Quimioinformática en Phyton

Algoritmos de clustering, métricas de validación y aplicaciones, y métricas de enriquecimiento, campañas de cribado retrospectivo y generación de señuelos

Alan Talevi

Laboratorio de Investigación y Desarrollo de Bioactivos (LIDeB)

Universidad Nacional de La Plata – CONICET

e-mail: alantalevi@gmail.com website: lideb.biol.unlp.edu.ar facebook: lideb.unlp twitter: @LIDeB_UNLP

Hoja de ruta

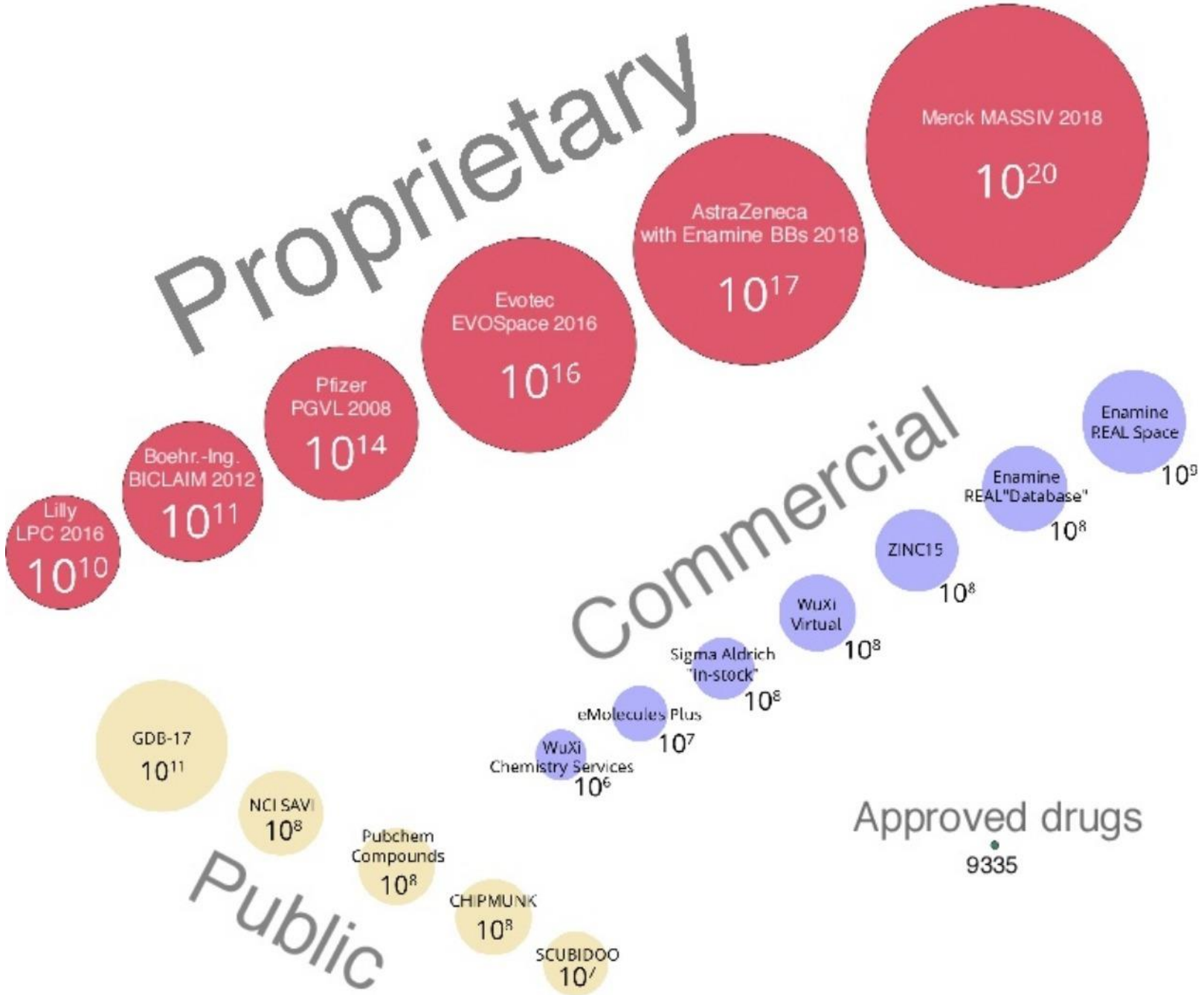
Contexto. Cribado virtual como estrategia eficiente para explorar el espacio químico.

Andamiaje activo y el hit-to-lead.

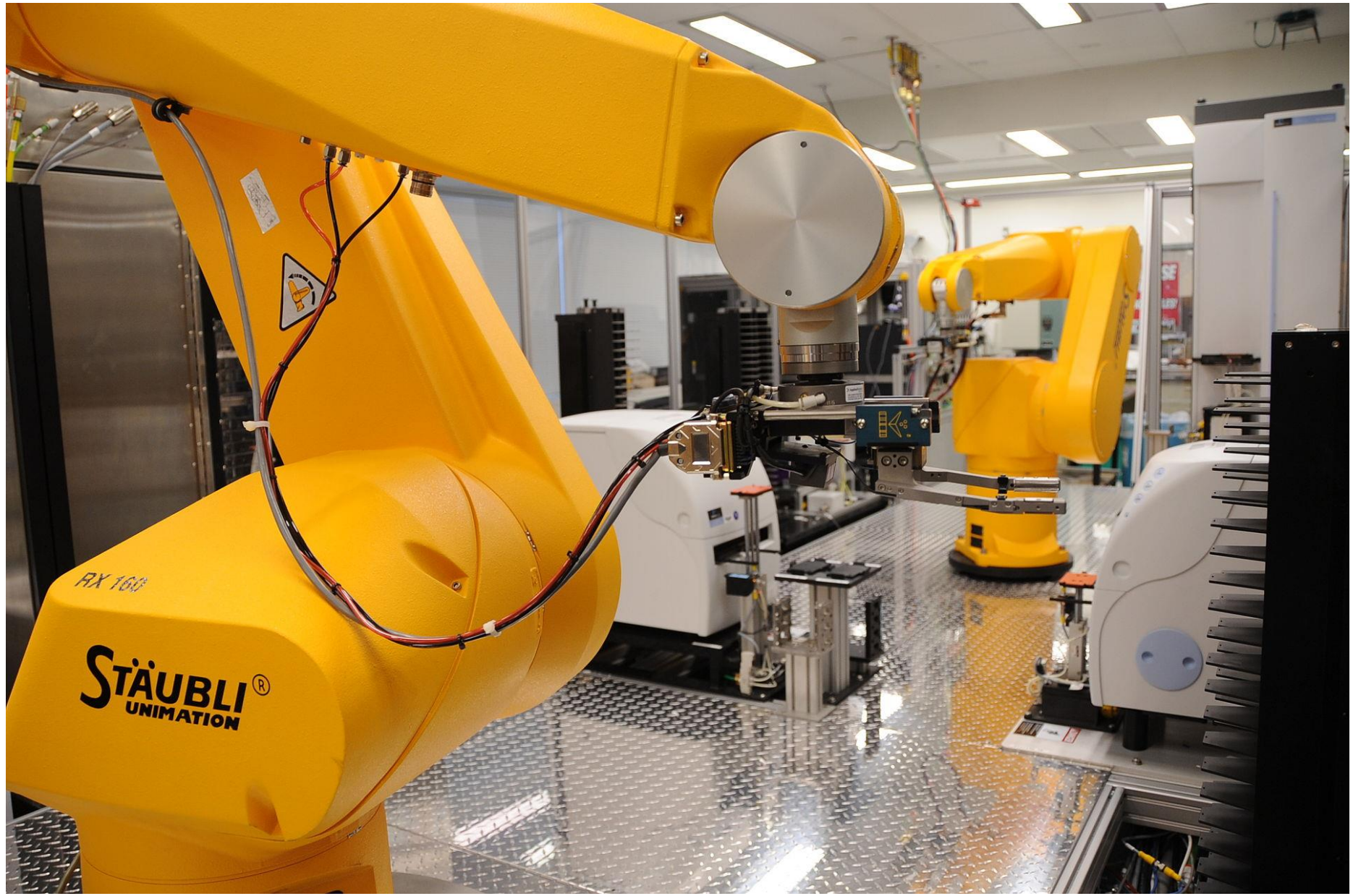
El problema del desbalance de clases en el cribado virtual. Importancia de las campañas de cribado retrospectivo para evaluar el desempeño. Métricas de enriquecimiento.

Generación de señuelos.

Clustering de pequeñas moléculas.



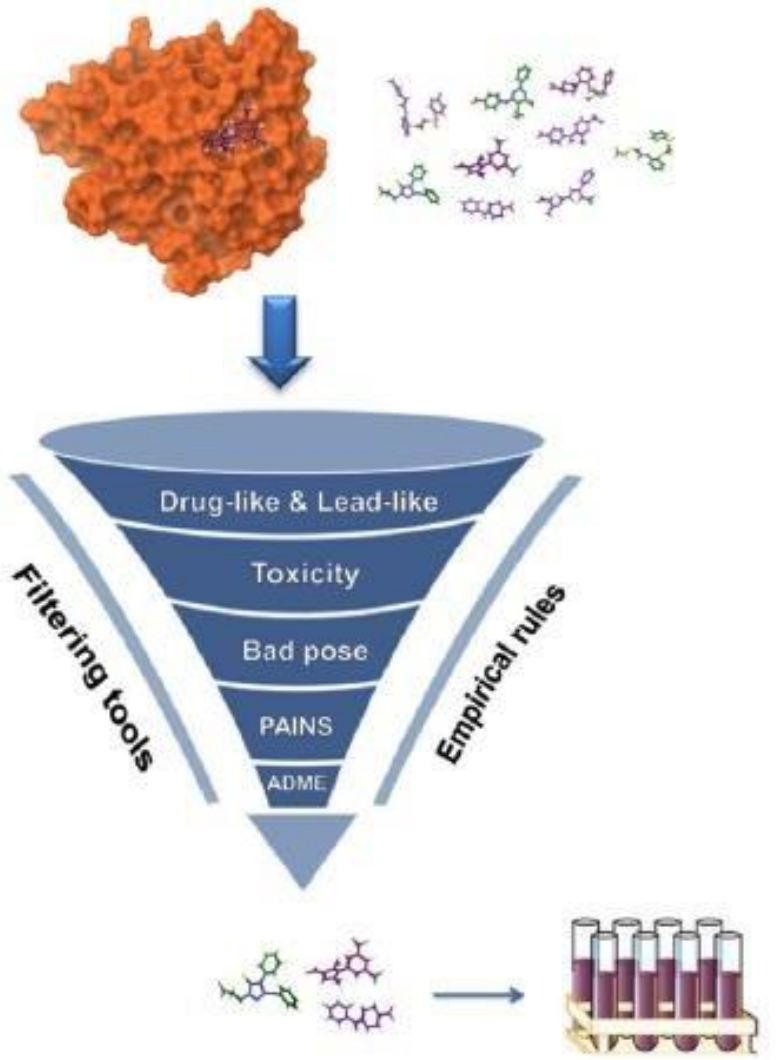
Hoffmann & Gastreich (2019) Drug Discovery Today 24:1148-56


RX 160
STÄUBLI
UNIMATION

virtual screening

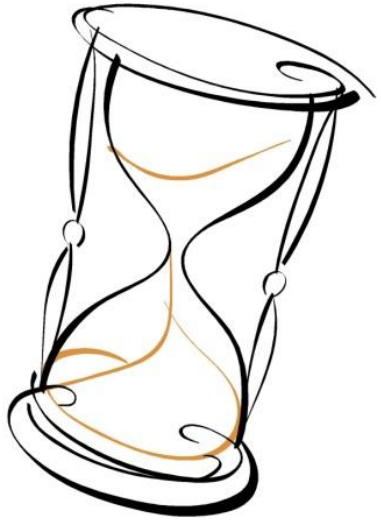
utilizar modelo(s) computacional(es)/algoritmo(s) para identificar en [grandes] quimiotecas digitales, compuestos que posean una o más propiedades [biológicas] de interés (y, ocasionalmente, una diversidad de propiedades farmacéuticamente relevantes).

Por lo general, el hit in silico confirmado se someterá a una campaña de optimización molecular (hit-to-lead).

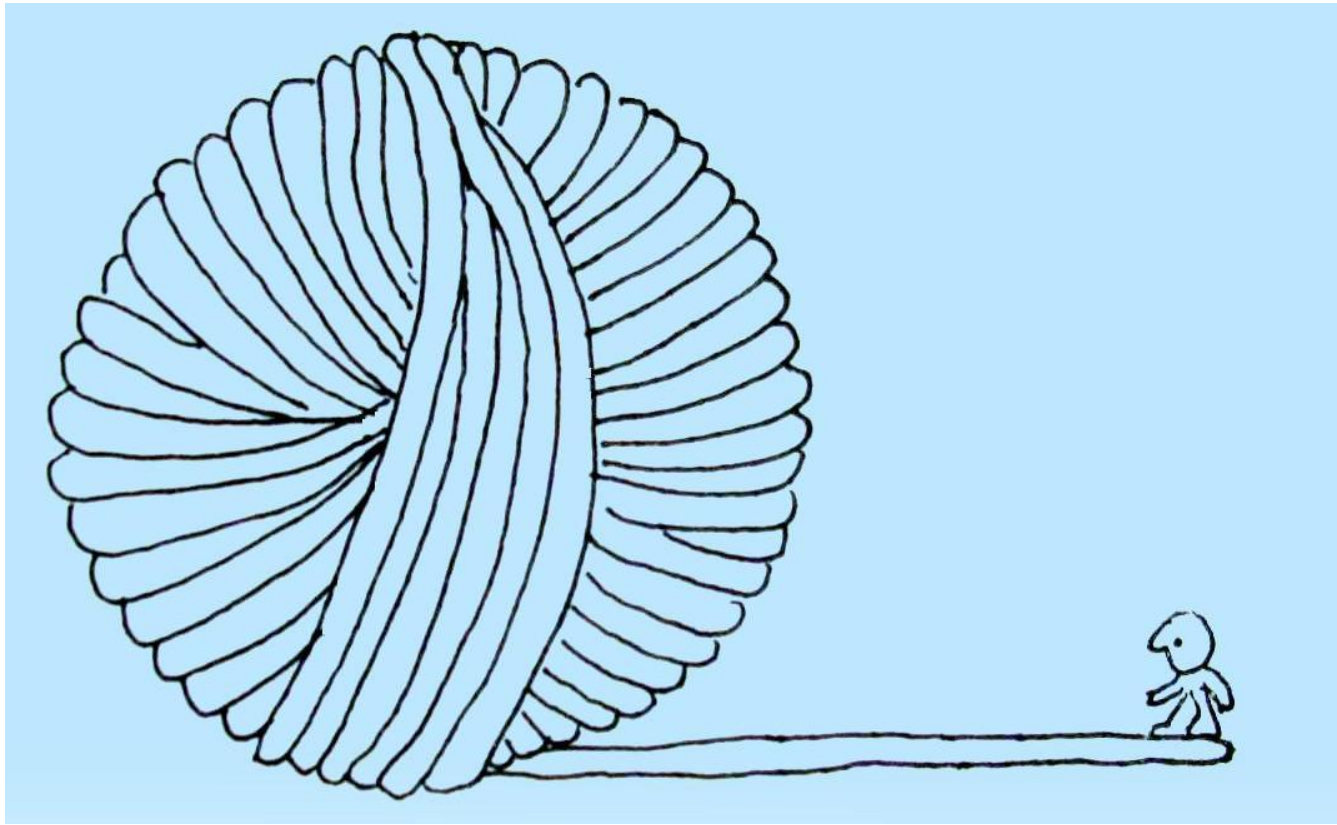


Surabhi & Singh (2018) Journal of Drug Delivery and Therapeutics 8:504-9

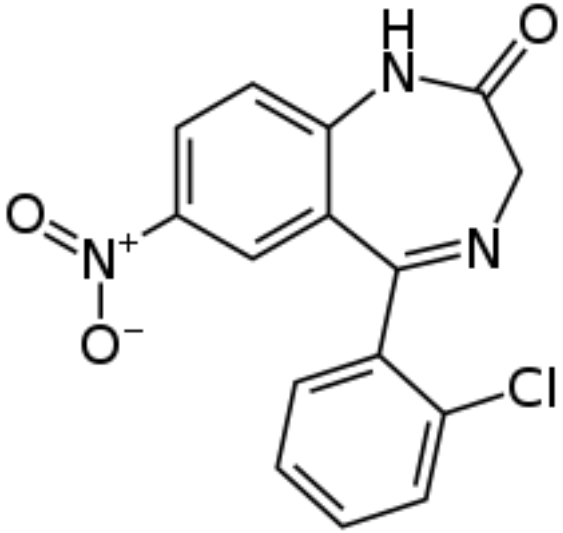
algunas ventajas del screening in silico



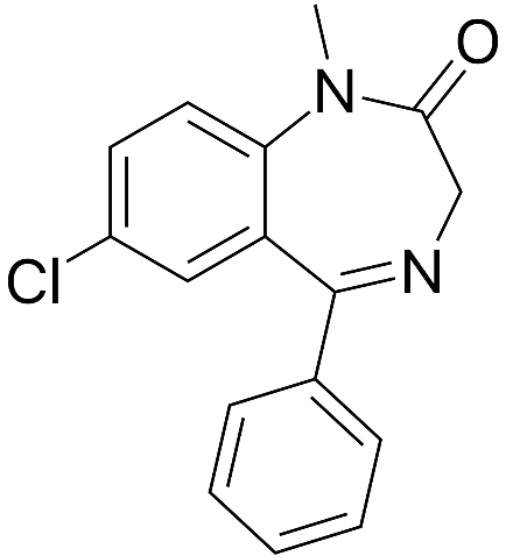
¿Cómo se descubre un nuevo fármaco?



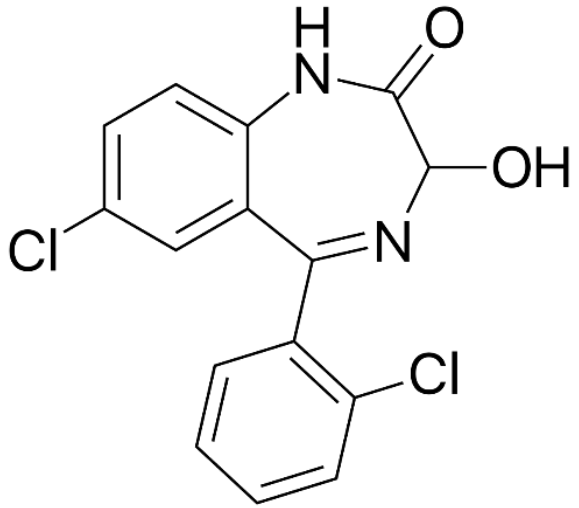
Andamiaje activo ("active scaffold"):
núcleo químico con determinada actividad biológica



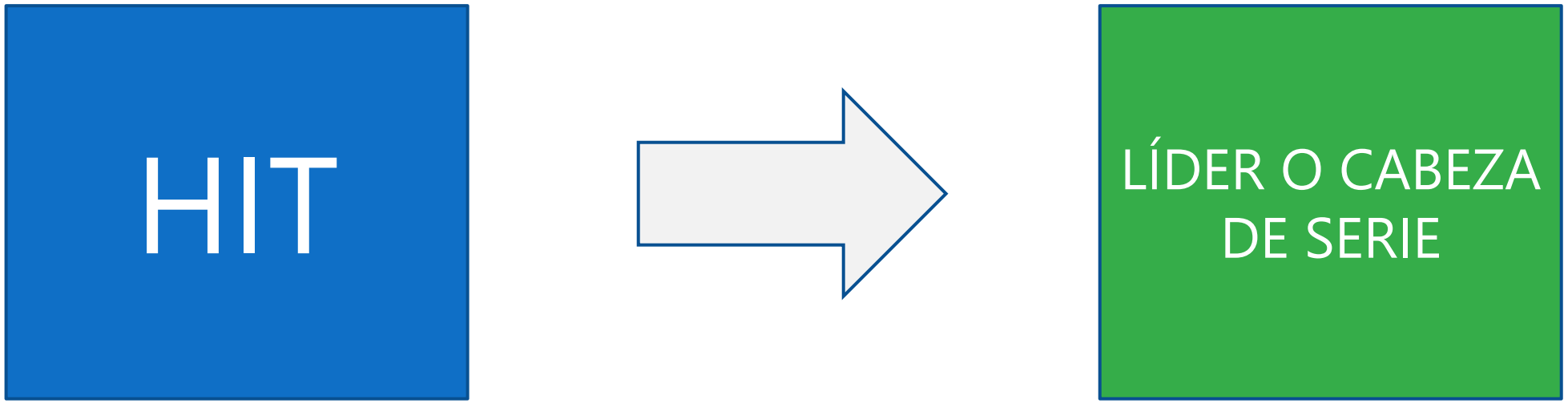
clonazepam



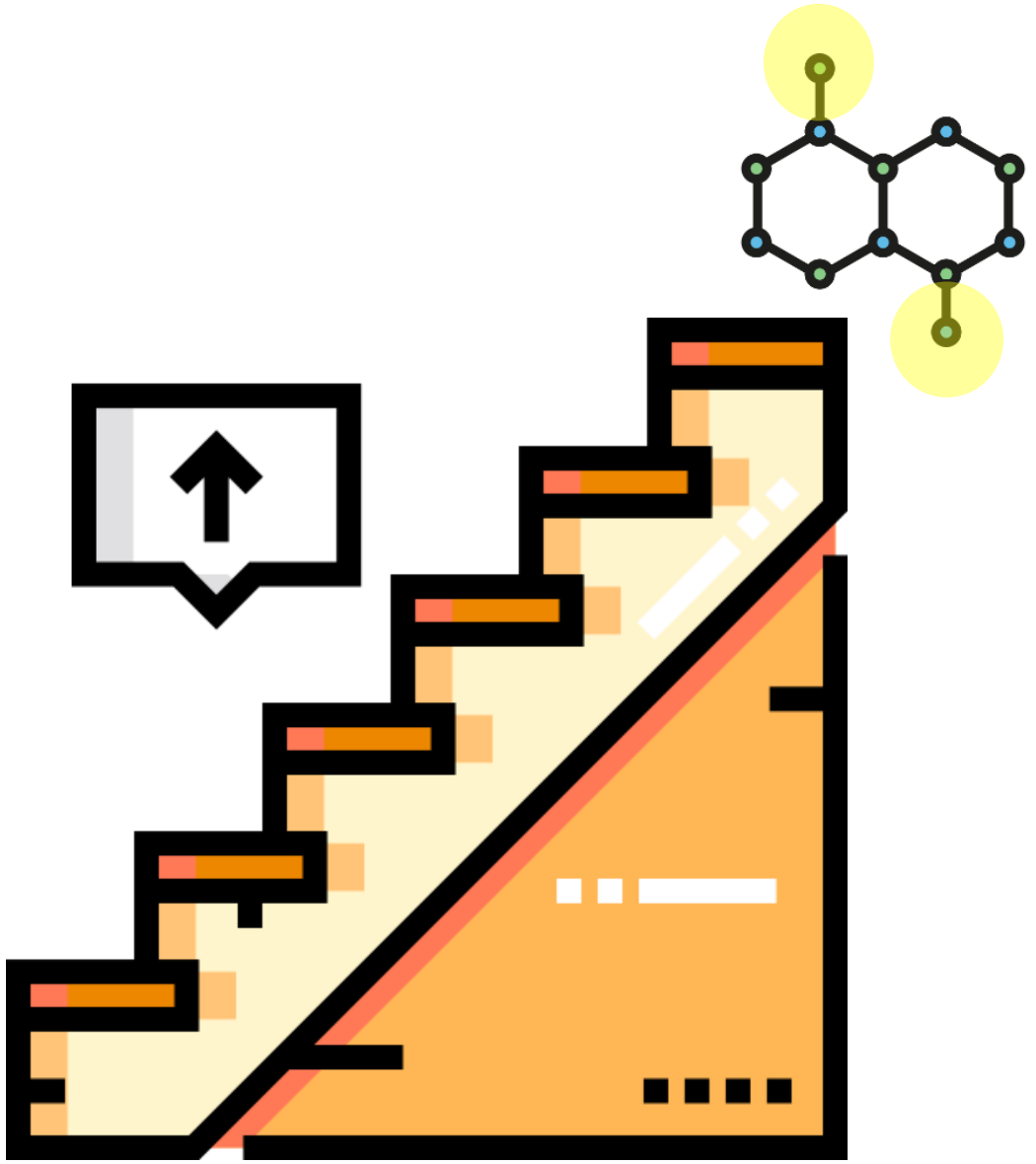
diazepam



lorazepam


HIT
LÍDER O CABEZA DE SERIE

Impacto de la optimización de hits y líderes



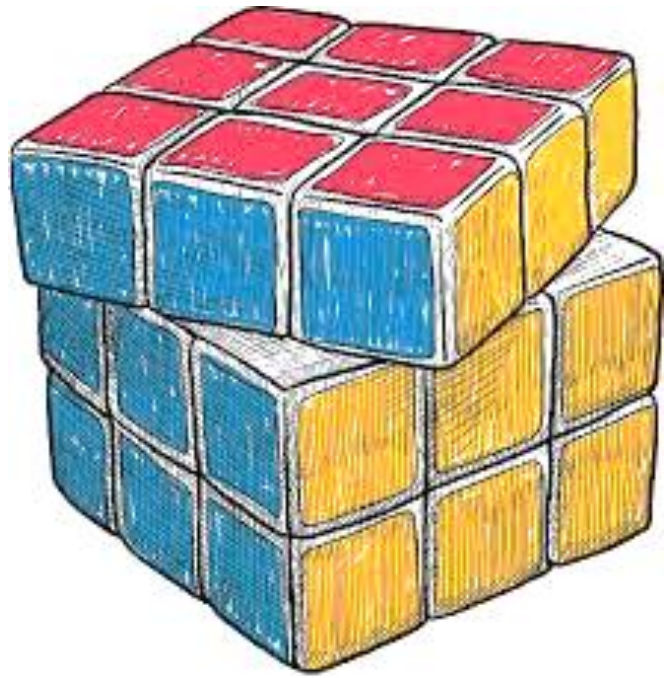
Incrementar la potencia / afinidad hacia el blanco farmacológico

Eliminación de efectos adversos

Mejorar biodisponibilidad

Aumentar la estabilidad química

Aumentar la novedad estructural



El problema del desbalance de clases

$$P(A) = (Se\ Ya)/[Se\ Ya + (1\text{-}Sp)\ (1\text{-}Ya)]$$

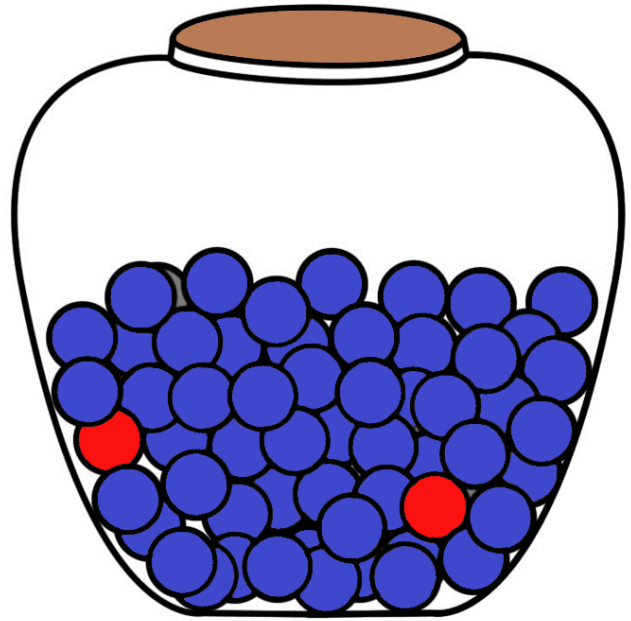
P(A): probabilidad de que un hit in silico confirmará experimentalmente la actividad predicha

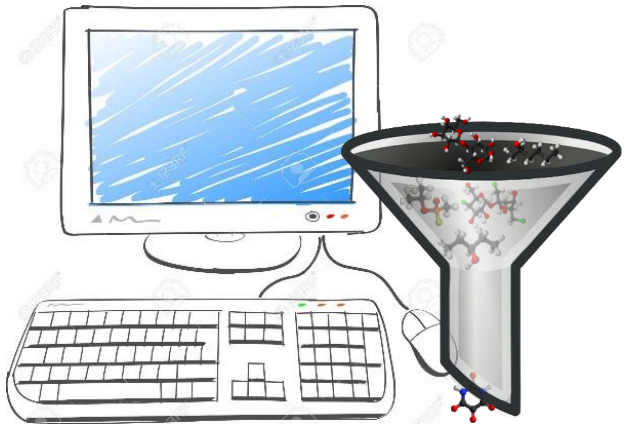
Se: Sensibilidad, tasa de verdaderos positivos

Sp: Especificidad, tasa de verdaderos negativos

Ya (o *Ra*): proporción de compuestos activos en la quimioteca sometida a screening (desconocida a priori)

Ya empírico 0,005 – 0,01
(para quimiotecas no enfocadas)



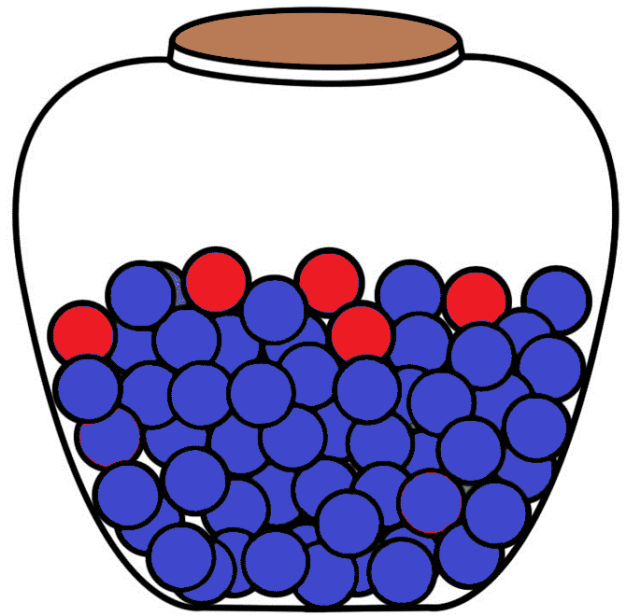


INACTIVO
INACTIVO
INACTIVO
ACTIVO
INACTIVO
INACTIVO
INACTIVO
INACTIVO
INACTIVO
ACTIVO
…

cribado virtual

ACTIVO
INACTIVO
ACTIVO
INACTIVO
ACTIVO
INACTIVO
ACTIVO
ACTIVO
INACTIVO
ACTIVO
…

si el único criterio de selección fuera el score del modelo, enviaríamos a confirmación experimental las primeras posiciones del ranking ordenado



La capacidad de enriquecimiento de un protocolo de cribado virtual se estima a través de experimentos retrospectivos, en los que se siembra un número relativamente bajo (pero diverso) de compuestos activos conocidos entre un número elevado de señuelos o *decoys* (compuestos inactivos verificados o presuntos/putativos). En el experimento retrospectivo se conocen la identidad y proporción (*Ya* / *Ra*) de los compuestos activos


7⁰
6⁶
6⁰
5⁶
5⁰
7⁰
6⁶
6⁰
5⁶
5⁰
4⁶
3⁰

Métricas de enriquecimiento utilizadas en experimentos de cribado retrospectivo

Métricas de enriquecimiento

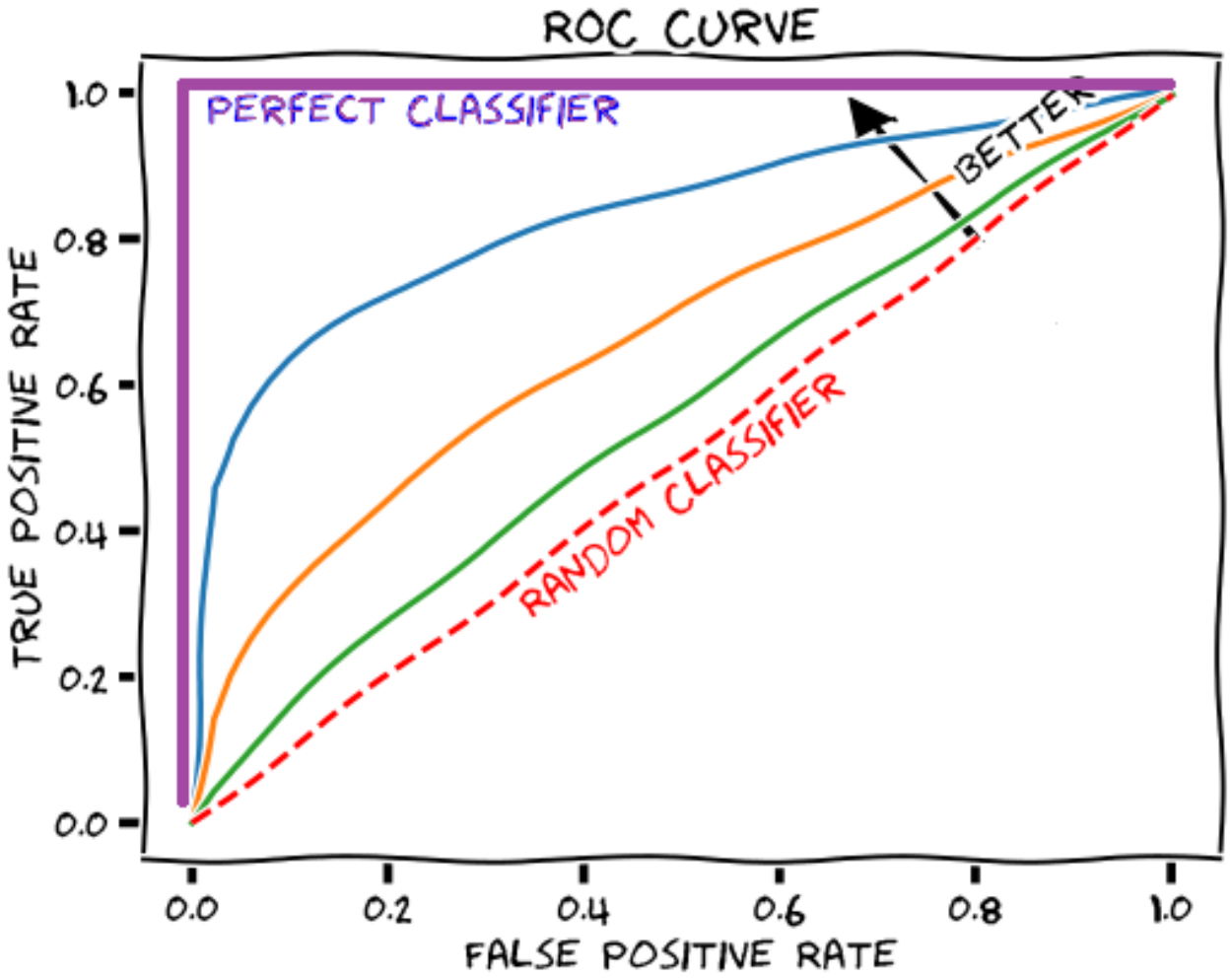
Permiten evaluar el desempeño de un modelo o protocolo de cribado virtual y comparar cuantitativamente / estadísticamente modelos o protocolos.

Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. J. Chem. Inf. Comput. Sci. 2001

Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO. 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. 2005

Truchon JF, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model. 2007

Área bajo la *Receiver Operating Characteristic* (ROC) curve - AUROC

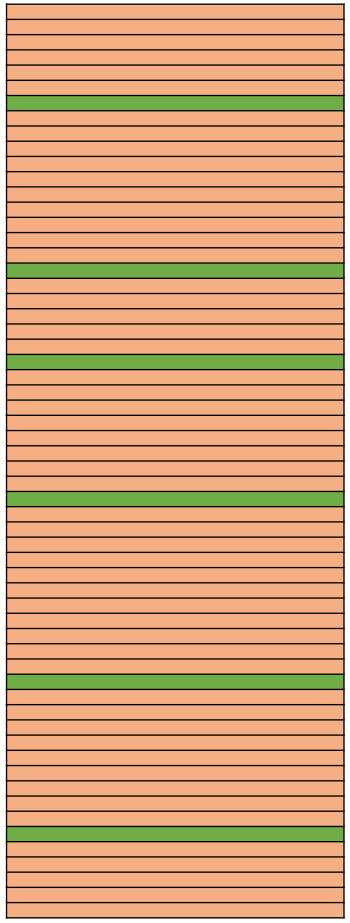


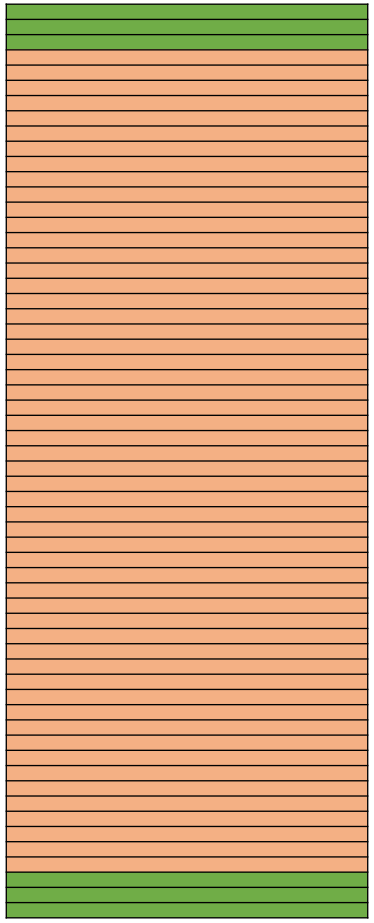
Área bajo la curva ROC (*AUROC*, *ROC*): limitaciones

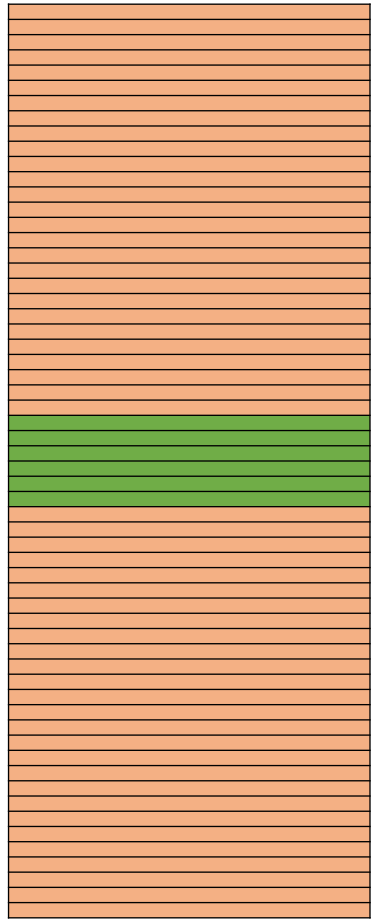
AUROC / ROC refleja la posición promedio de los verdaderos positivos en el ranking ordenado. Todas las posiciones del ranking contribuyen al valor de la métrica de igual manera.

Efecto de saturación.

AUROC: limitaciones







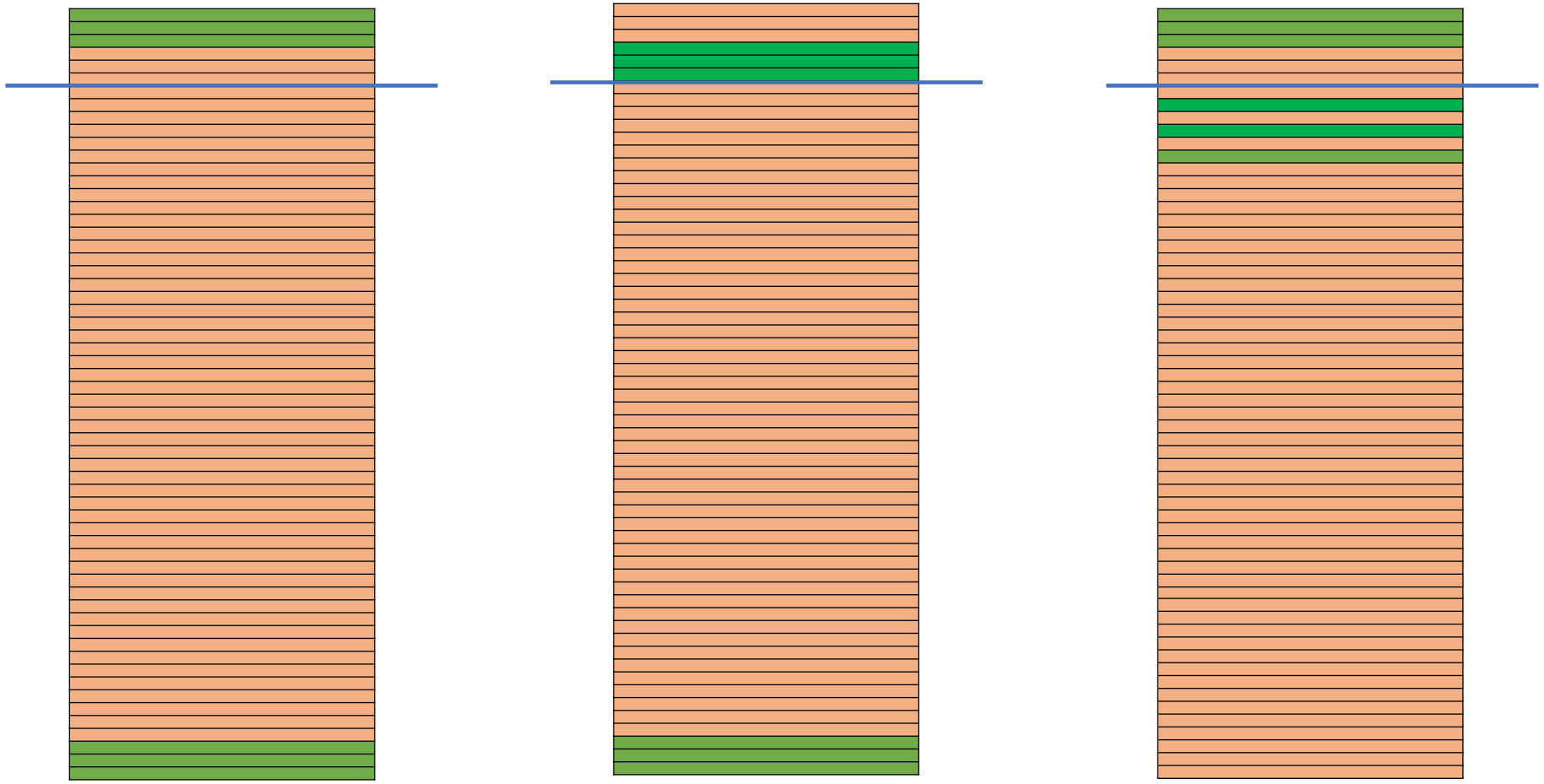
N = 60; n = 6; Ra = 6/60 = 0,1

Los tres rankings ordenados que se muestran resultan en una AUROC ≈ 0,5

Métricas enfocadas en el reconocimiento temprano: Factor de enriquecimiento (EF)

$$EF_{x\%} = \frac{VP \text{ en el } x\% \text{ de la quimioteca mejor rankeado por el modelo}}{VP \text{ en el } x\% \text{ mejor rankeado de la quimioteca ordenada al azar}}$$

ER: limitaciones



N = 60; n = 6; *Ra* = 6/60 = 0,1

¡En los tres casos el $ER_{10\%}$ es idéntico ≈ 5!

ER: limitaciones

No distingue diferencias entre distintos ordenamientos de los activos dentro del x% de la quimioteca en el que se pone el foco.

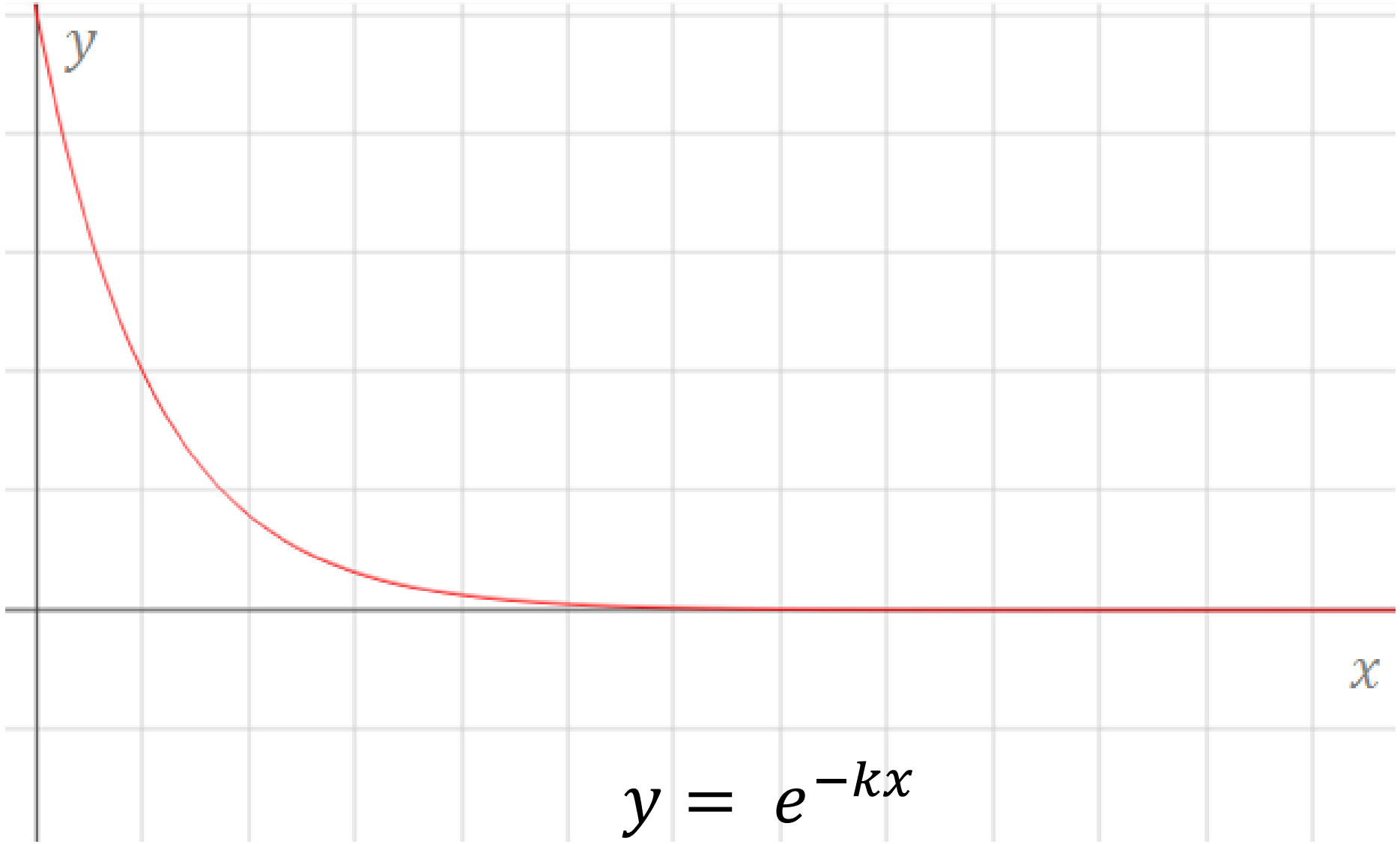
No distingue diferencias entre distintos ordenamientos de los activos en el 100-x% de la quimioteca en el que no se hace foco. Solo considera el x% mejor rankeado.

La cota superior de $EF_{x\%}$ depende de x, de n y de N. Es $1/x$ si $x \geq n/N$ y N/n si $x < n/N$.

Métricas enfocadas en el reconocimiento temprano: Robust Initial Enhancement (RIE)

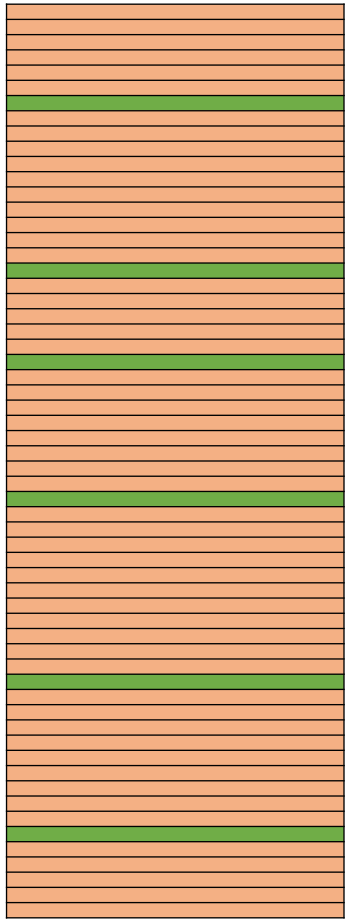
$$RIE = \frac{\sum_1^n e^{-\propto r_i/N}}{\langle \sum_1^n e^{-\propto r_i/N} \rangle r}$$

(ri = posición en el ranking del i-ésimo compuesto activo)

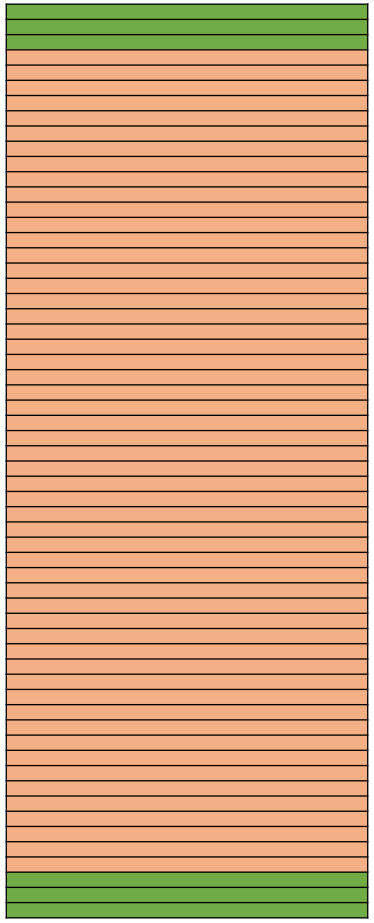


$$y = e^{-kx}$$

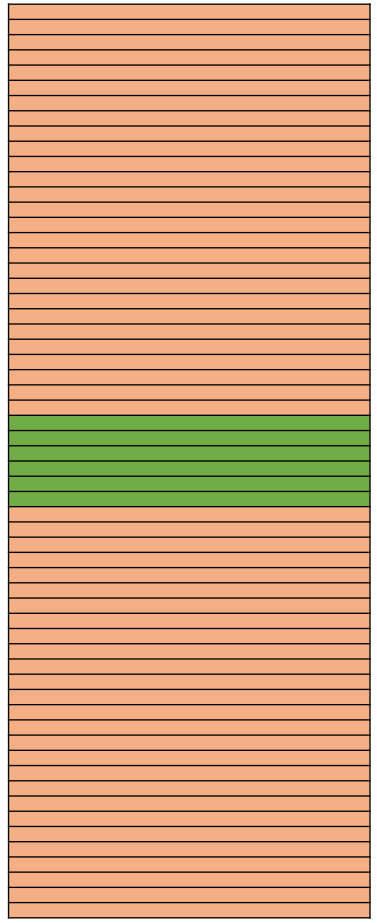
a)



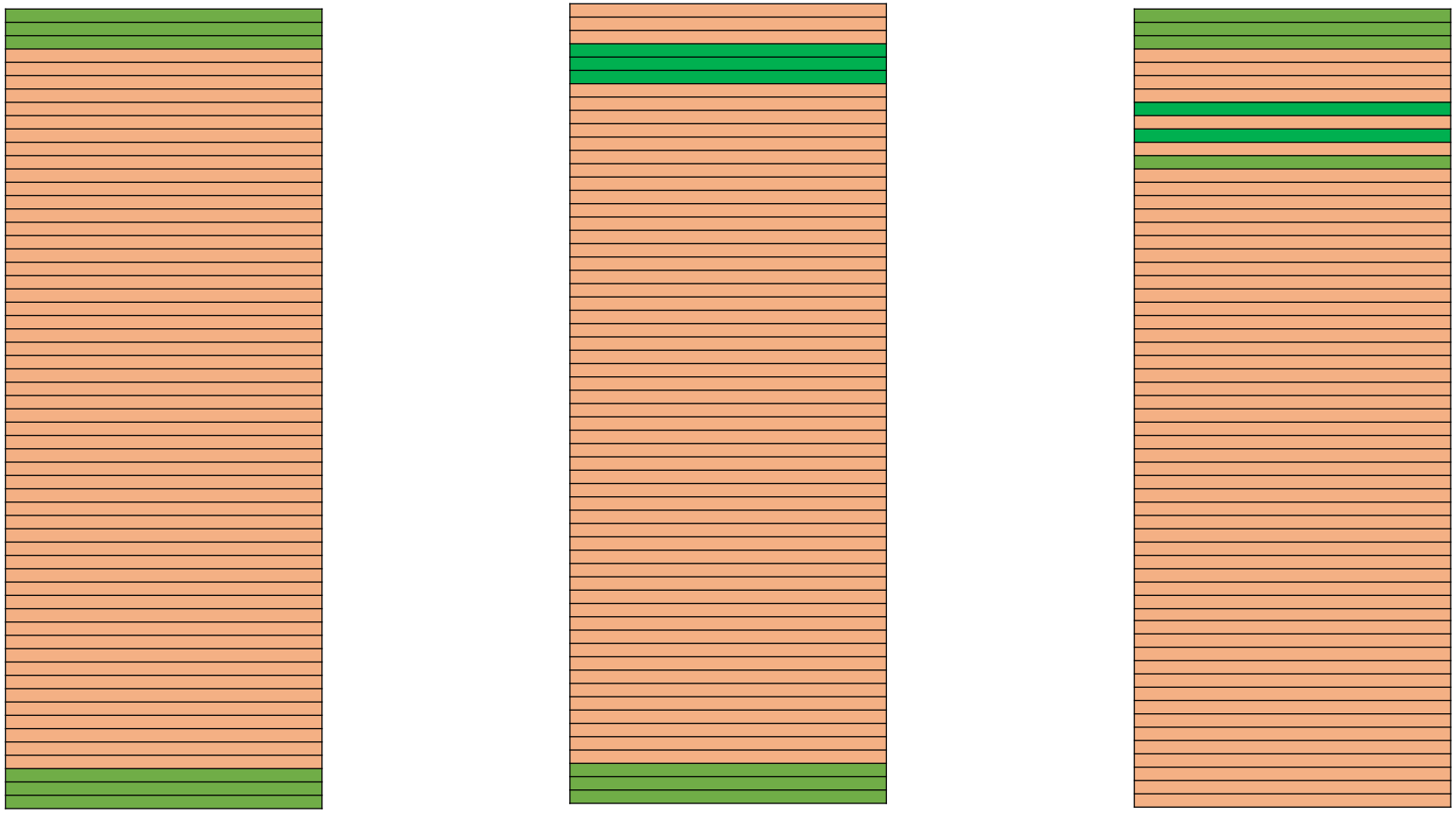
b)



c)



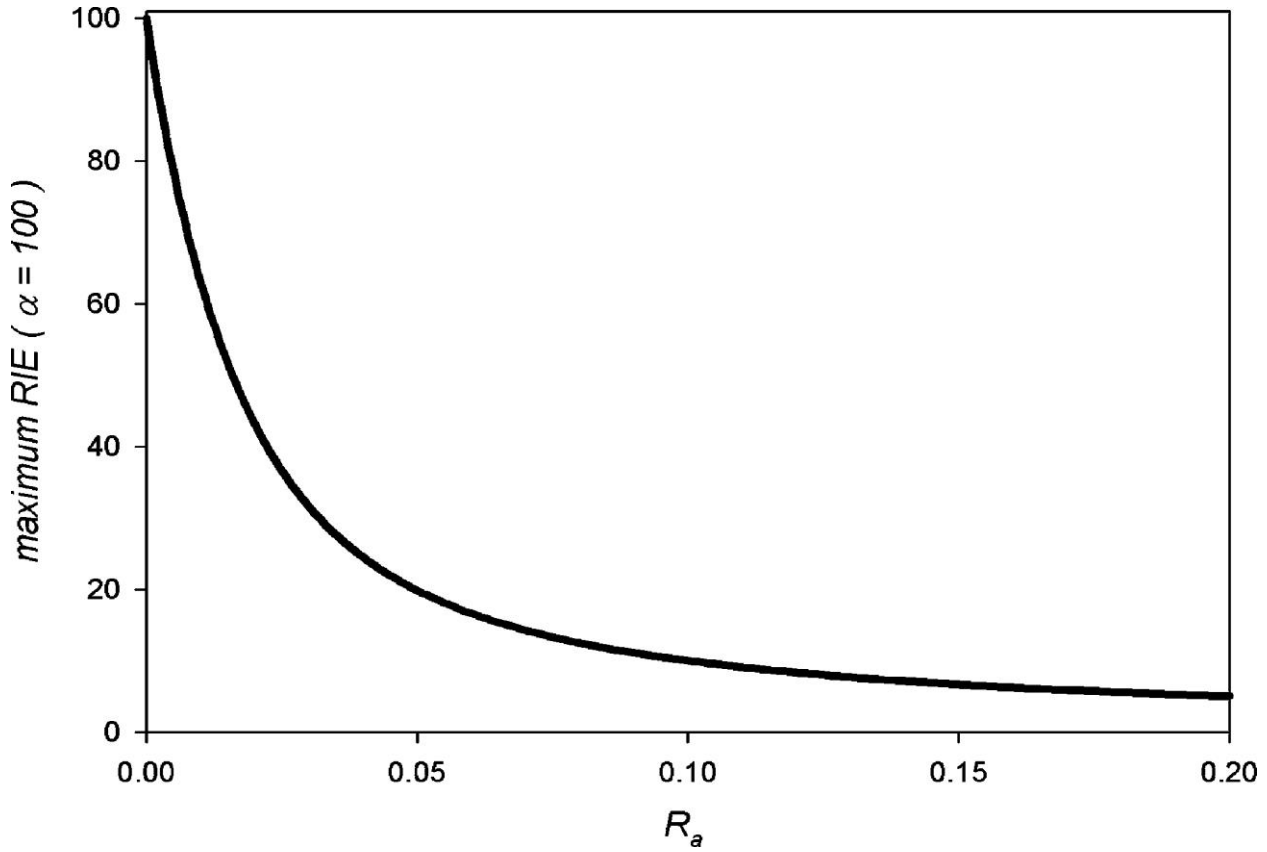
Estos tres ordenamientos posibles tendrían valores de RIE muy distintos. $RIE_b > RIE_a > RIE_c$



Estos tres ordenamientos posibles también tendrían valores de RIE distintos.

RIE: limitaciones

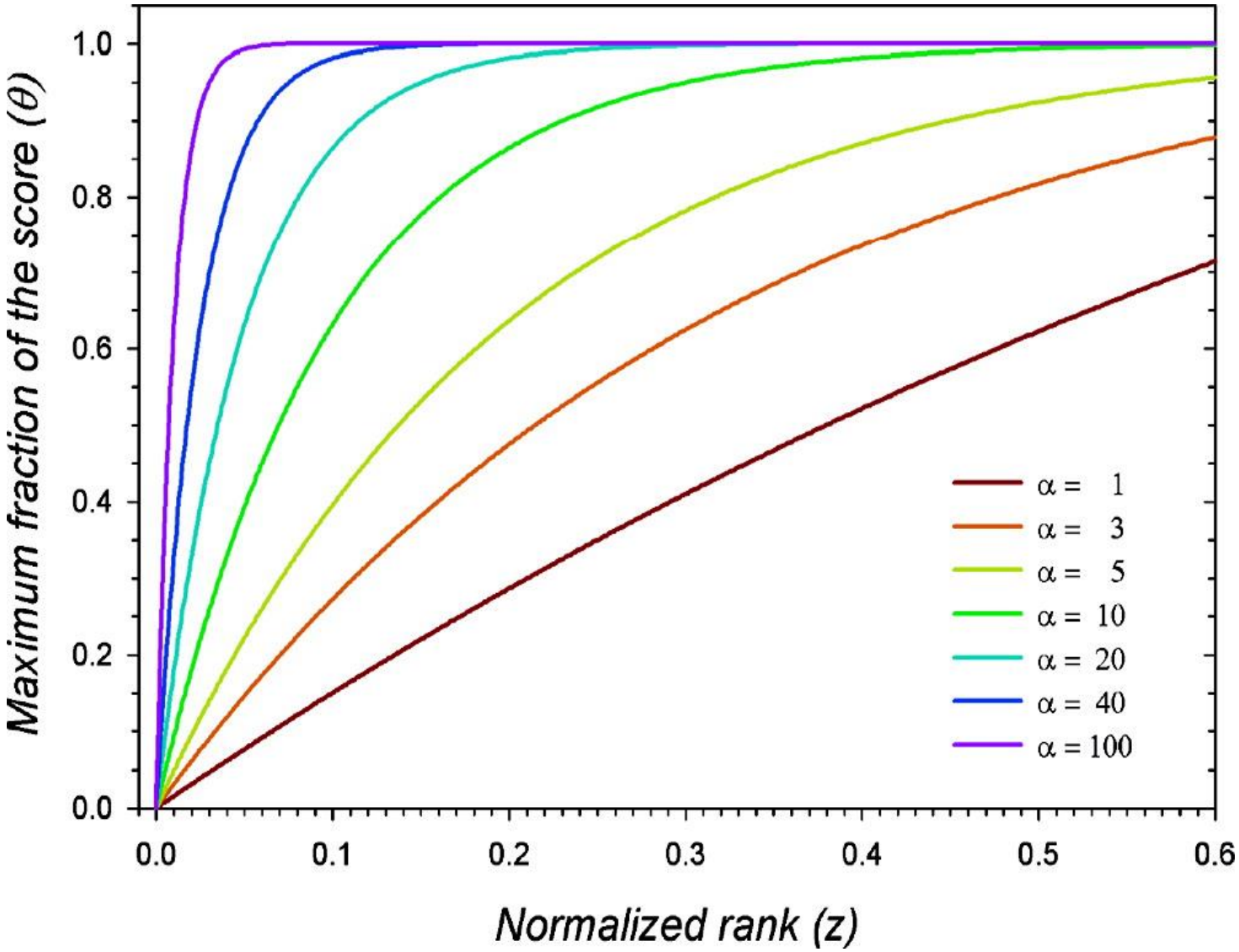
Los valores máximo y mínimo de RIE dependen de α, n y N.

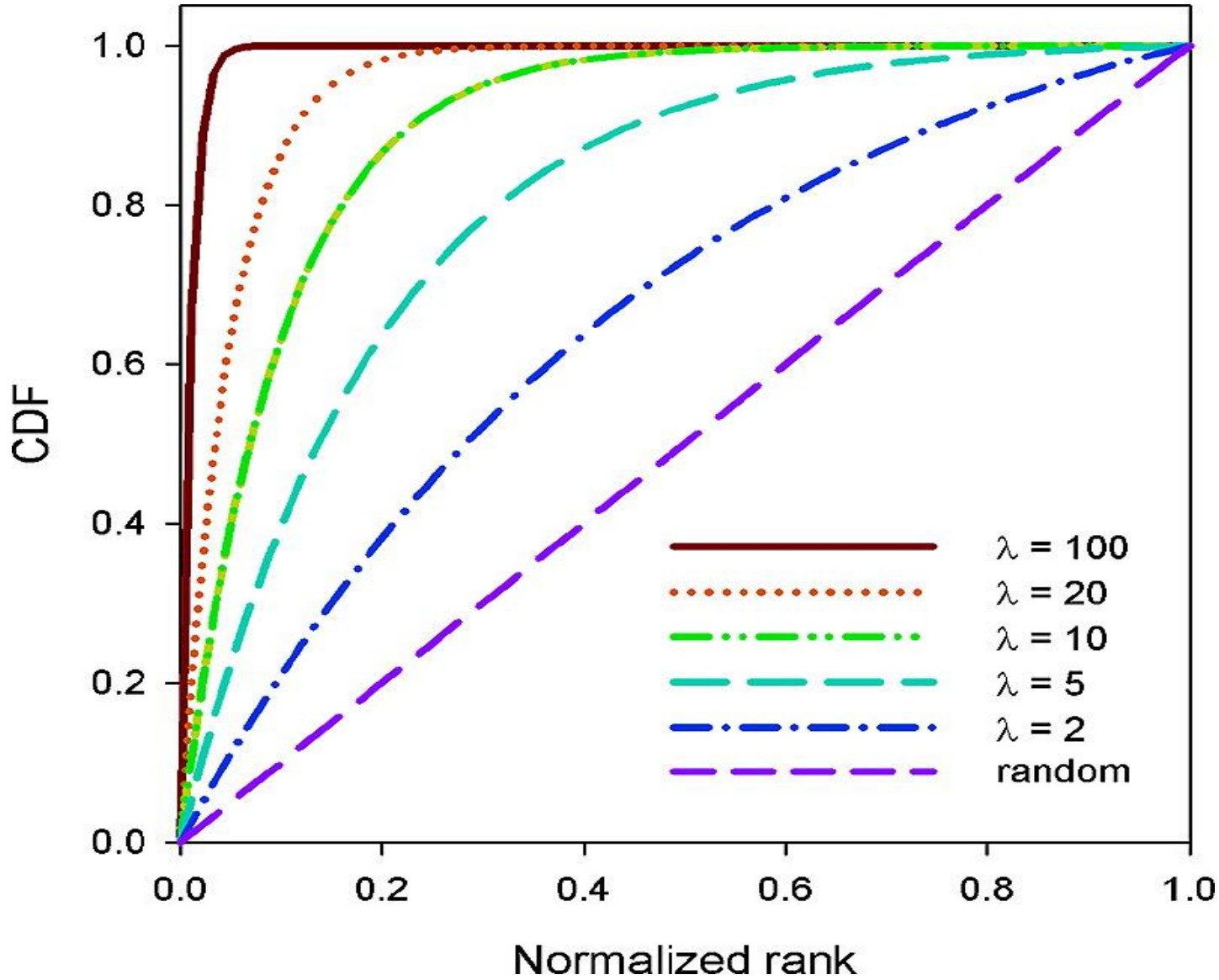


Métricas enfocadas en el reconocimiento temprano: Boltzmann-Enhanced Discrimination of ROC (BEDROC)

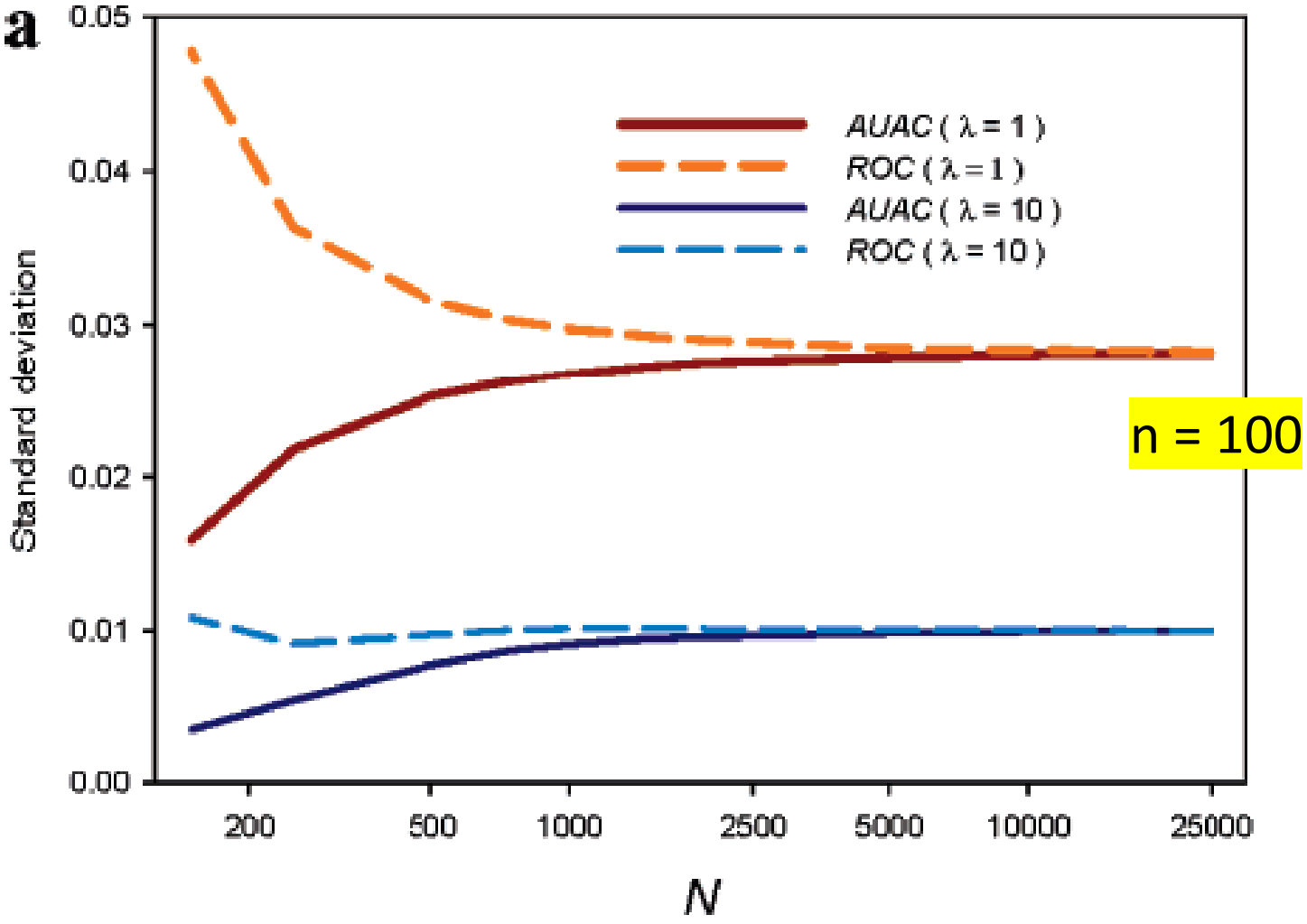
$$BEDROC = \frac{RIE - RIE_{min}}{RIE_{max} - RIE_{min}}$$

Relación entre el valor α y la fracción mejor rankeada de la lista ordenada que contribuye en mayor medida al valor de la métrica BEDROC. Por ejemplo, si α = 100, el 1% mejor rankeado contribuye aprox. con el



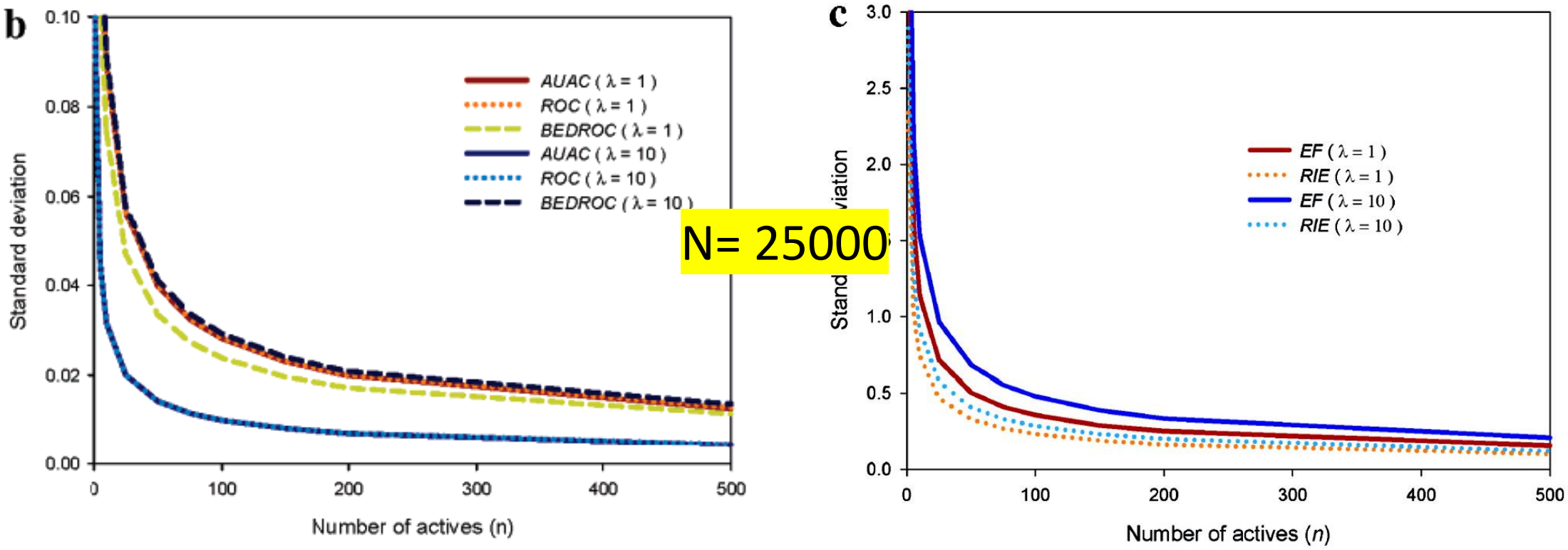

1.0
0.8
0.6
0.4
0.2
0.0
CDF
0.0
0.2
0.4
0.6
0.8
1.0
Normalized rank
λ = 100
λ = 20
λ = 10
λ = 5
λ = 2
random

Desviación estándar de ROC versus N, con n constante

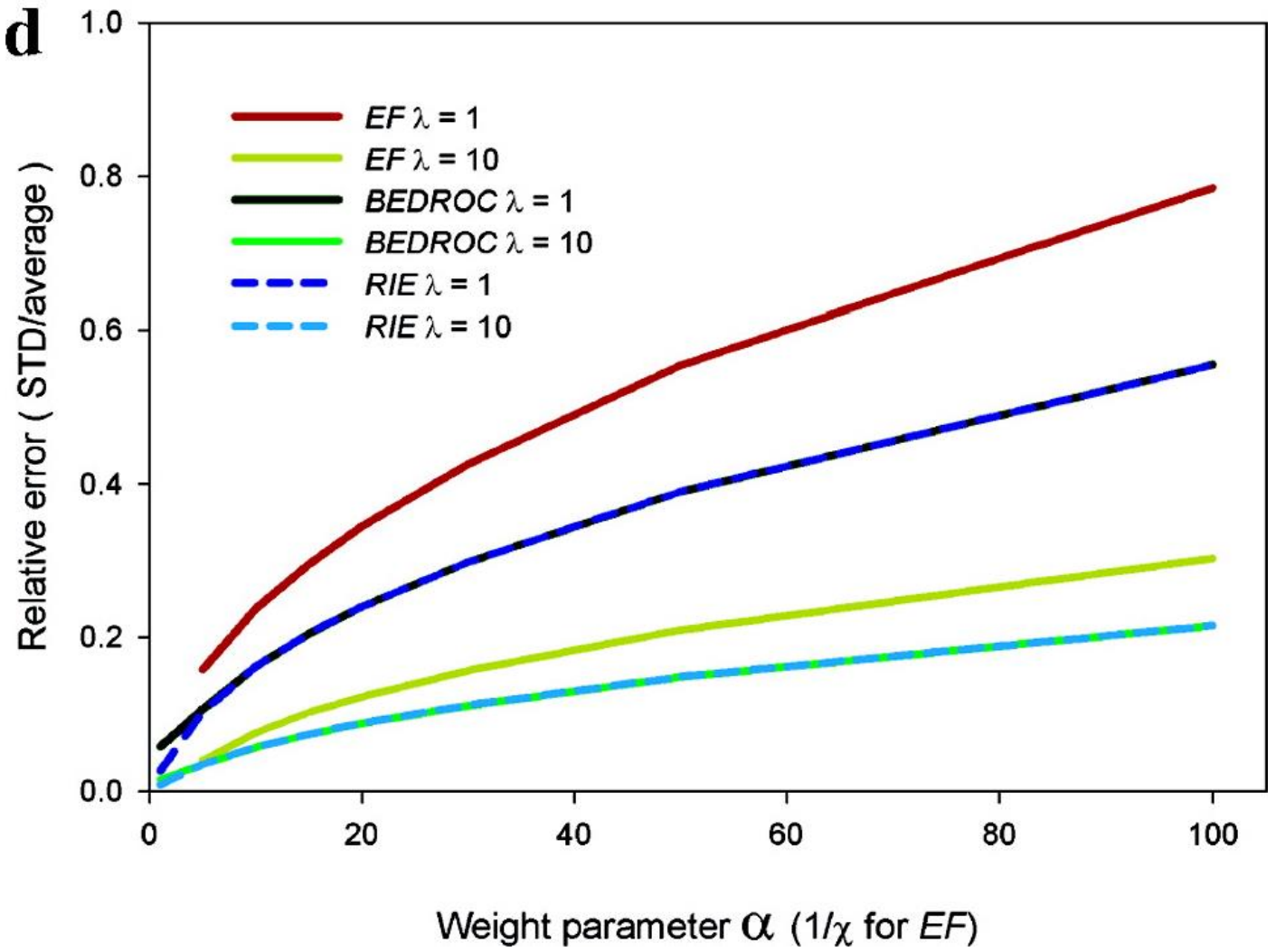


n = 100

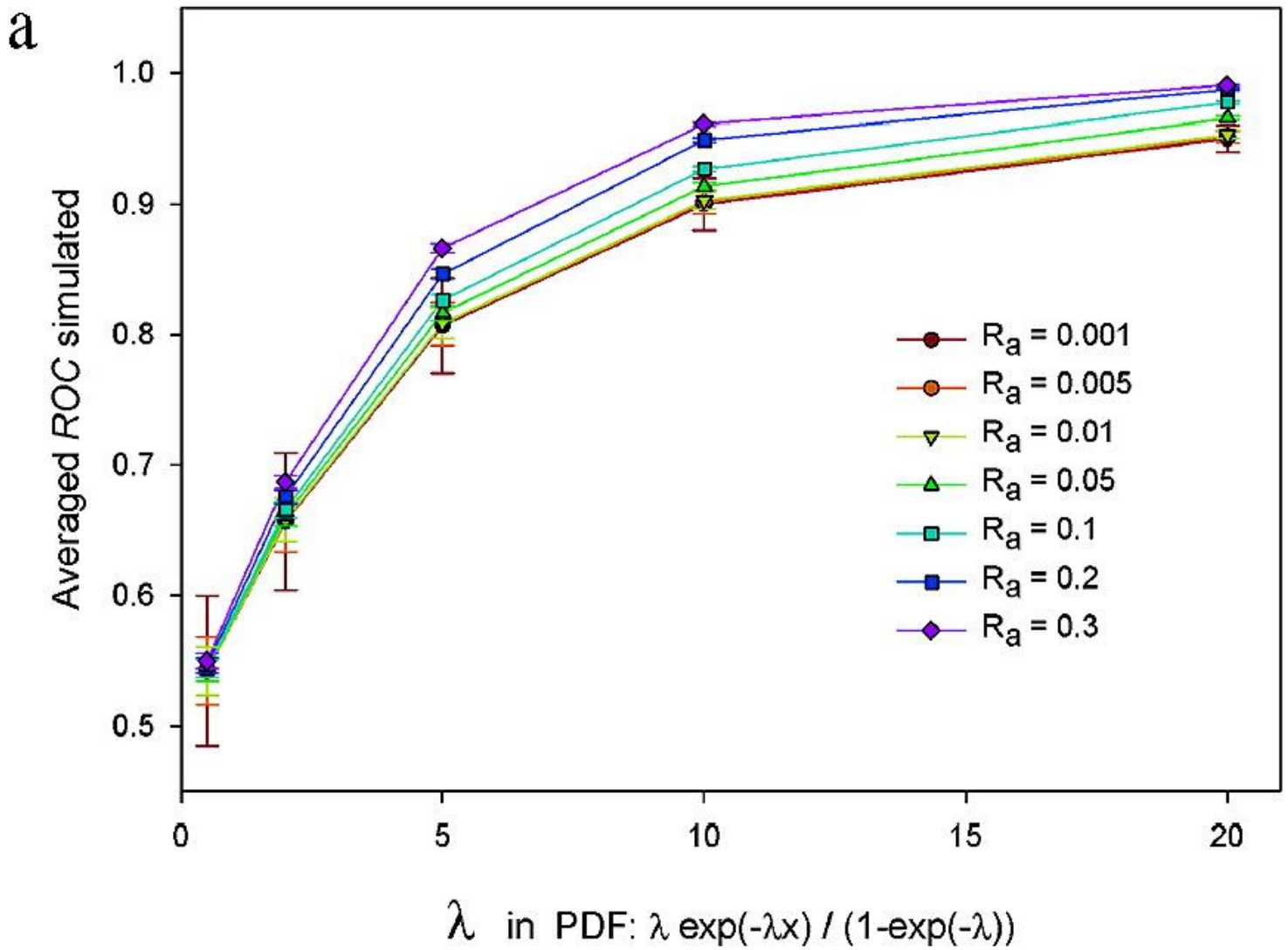
Desvío estándar de métricas de enriquecimiento versus n, N constante



Error relativo de métricas de enriquecimiento temprano versus α (RIE, BEDROC) o 1/x (EF). λ variable



Efecto de saturación



Para prevenir efecto de saturación:

ROC: trabajar con Ra <<1

RIE; *BEDROC*: *trabajar con* $\alpha Ra << 1$

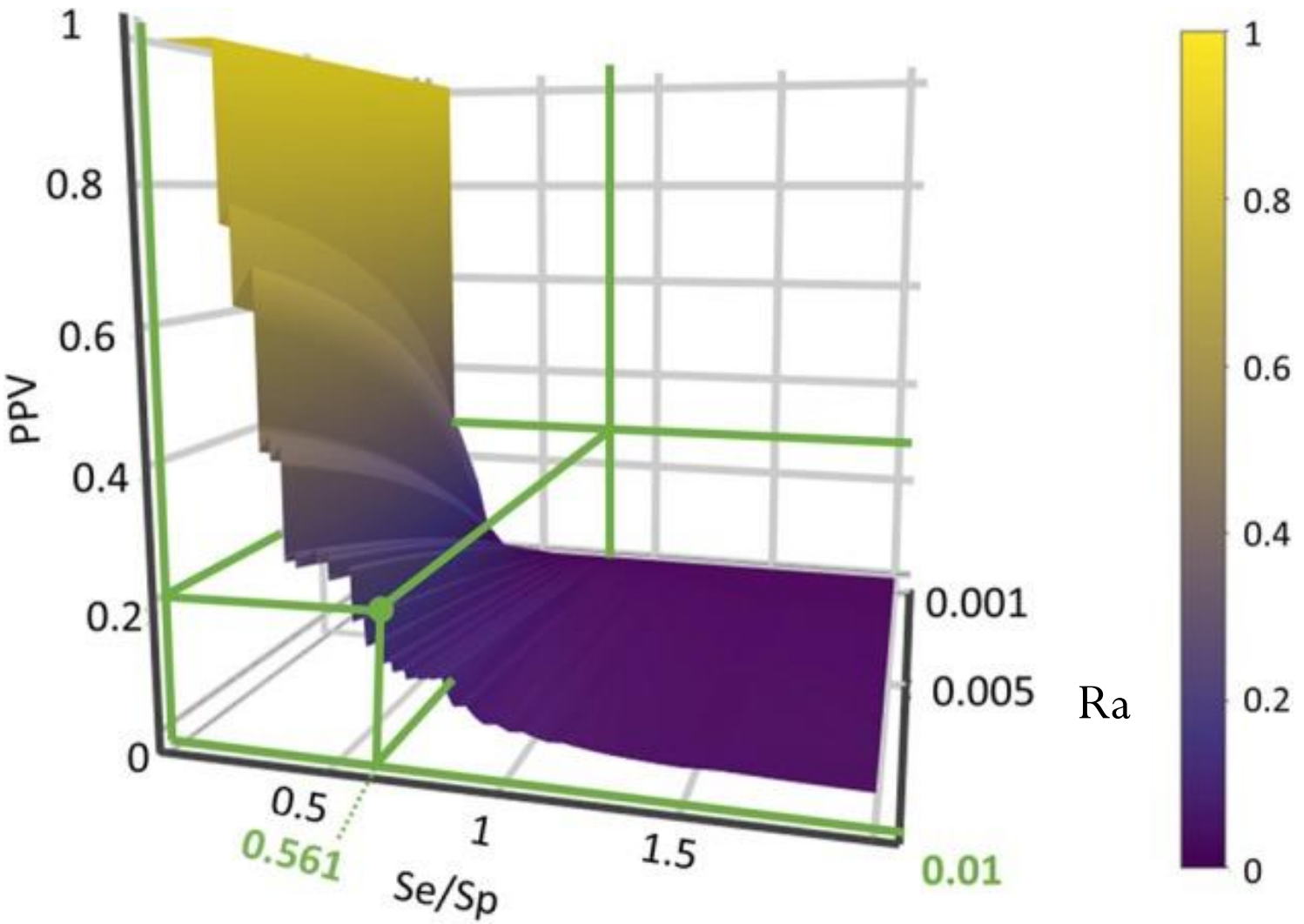
Una consideración práctica: cuántos hits in silico deberé evaluar para obtener un hit confirmado (experimentalmente)?

$$PPV = \frac{Se\ Ra}{Se\ Ra + (1 - Sp)(1 - Ra)}$$

En un experimento de cribado real, Ra no se conoce a priori

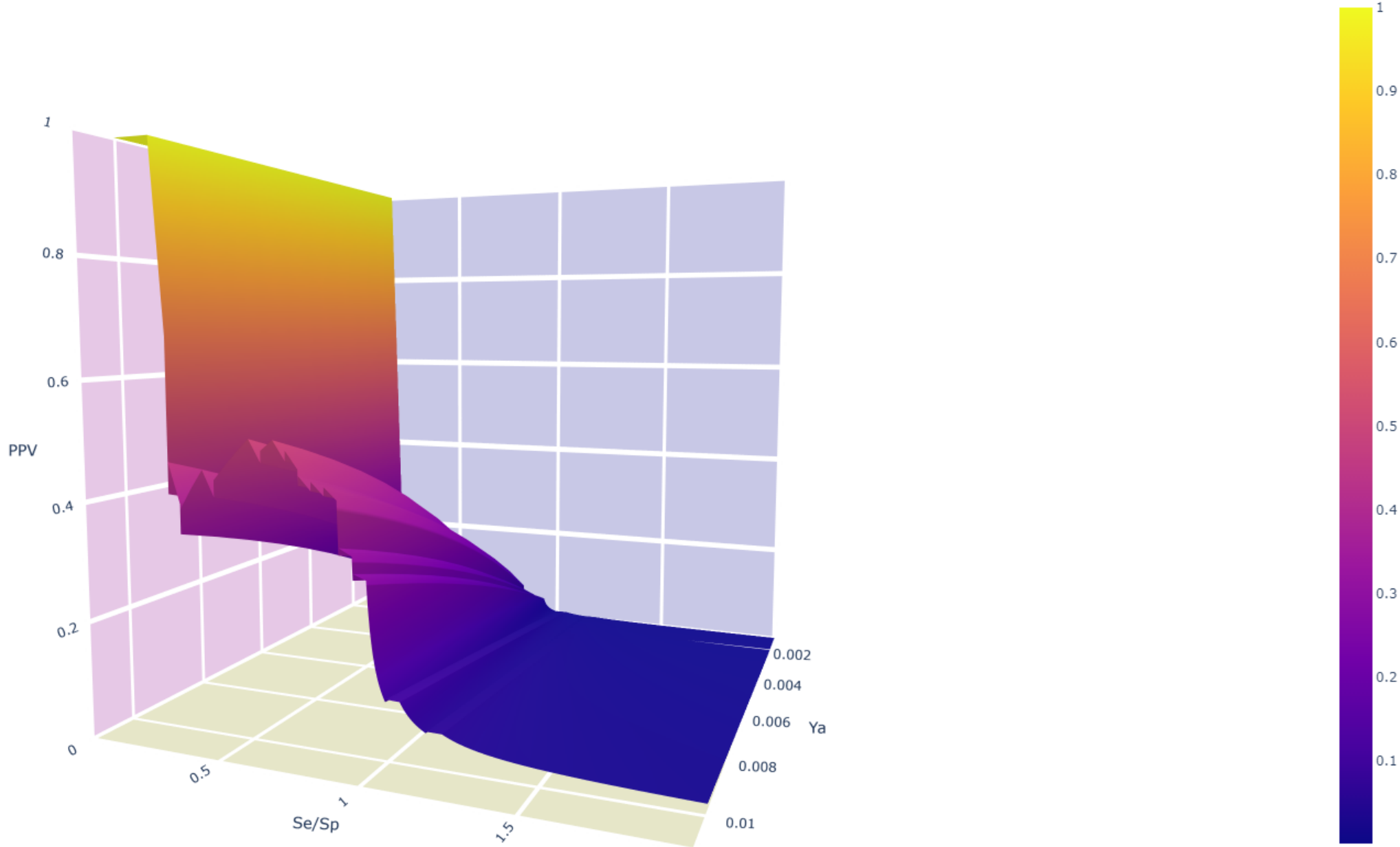
(suele variar, sin embargo, entre 0,01 y 0,005)

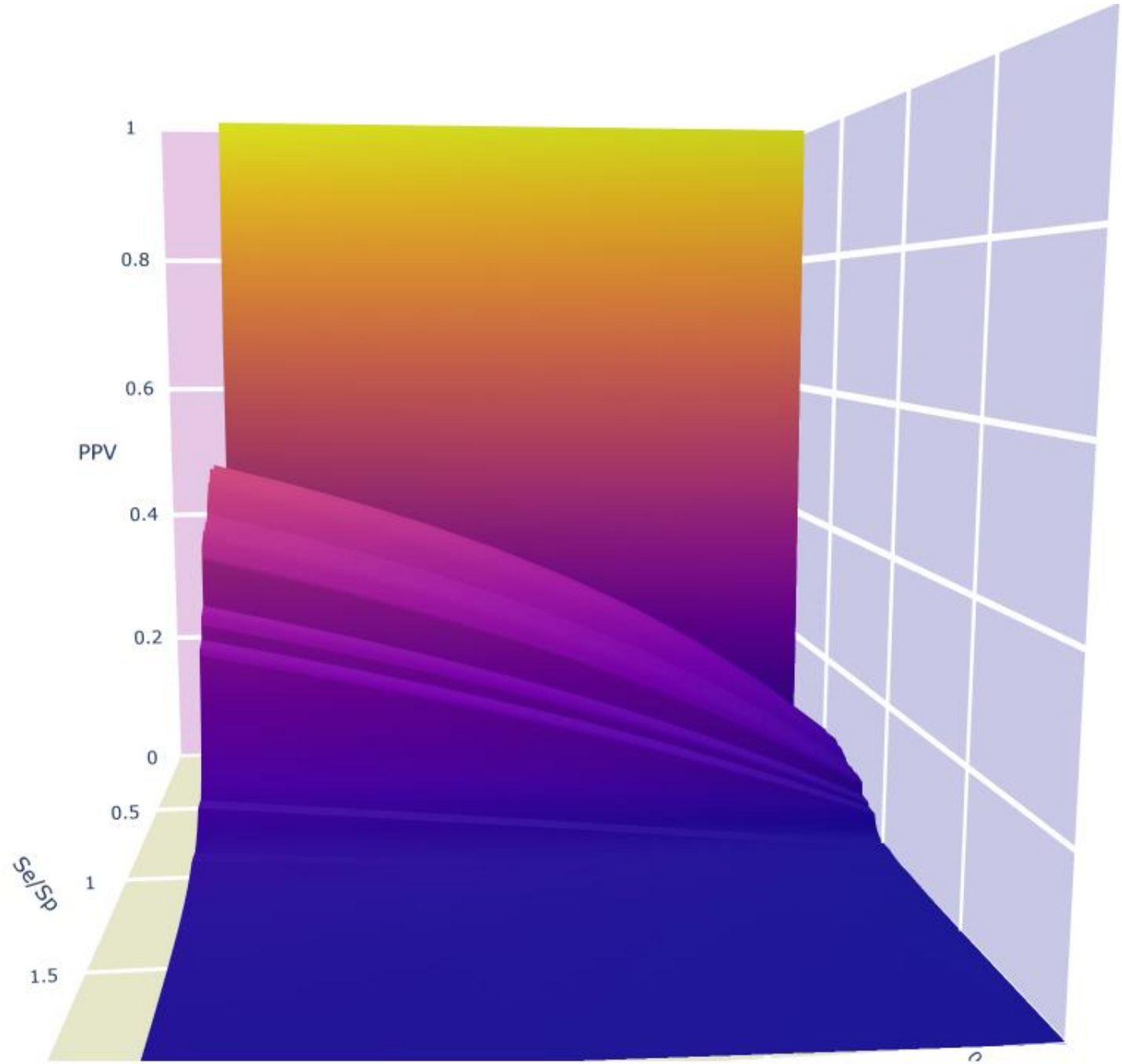
Superficies PPV para optimizar el valor de corte del score

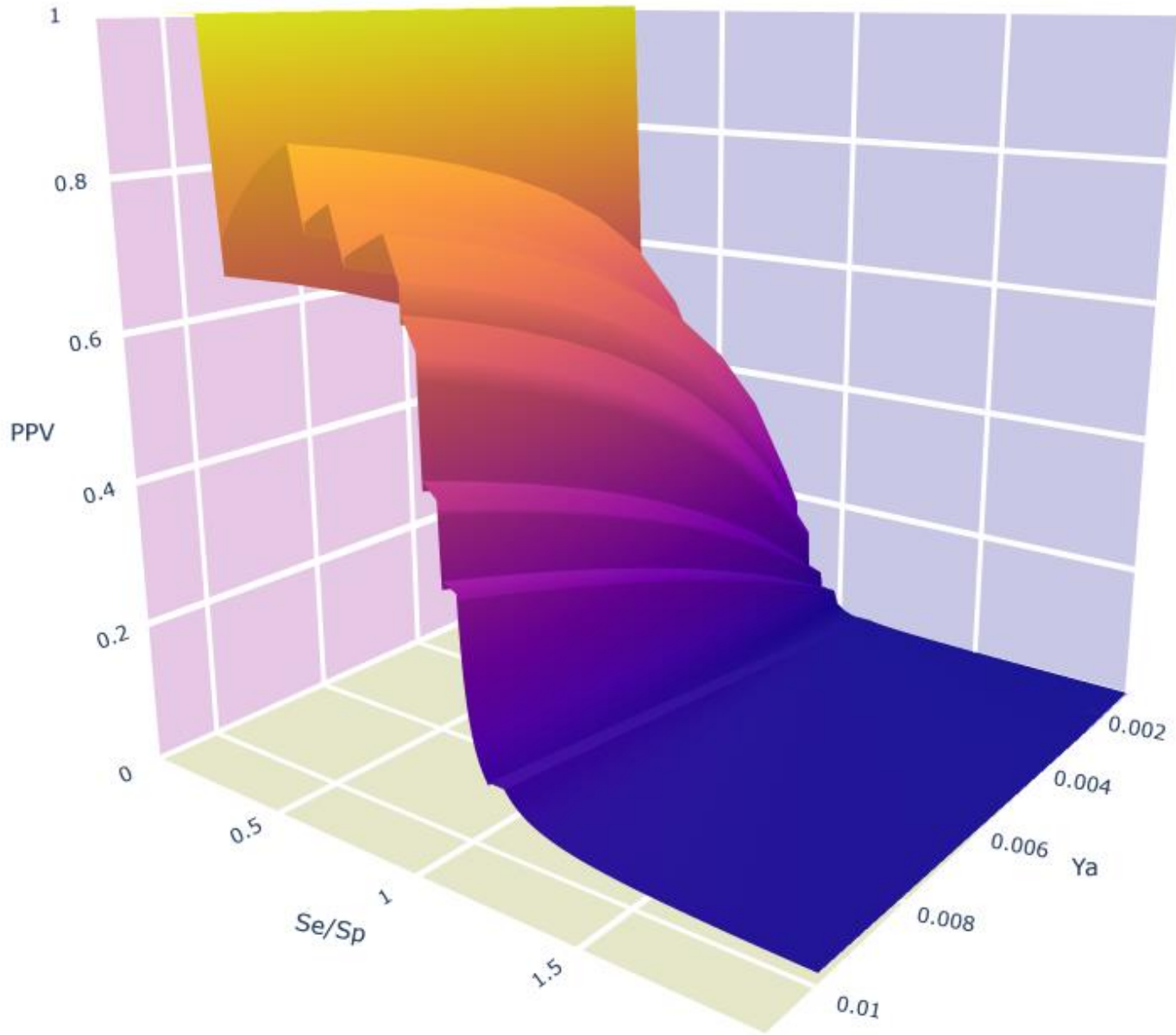


Alberca et al.
Front. Cell. Infect. Microbiol.
2018

Se asume que la relación Se/Sp observada en experimentos de cribado virtual retrospectivo adecuadamente diseñados se mantendrá aproximadamente en experimentos prospectivos


PPV
1
0.8
0.6
0.4
0.2
0
0.5
1
1.5
Se/Sp
0.002
0.004
0.006
0.008
0.01
Ya
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


1
0.8
0.6
PPV
0.4
0.2
0
0.5
1
Se/Sp
1.5

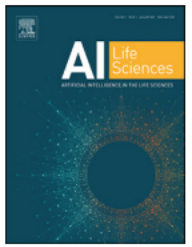

1
0.8
0.6
PPV
0.4
0.2
0
0.5
1
Se/Sp
1.5
0.002
0.004
0.006
Ya
0.008
0.01

Artificial Intelligence in the Life Sciences 2 (2022) 100049

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci



LIDeB Tools: A Latin American resource of freely available, open-source cheminformatics apps

Denis N. Prada Gori, Lucas N. Alberca, Santiago Rodriguez, Juan I. Alice, Manuel A. Llanos, Carolina L. Bellera, Alan Talevi*

Laboratory of Bioactive Compounds Research and Development (LIDeB), Faculty of Exact Sciences, University of La Plata (UNLP), La Plata, Buenos Aires, Argentina

LIDeB tools

Clustering de pequeñas moléculas

iRAPCA
Clustering

SOMoC
Silhouette Optimized Molecular Clustering

CHiCA
Comparative Hierarchical Clustering Algorithms

Generación de señuelos

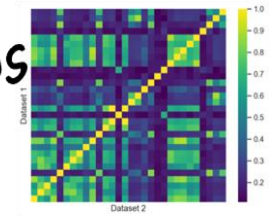
LUDe
LIDeB's Useful Decoys

Predicción de drogabilidad

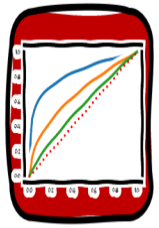
FADRA
FAst Druggability Assesment

Miscelánea

Heatmaps similarity



Metrics



LISTo Standardization

https://lideb.biol.unlp.edu.ar/

+54-0221-4235333 int 41 | lideb@biol.unlp.edu.ar

Facultad de Ciencias Exactas · Departamento de Ciencias Biológicas · UNIVERSIDAD NACIONAL DE LA PLATA

Laboratorio de Investigación y Desarrollo de Bioactivos

Facultad de Ciencias Exactas | Universidad Nacional de La Plata

Inicio | Institucional | Investigación | Docencia | Transferencia | LIDeB tools | Novedades | Contacto

Investigación

Principales Líneas de Investigación Desarrollo y cribado farmacológico de compuestos químicos con potencial terapéutico en el tratamiento de las epilepsias, con énfasis en el tratamiento de la epilepsia refractaria. Desarrollo...

¿Quienes Somos?

Laboratorio de Investigación y Desarrollo de Bioactivos

El Laboratorio de Investigación y Desarrollo de

Investigación

Principales Líneas de Investigación Desarrollo y

Publicaciones

Publicaciones últimos 5 años 2021 Llanos M..

LIDeB tools

About Us

We are a drug discovery team with an interest in the development of publicly available open-source customizable cheminformatics tools to be used in computer-assisted drug discovery. We belong to the Laboratory of Bioactive Research and Development (LIDeB) of the National University of La Plata (UNLP), Argentina. Our research group is focused on computer-guided drug repurposing and rational discovery of new drug candidates to treat epilepsy and neglected tropical diseases.

Coming soon more webapps!

Heatmaps- Similarity

About this App

Build a Heatmap of molecular similarity. These plots of inter-molecular similarity (computed as Tanimoto similarity coefficient using Morgan fingerprints and other molecular fingerprinting systems) allow for a fast, visual inspection of the molecular diversity of the datasets, and also preliminary detection of clusters within a dataset. The resulting plots are downloadable as .png files through a simple right click on your mouse!.

App Publication

iRaPCA-Clustering

About this App

iRaPCA Clustering is based on an iterative combination of the random subspace approach (feature bagging), dimensionality reduction through Principal Component Analysis (PCA) and the k-means algorithm. The optimal number of clusters k and the best subset of descriptors are selected from plots of silhouette coefficient against different k values and subsets. Different validation metrics can be downloaded once the process is finished. A number of graphs may be built and readily downloaded through a simple click.

App Publication

LuDe

SOMoC Clustering

- Importante financiamiento de Apoyo Dravet para el LIDeB
- Nuevos avances en la lucha contra la toxoplasmosis
- LIDeB tools
- 3rd International online congress on Dravet Syndrome and Refractory Epilepsy
- Seminario en IPMontevideo

Líneas de Investigación

Sitios de Interés

- Facultad de Ciencias Exactas
- Universidad Nacional de La Plata
- CONICET
- INDRE Network
- Departamento Ciencias Biológicas

Seguinos en:

Heatmaps- Similarity

About this App

Build a Heatmap of molecular similarity. These plots of inter-molecular similarity (computed as Tanimoto similarity coefficient using Morgan fingerprints and other molecular fingerprinting systems) allow for a fast, visual inspection of the molecular diversity of the datasets, and also preliminary detection of clusters within a dataset. The resulting plots are downloadable as .png files through a simple right click on your mouse!.

App Publication

iRaPCA-Clustering

About this App

iRaPCA Clustering is based on an iterative combination of the random subspace approach (feature bagging), dimensionality reduction through Principal Component Analysis (PCA) and the k-means algorithm. The optimal number of clusters k and the best subset of descriptors are selected from plots of silhouette coefficient against different k values and subsets. Different validation metrics can be downloaded once the process is finished. A number of graphs may be built and readily downloaded through a simple click.

App Publication

Seguinos en:

web app → App Publication → Artículo publicado

Upload your SMILES

Upload a TXT file with one SMILES per line

Drag and drop file here

Limit 200MB per file • TXT

Browse files

Example TXT input file

Check to change the default configuration

Physicochemical features limits

Molecular weight

5

0 10

LogP

0.50

0.00 0.50

Rotable bonds

1

0 3

Num of H Acceptors

1

0 2

Num of H Donors

1

0 2

LIDeB tools

LUDe LIDeB's Useful Decoys

ut Us

drug discovery team with an interest in the development of publicly available open-source customizable cheminformatics tools to be used in computer-assisted drug discovery. We belong to the Laboratory of Bioactive Research and Development (LIDeB) of the National University of La Plata (UNLP), Argentina. Our research group is focused on computer-guided drug re
diseases.

eB Tools - LUDe

DEB's Useful Decoys) is a WebApp that generates, from a set of active compounds, decoys (putative inactive compounds) which can be used to retrospectively validate virtual screening tools/protocols. Decoys are molecules that have not been tested against a molecular target of interest but due to their structural features are presumably not prone to bind the target wi
ith the known active compounds in relation to certain general physicochemical properties (e.g., molecular weight, log P, and others) but are topologically different from the query compounds. LUDe is conceptually similar to the Directory of Useful Decoys enhanced, but additional filters have been serially implemented to assure the topological dissimilarity between the d

leb App, decoys are obtained through four sequential steps:

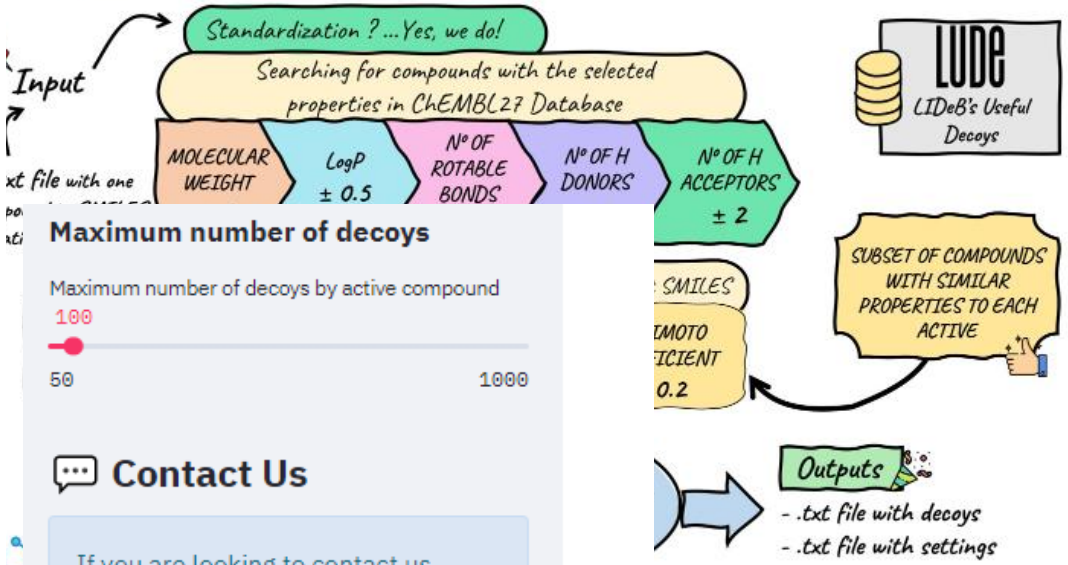
ching molecules with similar physicochemical properties of the input active molecules in a curated ChEMBL database.

ring the selected molecules by dissimilarity against each individual input molecule.

domly selecting a desired number of decoys for each individual input molecule.

ring the selected molecules by the dissimilarity against all the input molecules (set of active compounds used as query). The decoys will have low Tanimoto similarity with the input compounds, and also low Maximum Common Substructure (MCS) ratio and distinctive molecular frameworks (Murcko scaffold). All in all, these three serial filters assure that the decoys will
hermore, dissimilarity to the remaining active compounds is also checked. Finally, you can download a file with your decoys.

workflow summarizes the steps performed by this method:



Maximum number of decoys

Maximum number of decoys by active compound

100

50 1000

Contact Us

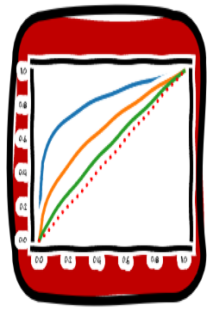
If you are looking to contact us, please or Twitter

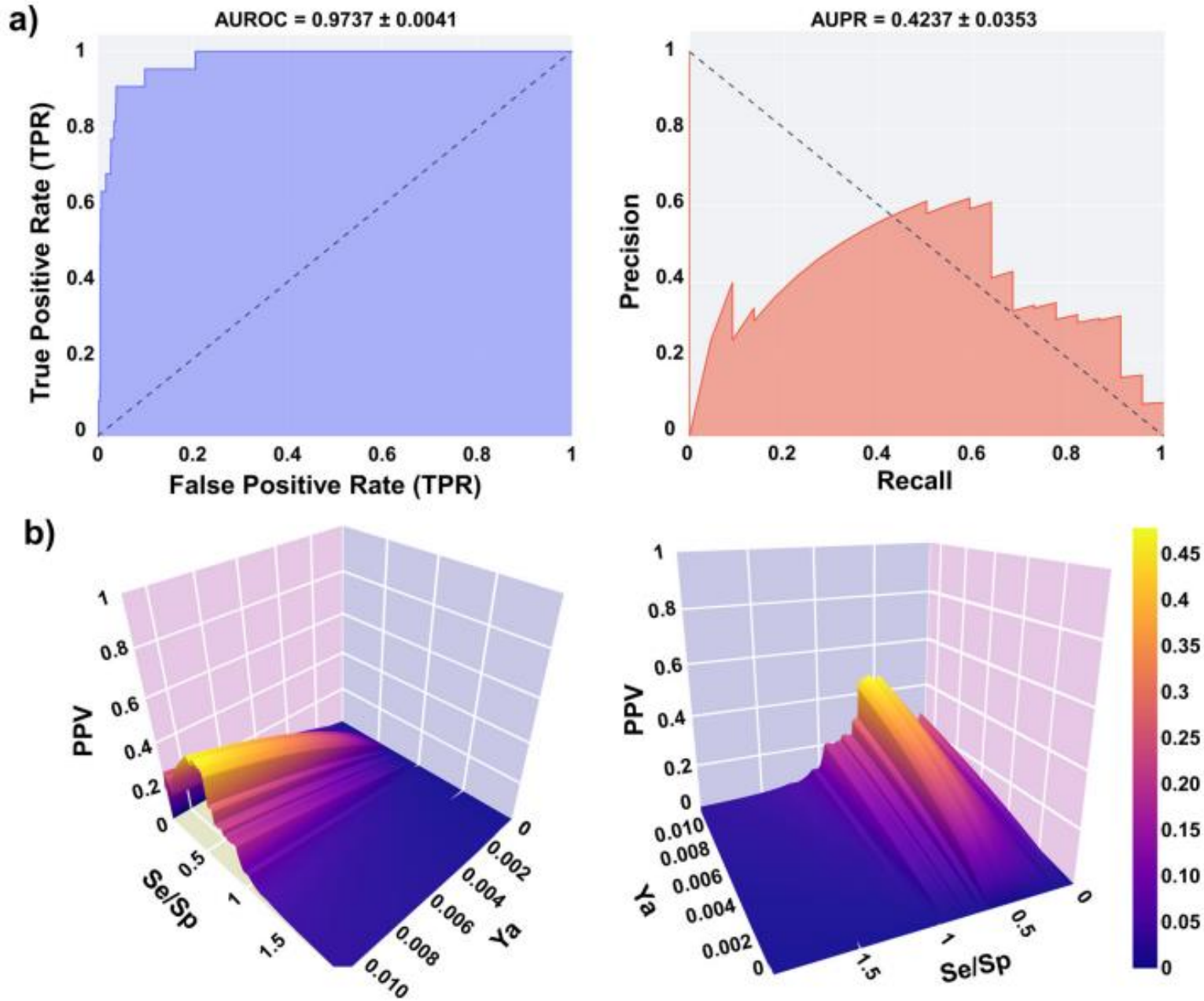
Streamlit

el código también está disponible en github.com

https://github.com/lideb

Metrics




a)
AUROC = 0.9737 ± 0.0041
True Positive Rate (TPR)
False Positive Rate (TPR)
AUPR = 0.4237 ± 0.0353
Precision
Recall
b)
PPV
Se/Sp
Ye
PPV
Se/Sp
Ye

Generación de señuelos

estos no son buenos decoys

estos tampoco son buenos decoys

los decoys son compuestos que en cierto sentido muy general son parecidos a los compuestos activos que se conocen contra determinado blanco farmacológico, pero que son topológicamente distintos

DUD-E

Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking

Michael M. Mysinger,[†] Michael Carchia,[†] John. J. Irwin,*[,†] and Brian K. Shoichet*[,†]

[†]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California 94158-2330, United States

DUD-Z

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim

Article

Property-Unmatched Decoys in Docking Benchmarks

Reed M. Stein, Ying Yang, Trent E. Balius, Matt J. O'Meara, Jiankun Lyu, Jennifer Young, Khanh Tang, Brian K. Shoichet,* and John J. Irwin*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 699–714

Read Online

DeepCoy

Bioinformatics, 37(15), 2021, 2134–2141
doi: 10.1093/bioinformatics/btab080
Advance Access Publication Date: 3 February 2021
Original Paper

OXFORD

Structural bioinformatics

Generating property-matched decoy molecules using deep learning

Fergus Imrie [1], Anthony R. Bradley[2] and Charlotte M. Deane [1,*]

DecoyFinder

BIOINFORMATICS APPLICATIONS NOTE

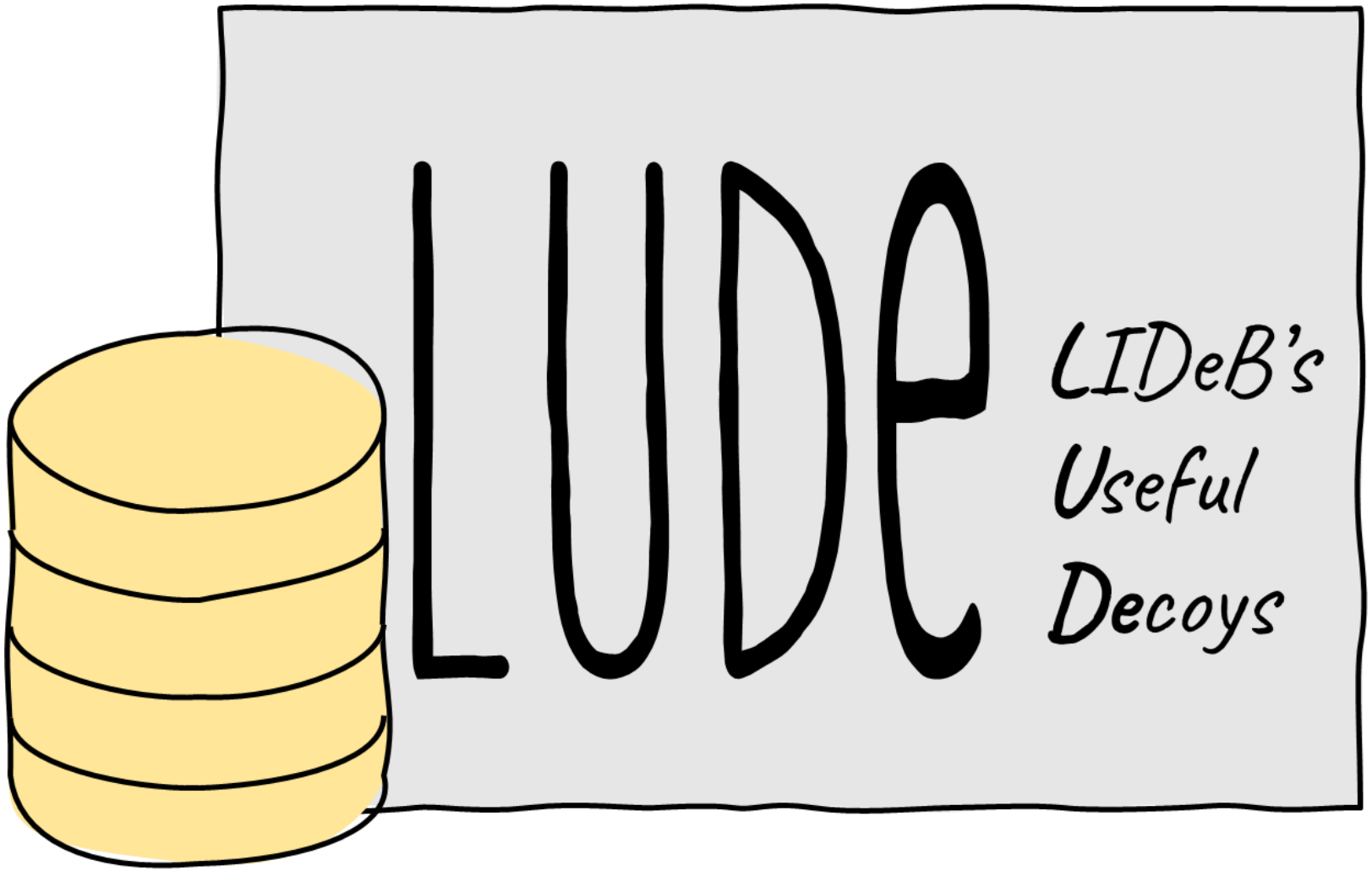
Vol. 28 no. 12 2012, pages 1661–1662
doi:10.1093/bioinformatics/bts249

Structural bioinformatics

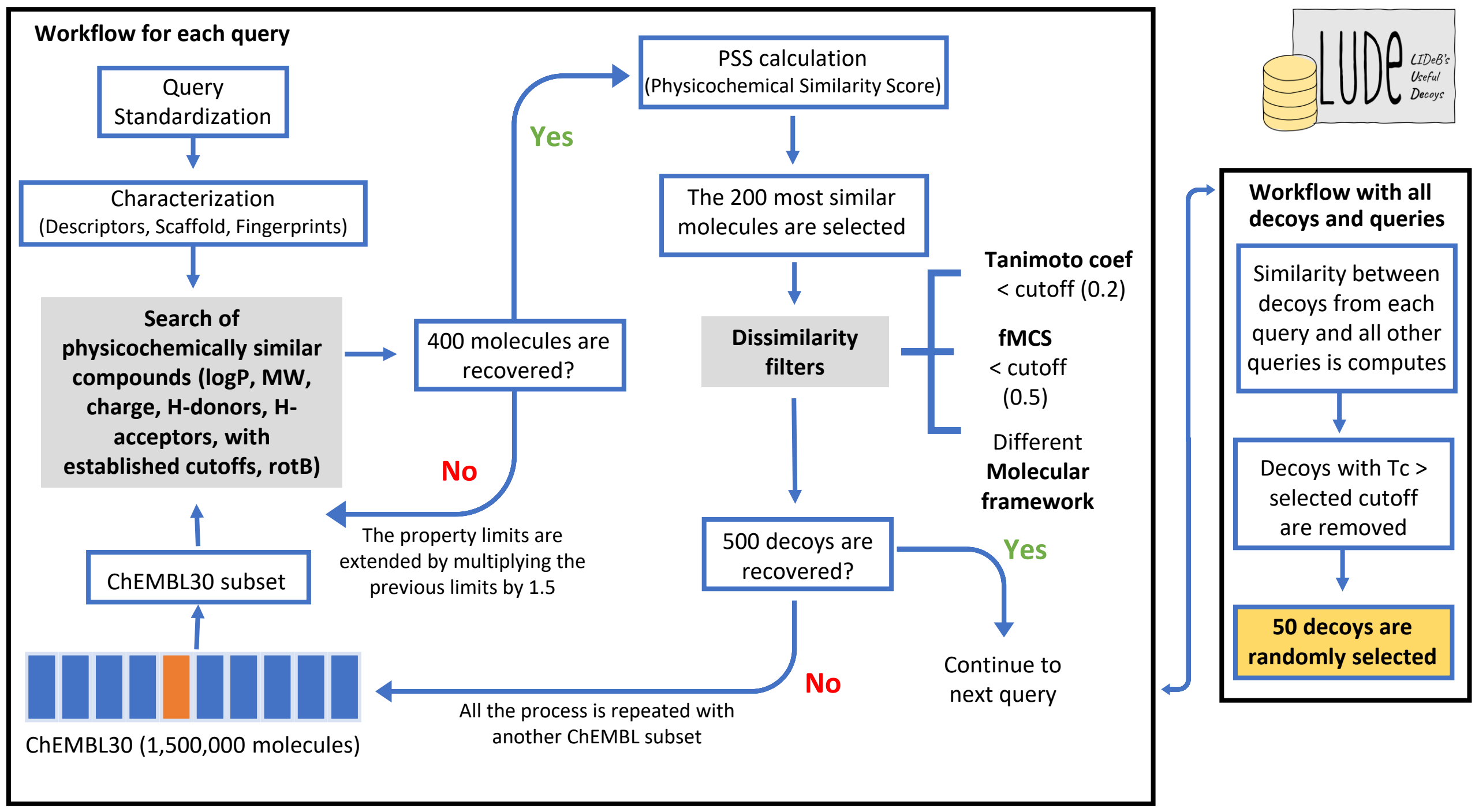
Advance Access publication April 26, 2012

DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets

Adrià Cereto-Massagué[1], Laura Guasch[1], Cristina Valls[1], Miquel Mulero[1], Gerard Pujadas[1,2] and Santiago Garcia-Vallvé[1,2,*]



LIDeB
tools


Workflow for each query
Query Standardization
Characterization (Descriptors, Scaffold, Fingerprints)
Search of physicochemically similar compounds (logP, MW, charge, H-donors, H-acceptors, with established cutoffs, rotB)
ChEMBL30 subset
ChEMBL30 (1,500,000 molecules)
400 molecules are recovered?
Yes
No
The property limits are extended by multiplying the previous limits by 1.5
PSS calculation (Physicochemical Similarity Score)
The 200 most similar molecules are selected
Dissimilarity filters
Tanimoto coef < cutoff (0.2)
fMCS < cutoff (0.5)
Different Molecular framework
500 decoys are recovered?
Yes
Continue to next query
No
All the process is repeated with another ChEMBL subset
LUDe
LIDeB's Useful Decoys
Workflow with all decoys and queries
Similarity between decoys from each query and all other queries is computes
Decoys with Tc > selected cutoff are removed
50 decoys are randomly selected

BENCHMARKING

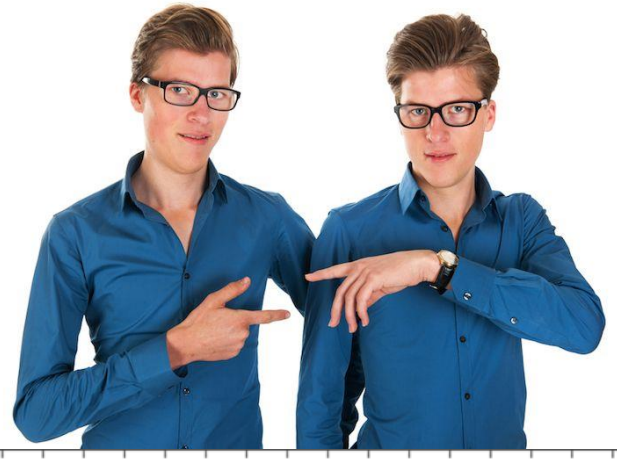
Validación de decoys

- **Sets de datos**:
  - 102 blancos farmacológicos extraídos de DUD-E
    - 50 decoys por compuesto activo
- **Métricas**:
  - DOE score: para cada query (activo) se calcula la distancia en el espacio de las 6 matching properties normalizadas, contra todos los demás (activos y señuelos). Se etiquetan como 1 a los activos y como 0 a los inactivos y se construye la curva ROC. Luego se calcula el promedio de *ROC* – 0,5 para todos los activos.
  - Doppelganger score. Mácima similitud de Tanimoto un part cualquiera de los queries y los decoys.

el DOE score fue más bajo para señuelos LUDe para el 65% de los blancos para señuelos LUDe

Doppelganger scores of the original DUD-E set (blue) compared to the LUDE generated decoys (red)

el Doppelganger score fue más bajo para el 85% de los blancos para señuelos LUDe; el Doppelganger score promedio de Doppelganger (para los 102 targets) fue de 0,23 para LUDe y 0,25 para DUD-E



Ejemplo: **Dataset fabp4** (Fatty Acid Binding Protein 4)

OUTPUT 2

Generated_decoys.csv

SMILE	Query
Clc1ccc(C(CCCOc2ccc(Br)cc2)Cn2ccnc2)c(Cl)c1	Query_1
CCOC(=O)CC[C@H](O[C@H]1O[C@@H]2O[C@]3(C)CC[C@H]4[C@H](C)CC[C@@H]([C@H]1C)[C@]42OO3)c1ccccc1	Query_1
CC(C)(F)c1noc(N2CCC(COC3CC=C(c4ccc(S(C)(=O)=O)cc4)CC3)CC2)n1	Query_1
COc1cc(c2ccc(COc3cc([C@@H](CC(=O)[O-])C4CC4)ccn3)cc2[C@@H]2CCCC2(C)C)c(F)cn1	Query_1
O=C([O-])CCc1ccccc1/C=C1/C2CCC(O2)C1c1nc(C(=O)NCCCCC2CCCCC2)co1	Query_1
O=C1[C@H](N2C(=O)c3ccccc3C2=O)CN1[C@H](COCc1ccccc1)c1ccc(OCc2ccccc2)cc1	Query_1
CCCc1cnc(N2CCC(C3Cc4cc(C5CCN(S(=O)(=O)CCC)CC5)ccc4O3)CC2)nc1	Query_1
Cc1ccc(C(C)C)c(SC2=C([O-])O[C@@](CCc3ccccc3)(c3ccccc3)CC2=O)c1	Query_1
O=C(CC(CC(=O)c1ccc(Br)cc1)c1cccnc1)c1ccc(Br)cc1	Query_1
CN(Cc1ccnc(N2CCCC2)c1)C(=O)c1cnc(Oc2ccc3c(c2)CCC(c2ccccc2)O3)s1	Query_1
Cc1ccc(c2nc3sc(C)nn3c2COCc2cn(Cc3ccc(F)cc3)nn2)cc1	Query_1
CC(C)c1ccc(C2CC(=O)c3ccc(OCc4cn(Cc5ccc(Br)cc5)nn4)cc3O2)cc1	Query_1
----------	Query_X

OUTPUT 3

Decoys_settings.csv

Decoys generated at:	8/8/2022
Physicochemical features limits:	
MW	+/- 20
logP	+/- 0.5
Num_rotable_bonds	+/- 1
Num_H_acceptors	+/- 1
Num_H_donors	+/- 1
Topological features ---> Dissimilarty conditions	
Morgan Fingerprints	
Fingerprint radio:	2
Fingerprint lenght:	1024
Limit of fraction of the Maximum Common Substructure:	0.5
Decoys with different framework:	True
Max Tc similarity between decoys and other actives:	0.2
Max number of decoys by loaded molecule:	50

NO USAR señuelos generados utilizando criterios de exclusión basados en similitud molecular para evaluar métodos basados en similitud molecular.

Los creadores de DUD-E indican que sus señuelos deben usarse exclusivamente para evaluar el desempeño de protocolos de docking molecular.

Clustering

Clustering: aplicaciones

Muestreo representativo de quimiotecas / bases de datos químicas.
Por ejemplo, para particionar representativamente un dataset en training, test, validation sets

Muestreo representativo de hits in silico

Aprendizaje automático no supervisado

Clustering: clasificación

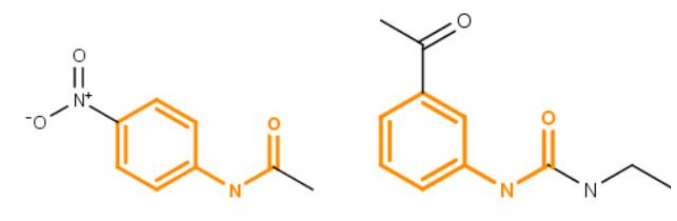
Jerárquico (métodos aglomerativos o divisivos): Single linkage, complete linkage, average linkage, Chemaxon's Library MCS.

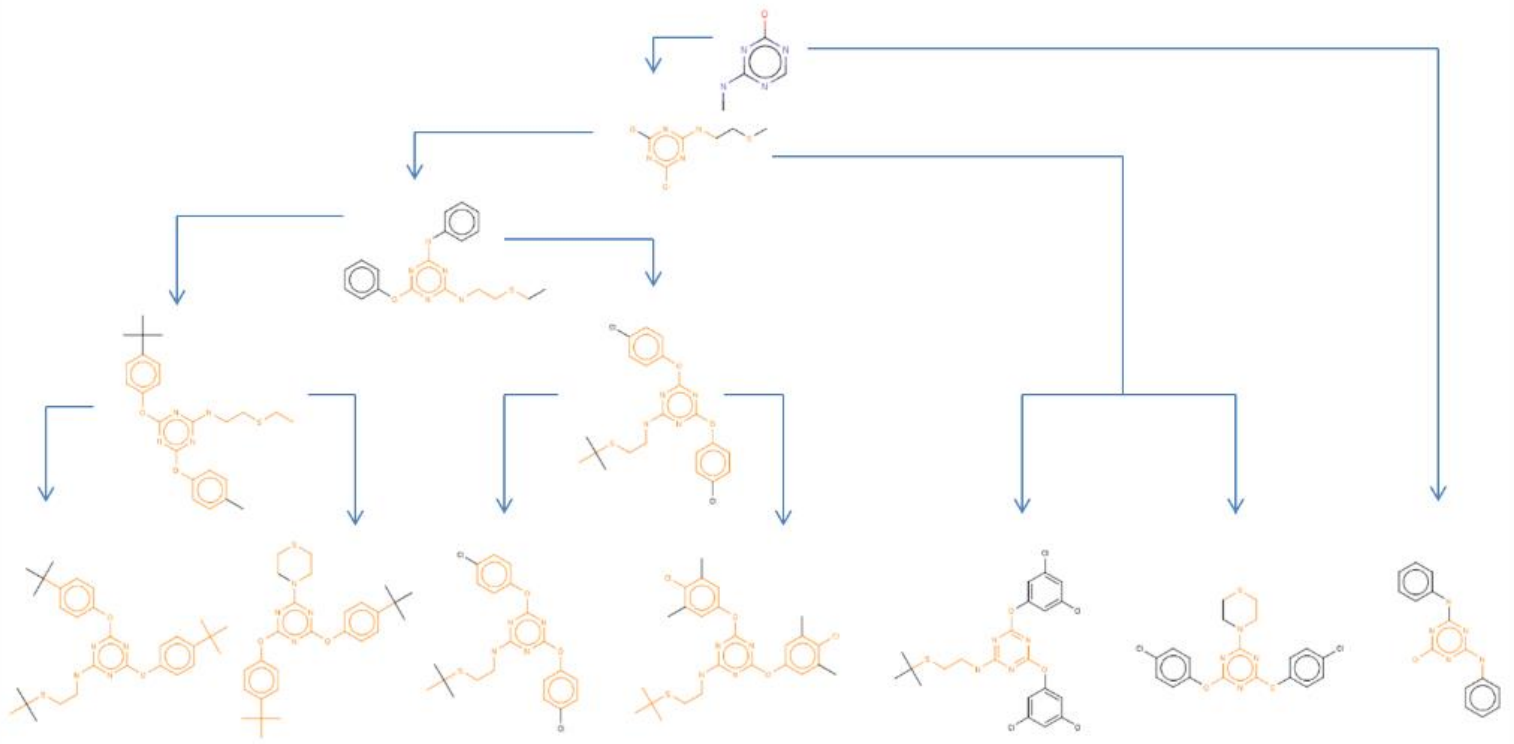
No jerárquico: Esferas de exclusión, K-means.

Híbridos: Library MCS + k-means.

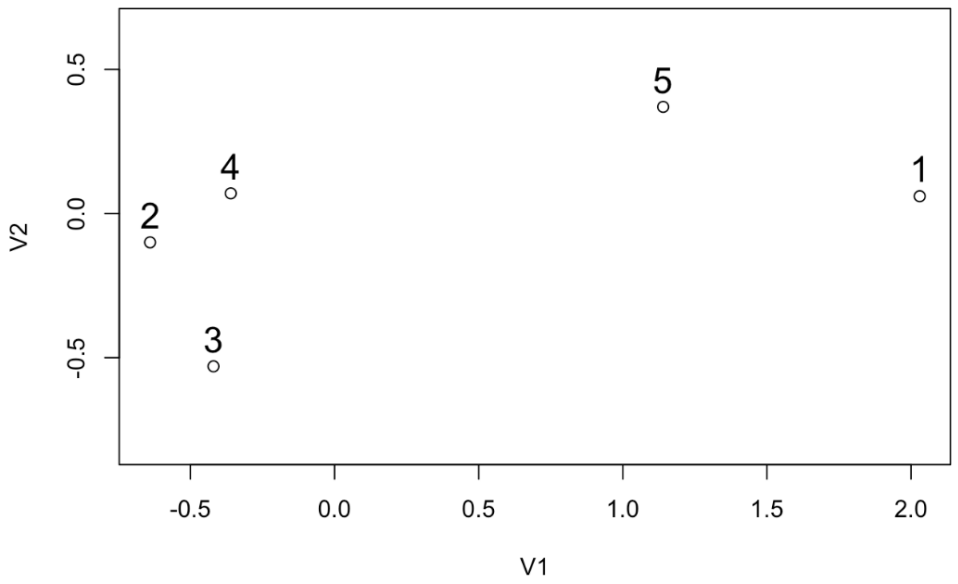
Library MCS

Se basa en agrupar las moléculas en función de su máxima subestructura común. El usuario puede definir el tamaño mínimo de la MCS y si los anillos se considerarán como tales o puede romperse. Se inicia agrupando las estructuras que presentan una MCS de mayor tamaño.





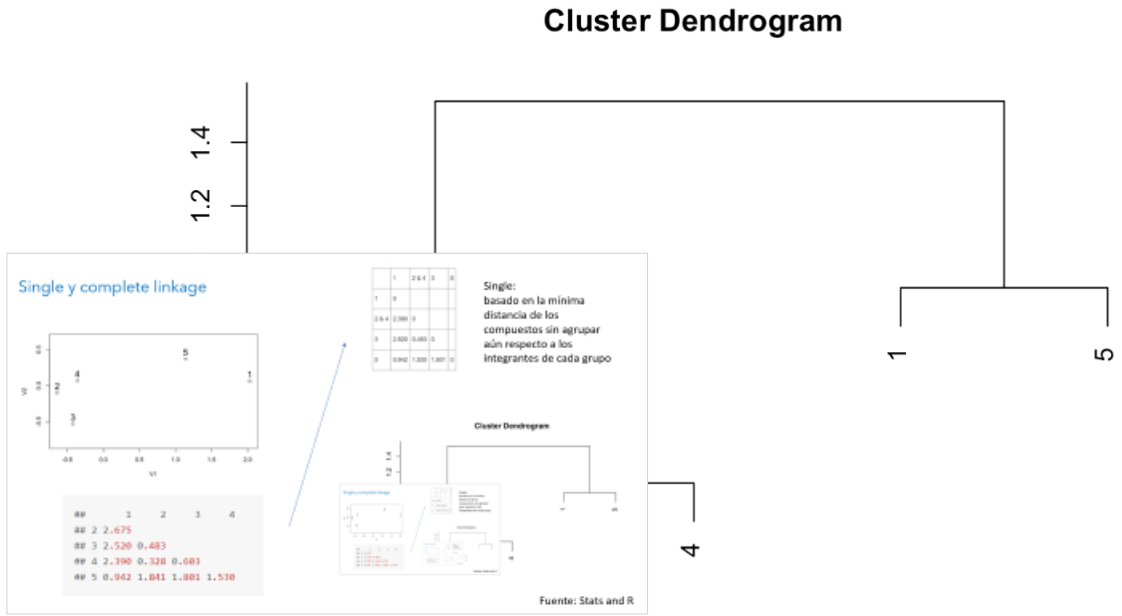
Single y complete linkage




```
##         1     2     3     4
## 2 2.675
## 3 2.520 0.483
## 4 2.390 0.328 0.603
## 5 0.942 1.841 1.801 1.530
```

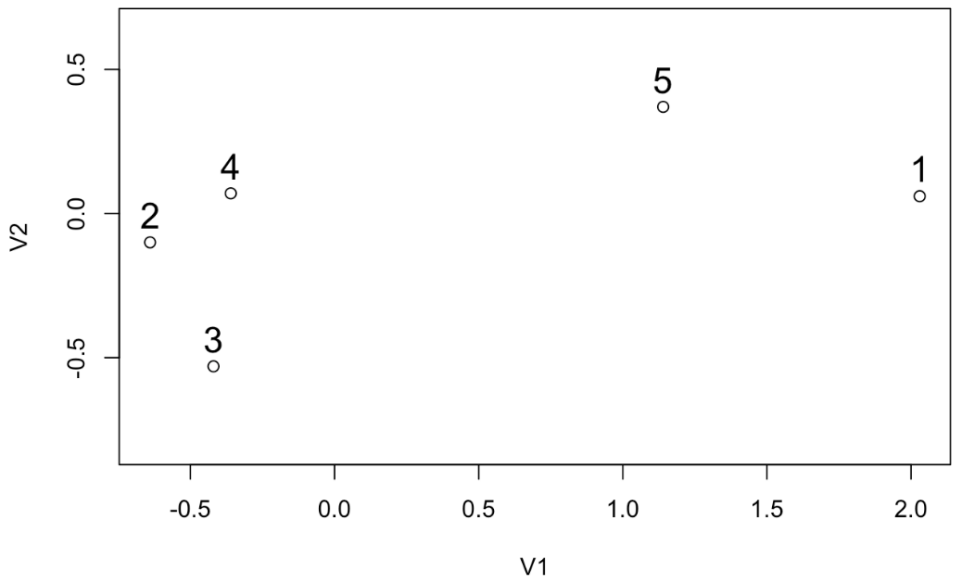
	1	2 & 4	3	5
1	0			
2 & 4	2.390	0		
3	2.520	0.483	0	
5	0.942	1.530	1.801	0

Single:
basado en la mínima distancia de los compuestos sin agrupar aún respecto a los integrantes de cada grupo



Fuente: Stats and R

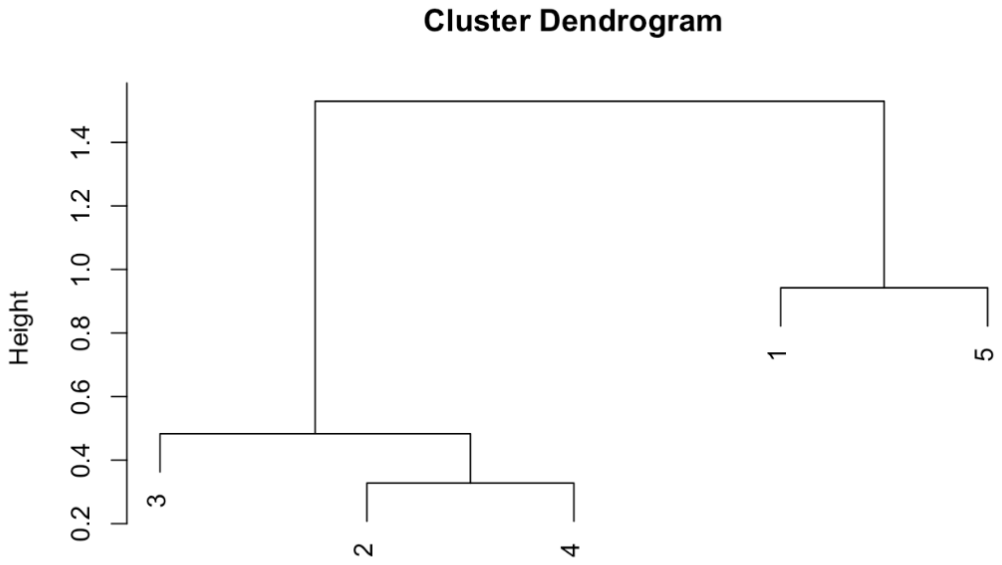
Single y complete linkage



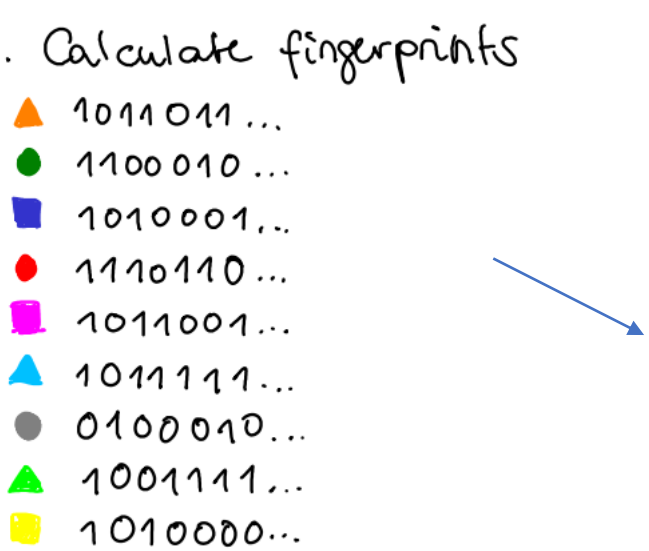

```
##       1     2     3     4
## 2 2.675
## 3 2.520 0.483
## 4 2.390 0.328 0.603
## 5 0.942 1.841 1.801 1.530
```

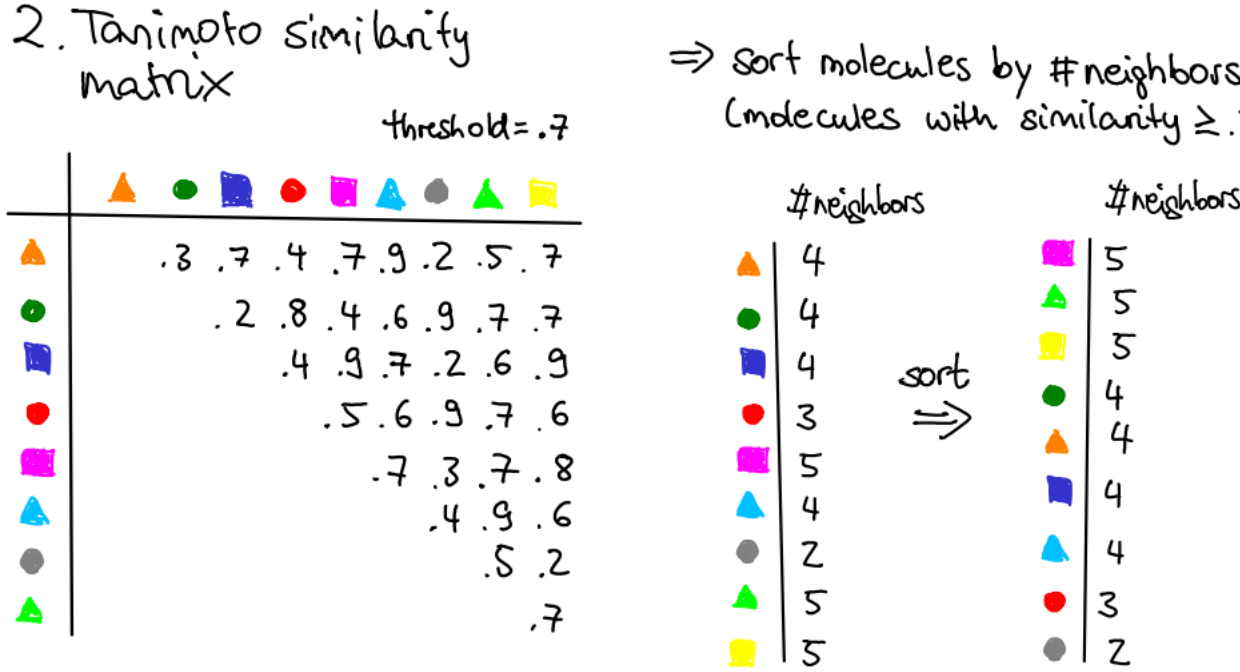
	1	2 & 4	3	5
1	0			
2 & 4	2.675	0		
3	2.520	0.603	0	
5	0.942	1.841	1.801	0

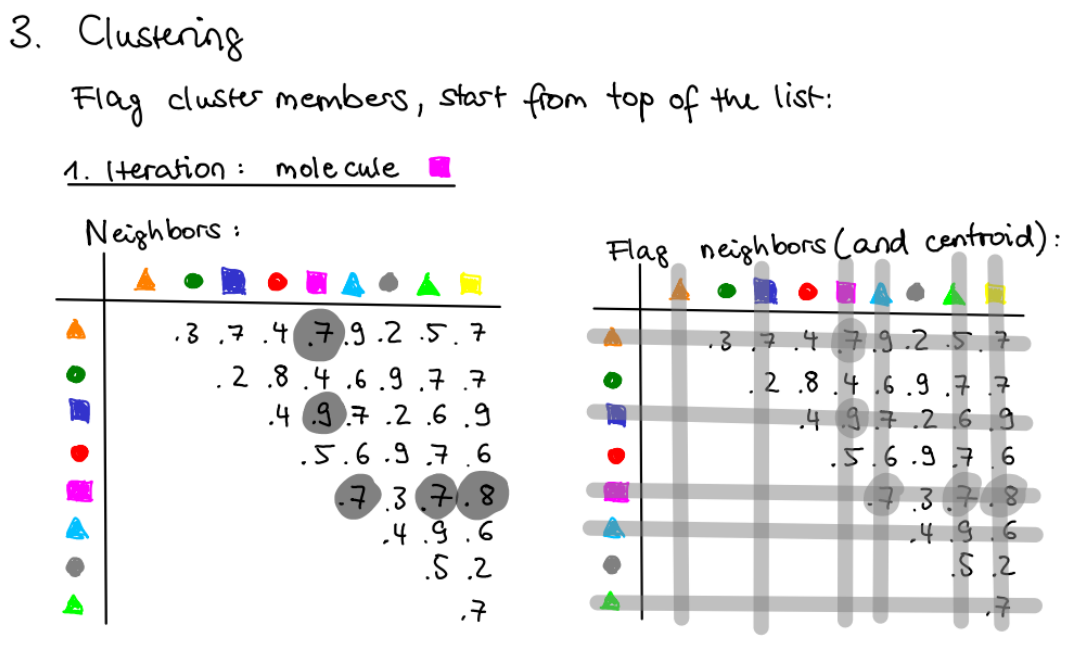
Complete:
basado en la máxima distancia de los compuestos sin agrupar aún respecto a los integrantes de cada grupo

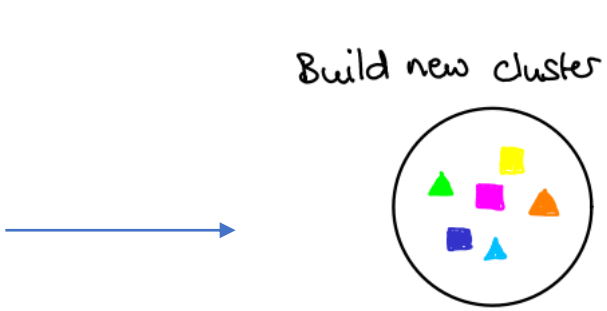


Butina



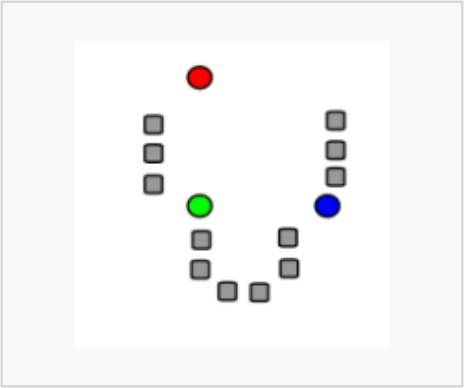




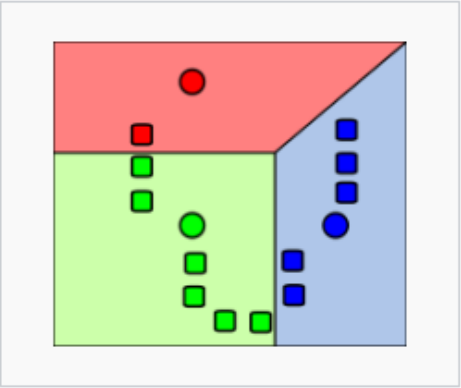


Fuente: TechOpenCADD

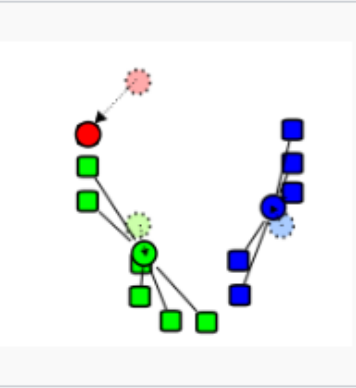
K-means



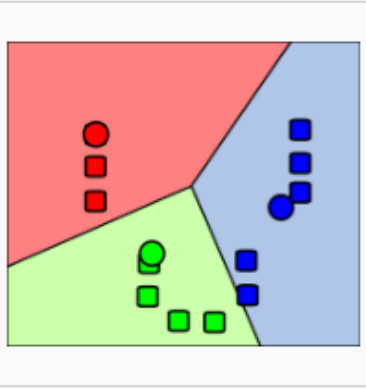
1) *k* centroides iniciales (en este caso *k*=3) son generados aleatoriamente dentro de un conjunto de datos (mostrados en color).



2) *k* grupos son generados asociándole el punto con la media más cercana. La partición aquí representa el diagrama de Voronoi generado por los centroides.



3) EL centroide de cada uno de los *k* grupos se recalcula.



4) Pasos 2 y 3 se repiten hasta que se logre la convergencia.

iRaPCA

Clustering

LIDeB tools

Iterative Random Subspace PCA clustering (iRaPCA)

- iRaPCA clustering se basa en una combinación iterativa de subespacios aleatorios (feature bagging), reducción de dimensionalidad mediante Análisis de Componentes principales (PCA) y el algoritmo K-means algorithm
- El valor medio de silhouette se utiliza habitualmente como métrica de validación del clustering. Para cada elemento del conjunto de datos, s(i) puede ser calculado como b(i)-a(i)/max{(a(i),b(i)}, donde a(i) es la distancia media entre i y todos los demás puntos de datos en el mismo cluster y b(i) es la distancia media más pequeña de i a todos los puntos en cualquier otro grupo.

iRaPCA Clustering

Input files

- .csv file with SMILES notation

Encoding

Standardization of structures with MolVS
Calculation of Mordred molecular **descriptors**

Feature exploration and dimensionality reduction

Removal of low variance descriptors
Generation of descriptor subsets by Random Subspace
Dimensionality reduction by correlation filters and PCA

Clustering

- **K-means** clustering in each subset using a range of K values
- Clustering evaluation by **SIL score**
- Selection of subset and K with the best SIL score

Iteration

The ratio of molecules per cluster/total number of molecules is computed
Those clusters that exceed an user-defined ratio are subjected to a new clustering round. Those clusters that do not exceed the chosen ratio are kept as they are

Finish process

The process ends when no cluster exceeds the specified ratio or when a user-defined number of iterations is completed.
The iterations may be avoided by choosing a number of iterations equal to one

Output files

- .csv files with all the molecules, their assigned clusters and the cluster distribution
- .html graphics (SIL vs K plot, scatterplot, sunburn graph, bar graph)

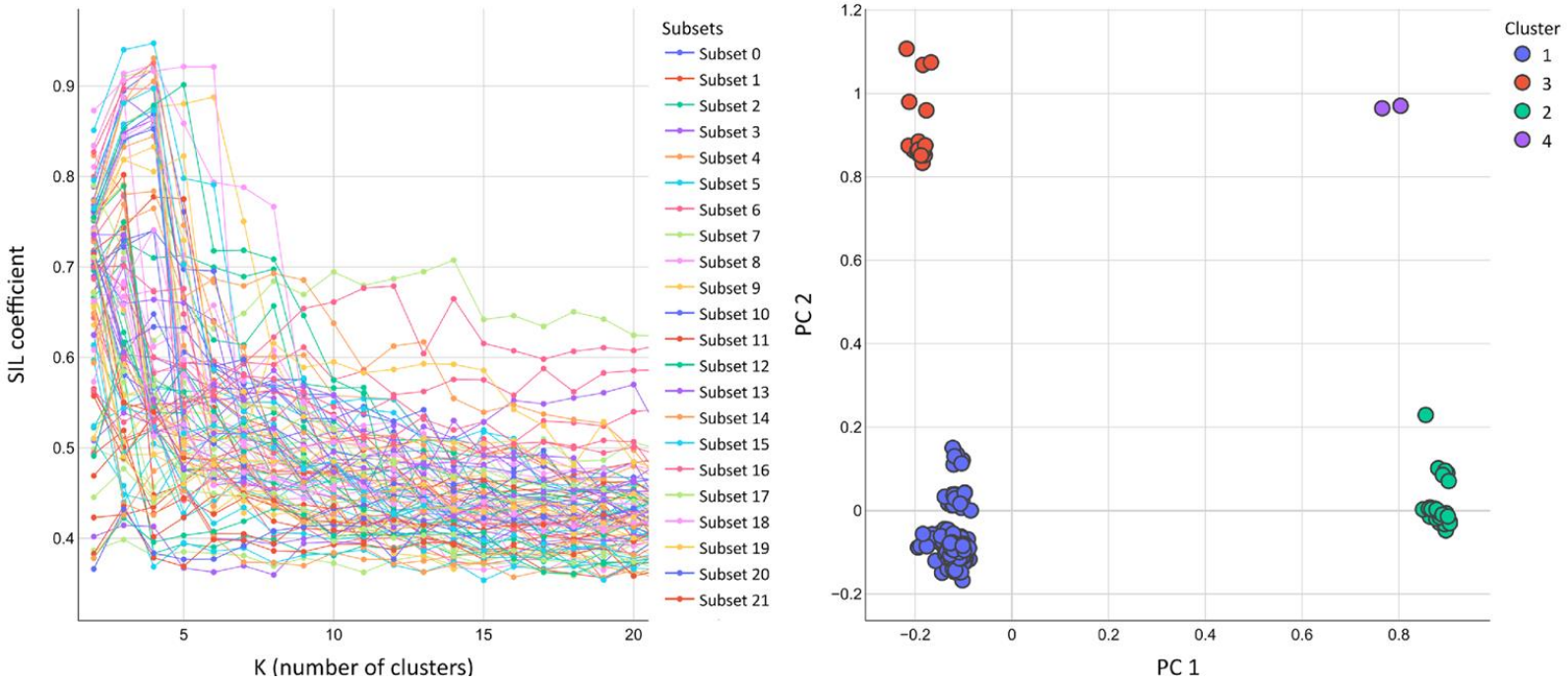
Ejemplo: **Dataset focal_adhesion**

INPUT

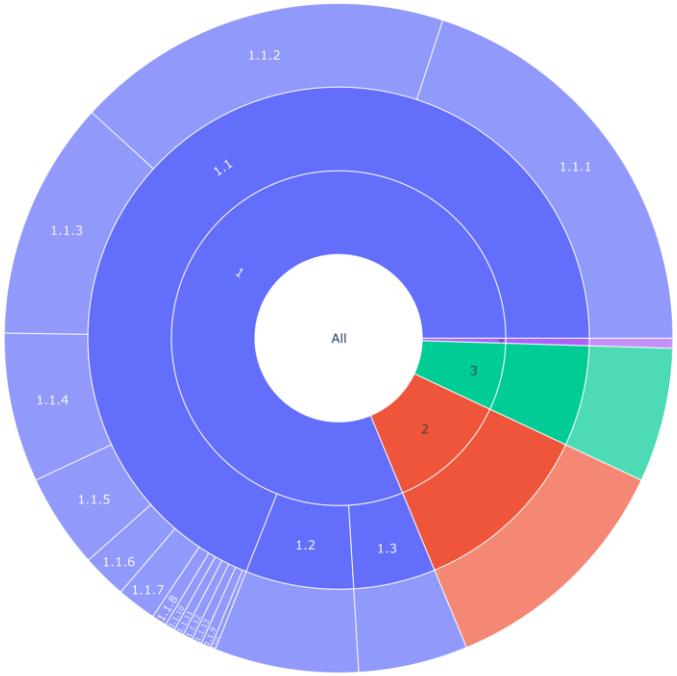
Archivo .csv con **416 moléculas** en anotación SMILES

OUTPUT 1

Plots interactivos:



These plot are generated for each clustering round

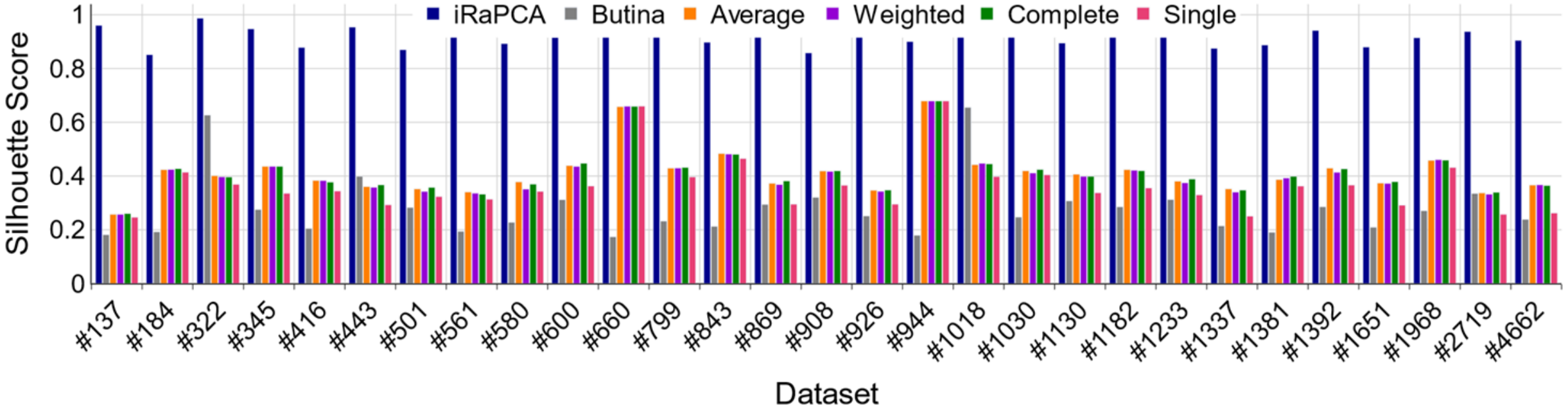


Sunburst graph for the final clusters

BENCHMARKING iRaPCA

Validation de iRaPCA

- **Databases**:
  - 29 from Cortes-Ciriano 2016
- **Métrica**:
  - Silhouette score



UNIVERSIDAD NACIONAL DE LA PLATA

CONICET LA PLATA

streamlit
scikit-learn
rdkit-pypi
molvs
mordred
plotly
seaborn
validclust
umap-learn
openbabel
scipy
biopython
pybiomed
mols2grid

Carolina Bellera

Lucas Alberca

Denis Prada

Manuel Llanos

Santiago Rodríguez

Maximiliano Fallico

Muchas GRACIAS!

www.lideb.biol.unlp.edu.ar

@lideb_unlp

lideb_unlp

https://github.com/LIDeB

questions?