Article:
Mazalan, L., Bell, A.J. orcid.org/0000-0002-8268-5853, Sbaffi, L. et al. (1 more author) (2017) Cross-classified multilevel modelling of the effectiveness of similarity-based virtual screening. ChemMedChem. ISSN 1860-7179

https://doi.org/10.1002/cmdc.201700487

Cross-Classified Multilevel Modelling of the Effectiveness of Similarity-Based Virtual Screening

Lucyantie Mazalan[b], Andrew Bell[c], Laura Sbaffi[d] and Peter Willett*[a]

[a] Prof. P. Willett, ORCID 0000-0003-4591-7173
Information School
University of Sheffield
Western Bank, Sheffield S10 2TN, UK
E-mail: p.willett@sheffield.ac.uk
[b] Ms. L.B. Mazalan, ORCID 0000-0002-3382-9099
Information School
University of Sheffield
Western Bank, Sheffield S10 2TN, UK
[c] Dr. A.J. Bell, ORCID 0000-0002-8268-5853
Sheffield Methods Institute
University of Sheffield
Western Bank, Sheffield S10 2TN, UK
[d] Dr L. Sbaffi, ORCID 0000-0003-4920-893X
Information School
University of Sheffield
Western Bank, Sheffield S10 2TN, UK

Abstract: The screening effectiveness of a chemical similarity search depends on a range of factors, including the bioactivity of interest, the types of similarity coefficient and fingerprint that comprise the similarity measure, and the nature of the reference structure that is being searched against a database. This paper introduces the use of cross-classified multilevel modelling as a way to investigate the relative importance of these four factors when carrying out similarity searches on the ChEMBL database. Two principal conclusions can be drawn from the analyses: that the fingerprint plays a more important role than the similarity coefficient in determining the effectiveness of a similarity search; and that comparative studies of similarity measures should involve many more reference structures than has been the case in much previous work.

Introduction

Similarity searching is one of the simplest, most widely used forms of ligand-based virtual screening in drug discovery programmes. The approach is based on the empirical observation – normally referred to as the 'similar property principle' - that molecules that are structurally similar to each other tend to have the same chemical, physical and biological properties.[1-6] Given a reference structure, *R*, with some desired biological activity and a database of structures that have not previously been tested for that particular bioactivity, a similarity search involves comparing *R* with each database structure in turn to determine the degree of inter-molecular structural similarity, and then returning those database structures that have the largest computed similarities to *R*. These nearest-neighbours are then candidates for biological screening, since the similar property principle indicates that they are the molecules with the highest *a priori* probabilities of exhibiting the desired activity. The effectiveness of the search can then be assessed by the extent to which the tested molecules do in fact prove to be bioactive.

At the heart of similarity searching is the measure that is used to quantify the degree of resemblance between two molecules. A measure has three components: the representation, or descriptor, that is used to characterise the two molecules that are being compared (with 2D fingerprints being by far the most common form of representation in current chemical information systems); the weighting scheme that is used to reflect the relative importance of different parts of the representation (though, as in the work reported here, most studies have considered binary, unweighted representations); and the similarity coefficient that is used to quantify the degree of resemblance between two appropriately weighted structural representations. Given these various factors, it is hardly surprising that many comparative studies have been reported that seek to identify those methods that yield the most effective searches.[7-12] These studies typically focus on the effect of variations in one particular characteristic, with other factors (such as the structures in the database that is being searched) held constant across a set of experiments. Such procedures enable the identification of, e.g., the 'best' similarity coefficient, but often only in the context of specified values of the other characteristics, and are usually unable to say anything as to the relative importance of the various factors. In this paper we report an alternative approach, in which a cross-classified multilevel model is developed that enables us to compare the importance of similarity coefficients and fingerprints (and other factors) in their effects on the overall screening effectiveness of a similarity search, and to find out the best and worst performing similarity coefficients and fingerprints.

Results and Discussion

The results from fitting the initial model in Eq. (1) (see Experimental Section) for all 46,500 similarity searches of the ChEMBL subset are listed in Table 1, which reports the parameter estimates and associated standard errors estimated

for β_0 and the variances of each of the four components (i.e., activity class, fingerprint, similarity coefficient, and residual error) in the model. Thus, the mean enrichment factor across all levels is estimated to be 12.725 with a standard error of 1.857; and the between-activity class variance is estimated to be 54.170 with a standard error of 23.635 *etc.* The same format has also been adopted for the results presented in Tables 2-4 discussed below, with the first pair of values in each row reporting the parameter estimate and the associated standard error for βo, and subsequent pairs the variance and associated standard error for each factor. It will be seen that the residual error in Table 1 is notably larger than the variances estimated for the other three factors, and of these the activity class variance is far greater than those for the fingerprint and the similarity coefficient (whose effects on the enrichment are of comparable magnitudes).

That the enrichment is strongly dependent on the type of activity class is to be expected, since a homogeneous set of active molecules is likely to yield high values for the enrichment factor in similarity searches, whereas this is unlikely to be the case with heterogeneous actives that are not strongly clustered together in chemical space. The large residual error variance in Table 1 describes the random variation from one search to another, suggesting that the nature of the individual reference structure also plays an important role in the enrichment that will be obtained. This is again not surprising since, even with a relatively homogeneous set of active molecules, the presence of an unusual sidechain or of a differently functionalized heterocycle can lead to marked variations in effectiveness, depending on the extent to which other actives are grouped around the chosen reference structure. Table 2 hence shows the results that were obtained when the basic model shown previously (Eq. 1) was extended by the inclusion of an additional term representing the effect of variations in the reference molecule that was used for the similarity searches in each activity class. It will be seen that there has been a substantial reduction in the magnitude of the residual error, with the activity class and the reference structure exhibiting by far the largest variances. Further sets of experiments were hence conducted to compare the relative roles of the fingerprint and the coefficient when the activity class and the reference structure respectively were held constant.

Tables 1 and 2 about here

First, 15 models were generated, each describing the 3,100 (i.e., 10 reference structures × 31 similarity coefficients × 10 fingerprints) distinct searches for one of the 15 activity classes. The results of these runs are shown in Table 3, where it will be seen that the residual error has been much reduced; instead, the fingerprint makes the largest contribution to the models with the sole exception of the phosphodiesterase searches, where the type of similarity coefficient is the largest contributor. We can hence conclude that the fingerprint component of a similarity measure in general has a greater influence on the effectiveness of screening than does the similarity coefficient. In the final set of experiments, 150 models were generated, each model describing the 310 (i.e., 31 similarity coefficients × 10 fingerprints) distinct searches for one of the ten reference structures in each of the 15 activity classes. The results for one of these sets of 10 models – those for the serotonin transporter (denoted by 5HT) activity class – are shown in Table 4. Inspection of this table shows that the influence of the residual error has been substantially reduced, and provides the largest contribution to the performance only once (in the ninth model); for the other nine models it always contributes less than does the fingerprint, but contributes more than the similarity coefficient in seven of them. It will be seen that the fingerprint contribution exceeds that of the similarity coefficient in all but the second model, and a similar pattern of behaviour was observed for all of the other activity classes: the fingerprint contribution exceeded that of the similarity coefficient in 136 of the 150 models; and the residual error provided the largest contribution in only 18 of the 150 models. These sets of experiments provide strong evidence for concluding that an appropriate choice of fingerprint is more important than the choice of similarity coefficient when constructing a similarity measure for ligand-based virtual screening.

Tables 3 and 4 about here

The relative degrees of importance of the various factors can be demonstrated graphically, as shown in Figure 1, which also shows the relative rankings of the various factors. The top-left part of the figure shows the effect of variations in the activity class. Here, each of the 15 points represents the mean enrichment factor when averaged over all of the 3,100 individual searches that involved a particular activity class. It will be seen that the enrichment factors are widely spread from the best-performing Type-1 Angiotensin II receptor searches (denoted here by AT1) down to the worst-performing 5HT searches at the right-hand end of the plot. There is a still greater degree of spread for the mean enrichment factors when averaged over the 310 searches that involved each of the 150 different reference structures, as shown in the bottom-right of the figure. The searches involving the ten different fingerprints (top-right of the figure) are also well dispersed (with the best results coming from use of the MorganR2 fingerprint that is analogous to the widely used ECFP4 fingerprint), but this is not the case with the searches involving the 31 different similarity coefficients shown in the bottom-left of the figure. Here, the identifier Bx denotes the x-th of the 51 different similarity coefficients studied by Todeschini *et al.*[11, 12], with the bext results coming from use of the Maxwell-Pilliner coefficient and with B3 in the figure being the Tanimoto coefficient that is the *de facto* standard for molecular similarity studies. It will be seen that, while there are a few poorly performing coefficients, the majority of them give broadly comparable mean enrichments. This implies that as long as one avoids the very weak performers here, a change in the similarity coefficient used is unlikely to have a significant effect on the screening ability of a similarity search system.

Figure 1 about here

As noted in the Introduction there are many comparative studies that consider the effect of variations in, e.g., the similarity coefficient on the effectiveness of similarity searching. Of these, the one most closely related to the present study is that described by Sastry *et al*., who conducted an extended series of similarity searches on a sample of the *MDL Drug Data Report* (MDDR) database and involved a systematic variation of the similarity coefficient, the fingerprint, the atom-typing and the weighting scheme.[9] Their comparison sought to identify the most generally useful settings for each of these parameters, and they concluded that an appropriate choice of parameter settings could result in high levels of screening effectiveness (as measured by the enrichment factor), though they also noted that no single combination of settings was ideal across the eleven activity classes from the MDDR database that were tested. Their study differs from that reported here in that they were seeking to identify the best combination(s) of parameter settings, whereas the present study has sought to establish the relative degrees of importance of the various parameters. That said, there is a fair measure of agreement between the two studies in that Sastry *et al*. found that the best overall fingerprint performance in their experiments was given by MOLPRINT2D, a circular fingerprint that is not markedly different in character from the fingerprints derived from the Morgan algorithm that are at the top-left of the fingerprint plot in Figure 1. They also noted that while there was much less variation in screening effectiveness between most of the twelve similarity coefficients that were tested, there were some that performed significantly worse than the others (as is clearly the case in the bottom-left portion of Figure 1).

The key role played by the individual reference structure that has been observed here provides a rationale for the findings of a study of the numbers of reference structures that were required to enable robust conclusions to be drawn as to the utilities of different types of similarity measure.[13] Arif *et al*. carried out similarity searches using 20 different similarity measures (based on five different binary fingerprints and four different similarity coefficients) with the aim of ranking these measures in order of decreasing effectiveness when averaged over multiple reference structures for each of six different MDDR activity classes. They found that rankings obtained using small samples of reference structures could be markedly different from those resulting from use of all of the available reference structures. This finding is just what would be expected if there are considerable differences in effectiveness between one search and another, and highlights a potential limitation of previous comparative studies (including many of those carried out in our laboratory) of similarity searching that have used only small numbers of reference structures. While showing that small samples of reference structures could yield misleading results, Arif *et al*. did not suggest any threshold number that should be employed, and this might hence usefully provide a focus for future studies of similarity searching

Conclusions

Similarity-based virtual screening is an important technique for use in the lead-discovery stage of pharmaceutical and agrochemical research programmes. There has thus been much interest in the development and evaluation of measures of inter-molecular structural similarity, in particular measures that are based on the use of binary association coefficients and of 2D fingerprint representations of molecular structure. Previous studies have evaluated many different coefficients and many different fingerprints to identify those that are most effective in identifying potential bioactive molecules. In the work reported here, we have considered both of these components of a similarity measure using cross-classified multilevel modelling and demonstrated that the type of fingerprint plays a much greater role in determining screening effectiveness than does the similarity coefficient. It is hence suggested that future work on the optimisation of similarity measures should prioritise the identification of the most appropriate type of fingerprint and that relatively less attention be devoted to the evaluation of similarity coefficients. Our results additionally suggest that comparative studies of similarity searching should use many more reference structures than has been the case in much previous research. Finally, it could be of interest to apply the cross-classified multilevel modelling approach introduced here to analyse other multi-factor chemoinformatics applications, such as the clustering of chemical databases and methods for flexible ligand-protein docking.

Experimental Section

Multilevel modelling is a statistical modelling technique that allows the structure of a dataset to be specified, with observations nested in one or more higher 'levels'.[14] The effects of these higher level factors can then be tested. Cross-classified modelling allows for situations when those factors are not nested exactly within each other.[15] The approach is completely general in nature, and has thus been applied to the analysis of a wide range of types of data, though principally thus far in the social and medical sciences. Examples include studies of the parental choices of secondary schools for their children,[16] of the relative importance of schools and neighbourhood effects on student attainment,[17] of the outcomes of criminal trials of indicted terrorists,[18] of women's reproductive behaviour,[19] and of bibliometric indicators describing the impact of academic research[20] *inter alia*. However the approach has not, to our knowledge, been applied to the analysis of virtual screening as considered here or, indeed, to problems in chemoinformatics more generally. The starting point for our work was a study by Bell *et al*. that analysed Formula One data to determine the relative importance of motor racing teams and of drivers in determining race success.[21] These authors were able to demonstrate that the individual racing team was generally more important than the individual driver in winning Formula One races (although the difference appeared to be reduced in wet weather and on street tracks), and it was this finding that spurred the study reported here where we have sought to determine the relative importance of factors affecting the success of similarity-based virtual screening.

Multilevel analyses are appropriate where one has a measured response variable that is the result of a set of influence variables, and where the

data is in some way structured. These structures can be a strict hierarchy, where the levels of the structure are nested within each other, or can be cross-classified, whereby observations are nested within two or more higher levels, but those levels are not nested within each other. The structure used in this paper is a combination of these two approaches, with the response variable being the enrichment factor as a measure of screening effectiveness. The resulting approach can perhaps be regarded as being analogous to multiple regression using nominal variables (e.g., fingerprint or coefficient as discussed below), but with the difference that the variance components in a model test the importance of the overall level (e.g., fingerprints as a whole), rather than/as well as the importance of individual identifiers within that level (e.g., individual types of fingerprint).

We are primarily interested in the relative importance of fingerprints and similarity coefficients as influence variables, but it is important to additionally account for the type of activity for which screening is being conducted. Much of the data in the ChEMBL dataset used here has come from drug-discovery programmes that have involved the synthesis and testing of close analogues, with the result that many of the active molecules have a relatively high-degree of similarity to each other; conversely, there are other types of bioactivity where the known actives are structurally heterogeneous, a factor that tends to make similarity searching less powerful than in the case of more homogeneous sets of active molecules. Given these three factors the basic model studied here is of the form shown below (though, as will be seen in Results and Discussion, further models were developed as the study progressed):

$$ef_i = \beta_0 + u_{class(j)} + u_{fp(k)} + u_{coef(l)} + e_i \quad (1)$$

Here, ef_i is the observed value of the enrichment factor for the top 1% of the ranked database for a given similarity search i ($1 \leq i \leq 46{,}500$), β_0 is the mean enrichment factor across all activity classes, fingerprints and similarity coefficients, $u_{class(j)}$ ($1 \leq j \leq 15$) represents the effect of similarity search i's activity class, $u_{fp(k)}$ ($1 \leq k \leq 10$) represents the effect of similarity search i's fingerprint, $u_{coef(l)}$ ($1 \leq l \leq 31$) represents the effect of similarity search i's similarity coefficient, and e_i is a term describing the residual error and incorporating the random variation from one search to another that can affect the enrichment value. The activity class, fingerprint, similarity coefficient and residual error are assumed to be statistically independent and to be normally distributed with zero means and constant variances that are estimated by the model. It is the variances that are of interest in the present context, since a small variance means that changes in an influence variable, e.g., the similarity coefficient, are unlikely to result in substantial changes in the effectiveness of screening (with the converse applying with a large variance).

The cross-classified models were run using MLwiN version 2.36.[22] This is a freely available software package that allows a user to create, fit and manipulate multilevel models, estimating the parameter variances using a Bayesian Markov chain Monte Carlo method.[23] In addition to its widespread availability, the Monte Carlo method used in MLwiN has the advantages over alternative maximum-likelihood-type techniques that it is relatively quick to run and that it does not suffer from the biases associated with analysing small numbers of units at each level (e.g., we consider here just ten different types of fingerprint).[24] MLwiN commences by making initial estimates of the various parameters and then iterates until a threshold number of iterations have taken place (for which the default value is 500 iterations); after this point, estimates are generated for a further number of iterations (500,000 in our experiments) and the summary statistics for this chain of estimates provide the mean and standard deviations for the model parameters. For further details of this estimation method, see Browne.[23]

Our dataset is based on the well-known, open-access ChEMBL dataset available from the European Bioinformatics Institute at https://www.ebi.ac.uk/chembl/. This contains a large number of drug-like bioactive compounds compiled from the published literature on a regular basis. The version used here was derived from those molecules in ChEMBL 18 that satisfied the following criteria: *homo sapiens* as the target organism; a pIC50 of at least 5.0; and a confidence score of 9. In view of the computational costs associated with the large number of similarity searches that were carried out, a systematic 1-in-10 sample of the database was used, this yielding a dataset containing a total of 134,362 molecules. A set of 15 activity classes were chosen from amongst those studied by Heikamp and Bajorath[25] so as to include examples of both structurally heterogeneous and structurally homogeneous sets of actives: the classes and the numbers of active molecules in each case are listed in the left-hand column of Table 3.

The molecules in the ChEMBL dataset were characterised using ten different 1024-bit 2D fingerprints that were generated using the RDKit software.[26] These fingerprints were AtomPair, Avalon, FeatMorganR1, FeatMorganR2, Layered, MorganR1, MorganR2, Pattern, RDKit and Torsion, as detailed by Landrum.[26] Similarities were computed using 31 different binary similarity coefficients, chosen from those studied by Todeschini *et al.*[11,12] after the removal of one coefficient from any pair of coefficients that were found to be fully monotonic with each other, and with the pairs of remaining coefficients showing Spearman rank-correlation values ranging from 0.99 down to -0.18. Taken together, the sets of fingerprints and coefficients yielded a total of 310 different similarity measures. Similarity searches based on each of these 310 measures were carried out using ten different, randomly selected reference structures for each of the 15 different activity classes, giving a total of 46,500 (i.e., 10×31×10×15) distinct searches. Whilst the reference structure is not particularly of interest in itself, it is important to include it because it is a key part of the structure of the data and Schmidt-Catran and Fairbrother have suggested that failing to include all relevant levels can lead to erroneous results.[27]

The effectiveness of each similarity search was evaluated using the enrichment factor for the top-1% of the ranked list of molecules, i.e., the ratio of the number of actives retrieved in the top 1% to the number of actives that would have been retrieved if molecules were picked from the database at random. Other evaluation criteria have been suggested in the literature but these tend to be closely correlated with each other. For example, Riniker and Landrum discuss the very close relationship between the enrichment factor and the BEDROC criterion, and note the greater comprehendibility of the former, as used here.[28]

Acknowledgements

LM thanks Majlis Amanah Rakyat (MARA) and Universiti Teknologi MARA (UiTM) for funding.

Keywords: Chemical fingerprint • Cross-classified multilevel modelling • Similarity coefficient • Similarity searching • Virtual screening

References:

[1] R. P. Sheridan, S. K. Kearsley, *Drug Discov. Today* **2002**, *7*, 903-911.

[2] H. Eckert, J. Bajorath, *Drug Discov. Today* **2007**, *12*, 225-233.

[3] P. Willett, *Ann. Rev. Inf. Sci. Technol.* **2009**, *43*, 3-71.

[4] G. Maggiora, V. Shanmugasundaram, *Methods Mol. Biol.* **2011**, *672*, 39-100.

[5] G. Maggiora, M. Vogt, D. Stumpfe, J. Bajorath, *J. Med. Chem.* **2014**, *57*, 3186-3204.

[6] P. Willett, *Mol. Informatics* **2014**, *33*, 403-413.

[7] R. P. Sheridan, *Expert Opin. Drug Discov.* **2007**, *2*, 423-430.

[8] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, J. W. Davies, *J. Chem. Inf. Model.* **2009**, *49*, 108-119.

[9] J. Duan, S. L. Dixon, J. F. Lowrie, W. Sherman, *J. Mol. Graph. Model.* **2010**, *29*, 157-170.

[10] M. Sastry, J. F. Lowrie, S. L. Dixon, W. Sherman, *J. Chem. Inf. Model.* **2010**, *50*, 771-748.

[11] R. Todeschini, V. Consonni, H. Xiang, J. D. Holliday, M. Buscema, P. Willett, *J. Chem. Inf. Model.* **2012**, *52*, 2884-2901.

[12] R. Todeschini, D. Ballabio, V. Consonni, *Distances and other dissimilarity measures in chemometrics*, at http://onlinelibrary.wiley.com/doi/10.1002/9780470027318.a9438/pdf.

[13] S. M. Arif, J. D. Holliday, P. Willett, *J. Inf. Sci.* **2013**, *39*, 7-14.

[14] H. Goldstein, in *Encyclopedia of Biostatistics Vol. 4* (Eds.: P. Armitage, T. Colton), Wiley, Chichester, **1998**, pp. 2725-2731.

[15] A. Fielding, H. Goldstein, *Cross-classified and Multiple Membership Structures in Multilevel Models: An Introduction and Review*, Department for Education and Skills, Birmingam, **2006**.

[16] H. Goldstein, *Sociolog. Methods Research* **1994**, *22*, 364-375.

[17] G. Leckie, *J. Royal Stat. Soc.: Series A* **2009**, *172*, 537-554.

[18] B. D. Johnson, *J. Quant. Criminol.* **2012**, *28*, 163-189.

[19] S. Zaccarin, G. Rivellini, *Stat. Methods Applicat.* **2002**, *11*, 95-108.

[20] L. Bornmann, R. Mutz, S. E. Hug, H.-D. Daniel, *J. Informetrics* **2011**, *5*, 346-359.

[21] A. Bell, J. Smith, C. E. Sabel, K. Jones, *J. Quant. Anal. Sports* **2016**, *12*, 99-112.

[22] J. Rasbash, F. Steele, W. J. Browne, H. Goldstein, *A User's Guide to MLwiN, v2.26*, Centre for Multilevel Modelling, University of Bristol, Bristol, **2012**.

[23] W. J. Browne, *MCMC Estimation in MLwiN, v2.32*, at http://www.bris.ac.uk/cmm/media/software/mlwin/downloads/manuals/2-32/mcmc-web.pdf.

[24] D. Stegmueller,. *Am. J. Pol. Sci.* **2013**, *57*, 748–761.

[25] K. Heikamp, J. Bajorath, *J. Chem. Inf. Model.* **2011**, *51*, 1831-1839.

[26] G. Landrum, *RDKit Documentation. Release 2016.03.1*, at http://www.rdkit.org/RDKit_Docs.current.pdf.

[27] A. W. Schmidt-Catran, M. Fairbrother, *Eur. Sociol. Rev.* **2015**, *32*, 23–38.

[28] S. Riniker, G. A. Landrum, *J. Cheminf.* **2013**, *5*, 26.

Table 1. Basic model of similarity searching

Mean enrichment factor		Activity class		Fingerprint		Similarity coefficient		Residual error	
12.725	1.857	54.170	23.635	4.689	3.042	4.222	1.772	92.473	0.607

Table 2. Model of similarity searching including consideration of the individual reference structures

Mean enrichment factor		Activity class		Fingerprint		Similarity coefficient		Reference structure		Residual error	
12.973	2.368	49.362	25.539	5.126	6.646	4.272	1.765	60.130	7.465	39.160	0.257

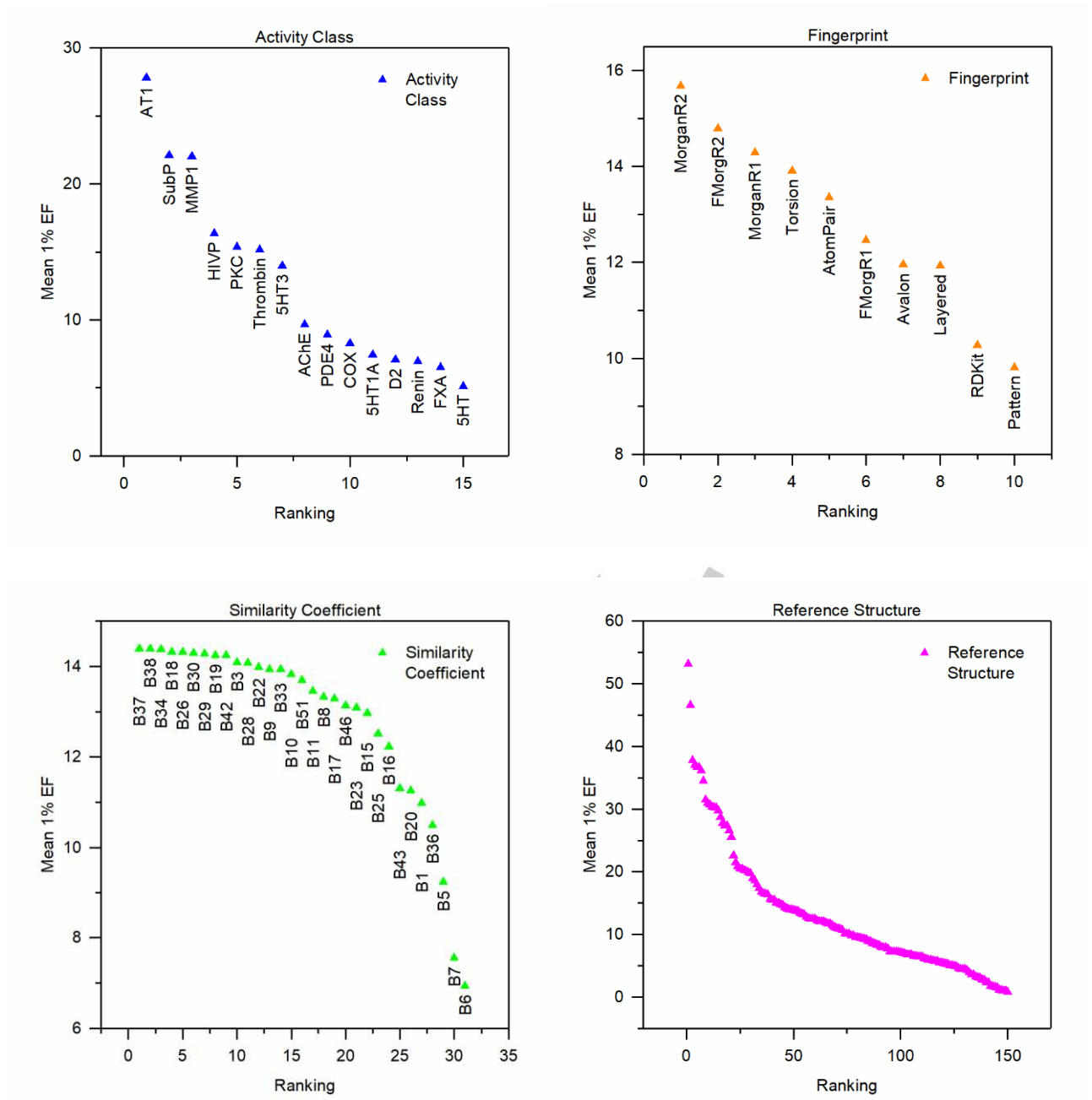
Table 3. Models of similarity searching for each of the 15 different activity classes

Activity class	Mean enrichment factor		Fingerprint		Similarity coefficient		Reference structure		Residual error	
Serotonin transporter (2447)	5.099	0.989	7.676	4.878	1.191	0.807	1.252	0.361	4.848	0.124
Serotonin 1a (5-HT1a) receptor (1483)	7.379	1.717	19.385	12.222	8.989	5.720	2.167	0.641	14.382	0.368
Serotonin 3a (5-HT3a) receptor (213)	13.801	4.368	170.275	106.560	13.740	8.754	3.355	1.008	27.571	0.706
Acetylcholinesterase (739)	9.577	2.125	37.322	23.471	5.136	3.526	1.612	0.477	10.468	0.268
Type-1 Angiotensin II receptor (106)	27.688	5.084	200.747	127.242	35.233	22.664	39.846	11.373	116.050	2.972
Cyclooxygenase-1 (139)	8.213	1.362	15.128	9.665	1.636	1.155	2.401	0.702	12.538	0.321
Dopamine D2 receptor (1858)	7.012	1.836	23.980	15.065	9.105	5.772	1.142	0.345	10.069	0.258
Coagulation factor X (1502)	6.439	1.592	22.049	13.900	1.172	0.859	2.051	0.588	6.899	0.177
HIV Type 1 protease (2157)	16.312	2.053	28.404	17.947	10.098	6.692	5.175	1.487	18.313	0.469
Matrix metalloproteinase-1 (395)	21.928	3.617	79.820	50.140	52.736	32.671	10.049	2.880	33.011	0.845
Phosphodiesterase 4a (254)	8.848	1.994	17.182	10.815	24.955	15.418	1.391	0.420	12.216	0.313
Protein kinase C Alpha (211)	15.058	4.964	241.093	149.078	5.106	3.290	1.391	0.438	18.675	0.478
Renin (982)	6.903	1.797	26.481	16.701	3.588	2.339	2.453	0.708	9.936	0.254
Neurokinin 1 receptor (847)	21.977	5.460	218.451	137.615	66.817	42.443	25.260	7.314	109.063	2.793
Thrombin (838)	15.093	2.334	39.749	25.084	9.531	6.292	8.422	2.384	17.613	0.451

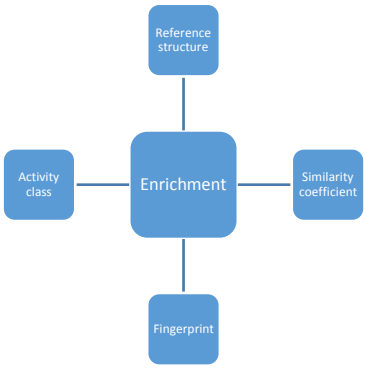
Table 4. Models of similarity searching for the ten reference structures in the serotonin transporter (5HT) activity class

Reference structure	Mean enrichment factor		Fingerprint		Similarity coefficient		Residual error	
1	4.437	1.155	13.365	8.322	0.668	0.213	0.969	0.084
2	5.465	0.503	1.729	1.123	2.484	0.727	1.334	0.116
3	1.594	0.558	3.117	1.960	0.139	0.059	0.698	0.061
4	4.243	0.600	3.286	2.094	1.029	0.327	1.463	0.127
5	3.105	0.500	2.390	1.518	0.363	0.128	0.946	0.082
6	7.702	0.907	7.574	4.790	2.303	0.707	2.437	0.212
7	7.007	0.782	5.408	3.425	2.389	0.716	1.887	0.164
8	9.180	1.026	9.165	5.795	4.587	1.359	3.092	0.269
9	6.155	0.639	3.582	2.383	1.130	0.442	4.438	0.387
10	2.324	0.481	2.299	1.446	0.136	0.054	0.571	0.050

Figure 1. Effect of the various components of a similarity search on the resulting enrichment factor.



Entry for the Table of Contents



This paper describes the use of cross-classified multilevel modelling to analyse the results of similarity-based virtual screening searches using 2D fingerprints. It is shown that the choice of fingerprint is more important than the choice of similarity coefficient, and that multiple reference structures need to be employed in benchmark studies such as this.