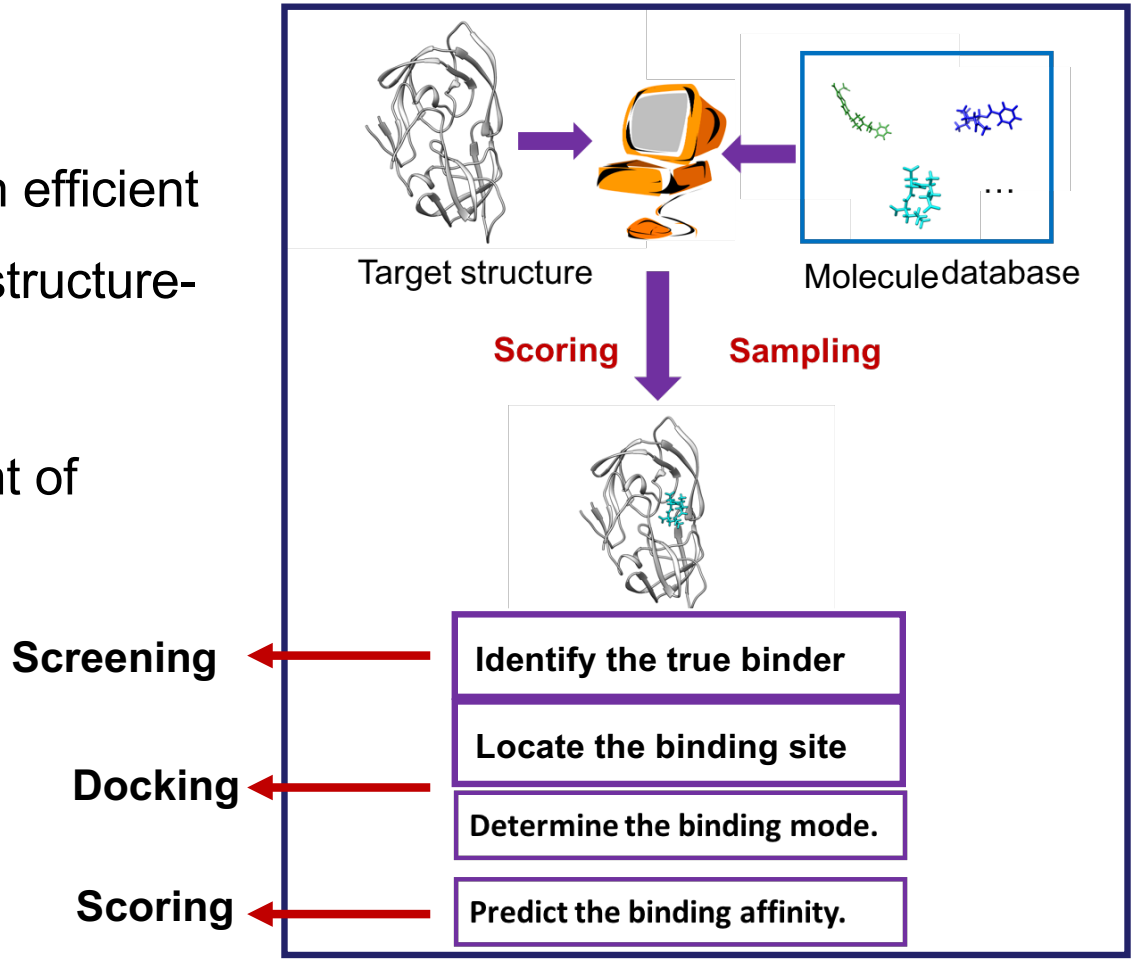
Improving Scoring-Docking-Screening Powers of Protein–Ligand Scoring Functions using Random Forest

Yingkai Zhang
Department of Chemistry, New York University

- Protein-ligand docking is an efficient computational approach in structure-based drug design.
- The most critical component of docking is scoring function.



Scoring methods

A fast and simplified estimation of binding energy

$$P + L \underset{k_d}{\overset{k_a}{\rightleftharpoons}} PL$$

$$K_a = K_d^{-1} = \frac{[PL]}{[P][L]}$$

Binding free energy

$$\Delta G_{bind} = -RT \ln K_a = RT \ln K_d$$

1 nm inhibitor: the free energy of binding = $0.5961 * \log(10^{-9}) = -12.4$ kcal/mol. $pK_d = 9$

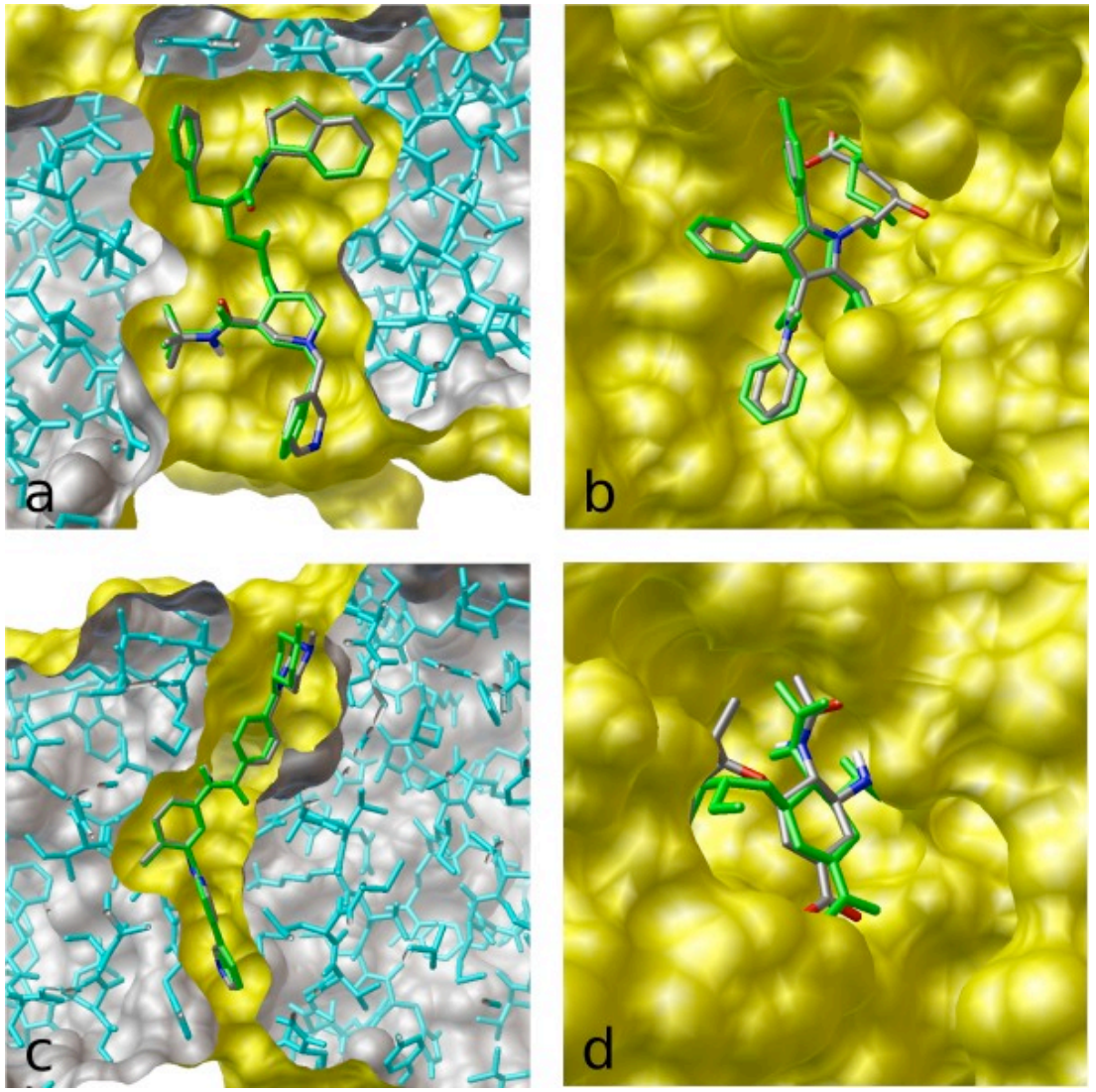
1 um inhibitor: the free energy of binding = $0.5961 * \log(10^{-6}) = -8.2$ kcal/mol. $pK_d = 6$

Classification of scoring functions

- Force Field-Based Scoring Function
  - Using non-bonded interaction terms from classical force field
  - Sometimes including solvation terms by GB/SA or PB/SA
- Empirical Scoring Function
  - Sum of several physical meaningful terms
  - Coefficients are derived from the regression analysis on experimental data
- Knowledge-Based Scoring Function
  - Statistical potential by using probability of finding atom pairs at a given distance between P and L
  - Require large number of terms
- Descriptor-Based Scoring Function
  - A pool of descriptors related to protein-ligand interaction
  - Machine learning algorithm to build the model

Liu, J.; Wang, R. *J. Chem. Inf. Model.* **2015,** 55, 475-482

AUTODOCK VINA



O. Trott, 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 (2010) 455-461

AutoDock Vina

- Gauss$_1$, Gauss$_2$, Repulsion, Hydrophobic, HBond, N$_{rot}$
- First five based on surface distance

$$d_{ij} = r_{ij} - R_{t_i} - R_{t_j}$$

$$c_{\text{inter}} = \sum_{i}^{\text{ligand}} \sum_{j}^{\text{protein}} (\omega_1 \text{gauss}_1(d_{ij}) + \omega_2 \text{gauss}_2(d_{ij}) + \omega_3 \text{Repulsion}(d_{ij})) + \sum_{i,i\in\text{HP}}^{\text{ligand}} \sum_{j,j\in\text{HP}}^{\text{protein}} \omega_4 \text{Hydrophogic}(d_{ij}) + \sum_{i,i\in\text{HB}}^{\text{ligand}} \sum_{j,j\in\text{HB}}^{\text{protein}} \omega_5 \text{HBond}(d_{ij})$$

$$g(c_{\text{inter}}) = \frac{c_{\text{inter}}}{1 + \omega N_{rot}}$$

$$\text{pK}_d\text{(Vina)} = -0.73349 * g(c_{\text{inter}})$$

Weight	Term
-0.0356	gauss$_1$ (ω_1)
-0.00516	gauss$_2$ (ω_2)
0.840	Repulsion (ω_3)
-0.0351	Hydrophobic (ω_4)
-0.587	Hydrogen bonding (ω_5)
0.0585	N$_{rot}$ (ω)

$$\text{gauss}_1(d) = e^{-(d/0.5)^2}$$

$$\text{gauss}_2(d) = e^{-((d-3)/2))^2}$$

$$\text{repulsion}(d) = \begin{cases} d^2 & d < 0 \\ 0 & d \geq 0 \end{cases}$$

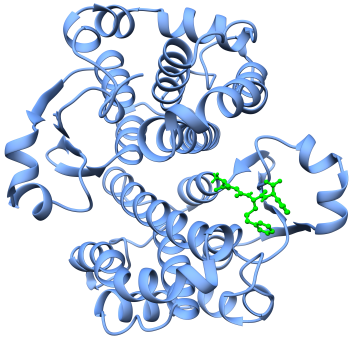
$$\text{Hydrophobic}(d) = \begin{cases} 1.0 & d < 0.5 \\ 1.5 - d & 0.5 \leq d \leq 1.5 \\ 0.0 & d > 1.5 \end{cases}$$

$$\text{HBond}(d) = \begin{cases} 1.0 & d < -0.7 \\ d/(-0.7) & -0.7 \leq d \leq 0 \\ 0.0 & d > 0 \end{cases}$$

Trott, O.; Olson, A.J.; *J. Comput. Chem.* **2010,** 31, 455-461

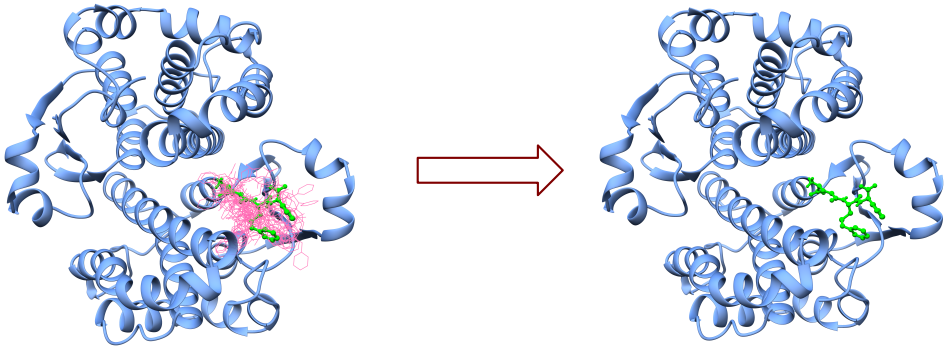
Scoring Function is the key in Protein-Ligand docking applications

- Binding affinity prediction

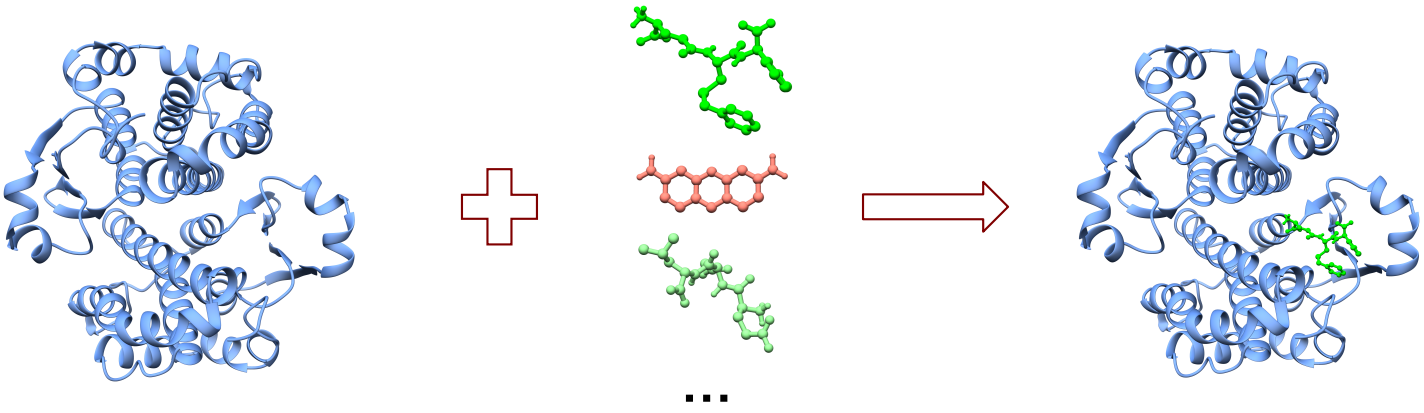


pK_d

- Binding mode identification



- Virtual screening



Evaluation Metrics of Scoring Functions

Comparative Assessment of Scoring Function (CASF) benchmark

Scoring power (binding affinity prediction)

- Linear correlation between predicted binding affinity and experimental binding affinity

Docking power (binding mode identification)

- Success rate of identifying the native binding mode among computer generated decoys

Screening power (Virtual screening)

- Success rate of Identifying the true binders to a given target protein among a pool of random molecules

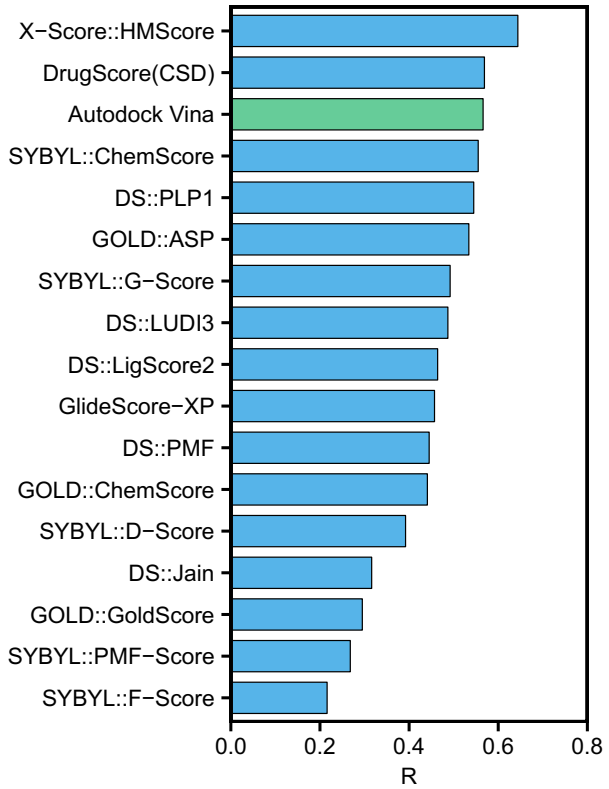
- CASF-2007: Scoring and docking powers
- CASF-2013: Scoring, docking and screening powers

Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R.; *J. Chem. Inf. Model.* **2009,** 49, 1079-1093
Li, Y.; Han, L.; Liu, Z.; Wang, R.; *J. Chem. Inf. Model.* **2014,** 54, 1717-1736

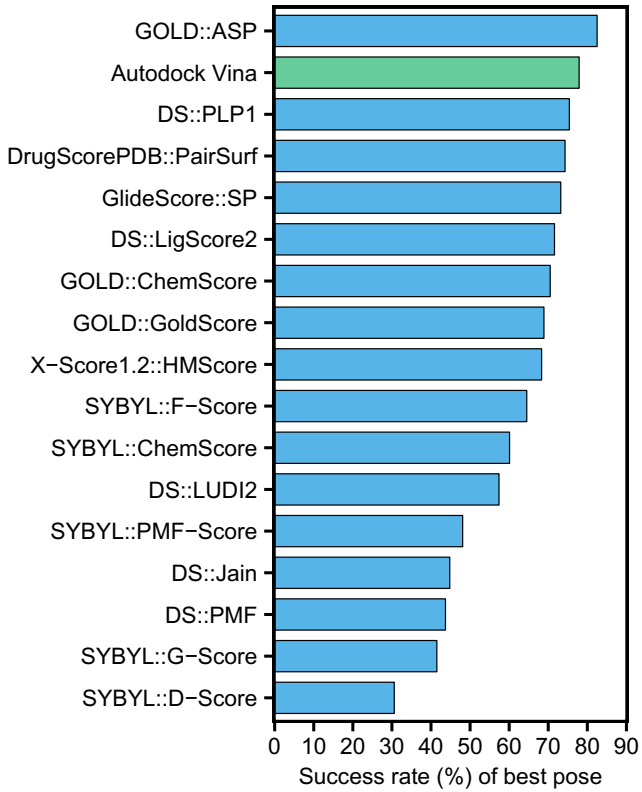
Scoring power is less satisfactory than docking/screening power

16 Scoring functions and Autodock Vina are evaluated in CASF-2007

? Scoring power
0.216 to 0.644
Autodock Vina: 0.566



✔☐ Docking power
30.6% to 82.5%
Autodock Vina: 77.9%



Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R.; *J. Chem. Inf. Model.* **2009,** 49, 1079-1093

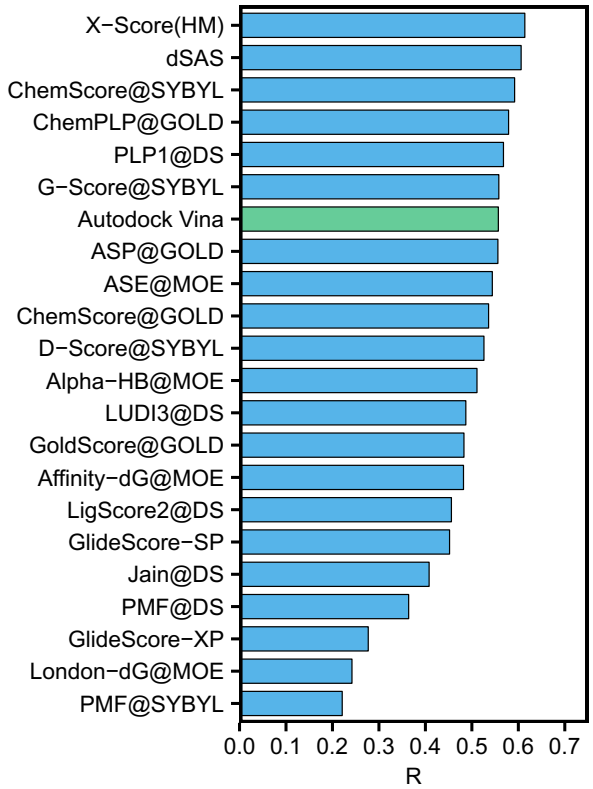
Scoring power is less satisfactory than docking/screening power

20 Scoring functions and Autodock Vina are evaluated in CASF-2013

? Scoring power (R)

0.221 to 0.614

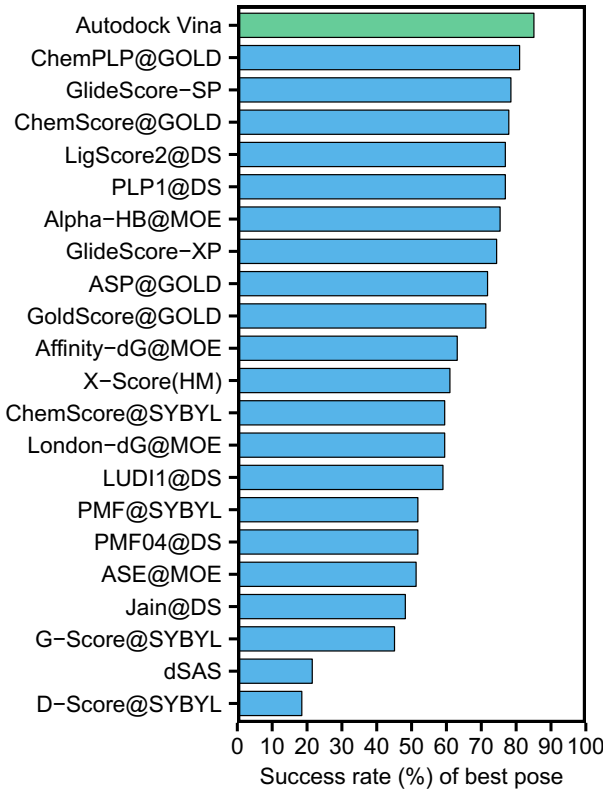
Autodock Vina: 0.557



✔ Docking power

18.5% to 85.1%

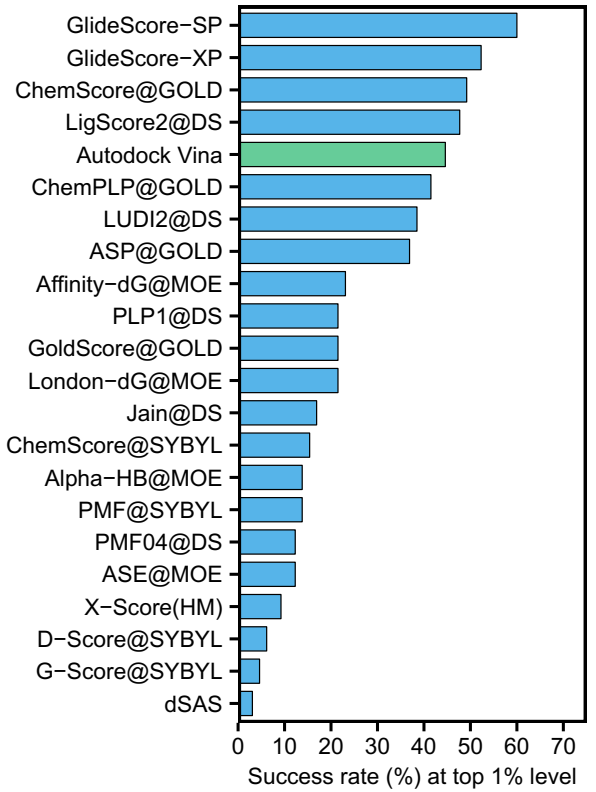
Autodock Vina: 85.1%



✔ Screening power

3.08% to 60.0%

Autodock Vina: 44.6%



Li, Y.; Han, L.; Liu, Z.; Wang, R.; *J. Chem. Inf. Model.* **2014,** 54, 1717-1736

RFbScores Achieve Excellent Scoring Power

Random Forest-based Scoring Function (RFbScore)

•Superior performance in predicting experimental protein-ligand binding affinity

CASF-2007

function	scoring power (R)
RF-Score::Elem-v2	0.803
RF-IChem	0.791
$SCFscore^{RF}$	0.779
$X\text{-}Score^{HM}$	0.644

CASF-2013

function	scoring power (R)
RF-Score::VinaElem	0.752
$X\text{-}Score^{HM}$	0.614

Ballester, P. J.; Mitchell, J. B. O. *Bioinformatics* **2010**, 26, 1169-1175
Ballester, P. J.; Schreyer, A.; Blundell, T. L. *J. Chem. Inf. Model.* **2014**, 54, 944-955
Li, H.J.; Leung, K.S.; Wong, M.H.; Ballester, P.J. *Molecules* **2015**, 20, 10947-10962
Zilian, D.; Sotriffer, C.A. *J. Chem. Inf. Model.* **2013**, 53, 1923-1933
Gabel, J.; Desaphy, J.; Rognan, D. *J. Chem. Inf. Model.* **2014,** 54, 2807-2815
Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R.; *J. Chem. Inf. Model.* **2009,** 49, 1079-1093
Gabel, J.; Desaphy, J.; Rognan, D. *J. Chem. Inf. Model.* **2014,** 54, 2807-2815

RFbScores Fail in Docking and Screening

Random Forest-based Scoring Function (RFbScore)

•Superior performance in predicting experimental protein-ligand binding affinity

•Fail in docking/screening tests

JOURNAL OF CHEMICAL INFORMATION AND MODELING

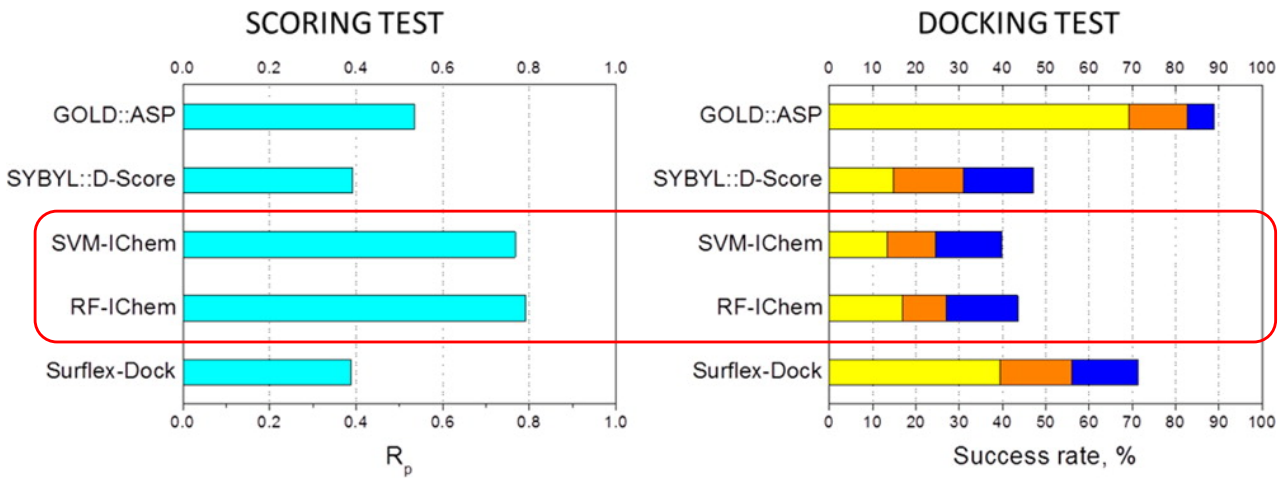
Article

pubs.acs.org/jcim

Beware of Machine Learning-Based Scoring Functions—On the Danger of Developing Black Boxes

Joffrey Gabel, Jérémy Desaphy, and Didier Rognan*

Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 74 route du Rhin, F-67400 Illkirch, France



Gabel, J.; Desaphy, J.; Rognan, D. *J. Chem. Inf. Model.* **2014**, 54, 2807-2815

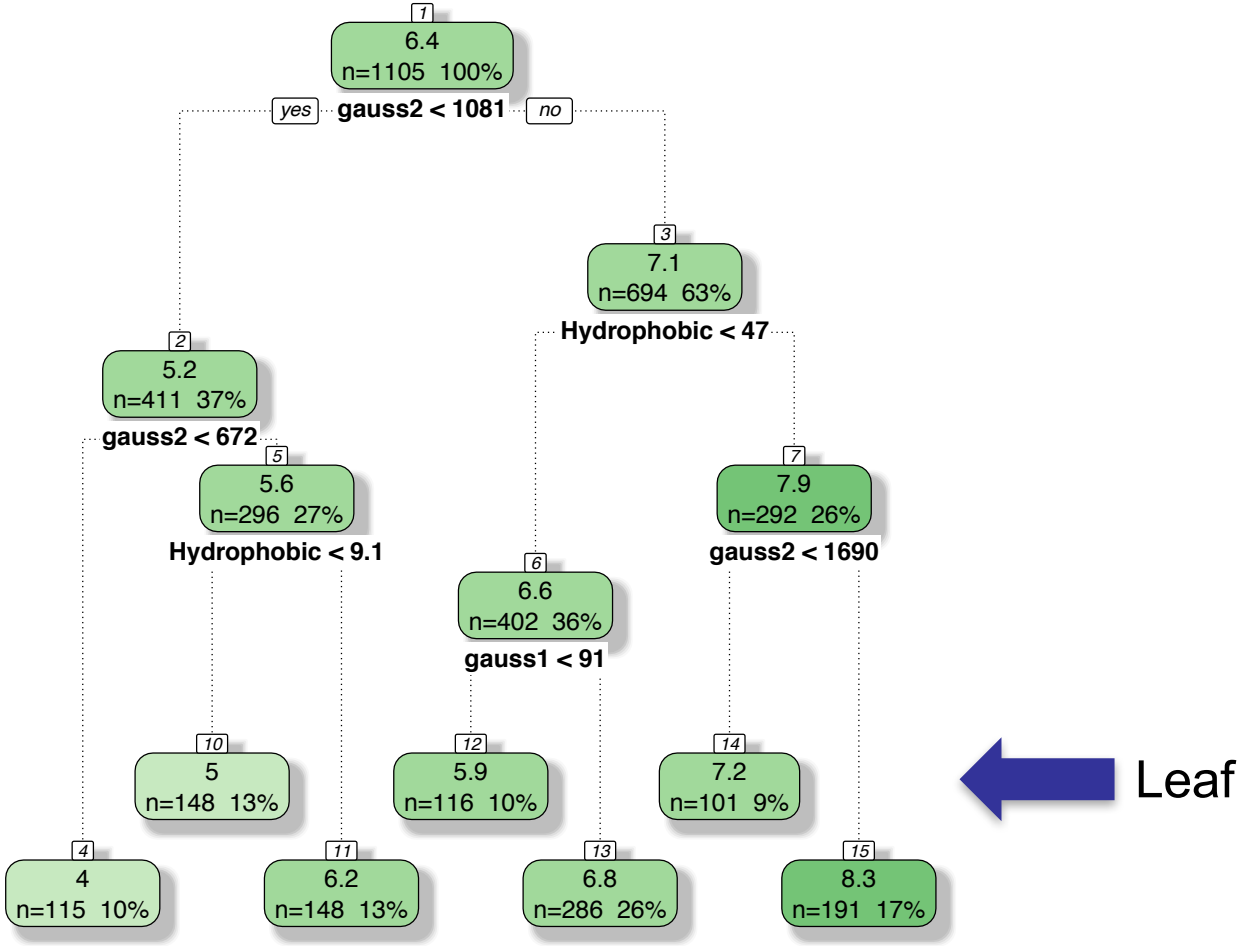
Random Forest

- An ensemble learning method based on the aggregation of numerous decision trees
- Performs remarkably well with very little tuning required
- Can handle a large feature set and correlated features
- Can also be used for assessing feature importance and feature selection.

Breiman, L. *Machine Learning* **2001**, 45, 5-32
Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer New York Inc.: New York, 2009
Wyner, A.J.; Olson, M.; Bleich, J.; Mease, D. **2015**, arXiv:1504.07676
Wager, S.; Walther, G. **2015**, arXiv:1503.06388

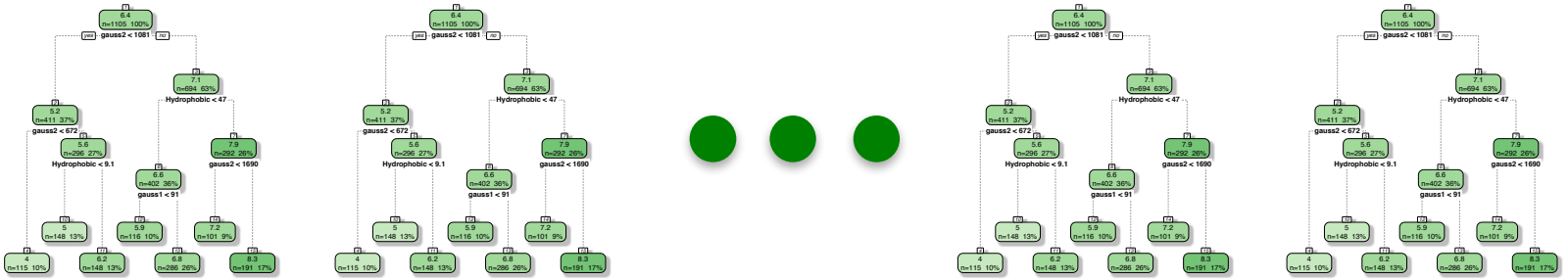
Random Forest – Interpolating

- Given input features (variable, predictor) $X^T = (X_1, X_2, \ldots, X_p)$
- Real-valued output Y_{train}
- The predicted Y_{pred} for each tree is in range [min(Y_{train}), max(Y_{train})]
- Each leaf in the tree is an average value of a Y_{train} subset.



Random Forest – Self-averaging

B Trees



Predict point x

$f^{*1}(x)$ $f^{*2}(x)$ $f^{*B-1}(x)$ $f^{*B}(x)$

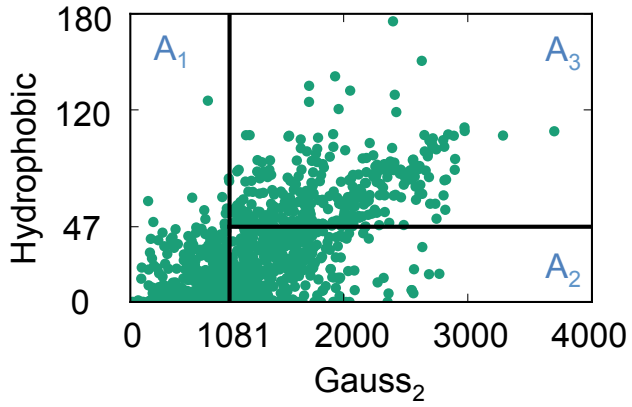
$$f_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(x)$$

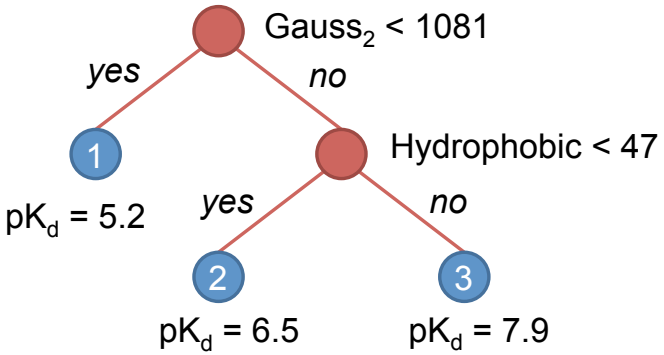
- The predicted Y_{pred} for each tree is in range [min(Y_{train}), max(Y_{train})]
- The predicted Y_{pred} for random forest is in range [min(Y_{train}), max(Y_{train})]

Predicted Value from Random Forest is Bounded by Training Set

Regression Tree Demo

- Each green point presents one training set complex from PDBBind v2007
- $Gauss_2$ and Hydrophobic are two features from Autodock Vina
- Each leaf node contains a subset of training set
- Averaged pK_d of subset complexes is used as predicted value





$$T(X; D_{train}^{*}) = \frac{1}{N_A} \sum_{i \in A} pK_d^{(i)}$$

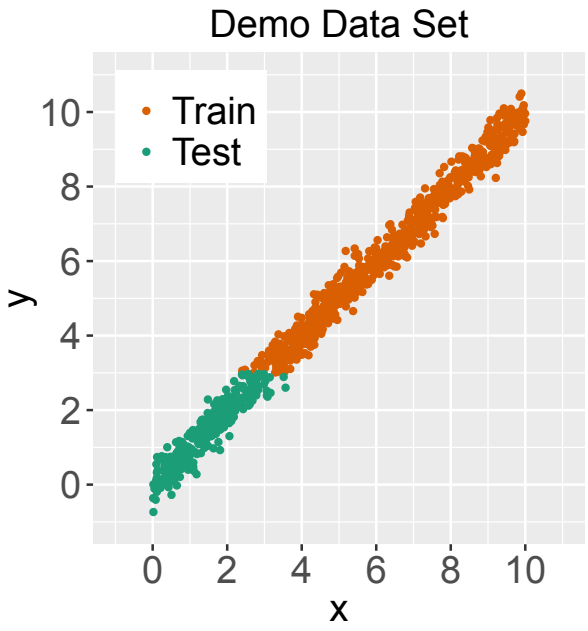
- The predicted $pK_{d\,pred}$ from each tree is in range [min($pK_{d\,train}$), max($pK_{d\,train}$)]
- The predicted $pK_{d\,pred}$ from random forest is in range [min($pK_{d\,train}$), max($pK_{d\,train}$)]

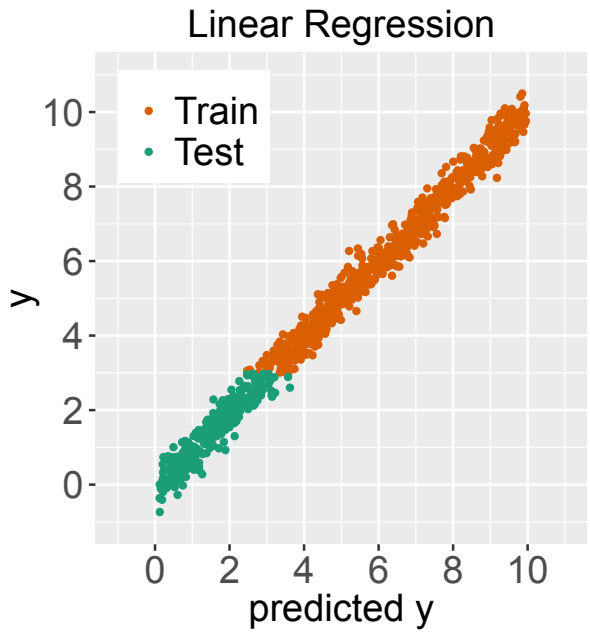
Wyner, A.J.; Olson, M.; Bleich, J.; Mease, D. *arXiv:1504.07676*
Wager, S.; Walther, G. *arXiv:1503.06388*

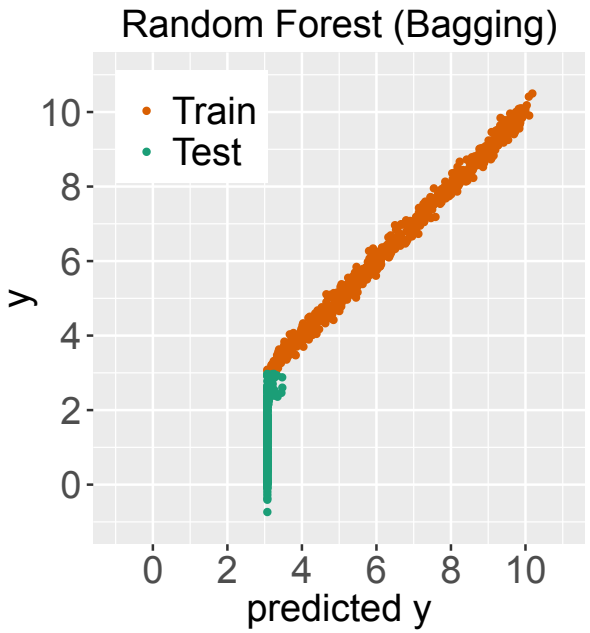
Random forest can only do **interpolation** and CANNOT do extrapolation

Example: y = x + *N*(0, 0.3), 1000 points

- Linear regression can do extrapolation
- Random forest can only predict data point in training space





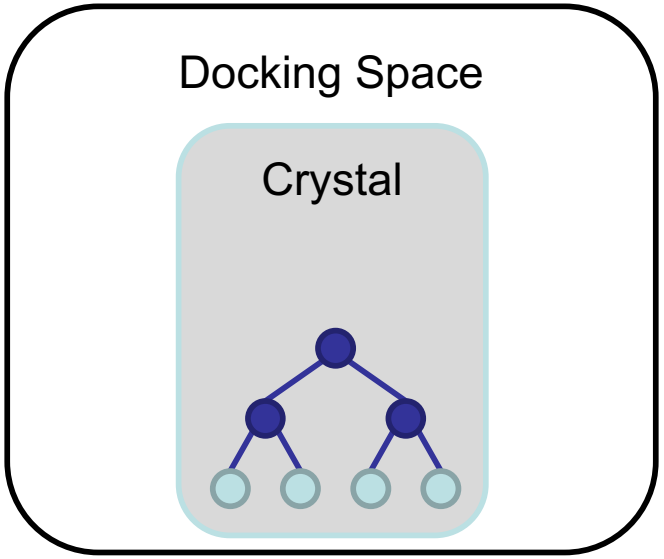


Wyner, A.J.; Olson, M.; Bleich, J.; Mease, D. *arXiv:1504.07676*
Wager, S.; Walther, G. *arXiv:1503.06388*

Extrapolation is Needed for Docking/Screening

- Random forest is designed to do interpolation and CANNOT do extrapolation
  - The predicted value from random forest is bounded by the training set

- Inferior performance of docking/screening for RFbScores comes from
  1. Only using crystal structure as training set
  2. Interpolation nature of Random Forest



Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R.; *J. Chem. Inf. Model.* **2009,** 49, 1079-1093
Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R.; *J. Chem. Inf. Model.* **2014,** 54, 1700-1716
Dunbar, J.B.; et al; *J. Chem. Inf. Model.* **2011,** 51, 2036-2046

Two-pronged Strategy

1. Expanding the training set
   - Experimental subset
   - Decoy subset

2. Δ_{vina}RF approach use RF to parameterize correction to Vina score to take advantage of
   - the excellent docking power of Vina
   - the strength of RF in improving scoring accuracy

$\Delta_{vina}RF_{20}$ is a scoring function based on Δ_{vina}RF approach with 20 features.

Ramakrishnan, Dral, Rupp, von Lilienfeld, J. Chem. Theory Comput. 2015, 11, 2087.
Wang, C.; Zhang, Y.K.; *J. Comput. Chem.* **2017,** 38, 169-177.

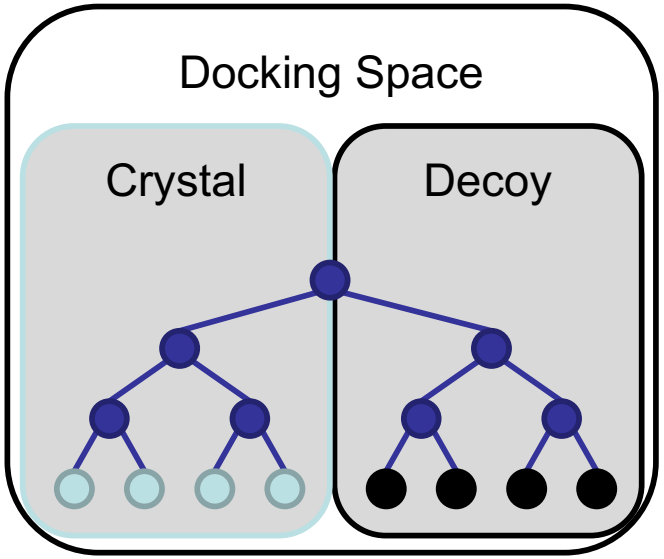
Expanding the Training Set

Two Subsets of Training Set

Experimental subset (3336)

Crystal structures with experimental binding affinity.

PDBbind-v2014



Decoy subset (3322)

Decoy structures generated by docking with binding affinity estimated by Vina.

CSAR-decoys

No overlap with CASF-2007 and CASF-2013

Dunbar, J.B.; et al; *J. Chem. Inf. Model.* **2011,** 51, 2036-2046
Huang, S.Y.; Zou, X.Q. *J. Chem. Inf. Model.* **2011,** 51, 2107-2114
http://www.csardock.org/downloads/DECOY_ALL.htm
Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R.; *J. Chem. Inf. Model.* **2014,** 54, 1700-1716
Wang, C.; Zhang, Y.K.; *J. Comput. Chem.* **2017,** 38, 169-177.

Δ_{vina}RF approach

Vina score as base scoring function.

Taking care of extrapolation & Good docking power of Vina.

$pK_d(\Delta_{vina}RF) = pK_d(Vina) + \Delta pK_d(RF)$

Correction to Vina score by random forest model

Taking advantages of RF in improving scoring accuracy.

Autodock Vina

- Gauss_1, Gauss_2, Repulsion, Hydrophobic, HBond, N_{rot}
- First five based on surface distance

$$d_{ij} = r_{ij} - R_{t_i} - R_{t_j}$$

$$\text{gauss}_1(d) = e^{-(d/0.5)^2}$$

$$\text{gauss}_2(d) = e^{-((d-3)/2))^2}$$

$$c_{\text{inter}} = \sum_{i}^{\text{ligand}} \sum_{j}^{\text{protein}} (\omega_1 \text{gauss}_1(d_{ij}) + \omega_2 \text{gauss}_2(d_{ij}) + \omega_3 \text{Repulsion}(d_{ij})) + \sum_{i,i\in\text{HP}}^{\text{ligand}} \sum_{j,j\in\text{HP}}^{\text{protein}} \omega_4 \text{Hydrophogic}(d_{ij}) + \sum_{i,i\in\text{HB}}^{\text{ligand}} \sum_{j,j\in\text{HB}}^{\text{protein}} \omega_5 \text{HBond}(d_{ij})$$

$$\text{repulsion}(d) = \begin{cases} d^2 & d < 0 \\ 0 & d \geq 0 \end{cases}$$

$$\text{Hydrophobic}(d) = \begin{cases} 1.0 & d < 0.5 \\ 1.5 - d & 0.5 \leq d \leq 1.5 \\ 0.0 & d > 1.5 \end{cases}$$

$$g(c_{\text{inter}}) = \frac{c_{\text{inter}}}{1 + \omega N_{rot}}$$

$$\text{pK}_\text{d}\text{(Vina)} = -0.73349 * g(c_{\text{inter}})$$

$$\text{HBond}(d) = \begin{cases} 1.0 & d < -0.7 \\ d/(-0.7) & -0.7 \leq d \leq 0 \\ 0.0 & d > 0 \end{cases}$$

Weight	Term
-0.0356	gauss_1 (ω_1)
-0.00516	gauss_2 (ω_2)
0.840	Repulsion (ω_3)
-0.0351	Hydrophobic (ω_4)
-0.587	Hydrogen bonding (ω_5)
0.0585	N_{rot} (ω)

Trott, O.; Olson, A.J.; *J. Comput. Chem.* **2010,** 31, 455-461

20 Features in $\Delta_{vina}RF_{20}$

10 Autodock Vina Features (source code)

5 Interaction Terms

- Non-hydrophobic
- Hydrogen bond
- Solvation from Autodock4
- Electrostatic term with x = 1 and x = 2

$$\frac{q_{a_1} \cdot q_{a_2}}{d^x}$$

5 ligand dependent Terms

- Number of heavy atoms
- Number of hydrophobic atoms
- Number of torsions
- Number of rotors
- Ligand length

10 Pharmacophore-based buried SASA Features

9 pharmacophore types

- Positive
- Negative
- Donor-Acceptor
- Donor
- Acceptor
- Aromatic
- Hydrophobic
- Polar
- Halogen

1 Total SASA

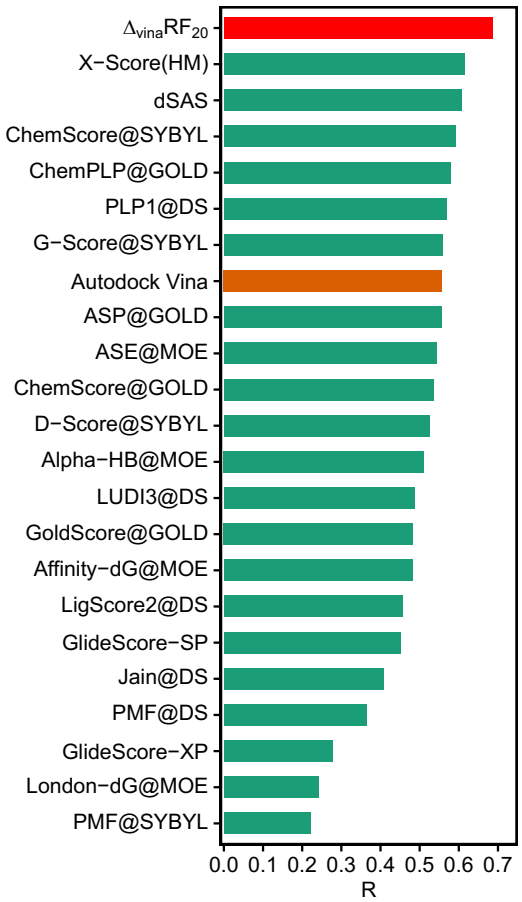
$\Delta_{vina}RF_{20}$ Performs Superior in CASF2013

Scoring power (R)

$\Delta_{vina}RF_{20}$: 0.686

Autodock Vina: 0.557

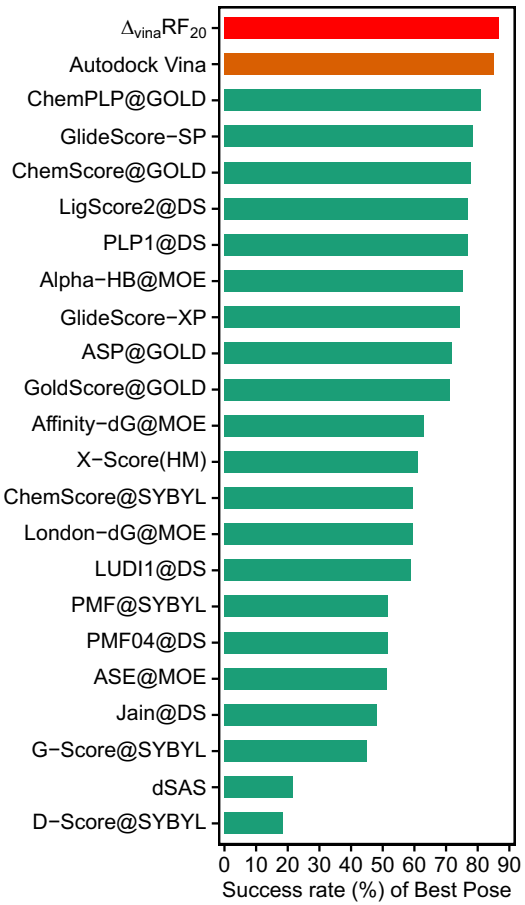
X-ScoreHM: 0.614



Docking power

$\Delta_{vina}RF_{20}$: 86.7%

Autodock Vina: 85.1%

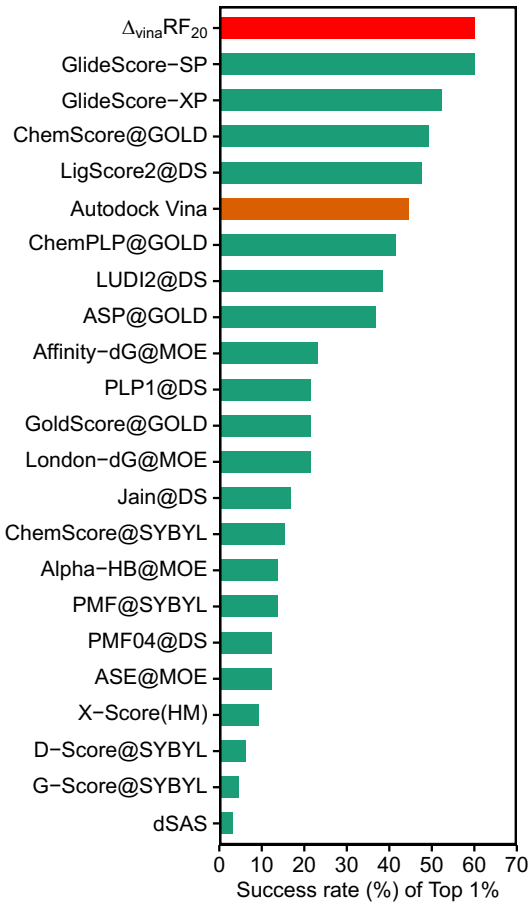


Screening power

$\Delta_{vina}RF_{20}$: 60.0%

Autodock Vina: 44.6%

GlideScore-SP: 60.0%



Li, Y.; Han, L.; Liu, Z.; Wang, R.; *J. Chem. Inf. Model.* **2014,** 54, 1717-1736

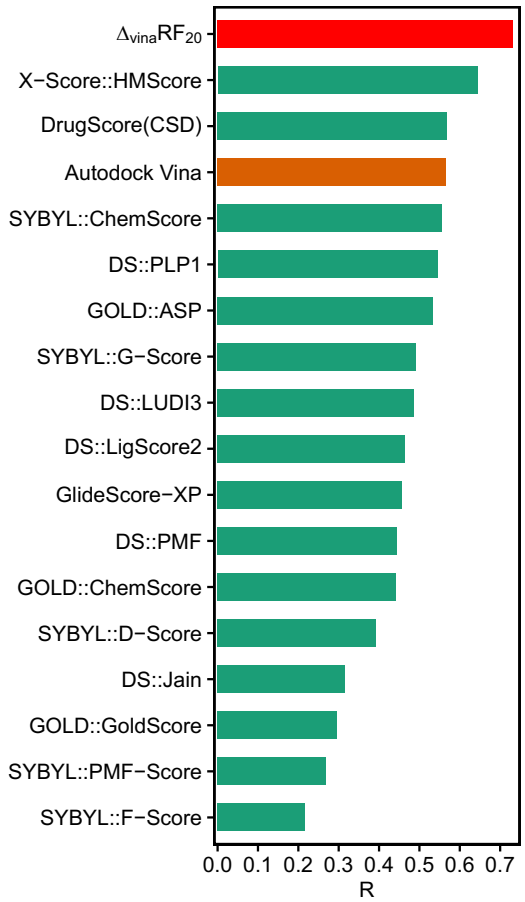
$\Delta_{vina}RF_{20}$ Performs Well in CASF-2007

Scoring power

$\Delta_{vina}RF_{20}$: 0.732

Autodock Vina: 0.566

X-ScoreHM: 0.644

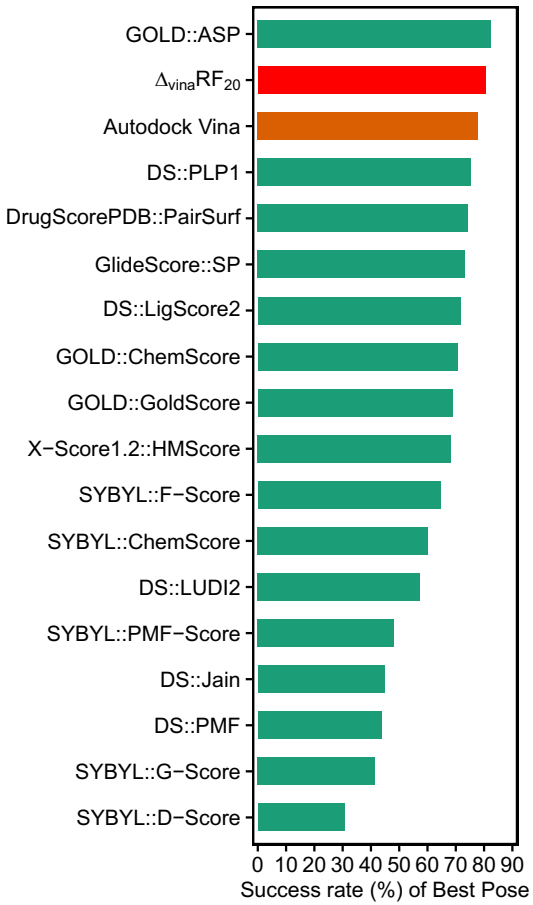


Docking power

$\Delta_{vina}RF_{20}$: 80.5%

Autodock Vina: 77.9%

Gold::ASP: 82.5%



Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R.; *J. Chem. Inf. Model.* **2009,** 49, 1079-1093

Summary

$\Delta_{vina}RF_{20}$ is a scoring function based on $\Delta_{vina}RF$ approach with 20 features achieves supeior performance in scoring, docking and screening power for CASF-2007 and CASF-2013 benchmarks in comparison with classical scoring functions.

- Expanding the training set
  - Experimental subset
  - Decoy subset
- $\Delta_{vina}RF$ approach
  - the excellent docking power of Vina
  - the strength of RF in improving scoring accuracy
- 20 Features
  - 10 Features from Autodock Vina Source Code
  - 10 Pharmacophore-based SASA

C. Wang and Y. Zhang, J. Comput. Chem. , 38 , 169-177 (2017).

Acknowledgement

Dr. Cheng Wang

http://www.nyu.edu/projects/yzhang/Practicals_YingkaiZhang.tar.gz