Suitability of GRIND-based principal properties for the description of molecular similarity and ligand-based virtual screening.

*Ángel Durán, Ismael Zamora, Manuel Pastor**

Research Unit on Biomedical Informatics (GRIB), IMIM/Universitat Pompeu Fabra, Dr. Aiguader 88, E-08003 Barcelona, Spain.

Received ...

*Corresponding author e-mail: manuel.pastor@upf.edu

Abstract

The information provided by the alignment-independent GRid Independent Descriptors (GRIND) can be condensed by the application of Principal Component Analysis, obtaining a small number of principal properties (GRIND-PP), more suitable for describing molecular similarity. The objective of the present study is to optimize diverse parameters involved in the obtention of the GRIND-PP and to validate their suitability for applications requiring a biologically relevant description of the molecular similarity. With this aim, GRIND-PP computed with a collection of diverse settings were used to carry out ligand-based virtual screening (LBVS) on standard conditions. The quality of

the results obtained was remarkable and comparable with other LBVS methods, and their detailed statistical analysis allowed to identify the method settings more determinant for the quality of the results and their optimum. Remarkably, some of these optimum settings differ significantly from those used in previous published applications, revealing their unexplored potential. Their applicability in large compound database was also explored by comparing the equivalence of the results obtained using either computed or projected principal properties. In general, the results of the study confirm the suitability of the GRIND-PP for practical applications and provide useful hints about how they should be computed for obtaining optimum results.

Introduction

For a long time it has been generally accepted that two compounds with a high degree of chemical similarity are likely to have similar biological properties. This statement, albeit true in the vast majority of cases, has striking exceptions, as it was shown in recent publications.[1] These exceptions are a consequence of the imperfect correspondence between the concepts of chemical similarity and biological similarity; two compounds might share many structural features, but show disparities in a few of them which play a critical role for their interaction with a biological target. Unfortunately, in most cases the identity of these biologically relevant features is unknown and as a consequence, chemical similarity does not guarantee biological similarity, in general terms.

Often, the only hint about how to obtain novel compounds with a certain biological property is a small set of compounds already exhibiting this property. In this situation it is reasonable to assume that the probability to obtain compounds with this biological property is higher in the compounds showing structural similarity with the active

compound.[2,3] This idea was at the basis of the ligand-based virtual screening (LBVS) methods, in which compounds of unknown activity are ranked according to their similarity with one or more known active compounds which are used as templates.[4]

One of the most critical aspects of LBVS methods is how to describe the compound similarity. Ideally, the molecular descriptors used should represent the aforementioned biologically relevant features, since a structural description, centered on describing the template chemotype, is likely to select hits from the same structural family (i.e. if the template is a beta-lactamic antibiotic every compound with a beta-lactamic ring will be selected as an antibiotic). This is inconvenient even when the selected hits are active, because LBVS aims to find compounds with some degree of novelty.[5,6] In this respect, 3D based molecular descriptors offer some advantages over 2D descriptors, since they make no direct use of the template 2D structure and are less likely to extract hits based on their chemotype. Moreover, when more than one template compound is used, 3D based methods can identify common 3D structural features not apparent in their 2D structure, and use them for the search. Therefore, the molecular descriptors used for similarity search in LBVS must represent biologically relevant properties but with the highest possible abstraction of the template chemotype or otherwise the hits will be "too similar" to the templates for being of interest.

The GRid-INdependent Descriptors (GRIND)[7] are an example of 3D based molecular descriptors. They were originally developed for 3D QSAR, but have been applied with success in other areas of drug design.[8] In few words, the GRIND are obtained starting from a collection of molecular interaction fields (MIF) computed using diverse chemical probes, which were discretized by finding the more representative positions

(hot spots). The relative position of these hot spots was encoded into a few arrays of values (correlograms), representing hot spots located at certain distance ranges. Hence, every GRIND variable represents couples of grid nodes, belonging to certain MIF types and separated by a certain distance range. For a certain compound, the GRIND variable is assigned a value of 0.0 when no such couple exists for the MIFs considered or the product of their energy values when such couple does exist. A detailed description of the algorithm used for computing the GRIND was described elsewhere[9] but, as can be seen, the GRIND have the advantage of providing a complete abstraction of the ligand chemotype and indeed, different 2D structures can produce very similar MIF and GRIND. For this reason, providing information relevant for representing the ligands molecular interaction potential without making use of their chemotype, the GRIND are attractive descriptors for LBVS applications. Indeed, a few applications of GRIND in LBVS have been reported[10,11] producing good results, often remarked to be original with respect to the results obtained with other methods.

One of the peculiarities of the GRIND is the intrinsic redundancy of the variables obtained. The presence of any structural feature is reflected in many variables simultaneously, often located in different correlograms.[8] This is not inconvenient for their application in 3D QSAR, since the regression method applied (Partial Least Squares) is highly insensitive to this problem. However, this characteristic could be detrimental for the quantification of compound similarity for two reasons; (i) the relative importance given to diverse structural features depends on the level of redundancy of the description and (ii) they are not efficient, both in terms of storage and of time required to compute similarity indexes involving so many variables. For these reasons it has been suggested[8] that the t-scores obtained with Principal Component

Analysis (PCA) would be more suitable for molecular similarity applications than the original variables and still retain the most relevant information related to their biological properties.

The use of t-scores, also called principal properties (PP), for replacing the original molecular descriptors is not new and early examples can be found in the characterization of heteroaromatics[12] or aminoacids.[13] A more recent example is the use of t-scores for building chemical spaces[14,15] where large numbers of compounds can be localized according to relevant physico-chemical properties, providing a global similarity metric. Such chemical spaces have been used with success for practical purposes, like the identification of drug-likeness regions or subset selection.

In the present work we aim to explore the suitability of principal properties obtained from GRIND (GRIND-PP) for applications which require a description of the molecular similarity. LBVS was chosen to illustrate such applications here, because it is a technique well known in the field of drug design and there are simple and reliable indexes to describe the quality of the results. With this aim, preliminary studies were carried out in order to investigate the effect of diverse method parameters (like the number of PCA components, the scaling and the size of the template sets) on the quality of the results obtained. Then, the method was applied again, with optimum settings, in order to evaluate the overall performance of the method in practical applications. However, it is important to remark that the main aim of the present work is not to benchmark the new descriptors for LBVS applications, but to obtain a preliminary quality assessment revealing their generic suitability for molecular similarity applications.

In addition, properties of such t-scores spaces relevant for their practical use will be tested, like their stability upon addition of new compounds or compounds sets and the suitability of projected t-scores spaces in LBVS applications.

Materials and methods

Data

The GRIND-PP were tested by carrying out LBVS on several standard compound databases: WOMBAT, DUD and ZINC.

WOMBAT (Word of Molecular BioAcTivity)[16] is a database containing molecules collected from articles published in medicinal chemistry journals since 1975. In this study the version 2007.v2, containing 203,924 chemical compounds, was used. In order to enrich the content of lead-like compounds[17] and to obtain more realistic results, compounds with a molecular weight lower than 450 and a computed logP value (AlogP) below 5.5 were removed,[17,18] thus obtaining a set of 123,370 compounds. We made use of the 2D to 3D conversion program CORINA 2.4[19] to generate a single 3D conformation for each compound, using default parameters. The conversion was possible for 123,276 compounds and produced diverse errors for 94 compounds, which were discarded from the study. Database annotations were used to extract active compounds (activity value over 6) for the following seven different targets: 5-HT_3 antagonists, 5-HT_{1A} agonists, D_2 antagonists, angiotensin II AT1 antagonists, thrombin, HIV protease inhibitors and protein kinase C inhibitors. In order to select a suitable set of templates for each target, the GRIND descriptors obtained for every set of actives were exported to GOLPE,[20] where subsets of 5% and 10% representative compounds

were extracted using two alternative algorithms: MDC (most descriptive compound)[21] and LMD (longest minimum distance).[22] In both cases, the selection was carried out in the PCA t-scores space using 2 PC.

The DUD (Directory of Useful Decoys) database[23] was originally developed to provide a realistic assessment of structure-based VS performance, even if it has been also used for ligand-based VS.[24] It contains 2950 active compounds against a total of 40 target proteins and every ligand has 36 decoy molecules that are physico-chemically similar but topologically distinct, leading to a database of 98,266 compounds. In this work the database was used to run separate LBVS for each target, using as templates a set of either 5% or 10% of the active compounds, which were selected using the MDC algorithm on the PCA t-scores space, as described above for the WOMBAT dataset. For each target the search was carried out on a database containing only the decoys and the active compounds not included in the template set.

ZINC is a free database of commercially-available compounds including around 8 million of compounds.[25] In this work, we used only the drug-like subset (subset #3, as defined in Zinc version 7), containing 2,066,906 "drug like" compounds, as an example of a large compound database.

GRIND

All the GRIND computed in this work correspond to the next generation (GRIND-2) of alignment-independent descriptors, developed around the idea of GRIND[7] but including some major improvements, like the use of AMANDA,[9] a novel MIF discretization algorithm which offers several advantages over the original GRIND

algorithm in terms of speed of calculation and quality of the hot spots extracted. However, for the sake of simplicity, they will be mentioned in the manuscript as GRIND and not as GRIND-2. These new descriptors are more suitable than the original GRIND for VS applications, since they can be obtained for series containing highly structurally dissimilar compounds without any manual adjustment of the algorithm and are much faster to compute. In this work, the GRIND-2 were obtained as implemented in program Pentacle,[26] using default settings (DRY, O, N1 and TIP probes, with 0.5Å grid step, dynamic parameterization, default AMANDA MIF discretization and default MACC2 with 0.8 smoothing window).

PCA

PCA is a multivariate analysis tool used for data supervision and dimensionality reduction. The method has been described elsewhere[27] but basically it works by computing an approximated representation of the original data matrix X, in terms of the product of two matrices; the matrix of objects T (scores) and the matrix of variables P (loadings). In the matrix T, every object is represented by a few number of new variables (Principal Components, PC), which are a linear combination of the original variables, chosen to explain as much as possible the variance present in X. In this work, PCA was applied on large matrices of GRIND descriptors using in-house software written in ANSI-C, implementing the original NIPALS algorithm.[27] In all applications, the X matrix was centered but not scaled.

Assessing the performance

The quality of the LBVS results obtained using diverse methods settings was quantified using standard metrics like ROC derived enrichment factor and, preferably,

the BEDROC. A description of these metrics together with a detailed discussion which can be found in recent reviews,[28] but they will be briefly described here. The enrichment factor (E)[10] used in this work was obtained from ROC curves, by describing the percentage of area under the curve (AUC) that is over the random ROC (a diagonal line), calculated as indicated by eq. 1

$$E = (AUC - AUC_R) / (AUC_T - AUC_R) \qquad \text{eq .1}$$

where AUC is the value of the area under the curve, AUC_T is the maximum value of the area under the curve (corresponding to the result in which the n active compounds correspond to the first n compounds selected) and AUC_R is the value of the area under the curve for a purely random identification (0.5 was used). All area values were normalized. According with this definition the E value will range between -1.0 and 1.0, the value of 0.0 corresponding to a random identification. This metric was reported here because is conceptually simple and fast to compute but has the disadvantage of being rather insensitive to the "early recognition" of active compounds.

The BEDROC[28] is a more sophisticated and reliable metric which emphasizes the early recognition of actives; a higher value is obtained when the actives are recovered early. The BEDROC is calculated using eq. 2

$$BEDROC = \frac{\sum_{i=1}^{n} e^{-\alpha r_i / N}}{\frac{n}{N}\left(\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}\right)} \times \frac{R_a \sinh(\alpha/2)}{\cosh(\alpha/2)-\cosh(\alpha/2-\alpha R_a)} + \frac{1}{1-e^{\alpha(1-R_a)}} \qquad \text{eq. 2}$$

where n is the number of known active structures, N is the number of inactive structures, r_i is the rank of the *ith* active structure, R_a is the ratio of active to inactive structures n/N. The coefficient α is a weighting factor which controls the weight assigned to the "early recognition", and higher α values displace the region of importance towards the beginning of the ranked list. In the present work the α value was set to 32.2, according to the recommendations in.[28]

Evaluating the similarity between different PCA spaces

The present work aims to validate the equivalence of the LBVS results yielded by GRIND-PP obtained after the application of PCA to the whole database and those obtained by projecting the GRIND of the whole database on a pre-calculated PCA model. The metric proposed here is based on the assessment of the similarity between the lists of hits obtained in VS searches and therefore GRIND-PP spaces are considered equivalent if the application of LBVS yields the same results. The procedure for comparing the results obtained with settings A and B works as follows: two lists with the 20 first compounds (an arbitrary small number) were extracted and the percentage of common elements present in the topmost positions of both lists was computed. As an additional indicator of the list similarity, the order of extraction of the common elements in both lists was compared by using Spearman correlation coefficient, obtaining values that range between 1.0 if the order of the common elements is identical (including non-common elements for calculating the ranks) and -1.0 if they were extracted in reversed order. During the application of this evaluation procedure we found a small number of highly similar compounds with identical GRIND-PP values, corresponding to database duplicates or equivalent compounds (i.e. tautomers or enantiomers), the position of which within the list is arbitrary. In these cases and, in

order to avoid artifactual results, elements which appear twice in the list with identical GRIND-PP were removed from the analysis. Please notice that both kind of indexes must be considered together, and the A and B results can be considered similar only if the percentage of common elements in top-most positions is high and the value of the Spearman correlation coefficient is positive and close to 1.

Here, and in order to obtain an exhaustive analysis, every single compound was used as a template in a separate LBVS run, so the calculated and the projected results were compared by using the aforementioned metrics averaged for all the compounds in the database. In these LBVS runs the following settings, identified as optimum in the previous tests, were used: 30 PC, minimum distance (the template is a single compound) and original scaling.

Results and discussion

Optimization of the method setting for LBVS

As stated before, LBVS results will be used as a tool for investigating the ability of the GRIND-PP to describe molecular similarity with an emphasis on biologically relevant features. However, before it can be applied with this purpose, the methodology used for computing GRIND-PP must be optimized. It must be remarked that, even if the application of the GRIND-PP in VS is not new and some applications have been published, no systematic study has ever explored the influence of some parameters involved in their computation. In particular, it is important to understand the effect of the t-scores space dimensionality and of the t-scores scaling in the quality of the results obtained. Other parameters that might also affect the results in LBVS (but not in other

molecular similarity applications) are those related with the template selection (% templates and method used for select them) and the method used to handle multiple templates. In this study we run a large number of LBVS queries, in which the parameters were systematically changed according to a full factorial pattern. For each run, the quality of the results was evaluated using the E and BEDROC parameters (computed as described in the Methods Section) and finally the effect of every method's settings on the quality was estimated using ANOVA.

LBVS queries were carried out trying to simulate realistic conditions. The initial round of tests were run on WOMBAT (Word of Molecular BioAcTivity)[16] using seven subsets of compounds with significant biological activity against known targets (see Table 1). Full details about the filtering of the database and the selection of the subsets were provided in the Methodology section.

GRIND descriptors were computed for all the compounds in the WOMBAT dataset, obtaining 930 variables. Principal Component Analysis (PCA) was applied to this matrix, extracting an excess of 50 principal components (PC) which explained more than 85% of the variance. For every target, a small subset of the active compounds (of about 5% or 10% of the original size) was extracted and used as template structures. The choice of the templates was made trying to include representatives from the diverse structural class present in the set of active compounds, by using two different subset selection algorithms; the longest minimum distance (LMD)[22] and the most descriptive compound (MDC).[21]

The query run using this template set extracted compounds from the database according to their similarity with the templates in terms of Euclidean distances in the t-

scores space. Since the template set contains more than one compound, this scoring distance can be computed as the minimum distance with any of the compounds in the template set or as the distance to the whole template set, represented by the centroid of all the compounds. Both options were tested in our study. In any case, compounds were selected according to distances, and therefore the scaling applied to the t-scores can be expected to have a large impact on the results. It has been previously hypothesized [8] that the use of the original PC scaling might produce unrealistic similarity scores due to the high degree of redundancy present in the GRIND descriptors. The application of a simple PC normalization (often called PC whitening[29]) will be able to remove this redundancy, and the Euclidean distances on the scaled scores will be equivalent to Mahalanobish distances. A more advanced alternative is to apply a scaling adjusted ad hoc for the series, using the following method: when the template set contains more than one compound, the PC_j can be scaled by a factor which reflects the ratio of the dispersion within the template set and the overall dispersion in the database (eq. 3)

$$Ratio_j = SD_{\text{j database}} / SD_{\text{j templates}} \qquad \text{eq. 3}$$

From its definition, the ratio will be large for PC which discriminate well template-like compounds from other compounds and therefore, when used as scaling factor, it will enhance the weight of such "good" PC in the computation of the distances. In the present study, the influence of the scaling in the quality of the LBVS results was tested by applying the three aforementioned scaling schemes to the PC: no-scaling, normalized and ratio.

The effect of the aforementioned parameters in the final quality of the results was tested by carrying out a systematic analysis in which the values of these parameters were set according to a full factorial experimental plan, summarized in Table 2, including a total of (2x2x2x3x2) LBVS queries for each of the seven targets listed in Table 1.

The quality of the results was quantified using E and BEDROC (see Methods section). In general, the quality indexes exhibit a large variability. For the BEDROC values, the average value for all the targets was 0.32, but with a large SD of 0.21, stressing the importance of using appropriate method settings for obtaining good results. In order to extract objective conclusions the data was imported to SPSS 12.0[30] and analyzed using ANOVA, with BEDROC as dependent variable and assuming a model with main effects only. The effects obtained were listed in Table 3, including the optimum values for all the settings studied.

The results show that the more important effect is the multiple templates handling method, followed in importance by the target set, the template selection method and the template set size. Remarkably, the effect of the PC scaling was not statistically significant and the number of PC ranked last in importance.

In this study, the settings related with how the template set is built seem to be critically important, showing that the best results were obtained when the template set was large (10%) and when it was selected using the MDC algorithm. In both cases, the best results are obtained when the settings increase the chances of incorporating in the template set of representatives for all the chemotypes present in the active set. This is not surprising in a ligand-based method and can be explained by the structural richness

present in the WOMBAT database, in which the active sets often contain several highly dissimilar structures.

The most statistically significant effect found in the study was the choice of method used to handle multiple templates; in all instances, the best results were obtained using the minimum distance scoring. This finding is consistent with previously reported results[31] and it can be justified here by the presence of well defined structural families in all the sets of active compounds. Algorithms based on the centroid of a set of bioactive structures could be expected to perform better when the templates are not too structurally diverse. Besides, the use of the centroid can serve to base the similarity search in common structural features, not necessarily associated to a certain chemotype and present in all the members of the template set. The results from this kind of search can be more useful in terms of the originality of the compounds found, but in a study like the present (with active compounds belonging to defined chemotypes), there is a high chance that the compounds extracted were not recognized as active and therefore rank low in terms of BEDROC or E.

With respect to the PC scaling method, contrarily to our expectation, the statistical analysis did not detect any significant relationship, thus indicating that no setting is producing consistently good results in all the runs. A detailed analysis shows that in most cases the best results were obtained with original scaling or ratio scaling, even if the differences in the BEDROC values obtained with the different methods were not large. It is remarkable that the results contradict our previous statement regarding the detrimental effect of the redundancy present in the original descriptors.

Also interesting, from a practical point of view is the effect of the number of PCs included in the search. In the analysis the best results were consistently obtained with 25 PC, in contrast with previous applications of GRIND-PP where a much smaller number of PC was used; 3 scaled PC in[10] and 2 unscaled PC in[11], probably justifying the discrete quality of the results reported. Based on these results we decided to investigate the effect of the number of PC in more detail, widening the range of PC explored, and carrying out additional runs using from 5 to 50 PC in 5 PC intervals, and setting the rest of the parameters to their optimal values (MDC algorithm, 10%, minimum distance score and ratio scaling) for three selected targets (5-HT3, thrombin and HIV-1 P). The results are shown in Figure 1a, representing the BEDROC versus the number of PC, which indicate that the optimum BEDROC values were obtained with 25-30 PC. Probably the BEDROC depends from the percentage of X variance explained by a certain number of PC, more than from the number of PC and therefore, in order to make the results more general, the same quality indexes were represented versus the variance explained in Figure 1b. This graph shows that the optimum results were obtained when the GRIND-PP explain between 75% and 80% of the variance present in the original matrix. It should be noticed, also, that the addition of more PC does not increase indefinitely the quality of the results; there is an optimum dimensionality for every target.

All the results obtained from this study can be used to define a set of optimum method settings. However, all the LBVS queries used to derive them were carried out in a single database and it would be sensible to check their general applicability by running additional queries on different databases. With this aim, our systematic study was extended by using the DUD database.[23] Unlike WOMBAT, DUD was specifically developed to test the performance of structure-based VS methods (see Methods section

for details). For this study, the template sets were built using the same methodology described for the WOMBAT dataset and the LBVS queries were run on a database containing the rest of the active compounds plus the decoys defined for the corresponding target (see the Methods section for details). The name of the targets, the number of active compounds and the size of the template sets were detailed in Table 4.

With respect to the method, some of the settings clearly identified as highly influential in the previous section (template selection method, and multiple template handling) were set fixed to their optimum values (MDC, and minimum distance). Conversely, the template sets size, the scaling and the number of PC was changed systematically, as summarized in Table 5, resulting in a full factorial design of 2x2x10 runs per target.

The results obtained in terms of BEDROC with the different settings for the 40 targets were analyzed using ANOVA, as described for the WOMBAT dataset. The statistical significance of the main effects, and the optimum values for each setting were shown in Table 6.

The variability observed for the different targets was rather high and the differences are statistically significant. The highest BEDROC values were obtained for dhfr (0.92). As obtained for WOMBAT, the template set size is an important method setting and the best results were obtained with the largest value (10%), probably for the same reasons discussed above. The PC scaling method follows in importance. Here, the effect is statistically significant and the best results were clearly obtained with the original scaling, instead of the ratio scaling. One of the possible reasons explaining the differences observed in WOMBAT and DUD with respect to this effect is the much smaller size of the template sets used in DUD (see Tables 1 and 4). The ratio scaling

was based on the dispersion of the PP values for the compounds belonging to the template set, and when this set is very short the scaling does not seem to behave as well as for larger template sets.

With respect to the number of PC, the effect observed was more statistically significant than the effect observed in WOMBAT, probably because here the experiments covered from the beginning a wider range of PC (from 5 to 50 in 5 PC intervals). Remarkably, there is also an optimum dimensionality and the quality, in terms of BEDROC, does not grow linearly with the number of PC, reaching a maximum between 20 to 30 PC. The relationship between the performance and the dimensionality was also studied using the percentage of X variance explained, as reported for WOMBAT, finding that the optimum values were obtained when the GRIND-PP explains between 70% and 80% of the original matrix variance. The results for the 10 targets with the highest BEDROC were represented in Figure 2.

These results further confirm the importance of choosing the right number of PC for obtaining good results as well as the relative stability of the optimum dimensionality. In general, for the datasets explored the optimum dimensionality is around 20 to 30 PC or the number required to explain between 70 to 80% of the X variance. Even if the optimum value can vary for different databases and queries, any value within this range guarantees a reasonable method performance, unlike the very short values used in prior GRIND-PP applications.

GRIND-PP performance test

The results obtained with optimum settings for the WOMBAT and DUD datasets, reported in the previous section, provided a first quantification of the GRIND-PP performance in LBVS applications (see Table 7 and Figure 3).

In WOMBAT, the BEDROC values rank from 0.41 for PKC to 0.81 for AT1, with an average value of 0.56. According with the meaning of the BEDROC scores, the method behaves well in terms of its early recognition capabilities, being able to extract a significant percentage of the active compounds in the top 5% of the hits.

In DUD, the average BEDROC value obtained for all the targets is very similar (0.55), thus allowing to draw similar conclusions even if the values exhibit a wider span, from 0.17 (cox1) to 0.92 (dhfr). This result is remarkable if we take into account the high content of decoy structures in this latter database, in particular for a ligand-based VS method. In both cases, the method compares well with other state-of-the-art methods,[32] even if the differences in the quality testing methodology do not allow a direct numerical comparison of the metrics.

In any case, as stated before, the aim of the present work is not to carry out a comparative analysis of the GRIND-PP with other methods in LBVS, but to evaluate in general the suitability of the GRIND-PP for representing molecular similarity. It is our belief that such analysis should not be limited to compare ROC curves and enrichment values but must also incorporate the potentialities of this 3D-based similarity method to extract original structures, and include a detailed analysis of the topmost structures obtained using diverse methods. This study is currently in progress in our group and will be submitted in due course.

Stability of the GRIND-PP spaces.

The applicability GRIND-PP in fields like LBVS is not related only with the quality of the results obtained and depends also of technological and practical issues. In this respect, it is important to consider if these descriptors can be used for characterizing extremely large databases. Today, databases containing several million compounds are not uncommon, particularly in corporate environments. Most PCA software is not ready to handle matrices with so many objects, and even if special software is applied, the process would be slow and not suitable for being applied after every database update. However, in most cases the addition of a few new compounds to a large X matrix is unlikely to change the results of the PCA, provide that the new compounds do not contain extremely different structural or physicochemical features. Indeed, in our experience, the PCA t-scores spaces obtained for large collections of compounds are relatively stable and the values of the t-scores assigned to the new compounds after the PCA are very similar to the projected t-scores ($\mathbf{T_p}$) which could be obtained very simply, as described by eq. 4.

$$\mathbf{T_p} = \mathbf{X}.\mathbf{P} \qquad \text{eq. 4}$$

where $\mathbf{T_p}$ are the projected t-scores, **X** is the matrix containing the (centered) descriptors for the new compounds and **P** the p-loadings of a pre-computed PCA model.

This observation suggests that it would be possible to build t-scores spaces by carrying out PCA on a "core dataset" of compounds and then expanding the space simply by applying eq. 4 to new compounds, without the need of recomputing the PCA for the whole database. A similar approach was used with success by Oprea for

obtaining a “chemspace map” (ChemGPS) with diverse applications in the field of drug discovery.[14]

As a part of the present study, we decided to validate the suitability of the projected t-scores for replacing PCA computed t-scores in LBVS applications. A numerical comparison of the t-scores would not be useful, since their equivalence for characterizing molecular similarity depends only on how well the relative distances between the compounds is preserved. For this reason, the test was based on carrying out LBVS using both kinds of GRIND-PP and comparing the results obtained, with an special emphasis on the top-most compounds.

The first test was run on a large subset of the Zinc database. The objective was to test the similarity of the results obtained with original GRIND-PP computed after a PCA on the whole set and GRIND-PP obtained by projecting the GRIND of core databases of diverse size. The procedure started by selecting randomly 100,000 compounds from the Zinc database (see Method section for details), computing GRIND and obtaining regular GRIND-PP (t-original) for them. Then, 5 subsets of 1000, 5000, 10000, 25000 and 50000 compounds were randomly selected, and 5 different core PCA models were obtained. These were used to project the GRIND of the 100,000 compound set, thus obtaining five projected GRIND-PP (t-1K, t-5K, t-10K, t-25K and t-50K). The potential equivalence of such t-spaces for being applied in LBVS applications was thoroughly tested by carrying out a VS query for every compound in the dataset, in which this compound was used as template and a ranked list of the 20 more similar compounds was extracted (see Method sections for details). For every compound, the lists obtained in the projected t-spaces were compared with those obtained in t-original and their

similarity was quantified using different indexes, described in detail in the Method section.

Table 8 shows that the results obtained with the projected t-scores are, in general, rather similar to those obtained in t-original. Even for the smaller subset (t-1K), the first hit is found as first or second hit in 95% of cases. The comparison of the results obtained projecting in t-10K (10% of the original size) shows that the first hit is identical in both lists in 92% of cases and first or second in a 99% of cases. The high percentages of common hits and the large Spearman correlation coefficient obtained also indicates that the correspondence is rather general for all the list members and does not affect only the topmost. All in all, these results confirm our prior experience, and avail the use of projected GRIND-PP for LBVS.

Against this conclusion it can be argued that the subsets were selected randomly from the database and therefore they can be considered random samples of the same chemical space. In pharmaceutical industry, compound databases are often enriched with batches of compounds belonging to novel projects and therefore the chemical space represented by the database will diverge more and more from the original one from which core datasets were extracted. In order to test this scenario a second test was run, in which the whole WOMBAT database was projected using subsets of 50,000 and 100,000 compounds extracted from the Zinc database. Figure 4 represents the WOMBAT database using the first and second PC of the t-original (4a) and the projected t-5K (4b). In both plots, the shape of the database is rather similar, but the projected t-scores appear slightly rotated with respect to the original. The effect of these differences in the relative distances between the compounds cannot be appreciated by visual inspection of the plots and require a comparison of the LBVS results.

The procedure used for the comparison of the original and the projected t-scores was identical to the method used in the previous run. The results were shown in Table 8, and are clearly worse than those obtained for compounds of the same database. However, even for the smaller dataset tested (50,000 compounds), the first hit is found as first or second hit in 93% of the cases, the mean percentage of common hits is over 83% and the Spearman correlation coefficient is rather high. In our opinion, the results indicate that the projected t-scores spaces are reasonably stable, and the projection methodology can be safely used as a fast method to update GRIND-PP databases, in most cases. However, the addition of novel compounds representing diverse chemical spaces can deteriorate progressively the quality of the values obtained and therefore, the core dataset must be updated from time to time before it wears out.

Conclusions

The results obtained confirm that the GRIND-PP are promising molecular descriptors for applications requiring a biologically relevant representation of compound similarity. The application of PCA compacts the original GRIND into a few number (20 to 30) of information rich PC, easy to store, and to apply in many computational methods. The study shows that optimum results in some typical applications can be obtained with a limited number of PC, which can be easily assessed by the percentage of variance explained (around 70-80%) and does not grow indefinitely. For LBVS applications the best results were obtained using GRIND-PP without any scaling or with an ad-hoc ratio scaling, contradicting previous statements recommending the application of normalization. These conclusions can probably be extended to other t-scores descriptors derived from other kinds of molecular descriptors.

With respect to the suitability of the GRIND-PP for being used in large collection of compounds, the results obtained shown that the PP can be obtained in a simple and fast way from projection on a core PCA model obtained for a small set of representative compounds. The properties of the projected descriptors were thoroughly compared with the PCA derived ones, in terms of the results obtained with both in VS applications. In general, the results support the use of the projected PP in practical applications, provided that the core PCA model was obtained using a set with enough compounds (around 10% or the full database size), which are reasonably similar to the projected compounds.

Most of the above investigations were carried out using LBVS, as a reference technique requiring an accurate description of the compound similarity. The testing allowed a fine tuning of the method parameters for obtaining good results in terms of high BEDROC and provided a first evaluation of the performance of the method. However, the true relevance of these results to assess the performance of GRIND-PP in LBVS has to be considered with care, because the standard quality indexes used here do not reflect the originality of the structures extracted and are biased by the fact that the active compounds often belong to the same chemotype than the templates. Therefore, the results reported here reflect mainly the ability of the GRIND-PP to describe generic molecular similarity, including non-biologically relevant chemotype features, which defines the bottom-line of the descriptors quality. Even under these non-favorable testing conditions, the t-scores space performed well, obtaining results which can be compared with other state-of-the-art methods. A more complete study, specifically

aimed to validate the usefulness of the proposed methodology for obtaining original hits is now in progress and the results will be reported in due term.

Future developments of this work will involve the testing and validation of GRIND-PP for diverse purposes. Among them, structure masking[33] appears like a highly interesting application, for which the GRIND-PP have unique properties. Good masking descriptors must encode chemical structure into biologically relevant descriptors, from which it would not be possible to guess the original structure. In this respect the GRIND-PP are not only highly relevant for representing biological properties, the peculiarity of being obtained using a PCA projection makes the resulting t-scores impossible to revert into the original GRIND without the p-loadings of the core PCA dataset. Therefore, as far as this information is kept confidential the t-scores cannot be reverse engineered to guess the compound structure. It must be emphasized that this is not only a technological barrier; during the projection only a certain percentage of the GRIND is retained (between 70% and 80% for optimum results), while the remaining information is irreversibly lost, thus making a backprojection virtually impossible.

All in all, the results of this work provide very useful information regarding the application of GRIND-PP for the description of molecular similarity and demonstrate that they produce results at least comparable with other state-of-the art methods in LBVS. This is a first step before they can be applied for drug design tasks requiring an accurate and biologically relevant description of the molecular similarity and exploiting the unique properties of these descriptors.

Acknowledgments

We thank Molecular Discovery Ltd. for supporting this research, including a grant to one of us (AD). The project also received partial founding from the Spanish Ministerio de Educación y Ciencia (project SAF2005-08025-C03) and the Instituto de Salud Carlos III (Red HERACLES RD06/0009). We also thank Tudor Oprea for kindly providing us the WOMBAT database.

Figure 1. BEDROC values obtained for a few representative WOMBAT targets (HIV-1 P, thrombin, 5-HT3) plotted against the number of PC (a) and the X variance explained (b).

Figure 2. BEDROC values obtained for 10 representative DUD targets (cox2, dhfr, gart, hsp90, me, p38, pnp, ppar_gamma, thrombing and trypsin) plotted against the number of PC (a) and the percentage of X variance explained (b).

Figure 3. Values of BEDROC obtained for diverse targets in database WOMBAT (top) and DUD (bottom). See text for details.

Figure 4. Scatterplot representing the WOMBAT databases using the first and second PC, obtained (a) from a complete PCA and (b) as a projection on model obtained with a core database of 50,000 Zinc compounds.

Table 1. Targets studied in WOMBAT database.

target name	num. actives	num. templates (10%)	num. templates (5%)
5-HT3	1166	117	56
5-HT1A	3501	351	176
D2	3350	335	168
AT1	894	90	45
thrombin	850	85	43
HIV-1 P	184	19	10
PKC	166	17	9

Table 2. Number of levels and values tested in the full factorial experimental plan used in WOMBAT.

Variable	num. levels	values
template selection method	2	MDC, LMD
template set size	2	5%, 10%
multiple template handling	2	minimum distance, centroid
PC scaling	3	no scaling, normalize, ratio
PC number	2	10, 25

Table 3. Statistical significance of the main effects and best value settings.

Variable	F	p	best value
target	201.6	<0.001	AT1
template selection method	154.7	<0.001	MDC
template set size	32.4	<0.001	10%
multiple template handling	698.7	<0.001	minimum
PC scaling	1.2	0.294*	ratio*
PC number	10.0	0.002	25PC

* non-significant effect at 95% CI

Table 4. Targets studied in DUD database.

name	num. actives	num. templates (5%)	num. templates (10%)
ace	49	3	5
ache	107	6	12
ada	39	2	4
alr2	26	2	3
ampc	21	2	3
ar	79	4	8
cdk2	72	4	8
comt	11	1	2
cox1	25	2	3
cox2	426	22	43
dhfr	410	21	41
egfr	475	24	48
er_agonist	67	4	7
er_antagonist	39	2	4

fgfr1	120	6	12
fxa	161	8	15
gart	40	2	4
gpb	52	3	6
gr	78	4	8
hivpr	62	4	7
hivrt	43	3	5
hmga	35	2	4
hsp90	37	2	4
inha	86	5	9
mr	15	1	2
na	49	3	5
p38	454	23	46
parp	35	2	4
pde5	88	5	9
pdgfrb	170	8	17
pnp	50	3	5
ppar_gamma	85	5	9
pr	27	2	3
rxr_alpha	20	1	2
sahh	33	2	4
src	159	8	16
thrombin	72	4	8
tk	22	2	3
trypsin	49	3	5
vegfr2	88	5	9

Table 5. Number of levels and values tested in the full factorial experimental plan used in DUD.

Variable	num. val	values
template set size	2	5%, 10%
PC scaling	2	no scaling, ratio
PC number	10	5 to 50, in 5 unit steps

Table 6. Statistical significance of the main effects and best values settings.

Variable	F	p	best value
target	171.3	<0.001	dhfr
template set size	533.0	<0.001	10
PC scaling	283.9	<0.001	original
PC number	18.0	<0.001	30

Table 7. Summary of the results obtained with the best method settings.

		WOMBAT	DUD
E	min.	0.54	0.02
	max.	0.95	0.92
	average	0.74	0.55
	median	0.75	0.58
	sd.	0.13	0.24
BEDROC	min.	0.41	0.17
	max.	0.81	0.92
	average	0.56	0.55
	median	0.55	0.54
	sd.	0.14	0.17

Table 8. Similarity between the results obtained with complete t-scores and projected t-scores obtained from core sets of diverse sizes.

database	core size	Spearman	%common	%first	%first-sec.
Zinc	1,000	0.819	85.5	83.8	94.9
	5,000	0.917	91.6	89.4	97.7
	10,000	0.946	93.8	91.9	98.5
	25,000	0.976	96.4	95.3	99.5
	50,000	0.986	97.5	96.8	99.8
WOMBAT	50,000	0.806	83.2	83.5	92.7
	100,000	0.823	84.5	84.9	93.6

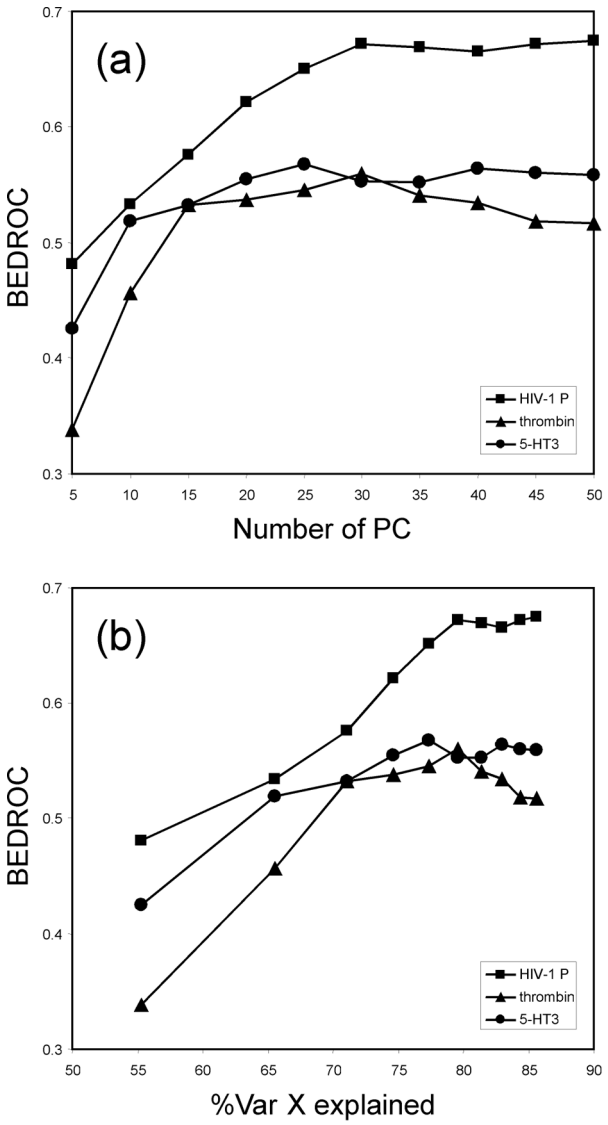
(a)
BEDROC
Number of PC
HIV-1 P
thrombin
5-HT3
(b)
BEDROC
%Var X explained
HIV-1 P
thrombin
5-HT3

Figure 1

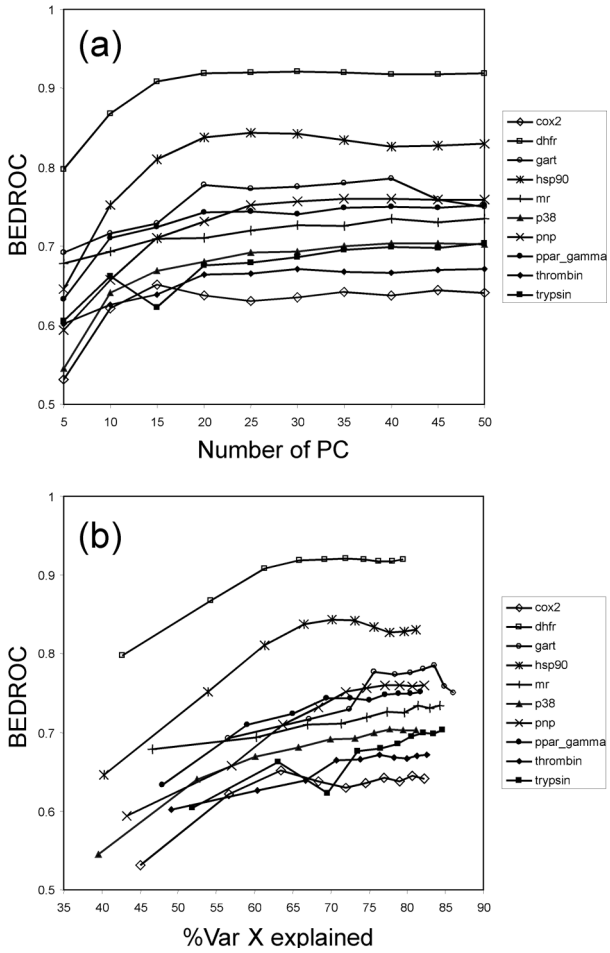
(a)
BEDROC
Number of PC
cox2
dhfr
gart
hsp90
mr
p38
pnp
ppar_gamma
thrombin
trypsin
(b)
BEDROC
%Var X explained
cox2
dhfr
gart
hsp90
mr
p38
pnp
ppar_gamma
thrombin
trypsin

Figure 2

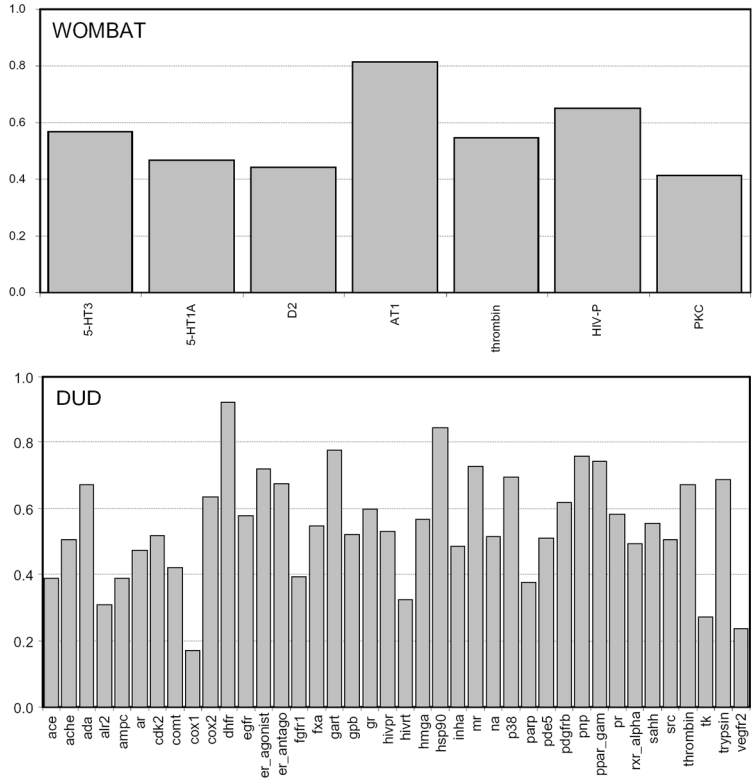

Figure 3

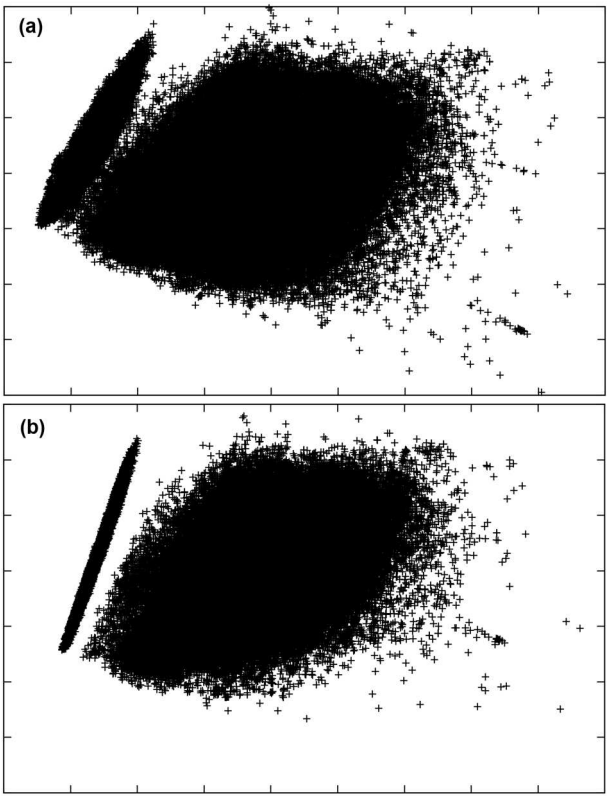

Figure 4

References

1. Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002***, 19,* 4350-4358.

2. Johnson, M.; Lajiness, M.; Maggiora, G. Molecular similarity: a basis for designing drug screening programs. *Prog. Clin. Biol. Res.* **1989**, *291*, 167-171.

3. Lajiness, M. S.; Johnson, M. A.; Maggiora, G. M. Implementing drug screening programs using molecular similarity methods. *Prog. Clin. Biol. Res.* **1989**, *291*, 173-176.

4. Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998***, 6,* 983-996.

5. Hristozov, D.; Oprea, T. I.; Gasteiger, J. Ligand-based virtual screening by novelty detection with self-organizing maps. *J. Chem. Inf. Model.* **2007***, 6,* 2044-2062.

6. Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* **2008***, 5,* 941-948.

7. Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000***, 17,* 3233-3243.

8. Pastor, M. Alignment-independent Descriptors from Molecular Interaction Fields. In *Molecular Interaction Fields, Applications in Drug Discovery and ADME Prediction*; Cruciani, G., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 117-143.

9. Duran, A.; Martinez, G. C.; Pastor, M. Development and validation of AMANDA, a new algorithm for selecting highly relevant regions in Molecular Interaction Fields. *J. Chem. Inf. Model.* **2008***, 9,* 1813-1823.

10. Gregori-Puigjane, E.; Mestres, J. SHED: Shannon entropy descriptors from topological feature distributions. *J. Chem. Inf. Model.* **2006***, 4,* 1615-1622.

11. Carosati, E.; Mannhold, R.; Wahl, P.; Hansen, J. B.; Fremming, T.; Zamora, I.; Cianchetta, G.; Baroni, M. Virtual screening for novel openers of pancreatic K(ATP) channels. *J. Med. Chem.* **2007***, 9,* 2117-2126.

12. Skagerberg, B.; Bonelli, D.; Clementi, S.; Cruciani, G.; Ebert, C. Principal Properties for Aromatic Substituents. A Multivariate Approach for Design in QSAR. *Quant. Struct.-Act. Relat.* **1989***, 1,* 32-38.

13. Hellberg, S.; Sjoestroem, M.; Skagerberg, B.; Wold, S. Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem.* **1987***, 7,* 1126-1135.

14. Oprea, T. I.; Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **2001***, 2,* 157-166.

15. Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. ChemGPS-NP: tuned for navigation in biologically relevant chemical space. *J. Nat. Prod.* **2007***, 5,* 789-794.

16. Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery;* Oprea, T. I., Ed.; Wiley-VCH: New York, 2004; pp 223-239.

17. Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *5,* 1308-1315.

18. Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput. Aided Mol. Des.* **2008**, *3-4,* 169-178.

19. *CORINA*, version 2.4; Molecular Networks GmbH: Erlangen, Germany, 2001.

20. *GOLPE*, version 4.6.0; Multivariate Infometric Analysis Srl.: Perugia, Italy, 2003.

21. Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J.; Osman, J. Parameter Based Methods for Compound Selection from Chemical Databases. *Quant. Struct.-Act. Relat.* **1996**, *4,* 285-289.

22. Marengo, E.; Todeschini, R. A new algorithm for optimal, distance-based experimental design. *Chemometrics Intellig. Lab. Syst.* **1992**, *1,* 37-44.

23. Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *23,* 6789-6801.

24. von Korff, M.; Freyss, J.; Sander, T. Comparison of Ligand- and Structure-Based Virtual Screening on the DUD Data Set. *J. Chem. Inf. Model.* **2009**, *49*, 209-231.

25. Irwin, J. J.; Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *1,* 177-182.

26. *Pentacle,* version 1.0.3; Molecular Discovery Ltd.; Perugia, Italy, 2009.

27. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometrics Intellig. Lab. Syst.* **1987**, *1-3,* 37-52.

28. Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007***, 2,* 488-508.

29. De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. L. The Mahalanobis distance. *Chemometrics Intellig. Lab. Syst.* **2000***, 1,* 1-18.

30. *SPSS*, version 12; SPSS Inc.: Chicago, IL., 2003.

31. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004***, 3,* 1177-1185.

32. Hristozov, D. P.; Oprea, T. I.; Gasteiger, J. Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. *J. Comput. Aided Mol. Des.* **2007***, 10-11,* 617-640.

33. Masek, B. B.; Shen, L.; Smith, K. M.; Pearlman, R. S. Sharing chemical information without sharing chemical structure. *J. Chem. Inf. Model.* **2008***, 2,* 256-261.

For Table of Contents use only

Suitability of GRIND-based principal properties for the description of molecular similarity and ligand-based virtual screening.

Ángel Durán, Ismael Zamora, Manuel Pastor*