Structural similarity based kriging for quantitative structure activity and property relationship modeling

Ana L. Teixeira[*,†,‡] and Andre O. Falcao[†,¶]

LaSIGE, Faculty of Sciences, University of Lisbon

E-mail: ateixeira@lasige.di.fc.ul.pt

1. Discriminate molecules with different property/activity value based on structural dissimilarity

1.1 Dataset A - Dihydrofolate reductase (DHFR) inhibitors activity

In order to analyze the difference in the activity value of each pair of molecules solely based in their structural distance, Figure 1 displays plots showing on the left side the distribution of the pairwise distance between the 237 compounds in the training set, totalizing 27966 unique pairs of structures (excluding self-distances), calculated using a) molecular descriptors, c) fingerprints and e) NAMS and the corresponding absolute difference in the pIC_{50} value. While on the right side, Figure 1 displays plots showing at each level of pairwise distance, using the b) molecular descriptors, d) fingerprints and f) Noncontiguous Atom Matching Structural Similarity (NAMS) for such calculation, the probability of finding a pair of compounds with a certain absolute difference in the pIC_{50} value.

[*]To whom correspondence should be addressed
[†]LaSIGE, Faculty of Sciences, University of Lisbon
[‡]CQB - Centro de Química e Bioquímica, Faculty of Sciences, University of Lisbon
[¶]Department of Informatics, Faculty of Sciences, University of Lisbon

In general, it is possible to verify that NAMS and fingerprints are able to discriminate the compounds in the training set according to their pairwise distance versus their difference in the pIC_{50} value, especially for most similar pairs of compounds, verifying the similarity principle,[1] since the plot of pairwise distance values versus absolute difference in the pIC_{50} values (Figure 1 - c) and e)) for the set of molecules exhibit a trapezoidal distribution, revealing a neighborhood behavior with a low frequency of pairs in the upper left triangle (very similar compounds with a high difference in the property value). In the probability distribution plot, it can be observed (Figure 1 - d) and f)) that for both Fingerprints and especially for NAMS there is an high probability for compounds that are very close to each other to have a small difference in the property value. Nevertheless, fingerprints have a higher number of similar pairs of compounds (even 100% similar) with a higher difference in the property values than NAMS which is contrary to the similarity principle. When using molecular descriptors, the relationship between the pairwise distance of the compounds and their difference in the property is not as clear, even though there is a tendency for pairs of very dissimilar compounds have higher differences in the property value (Figure 1 - a) and b)).

1.2 Dataset B - Aqueous Solubility

For the aqueous solubility dataset the distribution of the pairwise distance between the 1033 compounds in the training set is similar to the previous dataset. Figure 2 plots the 533028 different pairs of structures (excluding self-distances) calculated using a) molecular descriptors, b) fingerprints and c) NAMS and the corresponding difference in the aqueous solubility absolute value.

In general, it is possible to verify that NAMS and fingerprints are able to discriminate the compounds according to their pairwise distance versus their difference in the aqueous solubility value and verify the similarity principle,[1] since the plot of pairwise distance values versus absolute difference in the aqueous solubility values (Figure 2 - c) and e)) for the set

of molecules exhibit a trapezoidal distribution, revealing a neighborhood behavior with a low frequency of pairs in the upper left triangle (very similar compounds with a high degree of difference in the property value). In the probability distribution plot, it can be observed (Figure 2 - d) and f)) that for both Fingerprints and especially for NAMS there is an high probability for compounds that are very close to each other to have a small difference in the property value. While using molecular descriptors the relationship between the pairwise distance of the compounds and their difference in the property is not as clear and the discrimination of the compounds is more complicated, since there is a high density of pairs with high similarity values and high differences in the property value. The only tendency that is shown for molecular descriptors is that for high distance scores the property value is also dissimilar.

References

(1) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.

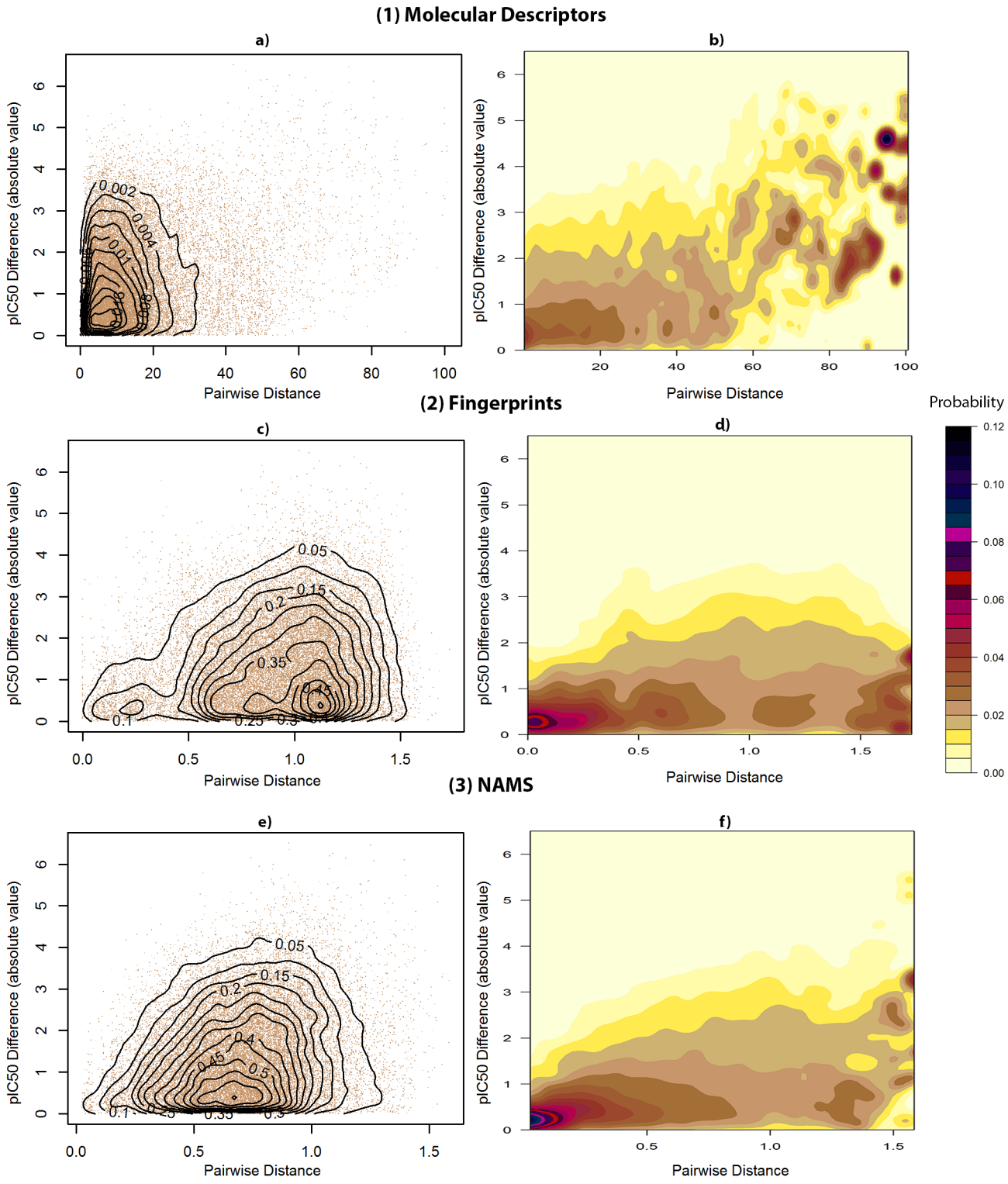

Figure 1: On the left side the plots represent the distribution of pairwise distance between pairs of compounds (training set A), calculated using *a)* molecular descriptors, *c)* fingerprints and *e)* NAMS and the corresponding absolute difference in the pIC_{50} value. The contour lines represent two-dimensional kernel density of the pairwise distance between pairs of compounds and the respective absolute difference in the pIC_{50} value. On the right side the plots show at each level of pairwise distance, using the *b)* molecular descriptors, *d)* fingerprints and *f)* NAMS for such calculation, the probability of finding a pair of compounds with a certain absolute difference in the pIC_{50} value.

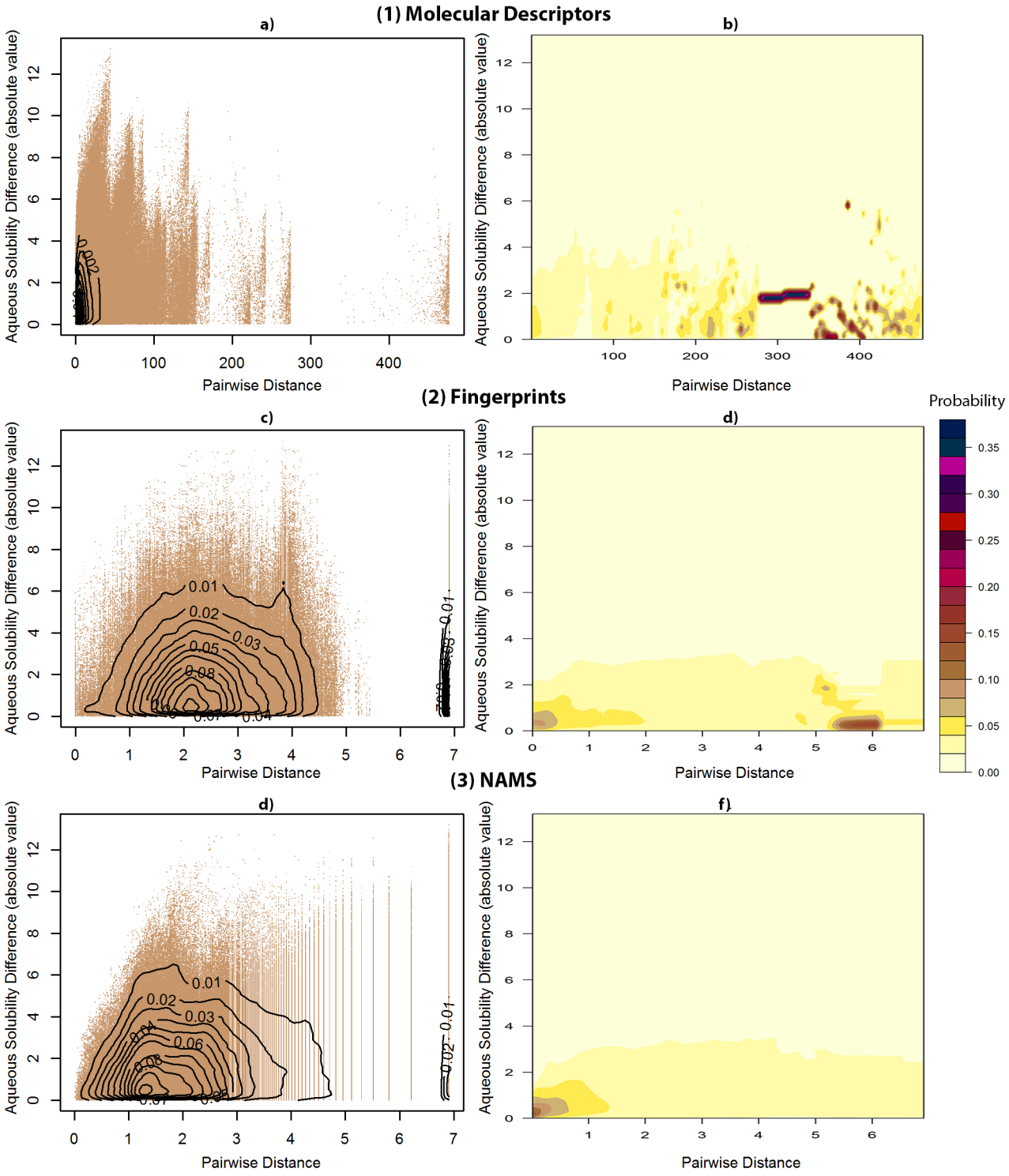

Figure 2: On the left side the plots represent the distribution of pairwise distance between pairs of compounds (training set B), calculated using *a)* molecular descriptors, *c)* fingerprints and *e)* NAMS and the corresponding absolute difference in the aqueous solubility value. The contour lines represent two-dimensional kernel density of the pairwise distance between pairs of compounds and the respective absolute difference in the aqueous solubility value. On the right side the plots show at each level of pairwise distance, using the *b)* molecular descriptors, *d)* fingerprints and *f)* NAMS for such calculation, the probability of finding a pair of compounds with a certain absolute difference in the aqueous solubility value.