Artificial Intelligence in the Life Sciences 7 (2025) 100129

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Research Article

LIDEB's Useful Decoys (LUDe): A freely available decoy-generation tool. Benchmarking and scope

Lucas N. Alberca [a,b], Denis N. Prada Gori [a,b], Maximiliano J. Fallico [a,b], Alexandre V. Fassio [c], Alan Talevi [a,b,*], Carolina L. Bellera [a,b]

[a] Laboratory of Bioactive Compounds Research and Development (LIDeB), Faculty of Exact Sciences, University of La Plata (UNLP), La Plata, Buenos Aires, Argentina
[b] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), CCT La Plata, La Plata, Buenos Aires, Argentina
[c] São Carlos Institute of Physics, University of São Paulo, São Carlos, São Paulo 13563-120, Brazil

ARTICLE INFO

Keywords:
Decoys
Decoy generation
Retrospective screening
Cheminformatics
Open-source

ABSTRACT

In the field of chemoinformatics, and in particular, when developing models to be applied in virtual screening campaigns, it is essential to run retrospective virtual screening experiments that evaluate the performance of such models in a scenario similar to the real one. That is, the ability to recover a small number of active compounds dispersed among a much larger number of compounds without the desired activity. However, such a retrospective experiment is often limited by the relative scarcity of known inactive compounds against the pharmacological target of interest. In these cases, automatic decoy (putative inactive compound) generation tools are often of great importance. Their basic goal is to generate decoys that are similar enough to the known active compounds to challenge the models, but different enough so that the probability that the decoys modulate the molecular target of interest is small.

In this article, we report the latest version of our open-source decoy generation tool LUDe, inspired by the well-known DUD-E but designed to reduce the probability of generating decoys topologically similar to known active compounds. We have carried out a benchmarking exercise against DUD-E through 102 pharmacological targets, using the DOE score and the Doppelganger score as comparison criteria. LUDe decoys obtained better DOE scores across most of the targets, indicating a lower risk of artificial enrichment. The mean Doppelganger score, in contrast, was similar for LUDe and DUD-E decoys, exhibiting a slight improvement for LUDe decoys for most of the targets. Simulation experiments were performed to verify whether the generated decoys are unsuitable to validate ligand-based models. Our results suggest that LUDe decoys are apt to be used to validate and compare machine learning ligand-based screening approaches. Importantly, LUDe may be used locally, independently from external server availability, and is thus suitable to obtain decoys from large datasets. It is available as a Web App (at https://lideb.biol.unlp.edu.ar/?page_id=1076) and as Python code at (https://github.com/LIDeB/LUDe.v1.0)

1. Introduction

Virtual screening involves using one or more computational models or algorithms to screen large (or ultra-large) libraries of chemical compounds and identify those with the greatest chance of modulating a molecular target of interest. To decide which model(s) and protocols will perform best in prospective virtual screening campaigns, the average and early enrichment capacity of these models is often estimated in retrospective screening experiments [1], usually resorting to enrichment metrics such as the area under the Receiver Operating Characteristic (ROC) curve (AUC ROC) [2], the Boltzmann-enhanced discrimination of ROC (BEDROC) [3], and others.

Ideally, such retrospective screening should be done by confronting models against libraries in which a small number of (known) active compounds are dispersed among a much larger number of (also known) inactive compounds. Naturally, however, this is infeasible when working with recently reported/validated molecular targets associated with data scarcity on known ligands and non-ligands. On the other hand, the scientific community has historically shown a bias towards the collection and publication of positive results (a trend that, fortunately, has

* Corresponding author.
E-mail addresses: alantalevi@gmail.com, atalevi@biol.unlp.edu.ar (A. Talevi).

https://doi.org/10.1016/j.ailsci.2025.100129
Received 14 December 2024; Received in revised form 6 February 2025; Accepted 6 February 2025
Available online 7 February 2025
2667-3185/© 2025 The Authors. Published by Elsevier B.V. CCBYLICENSE This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

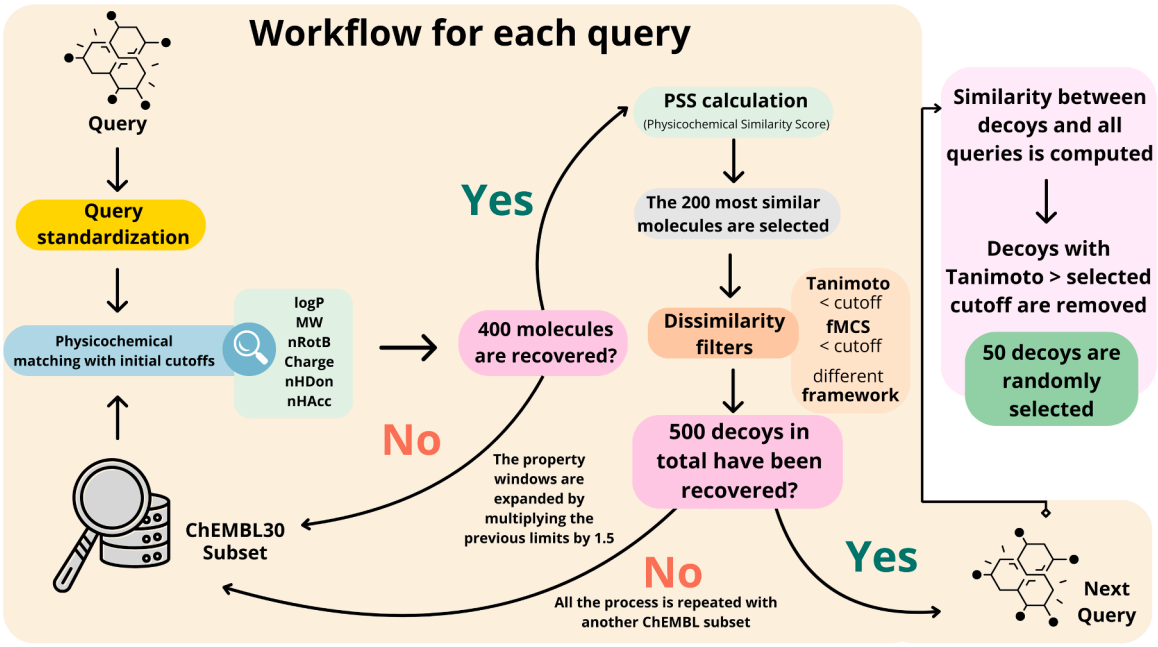

Fig. 1. LUDe workflow. The steps for each query compound can be traced from the upper left corner. It starts with the computation of physicochemical properties used to match decoys and active compounds and then excludes topologically similar decoys. After that, the cross-similarity of each decoy versus all the other queries is checked, and decoys that are topologically similar to any of the active compounds used as queries are disregarded (box on the right).

begun to reverse in recent years) [4,5], another reason that could explain, in certain cases, the relative scarcity of negative data.

Under such circumstances, various tools can be used to generate decoys [6–11], i.e., putative inactive compounds (negative data). Most of these resources, including the well-known Directory of Useful Decoys, Enhanced (DUD-E), are based on recovering or constructing decoys similar to known active compounds with respect to a series of general physicochemical properties while topologically different from these molecules. This practice is based on the assumption that the probability of decoys binding to the selected molecular target's binding site is low, thereby categorizing them as synthetic negative data. It has been noted that the construction of active/decoy datasets remains a fundamental determinant of measured performance in benchmark of virtual screening protocols [12]. Nicholls has compared retrospective screening with police lineup as a technique to validate virtual screening protocols [13]. This analogy is quite productive: as in a police lineup, the ability of the screening protocol to distinguish active compounds (true suspects) from decoys (false suspects) will provide evidence about the protocol's reliability. On the other hand, there are rules for choosing false suspects: an identical twin would not be a good decoy, but neither would a decoy so different from the true suspect that identification would be trivial.

In this article, we report the latest version of our open-source decoy generation tool LUDe (LIDeB's Useful Decoys). LUDe is inspired by the well-known DUD-E with three potential advantages. First, it has been conceived to reduce the probability of generating decoys topologically similar to known active compounds, thus reducing false negative bias. Second, LUDe is available as a Web App (at https://lideb.biol.unlp.edu.ar/?page_id=1076) and as code (at https://github.com/LIDeB/LUDe.v1.0), which can be used as a standalone version (and it also included as Supplementary material). In contrast, DUDE-E and its latest version, DUDE-Z, are based on a remote job queue, and the rate of job processing often depends on server availability. Finally, LUDe's output includes metrics that reflect the quality of the decoys, which may help users decide whether the generated decoys are useful to their specific needs or to re-run our algorithm with different settings. We have carried out a benchmarking exercise against DUD-E through 102 pharmacological targets, using the deviation from optimal embedding (DOE) score and the Doppelganger score as comparison criteria [6]. Moreover, simulation experiments were performed to verify whether the generated decoys are, as usually proposed, trivial to be used in the validation of ligand-based models.

2. Methods

2.1. LUDe workflow

LUDe retrieves decoys from a curated ChEMBL30 database [14]. To reduce artificial enrichment, these decoys match active compounds (used as queries) in a multiple physicochemical properties space. To reduce false negative bias, several topological filters have been serially implemented in LUDe to ensure that queries and decoys do not share molecular scaffolds. LUDe is written in Python (v.3.10.4), and its Web App version is deployed under the Streamlit framework (https://streamlit.io/). As input, the *in-house* script requires a .txt file with molecular representations of the query compounds (known active compounds) in the SMILES notation, one by line. The general workflow to retrieve decoys is shown in Fig. 1. Below, we include a detailed description of each step of the workflow.

2.1.1. Standardization and physicochemical characterization of query compounds

Query compounds are standardized using the Standardize module of MolVS 0.1.1 package (https://molvs.readthedocs.io/en/latest/). This process identifies the largest organic covalent unit within the molecule, replaces all atoms with the most abundant isotope of each element, and removes any charges using the fragment_parent(), isotope_parent() and charge_parent() functions, respectively. Additionally, the ionization state at pH 7.4 is calculated using the Openbabel module [15]. Once the molecule is standardized, six molecular descriptors are calculated employing the Chem module from the RDKIT package (rdkit-pypi 2022.3.3) [16]: molecular weight (*MW*), octanol-water partition coefficient (*LogP*), number of rotatable bonds (*nRotB*), number of H-bond acceptors (*nHAcc*), number of H-bond donors (*nHDon*), and formal charge (*Charge*). Additionally, Morgan fingerprints and Murcko scaffolds are generated using the Chem module.

2.1.2. Matching by physicochemical properties

Decoys are retrieved from subsets of a curated ChEMBL30 database containing approximately 150,000 molecules each. The full database, containing approximately 2.1 million compounds, was standardized in the same way as the queries (identifying the largest organic covalent unit within the molecule, replacing all atoms with the most abundant isotope of each element, neutralizing charges in the fragment_parent, and then re-ionizating at pH 7.4 with Openbabel) and duplicated molecules were removed, resulting in a final set of 1.9 million compounds, which were distributed along 13 subsets to be easily accessible. To remove duplicated molecules, the standardized SMILES, along with the corresponding values of the 6 physicochemical properties used to match queries and decoys, were tabulated, and when two rows in the table present the same SMILES annotation and identical values of the 6 physicochemical properties, only one of them is retained and the rest are deleted.

A physicochemical similarity search, with established ranges for each physicochemical property, is carried out in one randomly selected ChEMBL subset. By default, the retrieved decoys will present *MW*±20 Daltons, Log*P*±0.5 log units, *nRotB*±1 bonds, *nHAcc*±1 bonds, *nHDon* ±1 bonds, and *Charge*±1 units than the query molecule, although the tolerance ranges can be tuned by the user. If for a given query compound <400 molecules are recovered, the property limits are automatically extended by a 1.5 factor up to a maximum of five times. Compounds that present extreme values of these descriptors were removed; only those between the following limits are considered: $100 < MW < 1000$, $-5 < \log P < 10$, $nRotB < 20$, $nHAcc < 20$, $nHDon < 20$, and $-10 < Charge < 10$.

Decoys that are physicochemically similar to query compounds will be the most challenging for classification models. The physicochemical similarity score (PSS) between the query molecule and each decoy is calculated to estimate their physicochemical similarity [6]. For that purpose, for every of the six matching physicochemical properties, the decoys with the minimal and maximal values of the property j ($x_{j,\min}$ and $x_{j,\max}$, respectively) are found in the corresponding pool of decoys. For every active compound, the difference between the property value for the active compound ($x_{j,ref}$) and $x_{j,\min}$ or $x_{j,\max}$ is computed. The bigger of these differences is used to normalize the difference between the property value of every decoy i ($x_{j,ref}$) and $x_{j,ref}$ to a value ranging from 0 to 1. The resulting relative distance is converted into a similarity score $\delta_{i,j}$ by inverting the scale (0 = lowest similarity, 1 = highest similarity). Finally, the PSS score for decoy i is the arithmetic mean of the six similarity scores computed (one for each matching property). LUDe will then select the 200 decoys with higher PSS, as these will be better embedded in the space of the six matching properties.

2.1.3. Dissimilarity filters

Three successive topological similarity filters are applied to the previously retrieved decoys to select those that are less topologically similar to the query compound, to reduce the chances of false negative bias. Based on the similarity principle, which states that similar compounds are likely to have similar properties, we hypothesized that ensuring dissimilarity between queries and decoys is key to reducing the probability of retaining decoys that could be active (false negatives).

Our dissimilarity filters are applied in the following order:

a. The molecular fingerprints of the query compound and each potential decoy are compared using the Tanimoto similarity coefficient. By default, only decoys with a maximum similarity coefficient of 0.2 are kept and submitted to the next filter. This is a substantial difference in comparison with DUD-E, where the most dissimilar 25 % of the decoys are retained through the dissimilarity filter (regardless of the similarity values of those 25 % that are kept).
b. The maximum common substructure (MCS) between the query compound and each decoy is determined. The ratio between the number of atoms in the MCS and the total number of atoms in the query compound (fMCS) is calculated for each decoy. The decoys with an fMCS below a user-defined value (by default, 0.5) are retained and submitted to the next step.
c. The Bemis-Marcko scaffold [17] of the query compound is compared with those of the potential decoys and only those that present different scaffolds than the query are kept.

Steps 2.1.2 and 2.1.3 are repeated using a different ChEMBL subset until 500 decoys are obtained.

2.1.4. Cross-similarity

After performing the workflow for all query molecules, the molecular fingerprints of the resulting list of decoys are compared to the fingerprints of all query compounds; only those decoys with a Tanimoto similarity below 0.2 (by default, but also tunable) to their nearest neighbor among the queries are retained. Up to 50 decoys per query compound are retrieved by default, although this number can be customized.

2.1.5. Output

Three files are generated as the output of each LUDe run. The first of them is named "*Generated_decoys.csv*" and contains all the SMILES of the decoys and a label that indicates from which query molecule each decoy was generated. The second file, called "*Decoys_analysis.csv*", summarizes how many molecules passed each successive filter in the workflow. The third file, named "*Decoys_setting.csv*", contains all the settings used in the run, for the user's records. It also informs the DOE score (best possible DOE score = 0, worst possible score = 0.5), the mean Doppelganger score (best possible score = 0, worst possible score = 1), and the maximal Doppelganger score (best possible score = 0, worst possible score = 1) (the definitions are provided in Section 2.2.2).

2.2. Benchmark

2.2.1. Datasets used for benchmarking

DUD-E [8] contains one of the largest and most popular sets of decoys for benchmarking experiments, comprising 102 sets of ligands and decoys for different molecular targets. The performance of LUDe was compared with that of DUD-E using these 102 sets. For each set of active compounds retrieved from DUD-E, a new set of decoys was generated using LUDe. These sets will be referred to as *LUDe decoys* from now on.

An alternative protonation strategy similar to the one employed by DUD-E was also considered, i.e., we calculated the ionization state at pH 7 for the 102 DUD-E subsets and the ChEMBL30 database with an *in-house* script employing the ChemAxon package v. 5.3.8. Only tautomer species with a minimum distribution of 20 % were retained. Stereoisomers were also calculated for each valid tautomer using the ChemAxon package. For those stereoisomers having the same charge value, only one of them was randomly retained. All the stereoisomers with different charge values were kept. The rest of the workflow was not modified. These sets will be referred to as *LUDe ChemAxon decoys*.

2.2.2. Decoy quality assessment

It is important to consider three potential types of bias when a decoy set is used to evaluate a virtual screening protocol [18]: *artificial enrichment, analog bias, and false negative bias*. Artificial enrichment occurs when the active compounds and the decoys occupy very different regions of the chemical space so that differentiating them is a trivial task, which could lead to overoptimistic enrichment metrics in retrospective experiments. The analog bias occurs when the dataset of active molecules is structurally limited, which may also lead to overoptimistic enrichment by allocating in the compound set used for retrospective screening active compounds with similar chemotypes to those used to calibrate the virtual screening protocol. Finally, the false negative bias refers to the risk of having decoys that are real active compounds (false negatives), which would underestimate the performance of the protocol.

To assess the probability of artificial enrichment, we have

determined the level of embedding between the chemical space defined by the active compounds and the decoys by calculating the DOE score [6]. Total embedding of active compounds' and decoys' chemical spaces would result in a DOE score similar to 0, while poor embedding would result in a DOE score similar to 0.5, indicating large distances between active compounds and decoys in the property space [10].

Moreover, we assessed the effectiveness to differentiate ligands from decoys by machine learning models trained with the six individual physicochemical properties used to align active compounds and decoys (enumerated in Section 2.2.1) and all the possible combinations of 1 to 6 of these properties. As described by Sieg et al. [19], these experiments can highlight poor embedding between queries and decoys. We evaluated two algorithms: the 1-nearest neighbor (1-NN) with default parameters and random forests (RF) with n_estimators=400 and random_state=42. All models were implemented in Python (v.3.10.4) using the scikit-learn package (version 1.2.2) [21]. The AUC ROC was computed for each of the obtained models and the distribution of the AUC ROC values was verified for each set of models (single variable models, 2-variable models, 3-variable models, and so on). The mean AUC ROC values obtained using *DUD-E, LUDe* and *LUDe ChemAxon decoys* were compared pairwise using Mann-Whitney U rank test [22] for non-normal homoscedastic distributions and Yuen test [23] for non-normal non-homoscedastic distributions.

On the other hand, to reduce the false negative bias, decoys should be dissimilar from the active compounds. Thus, the quality of decoys in terms of the risk of introducing latent active compounds in the decoy set (LADS) was determined through the calculation of the Doppelganger score [6]. Morgan fingerprints (radius 3 and length 2048) were computed for all queries and decoys using RDKit (rdkit-pypi 2022.3.3) [16], and the Tanimoto similarity was calculated for every possible ligand-decoy pair. For each decoy, the similarity across all active compounds was calculated. Two Doppelganger scores were computed: the maximal Doppelganger score, i.e., the highest similarity between the decoy and any of the active compounds; and the mean Doppelganger score, i.e., the average similarity between the decoy and all active compounds.

2.3. Classification models: descriptor-based classification models vs. similarity-based classification

As previously emphasized, decoy generation tools usually employ a similarity filter to retain only those decoys that are dissimilar to the known active compounds used as query. Decoys generated in this manner are not expected to be of use when assessing the performance of ligand-based virtual screening [8,18]. This supposed limitation obviously applies to similarity-based ligand-based virtual screening. We wanted to study whether this limitation applies when other ligand-based approximations (e.g., descriptor-based approaches, e.g., QSAR) are employed.

To determine if *LUDe* and *DUD-E decoys* may be useful (i.e., non-trivial decoys) when using ligand-based machine learning (ML) models, we performed a benchmarking exercise comparing similarity-based screening and four ML algorithms: Linear Regression (LR), RF, Support Vector Machine with linear kernel (SVM_L), and Support Vector Machine with RBF kernel (SVM_RBF), via two enrichment metrics (AUC ROC and BEDROC), over a selected group of the 102 DUD-E Subset Targets (https://dude.docking.org/subsets/all).

For such a purpose, we first obtained activity tables for the 102 targets used for benchmarking exercises from ChEMBL30. Compounds were classified as active if they presented a Ki or $IC_{50} < 1$ μM, and as inactive if they presented a Ki or $IC_{50} > 30$ μM, which are the same cutoff values previously used by Mysinger et al. [8]. Following these criteria, among the 102 DUD-E datasets, we selected only those with at least 100 active molecules and 50 inactive molecules reported. For each of the resulting subsets, a balanced training set was generated by sampling 80 % of the molecules that belong to the category (active or inactive compounds) with the smallest number of molecules in each data set. Also, three retrospective screening sets were compiled containing the remaining active molecules for each subset (at least 50) and the decoys generated using DUD-E, LUDe, or LUDe Chemaxon and all active compounds of each dataset as queries.

2.3.1. Descriptor-based classification models

We computed 1612 conformational-independent molecular

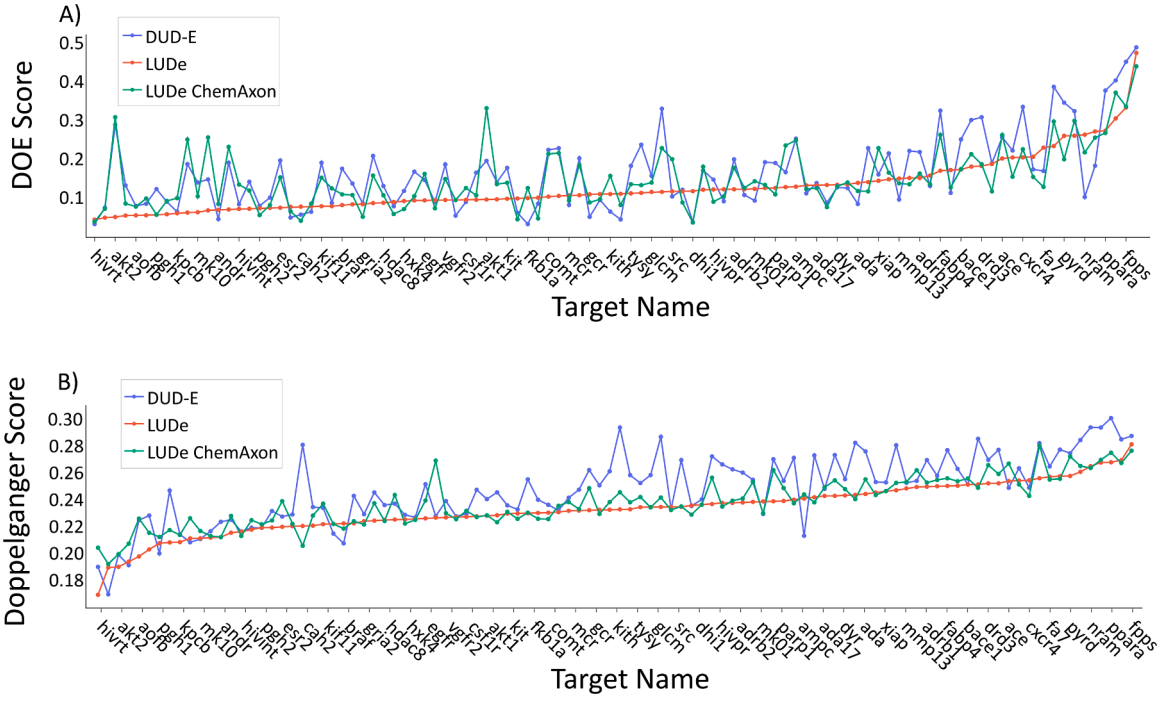

Fig. 2. (A) Comparison of decoy quality (in terms of DOE score) between *DUD-E decoys* and *LUDe decoys* across 102 targets. (B) Comparison of decoy quality (in terms of Doppelganger score) between *DUD-E decoys, LUDe decoys* and *LUDe Chemaxon decoys* across 102 targets.

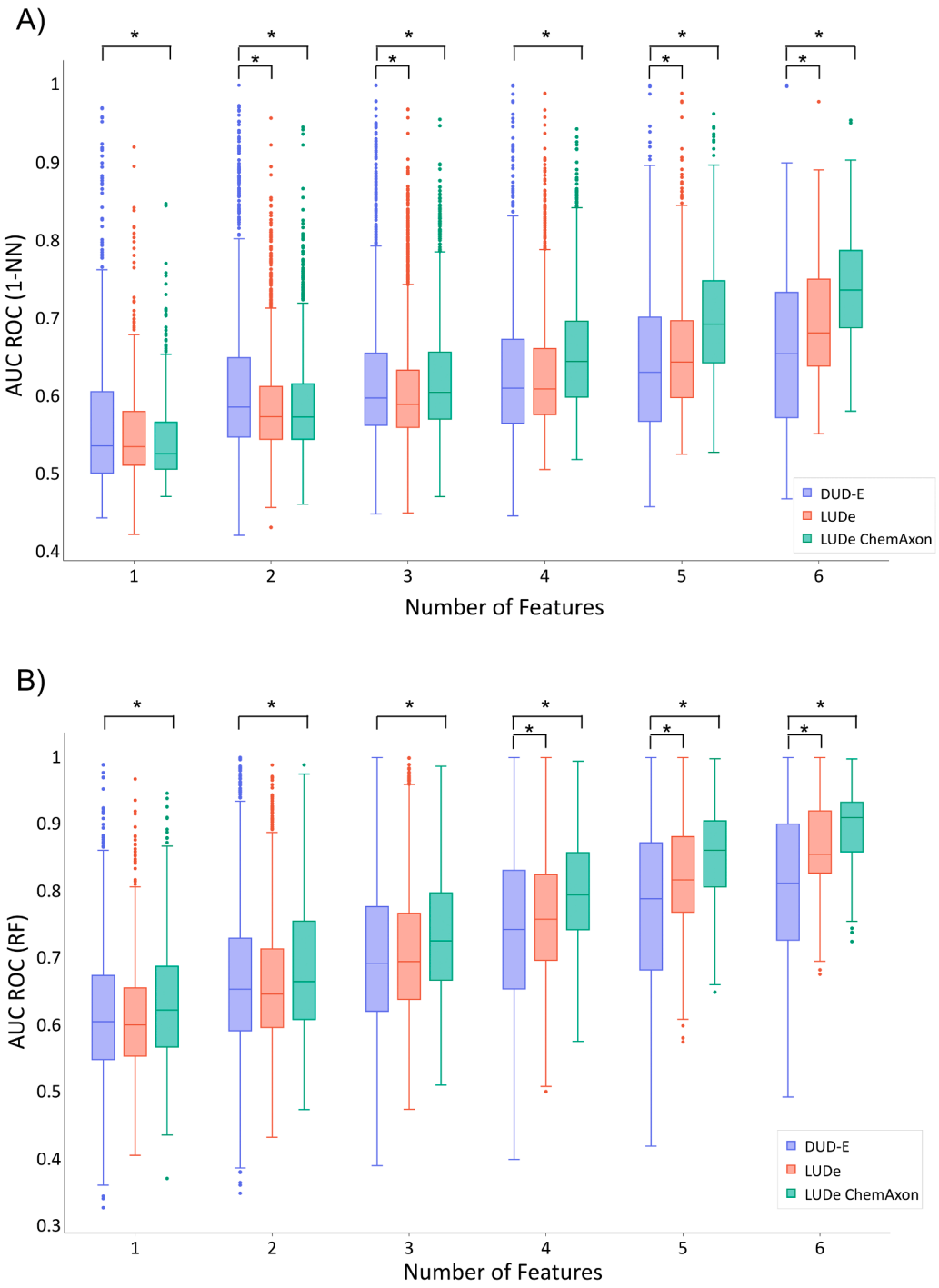

Fig. 3. Distribution of the AUC ROC obtained for the benchmark 102 data sets using 1-NN (A) and RF (B) and considering 1- to 6-physicochemical (matching) property spaces. An asterisk indicates statistically significant differences in the mean AUC ROC (*$p < 0.05$).

descriptors with the Mordred package [20] for each of the training sets and decoys sets. A dimensionality reduction strategy was then applied to each training set before the generation of classification models. First, low-variance descriptors were removed using the Variance Threshold function of the scikit-learn package [21]; then, a random subspace strategy (feature bagging) was applied to obtain 100 subsets of 200 descriptors each, and a correlation filter was applied on each subset to remove those descriptors that were highly correlated (Pearson coefficient > 0.85). Finally, the most important descriptors were selected using a sequential feature selection (SFS) approach using the forward selection algorithm of the scikit-learn module. This method incrementally builds a feature subset using a greedy approach. At each step, it selects the optimal feature to add or remove by evaluating the cross-validation accuracy score of the chosen feature.

The former descriptor subsets were used to generate classification models using four different algorithms from scikit-learn: LR, RF, SVM_L, and SVM_RBF. The predictive ability of the classification models was determined by calculating the AUC ROC and BEDROC [3].

2.3.2. *Similarity-based discrimination*

As a comparator, we evaluated the validity of decoys for the application of similarity-based virtual screening. Since decoys from DUD-E,

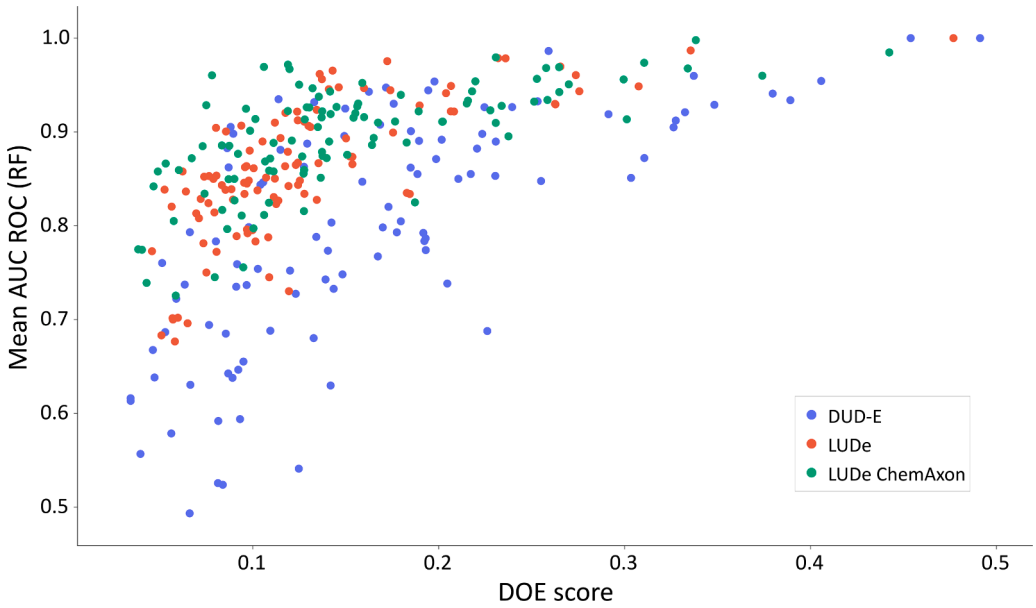

Fig. 4. Correlation between the mean AUC ROC obtained with RF vs. the DOE score.

LUDe, and LUDe ChemAxon are conceived to be molecularly dissimilar to active compounds and a fingerprint-based similarity filter is used for this purpose, we expect that similarity-based approaches may result in perfect discrimination between decoys and active compounds, demonstrating its invalidity to be used in this type of ligand-based virtual screening.

Six different similarity quantification approaches were carried out for each of the selected data sets. Three of them involved the generation of Morgan Fingerprints for every molecule (radius 3 and length 2048) and the calculation of Tanimoto, Dice, and Cosine coefficients between each molecule of the retrospective set versus all the active molecules of the training set. The other three approaches involved the characterization of the molecules by MACCS Fingerprints and the evaluation of the previously mentioned similarity coefficients. For every fingerprint/similarity coefficient combination, the average of all similarity values (*MEAN SIM*) and the maximum value of similarity (*MAX SIM*) between each molecule of the retrospective set and the active molecules in the training set were calculated. These values were used to determine the ability of the similarity search to discriminate active compounds from decoys.

3. Results and discussion

3.1. Benchmark

3.1.1. Artificial enrichment and false negative bias

To analyze the embedding of active molecules and decoys in the chemical space defined by the six physicochemical properties used to match ligands and decoys, the DOE score was calculated for the 102 target subsets with DUD-E, LUDe, and LUDe ChemAxon decoys. The lower values of the DOE score indicate a better embedding of the decoys and the active molecules, thus reducing the chances of artificial enrichment. Fig. 2A shows how the DOE score per target improved in 65 % and 61 % of the subsets with *LUDe decoys* and *LUDe ChemAxon decoys*, respectively, compared to *DUD-E*. 86 % of the subsets with *LUDe decoys* achieved a DOE score below 0.2, compared to 75 % of the subsets with *LUDe ChemAxon decoys* and 74 % of the subsets of *DUD-E decoys*.

Fig. 3 compares the distribution of AUC ROC values when resorting to 1-NN and RF and considering every possible feature or feature combination (from 1 to 6) of the physicochemical properties used to match queries and decoys. In the case of 1-NN (Fig. 3A), no statistical differences were observed between the mean AUC ROC values of the 1- and 4-variable models when comparing either *DUD-E decoys* vs. *LUDe decoys* or *DUD-E decoys* vs. *LUDe ChemAxon decoys*. For the 2- and 3-variable models, the mean AUC ROC values obtained with *LUDe decoys* were statistically lower than those obtained with *DUD-E*. Contrariwise, the mean AUC ROC values were lower for the *DUD-E decoys* than for *LUDe decoys* in the case of 5- and 6-variable models (Fig. 3A). In all cases except for *LUDe ChemAxon* in the 6-feature space, the median AUC ROC using 1-NN is rather low (below 0.7), although many models show ideal or almost ideal AUC ROC, indicating that the degree of embedding between queries and decoys is highly target-dependent. In the case of the RF models (Fig. 3B), *LUDe decoys* resulted in statistically similar AUC ROC values than *DUD-E* for models incorporating 1 to 3 features, and higher AUC ROC values for 4- to 6-feature models. *LUDe ChemAxon decoys*, in this case, invariably produced higher AUC ROC values. Higher median AUC ROC values are observed for RF models compared to 1-NN, especially for 5- and 6-feature spaces. The results suggest that LUDe provides decoys of higher quality than LUDe Chemaxon.

Interestingly, when plotting the mean AUC ROC obtained for each of the 102 data sets using RF versus the DOE score (Fig. 4), a clear positive correlation can be observed, with data sets that obtained high DOE (that is, less embedding between queries and decoys) invariably leading to high mean AUC ROC. This is particularly interesting, as it confirms that high DOE is associated with rather trivial (low-quality) decoys. This is relevant actionable observation: if a decoy generation workflow leads to high DOE (e.g., higher than 0.2), the workflow should be optimized to achieve higher embedding (e.g., by narrowing the allowed ranges for each matching physicochemical property). Importantly, the last version of LUDe automatically computes the DOE and the Doppelganger scores to provide indicators of the decoy quality and re-run the workflow using different settings if needed.

The false negative bias was studied by calculating the Doppelganger score. For each target, we report the mean and maximum Doppelganger score over all decoys. Compared to *DUD-E decoys*, the Doppelganger score was lower for 85 % and 75 % of the target subsets obtained with *LUDe decoys* and *LUDe ChemAxon decoys,* respectively. The average Doppelganger score for *LUDe decoys* was 0.23, compared with 0.24 for *LUDe ChemAxon decoys* and 0.25 for *DUD-E decoys* (Fig. 2B). The lower maximum Doppelganger score values obtained with *LUDe decoys* indicate higher dissimilarity of the decoys against the active molecules, thus decreasing the false negative bias.

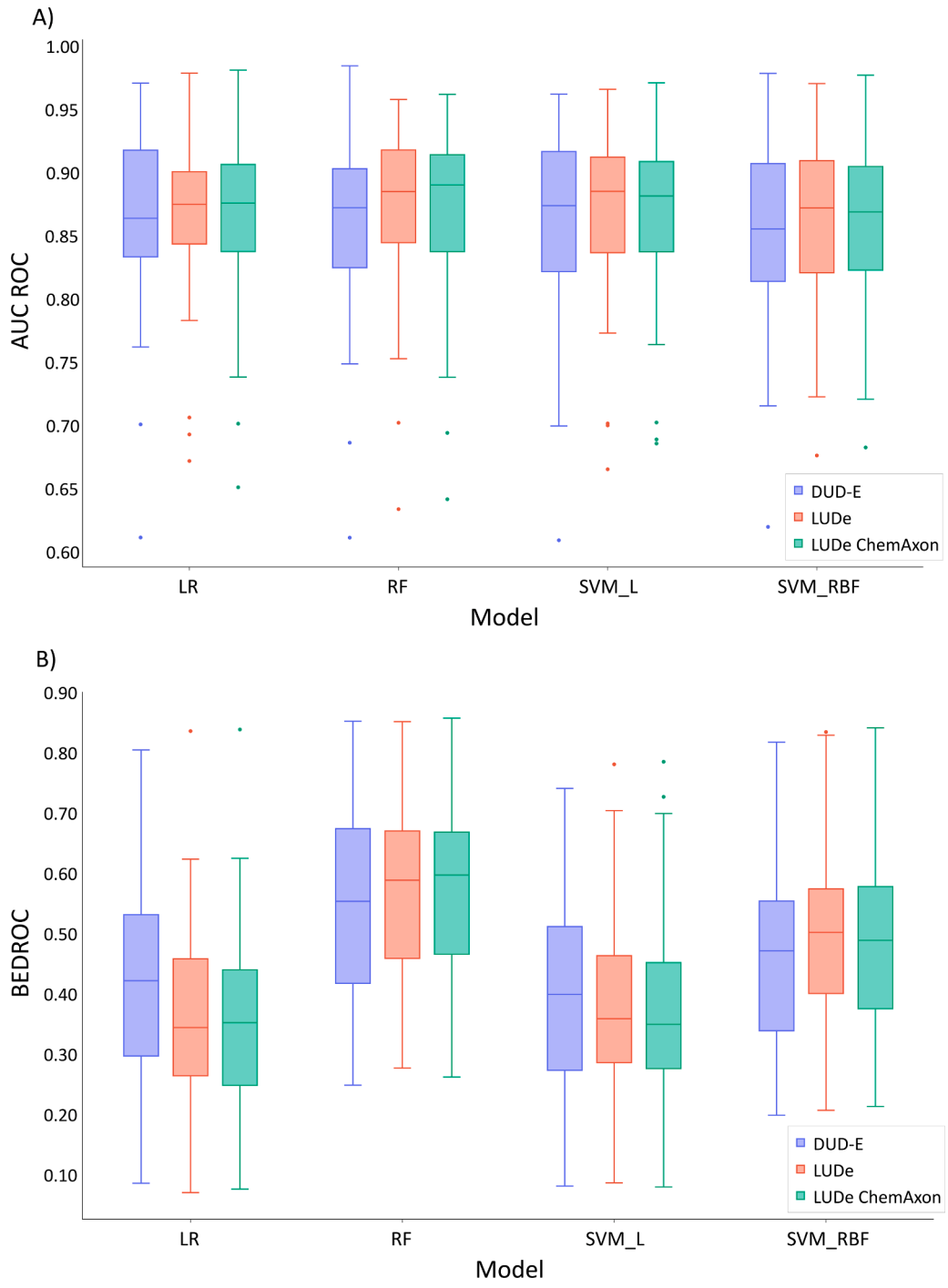

Fig. 5. Distribution of best AUC ROC (A) and BEDROC (B) values across the selected 46 molecular targets using different machine learning approximations.

3.2. 2D ligand-based machine learning vs similarity search

2D descriptor-based models were generated from 46 DUD-E benchmark subsets with at least 100 active and 50 inactive molecules. The performance of these models was compared with the performance of different strategies of discrimination between active compounds and decoys based on similarity comparisons, which are expected to perform virtually perfectly, considering that during the decoy generation processes by both DUD-E and LUDe, only decoys with low Tanimoto similarity to the active compounds used as query should be retained.

For descriptor-based models, we obtained for every target, 100 LR models, 100 RF models, 100 SVM_L models, and 100 SVM_RBF models (one model per modeling technique and per randomly generated subset of descriptors). For comparison purposes, we selected the best AUC ROC and BEDROC obtained in each retrospective screening set for each model type and target (Fig. 5). In the boxplots from Fig. 5A, the median AUC ROC for DUD-E, LUDe, and LUDe ChemAxon retrospective sets present values around 0.87, which indicates that, while descriptor-based models can discriminate between the active compounds and the decoys, they are far from perfect behavior (AUC ROC = 1). Regarding the boxplots for BEDROC (Fig. 5B), the median values are between 0.3 and 0.5 for LR, SVM_L and SVM_RBF, and 0.6 for the RF models, also showing that they are far from perfect behavior (BEDROC = 1). Naturally, the median value of the metrics is even lower if we analyze the performance across the 46 datasets of all the models generated in the retrospective screening, instead of the distribution of the best-performing models

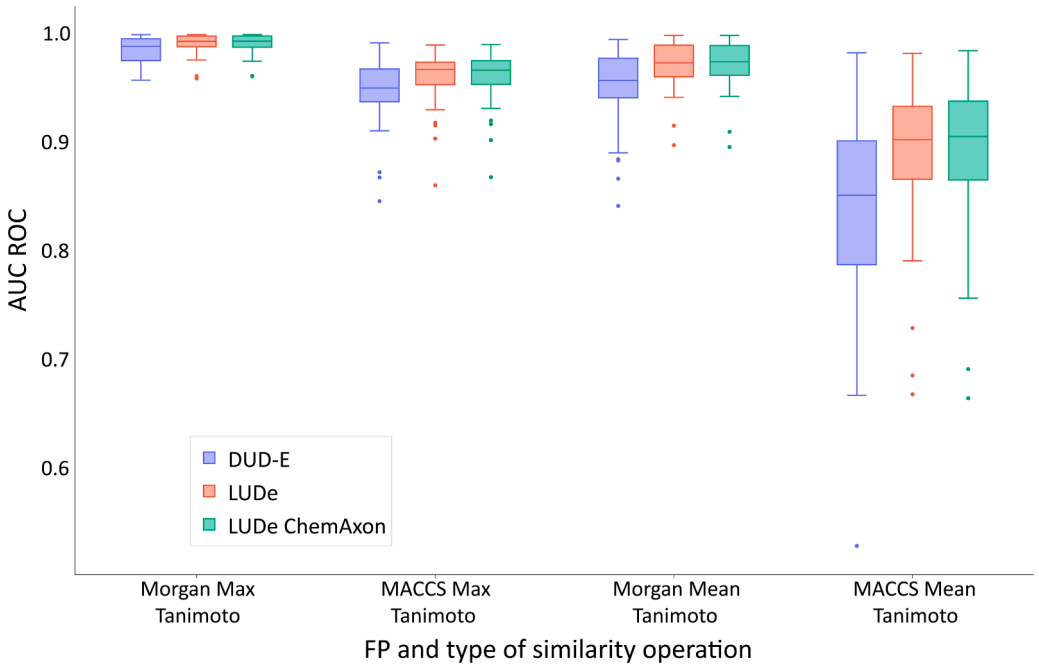

Fig. 6. Distribution of AUC ROC values across the selected 46 molecular targets using similarity coefficients as a criterion to retrieve active compounds in the retrospective screening experiments.

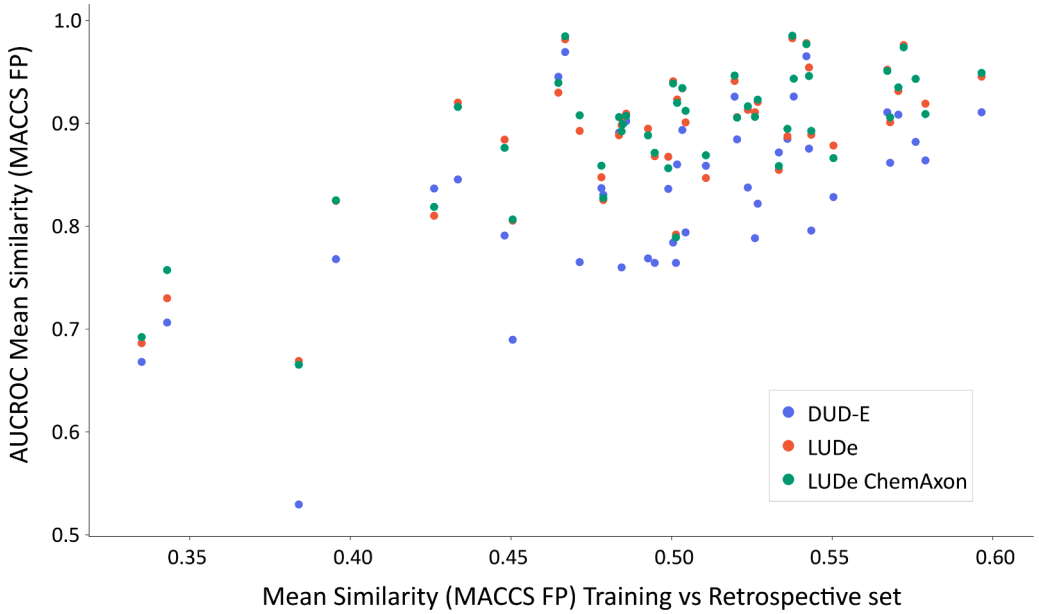

Fig. 7. Correlation between AUC ROC obtained for each target, and mean similarity between the active compounds in the training and retrospective sets.

(median AUC ROC values below 0.8 and median BEDROC values below 0.35 for the best-performing ML approach, RF, and a much poorer performance for all the other approaches) (Fig. S1). These results suggest that decoys are not trivial for descriptor-, ligand-based machine learning models.

Additionally, we used similarity comparisons between the compounds in the retrospective screening libraries and the active compounds in the training sets of each of the selected 46 data sets, to evaluate the usefulness of DUD-E and LUDe decoys to assess the performance of similarity-based approximations. Fig. 5 shows the results of using the mean similarity to the active compounds in the training set and the maximal similarity to the active compounds in the training set as criteria for discrimination between active compounds and decoys. In both cases, we have graphed the results obtained using the Tanimoto coefficient for both fingerprinting systems (Morgan and MACCS).

It is expected that the similarity between the decoys of the retrospective set vs. the active molecules of the training set will be low as this is a precondition to retrieving decoys; contrariwise, the similarity between the active molecules of the retrospective set vs. the active molecules of the training set should be higher. Accordingly, differentiating decoys from active compounds using similarity-based approximations should pose no challenge, and the AUC ROC is expected to be close to 1, indicating perfect discrimination. Fig. 6 shows that, in general terms, the results met the expectations. The median AUC ROC is close to 1 (i.e., perfect classification) when using the maximal Tanimoto coefficient to the active compounds in the training set to discriminate between active

compounds and decoys, both with Morgan and MACCS fingerprints. A similar result was observed when considering the mean Tanimoto coefficient using Morgan fingerprints and with Dice and Cosine similarity coefficients (not shown). However, the AUC ROC for the mean similarity using MACCS fingerprints is rather low compared to the other similarity-based approaches and comparable to the values obtained with the descriptor-based models.

To explain this rather unexpected behavior, we explored the correlation between the AUC ROC obtained with the mean Tanimoto similarity to the active compounds in the training set (MACCS fingerprints) versus the mean similarity between the active compounds in the training and the active compounds in the retrospective sets (Fig. 7). We observed a clear positive correlation between the AUC ROC values and the mean training versus retrospective set similarity. This makes sense: in those data sets where the mean similarity between the active compounds from the training and the retrospective set is low, at least some active compounds in the retrospective library exhibit low similarity to the active compounds in the training set, thus decreasing the AUC ROC.

4. Conclusions

Here, we report LUDe, an *in-house* open-source decoy generation tool inspired by the DUD-E workflow, with some modifications in the instance that exclude decoys that are topologically similar to the active compounds used as queries. In benchmark experiments, LUDe tends to provide lower DOE and Doppelganger scores than DUD-E for most of the 102 targets considered for comparison purposes, suggesting better embedding between decoys and active compounds and lower false negative bias. Interestingly, computing the protonation state with OpenBabel provides better results than Chemaxon with a lower computational cost. Another relevant advantage of LUDe is that it is available as code and may then be used locally as a standalone tool, independent from external server availability.

Interestingly, we have demonstrated that descriptor-based machine learning models provide suboptimal active enrichment using either DUD-E or LUDe decoys, which challenges the installed notion that DUD-E decoys are not valid to evaluate the performance of 2D ligand-based virtual screening protocols. Inversely, similarity-based virtual screening, as expected, results in almost perfect separation between active compounds and decoys, confirming that DUD-E (and LUDe) decoys should not be used to evaluate similarity-based approaches.

CRediT authorship contribution statement

Lucas N. Alberca: Writing – review & editing, Writing – original draft, Validation, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Denis N. Prada Gori:** Writing – review & editing, Writing – original draft, Validation, Software, Investigation, Formal analysis, Data curation, Conceptualization. **Maximiliano J. Fallico:** Validation, Investigation. **Alexandre V. Fassio:** Writing – review & editing, Writing – original draft, Resources, Investigation, Formal analysis. **Alan Talevi:** Writing – review & editing, Writing – original draft, Supervision, Resources, Funding acquisition, Formal analysis, Conceptualization. **Carolina L. Bellera:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Funding acquisition.

Declaration of competing interest

I declare that the corresponding author and two other co-authors, Dr. Alberca and Dr. Bellera, are co-founders of a startup (Boolzi SA) dedicated to providing b2b services to biomedical and pharmaceutical companies. To the best of my knowledge, our role in the company poses no conflict of interest with the content of this manuscript.

Acknowledgments

The authors are funded by UNLP (Incentivos UNLP) and Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT), via grants PICT-2021-00478, PICT 2019-1075 and PICT 2021-0404. A.T., L.N.A., and C.L.B. are members of the Research Career from the Argentinian National Research Council (CONICET, Argentina). The authors gratefully acknowledge the support of Streamlit.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2025.100129.

Data availability

Data will be made available on request.

References

[1] Xu H. Retrospect and prospect of virtual screening in drug discovery. Curr Top Med Chem 2002;2(12):1305–20. https://doi.org/10.2174/1568026023392869.
[2] 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;48(7):2534–47. https://doi.org/10.1021/jm049092j.
[3] Truchon JF, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model 2007;47(2):488–508. https://doi.org/10.1021/ci600426e.
[4] Mlinarić A, Horvat M, Smolčić VS. Dealing with the positive publication bias: why you should really publish your negative results. Biochem Med 2017;27:3. https://doi.org/10.11613/BM.2017.030201.
[5] Sharma H, Verma S. Is positive publication bias really a bias, or an intentionally created discrimination toward negative results? Saudi J Anaesth 2019;13(4):352–5. https://doi.org/10.4103/sja.SJA_124_19.
[6] Vogel SM, Bauer MR, Boeckler FM. DEKOIS: demanding evaluation kits for objective in silico screening—a versatile tool for benchmarking docking programs and scoring functions. J Chem Inf Model 2011;51(10):2650–65. https://doi.org/10.1021/ci2001549.
[7] Cereto-Massagué A, Guasch L, Vall C, Mulero M, Pujadas G, Garcia-Vallvé S. DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets. Bioinformatics 2012;28(12):1661–2. https://doi.org/10.1093/bioinformatics/bts249.
[8] Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem 2012;55(14):6582–94. https://doi.org/10.1021/jm300687e.
[9] Wang L, Pang X, Li Y, Zhang Z, Tan W. RADER: a RApid DEcoy Retriever to facilitate decoy based assessment of virtual screening. Bioinformatics 2017;33(8):1235–7. https://doi.org/10.1093/bioinformatics/btw783.
[10] Imrie F, Bradley AR, Deane CM. Generating property-matched decoy molecules using deep learning. Bioinformatics 2021;37(15):2134–41. https://doi.org/10.1093/bioinformatics/btab080.
[11] Stein RM, Yang Y, Balius TE, O'Meara MJ, Lyu J, Young J, Tang K, Shoichet BK, Irwin JJ. Property-unmatched decoys in docking benchmarks. J Chem Inf Model 2021;61(2):699–714. https://doi.org/10.1021/acs.jcim.0c00598.
[12] Chaput L, Martinez-Sanz J, Saettel N, Mouawad L. Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. J Cheminform 2016;8:56. https://doi.org/10.1186/s13321-016-0167-x.
[13] Nicholls A. What do we know and when do we know it? J Comput Aided Mol Des 2008;22:239–55. https://doi.org/10.1007/s10822-008-9170-2.
[14] Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR. The ChEMBL database in 2017. Nucl Acid Res 2017;45(D1):D945–54. https://doi.org/10.1093/nar/gkw1074.
[15] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. J Cheminform 2011;3:33. https://doi.org/10.1186/1758-2946-3-33.
[16] Landrum G, Tosco P, Kelley B, Vianello R, Schneider N, Kawashima E, Dalke A, Cosgrove D, Jones G, Cole B, Swain M, Turk S, Savelyev A, Vaucher A, Wójcikowski M, Gavid D. rdkit/rdkit: 2022_03_3 (Q1 2022) release (Release_2022_03_3). 2022. https://doi.org/10.5281/zenodo.6605135.
[17] Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. J Med Chem 1996;39(15):2887–93. https://doi.org/10.1021/jm9602928.
[18] Réau M, Langenfeld F, Zagury JF, Lagarde N, Montes M. Decoys selection in benchmarking datasets: overview and perspectives. Front Pharmacol 2018;9:11. https://doi.org/10.3389/fphar.2018.00011.

[19] Sieg J, Flachsenberg F, Rarey M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. J Chem Inf Model 2019; 59(3):947–61. https://doi.org/10.1021/acs.jcim.8b00712.
[20] Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. J Cheminform 2018;10:4. https://doi.org/10.1186/s13321-018-0258-y.
[21] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(85):2825–30.
[22] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist 1947;18(1):50–60. https://www.jstor.org/stable/2236101.
[23] Yuen KK. The two-sample trimmed *t* for unequal population variances. Biometrika 1974;61(1):165–70. https://doi.org/10.1093/biomet/61.1.165.