores de criblage et activité des composés, la courbe de prédictivité correspondrait à une ligne horizontale au niveau de la prévalence de l'activité des composés dans le jeu de données (Fig. 1, ligne discontinue).

#### **6 Exemple d'application**

Nous choisissons ici l'exemple d'expériences de criblage rétrospectives réalisées sur la protéine cible poly(ADP-ribose) polymérase (PARP). Pour chaque expérience, les courbes de ROC suggèrent que la reconnaissance précoce des composés actifs est accomplie (Fig. 2).

Les ROC AUC obtenues sont proches (AUC - Surflex-dock : 0.738, ICM : 0.755, Autodock Vina : 0.728). Les AUC partielles à 1 % du jeu de données permettent une première discrimination des performances des méthodes (pAUC1 % - Surflex-dock : 0.163, ICM : 0.009, Autodock Vina :

0.059). L'analyse des courbes de ROC suggère que Surflex-dock est la méthode de criblage la plus efficace pour ordonner notre jeu de données dans les conditions de cette expérience.

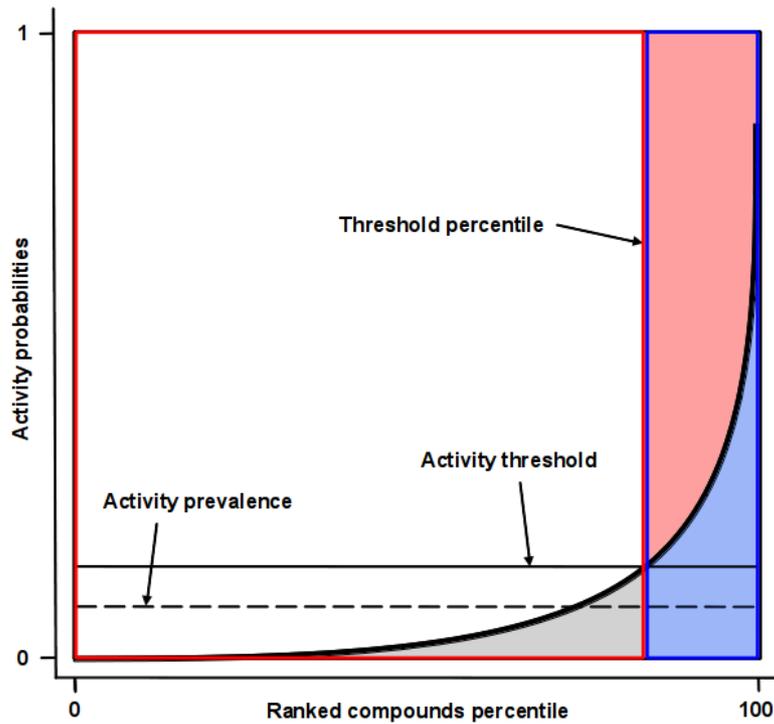
La visualisation des graphes de prédictivité suggère que Surflex-dock dépasse les performances d'ICM et Autodock Vina pour la reconnaissance précoce, en termes de calcul des scores d'affinité et de classement des composés actifs. Le gain total partiel à 1 % du jeu de données permet de quantifier cette performance (pTG1 % - Surflex-dock : 0.356, ICM : 0.181, Autodock Vina : 0.151).

Le graphe de prédictivité nous permet de visualiser une nette variation des probabilités d'activité pour Surflex-dock (Fig. 2, ligne noire discontinue). Cette variation, observée à 1.08 % du jeu de données ordonné, dénote une forte association des scores obtenus avec l'activité des composés criblés (pTG : 0.332, EF : 17.3, score de Surflex-dock : 7.12).

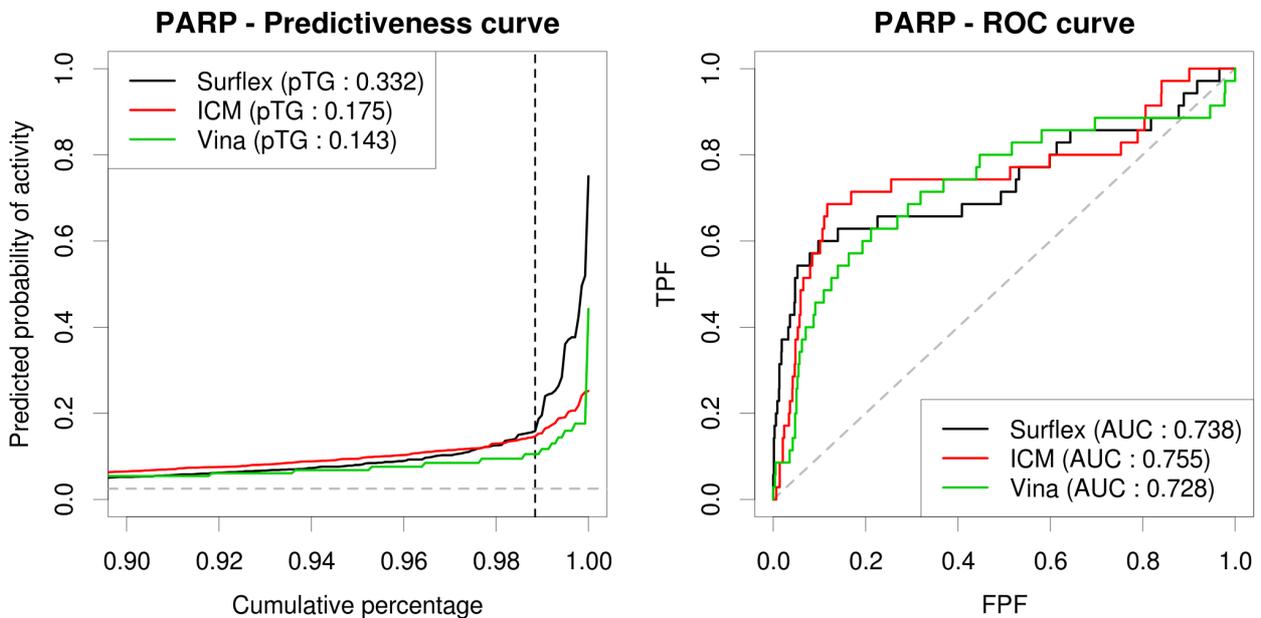
Lorsque des études prospectives permettent d'observer de telles variations des probabilités d'activité sur les graphes de prédictivité, il est intéressant de considérer les seuils de score de criblage associés comme optimaux. Dans ces cas, les seuils de score optimaux pourront être considérés comme reproductibles et utilisés pour la sélection de sous-ensemble enrichis en actifs dans des expériences de criblage virtuel prospectives.

## Bibliographie

- [1] Alvarez, J. C. (2004), High-Throughput Docking as a Source of Novel Drug Leads, *Curr. Opin. Chem. Biol.*, 8, 365–370.
- [2] Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.; Bertrand, H. (2005), Virtual Screening Workflow Development Guided by the “Receiver Operating Characteristic” Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4, *J. Med. Chem.*, 7, 2534–2547.
- [3] McClish, D. K. (1989), Analyzing a Portion of the ROC Curve, *Med. Decis. Mak.*, 9, 190–195.
- [4] Truchon, J.; Bayly, C. (2007), Evaluating Virtual Screening Methods: Good and Bad Metrics for the “early Recognition” Problem, *J. Chem. Inf.*, 47, 488-508.
- [5] Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. (2001), Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches, *J. Chem. Inf. Model.*, 41, 1395-1406.
- [6] Sachs, M. C.; Zhou, X. H. (2013), Partial Summary Measures of the Predictiveness Curve, *Biometrical J.*, 55, 589–602.



**Figure 1.** Diagramme schématisant les relations entre métriques de performance et courbe de prédictivité. Visualiser la PC permet de déterminer intuitivement des seuils de probabilité d'activité optimaux. Les métriques de performances sont visuellement représentées par la courbe. Valeur prédictive positive (PPV) : aire bleue/cadre bleu. Valeur prédictive négative (NPV) : aire blanche/cadre rouge. Taux de vrais positifs (TPF) : aire bleue/aire sous la ligne discontinue. Taux de faux positifs (FPF) : aire rouge/(1 - aire sous la ligne discontinue).



**Figure 2.** Graphes de prédictivité et ROC pour la protéine cible poly(ADP-ribose) polymérase (PARP) obtenus avec les méthodes Surflex-dock, ICM et Autodock Vina. Les lignes grises discontinues indiquent respectivement les niveaux de la prévalence des composés actifs et du tirage aléatoire de composés dans le jeu de données. La ligne noire discontinue montre un point de variation nette des probabilités d'activité associées aux scores obtenus par Surflex-dock. Le gain total partiel a été calculé après ce seuil pour chaque méthode de criblage.