Journal of
CHEMICAL INFORMATION
AND MODELING

Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS)

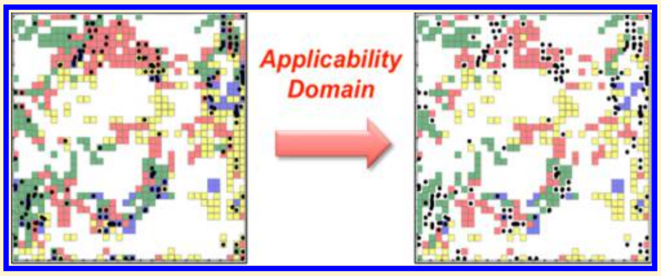
Héléna A. Gaspar,[†] Gilles Marcou,[†] Dragos Horvath,[†] Alban Arault,[‡] Sylvain Lozano,[‡] Philippe Vayer,[‡] and Alexandre Varnek*,[†]

[†]Faculté de Chimie, Université de Strasbourg, UMR 7140−Laboratoire de Chémoinformatique, 1 rue Blaise Pascal, 67000 Strasbourg, France

[‡]Technologie Servier, 25-27 rue Eugène Vignat, 45000 Orléans, France

S *Supporting Information*

ABSTRACT: Earlier (Kireeva et al. *Mol. Inf.* **2012**, *31*, 301−312), we demonstrated that generative topographic mapping (GTM) can be efficiently used both for data visualization and building of classification models in the initial *D*-dimensional space of molecular descriptors. Here, we describe the modeling in two-dimensional latent space for the four classes of the BioPharmaceutics Drug Disposition Classification System (BDDCS) involving VolSurf descriptors. Three new definitions of the applicability domain (AD) of models have been suggested: one class-independent AD which considers the GTM likelihood and two class-dependent ADs considering respectively, either the predominant class in a given node of the map or informational entropy. The class entropy AD was found to be the most efficient for the BDDCS modeling. The predominant class AD can be directly visualized on GTM maps, which helps the interpretation of the model.



1. INTRODUCTION

The BioPharmaceutics Drug Disposition Classification System (BDDCS, see Figure 1) based on the compound solubility and degree of metabolism, was introduced in 2005[1] as a complement to the BioPharmaceutics Classification System (BCS) based on the compound solubility and intestinal absorption. FDA (Food and Drug Administration) and EMA (European Medicine Agency) use them as decision tools for granting biowaivers, i.e., waivers of clinical bioequivalence studies. Since biowaivers allow to use in vitro drug dissolution data instead of human bioequivalence data to authorize a drug as orally available, BDDCS has attracted a lot of interest in the drug discovery area.[2,3] Thus, this classification was used to anticipate drug disposition,[4,5] drug−drug interactions,[2] transporter−enzyme interplay,[6,7] and relevance of genetic variants of transporters/enzymes.[8] It was also recently observed that the BDDCS class of a compound was a relevant parameter in predicting its blood−brain barrier permeability.[9] Furthermore, a correlation between the BDDCS class of a compound and Torsade de Pointes risks (polymorphic ventricular tachycardia identified on electrocardiograms) was observed.[10] The usefulness of considering the BDDCS class of a compound for microdosing results has been noted.[11]

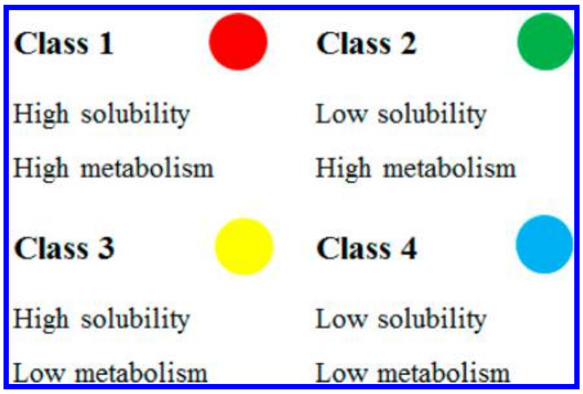


Figure 1. BDDCS classification system.

Several structure−property studies of BDDCS have been reported in the literature. Thus, using a data set of 927 drugs, Benet et al.[12] analyzed the capacity of different parameters to distinguish BDDCS classes. It has been shown that aqueous solubility and human intestinal permeation predicted with QSAR models are not good enough for this purpose. Khandelwal et al.[13] used different machine-learning methods (recursive partitioning (RP), random forest (RF), and support vector machine (SVM)) and descriptors (VolSurf and MolConnZ) to build classification models on a small set of 165 compounds. Broccatelli et al.[2] reported a SAR model to predict BDDCS classes. The modeling involved predicted FDA

Received: July 19, 2013
Published: December 2, 2013

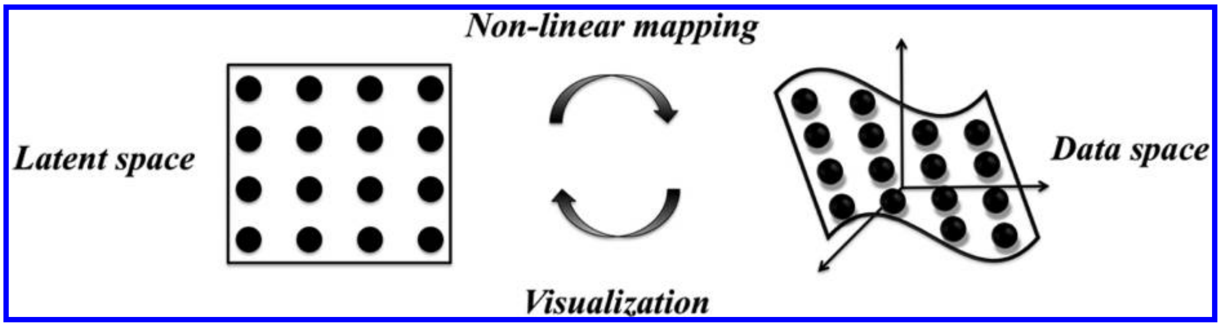


Figure 2. Schematic representation of the GTM algorithm: nonlinear mapping of grid nodes in 2D latent space onto a manifold in *D*-dimensional data space.

Solubility and Extent of Metabolism and VolSurf descriptors used in combination with different data mining algorithms (LDA, SVM, RF, naïve Bayes, *k*NN, RP). The models were trained on about 300 drugs and validated on an external test set of about 369 drugs from the data set of Benet et al.[12] Weak predictive performance of these models could be explained both by uncertainties in experimental data and also by the fact that the applicability domain of models was not taken into account.

In this paper, we apply generative topographic mapping (GTM) in order to rationalize the chemical space spanned by the compounds from the data set of Benet et al.[12] and to build classification models. Three new definitions for the applicability domain (AD) of the GTM-based models were suggested. Below, we demonstrate how the AD studies help to interpret the GTM models.

2. GENERATIVE TOPOGRAPHIC MAPPING

2.1. GTM Algorithm. In GTM,[14−16] each point in the low-dimensional (usually 2D) latent space (LS) is mapped onto the manifold embedded in a high-dimensional data space (input space, IS), as shown in Figure 2.

The manifold is defined by a mapping function $\mathbf{y}(\mathbf{x}; \mathbf{W})$ assessed with the help of *m* radial basis functions (RBFs) of width *w* regularly distributed in LS. The latent space is covered by a mesh containing *k* nodes each of which corresponding to a normal probability distribution (NPD) centered on the manifold in IS. Ensemble of NPD is used to compute a posterior probability for a data point t_n in *D*-dimensional IS to be projected onto a node x_k:

$$p(x_k|t_n, \mathbf{W}, \sigma) = \frac{p(t_n|x_k, \mathbf{W}, \sigma)}{\sum_k p(t_n|x_k, \mathbf{W}, \sigma)} \tag{1}$$

where **W** is a parameter matrix and σ the variance of the distribution of ***t***:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \sigma) = \left(\frac{1}{2\pi\sigma}\right)^{D/2} \exp\left\{-\frac{1}{2\sigma}\|\mathbf{y}(\mathbf{x}; \mathbf{W}) - \mathbf{t}\|^2\right\} \tag{2}$$

The log likelihood of the whole data set is calculated according to eq 3:

$$\mathcal{L}(\mathbf{W}, \sigma) = \sum_n \ln\left\{\frac{1}{K}\sum_k p(t_n|x_k, \mathbf{W}, \sigma)\right\} \tag{3}$$

The GTM is optimized with an expectation-maximization (EM) algorithm using data likelihood ($\mathcal{L}$) as the objective function (the best GTM map corresponds to the highest $\mathcal{L}$). The mapping depends on four parameters: the number *m* of RBF, the grid resolution *k*, the RBF width *w*, and the weight regularization coefficient *l*. The latter is used for re-estimating the **W** parameter matrix and influences the flexibility of the manifold. Notice that a too flexible manifold, although nicely approximating the training data, may lead to overfitting when, i.e., the manifold may wrongly describe new data points projected onto the map. Thus, weight regularization coefficient *l* is a trade-off between training data approximation and predictive performance of the map.

2.2. GTM-Based Classification Models. An interesting feature of the GTM is that probability density functions (PDF) calculated for a given data set both in the input and the latent spaces can feed a Bayesian classification model. Earlier, we reported classification models built with PDF in the input space.[17] Here, only PDF in 2D latent space are considered, and therefore, the obtained classification models can be linked to the 2D visualization. The posterior probability $p(x_k|\mathbf{t}, \mathbf{W}, \sigma)$ that the given point *t* in the data space is generated from the *k*th node (so-called *responsibility* of the *k*th node for data point *t*) is computed using Bayes' theorem:

$$\begin{aligned} R_{tk} &= p(x_k|\mathbf{t}, \mathbf{W}, \sigma) \\ &= \frac{p(\mathbf{t}|x_k, \mathbf{W}, \sigma)p(x_k)}{\sum_{k'=1}^{k} p(\mathbf{t}|x_{k'}, \mathbf{W}, \sigma)p(\mathbf{x}_{k'})} \\ &= \frac{\exp\left\{-\frac{1}{2\sigma}\|\mathbf{y}(x_k; \mathbf{W}) - t\|^2\right\}}{\sum_{k'=1}^{k} \exp\left\{-\frac{1}{2\sigma}\|\mathbf{y}(x_k; \mathbf{W}) - t\|^2\right\}} \end{aligned} \tag{4}$$

Equation 4 shows that unlike Kohonen self-organized maps (SOM),[18] in which only a single node is "responsible" for a data point, in GTM all nodes share such a "responsibility". This means that a data point *t* is mapped onto each node with a certain probability. For this reason, the GTM method is often considered as a fuzzy analog of SOM with the membership function for mapping data to different nodes defined by node responsibilities. The mean position $x^{\text{mean}}(t)$ in the latent space is calculated by averaging over all nodes taking the responsibilities as weighting factors:

$$x^{\text{mean}}(t) = \sum_{k=1}^{k} x_k R_{tk} \tag{5}$$

Averaging the responsibilities $R_{tk}(C_i)$ over N_{C_i} training set compounds belonging to the *i*th class in the latent space gives direct access to the conditional probability $P(k|C_i)$ of finding a new instance close to a node *k*:

$$P(k|C_i) = \frac{\sum_t R_{tk}(C_i)}{N_{C_i}} \tag{6}$$

Using Bayes' theorem, one can find $P(C_i|k)$—the conditional probability of class C_i given node k:

$$P(C_i|k) = \frac{P(k|C_i) \times P(C_i)}{\sum_{C_j} P(k|C_j) \times P(C_j)} \tag{7}$$

where $p(C_i) = N_{C_i}/N_{\text{tot}}$ with N_{tot} being the total number of compounds in the training set. These conditional probabilities are then used for the qth test set compound to estimate the ith class probability $P(C_i|q)$:

GTM−Bayesian

$$P(C_i|q) = \sum_k P(C_i|k) \times R_{qk} \tag{8}$$

Finally, the class with the largest probability is assigned to the given compound. Instead of using eq 8, one can simply assign to the qth molecule the predominant class attributed to its nearest node k on the 2D map according to the conditional probability $P(k|C_i)$:

GTM−kNNd

$$P(C_i|q) = P(k|C_i) \tag{9}$$

Below, we refer to this method as kNNd or the k-nearest node approach.

2.3. Applicability Domain of Models. Below, we consider three definitions of the models applicability domain (AD) based on the GTM approach. One of them—the likelihood-based AD—is class-independent, whereas the two others, the predominant class AD and class entropy AD, are class-dependent.

(i) Class-Independent AD. This AD definition is based on consideration of the likelihood parameter. From eq 3, it follows that the log likelihood of the nth data point L_n depends on the distance between this point and the manifold. One could reasonably suggest that the data points in the input space situated too far from the manifold are poorly described by GTM. Therefore, rejecting the molecules for which the likelihood is larger than a certain threshold (L_{LF}):

$$L_{\text{LF}} < L_n \tag{10}$$

may improve the overall performance of the model. Notice, that this AD definition is "unsupervised" because it does not depend on activity classes. In our calculations, L_{LF} (likelihood factor) is the LFth percentile of the training set likelihood, so that L_n is larger than LF% of the molecules of the training set. LF varied from 5 (almost all compounds accepted) to 100 (all compounds rejected) in increments of 5 in order to observe the variation of the performance of the model as a function of LF.

(ii) Predominant Class AD. The accuracy of the classification model depends on the difference between probabilities of different classes ($P(k|C_i)$ and $P(k|C_j)$) in the grid nodes (see eqs 6 and 7). A too small difference could lead to uncertainty of predictions. Thus, the idea is to consider only the nodes for which the probability of the predominant class is definitely larger than that of any other class. This could be realized by introducing a class prevalence factor:

$$\text{CPF} < \frac{\max_c P(k|C)}{P(k|C_i)}, \qquad \forall\ C_i \neq C \tag{11}$$

Here, $\max_C\ P(k|C)$ and $P(k|C_i)$ are, respectively, the probabilities of predominant and any other class. Thus, the classes are assigned only to the nodes that follow eq 11. We suggest that only these nodes delineate the AD, i.e. a molecule situated near an unlabeled node would be considered to be outside the AD. In 2D latent space, the nearest node k is determined by Euclidean distances between the mean position of the molecule and the node's coordinates. Increase of CPF should improve the predictive performance of the model because "noisy" nodes are discarded. In this work, the CPF was systematically varied from 1 (all compounds accepted) to 10^6 (almost all compounds rejected). The probability distribution of classes could be visualized by attributing to each node the class having the largest probability (this is a typical way of data visualization in grid-based methods, e.g., Kohonen maps). In this way, for a given CPF value, one may visualize the latent space regions corresponding to the applicability domain of the model.

(iii) Class Entropy AD. Another class-based AD definition can be related to the class entropy of the qth molecule calculated as

$$S_q = -\sum_i P(C_i|q) \log(P(C_i|q)) \tag{12}$$

This value varies between 0 (the probability of one of the classes is 1, i.e., case of nonambiguous class attribution) and its maximal value $S_{\max} = \log(N_C)$ where N_C is the number of classes (all class probabilities are equal to $1/N_C$). Thus, the performance of the model may be improved if molecules with a high entropy value are discarded. In this case, the molecule is within the AD if

$$\text{CLF} > \frac{S_q}{S_{\max}} \tag{13}$$

where CLF (class-likelihood factor) is a user-defined parameter. In this work, CLF was varied from 0 (all compounds out) to 1 (all compounds in) in increments of 0.05.

3. COMPUTATIONAL PROCEDURE

3.1. Data Preparation. The data set was extracted from the collection of 927-marketed drugs of Benet et al.[12] (Table 1).

Table 1. BDDCS Data Set Used in the Modeling

	initial data set from Benet et al.[12]	curated data set
class 1	351	341
class 2	265	264
class 3	247	238
class 4	53	50
total	**927**	**893**

Several compounds/substances were discarded for the following reasons:

- It was impossible to assign unambiguously a BDDCS class to a compound from a substance (as for instance for potassium chloride).
- The active substance was a peptide (and therefore out of the scope of this study).

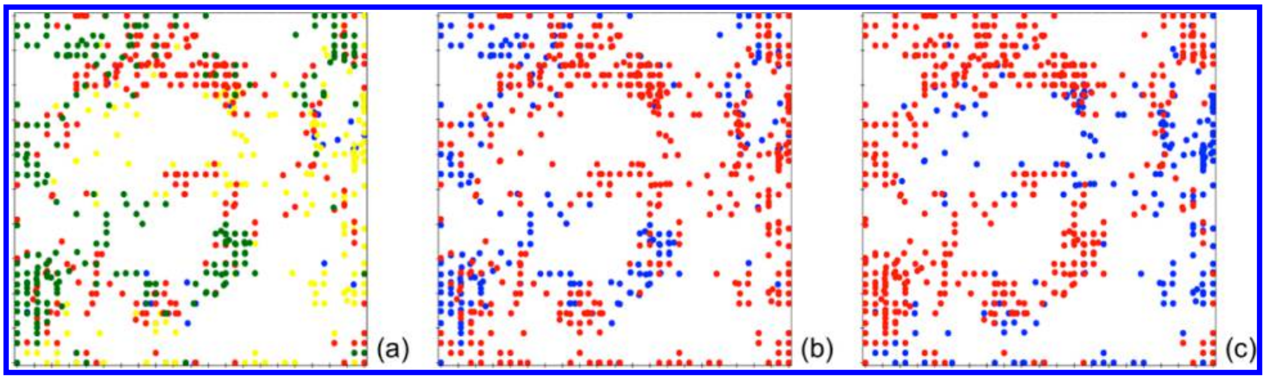


Figure 3. GTM map of the entire set of 893 molecules in the space of VolSurf descriptors ($m = 36$, $k = 1296$, $w = 1.5$, $l = 1.0$). Each point refers to a compound; the color stands for (a) the BDDCS class it belongs to (see Figure 1), (b) solubility "high/low" (red/blue) classes, (c) metabolism high/low (red/blue) classes. The maps have been built with the ISIDA/GTM program; see section 3.4.

- A substance appeared several times as different stereoisomer or as a mixture of stereoisomers (e.g., norgestrel and levonorgestrel).
- A substance appeared several times differing by the formulation (e.g., erythromycin).
- A substance appeared as an alcohol or as a ketone (e.g., dolasetron and hydrodolasetron).

For a mixture, the BDDCS class was determined for the largest compound. However, this hypothesis is not always correct because the attribution of the compound to the BDDCS class may depend on formulation. For instance, erythromycin belongs to class 4 if it is delivered with a fatty acid (stearate) and to class 3 otherwise. Such ambiguous compounds were, therefore, excluded. Finally, the remaining data set containing 893 compounds (Table 1) was used for the modeling. The list of excluded compounds is given in the Supporting Information.

3.2. Descriptors. The following VolSurf descriptors[19] were used in the modeling: AllVS=V, S, R, G, W1−W8, D1−D8, WO1−WO6, WN1−WN6, IW1−IW4, CW1−CW8, ID1−ID4, CD1−CD8, HL1, HL2, A, CP, POL, MW, FLEX, FLEX RB, NCC, DIFF, LOGP n Oct, LOGP c Hex, PSA, HSA, PSAR, PHSAR, LgD5−LgD10, AUS74, FU4−FU10, triangular pharmacophore, SOLY, LgS3−LgS11, PB, VD, CACO2, SKIN, LgBB, MetStab, HTSflag, L0LgS−L4LgS, DD1−DD8. Two subsets of AllVS were studied, those containing ADME-like descriptors (admeVS=SOLY, LgS3−LgS11, PB, VD, CACO2, SKIN, LgBB, MetStab, HTSflag, L0LgS−L4LgS) and the others (NOTadmeVS = AllVS − admeVS). All descriptors were normalized.

3.3. Calculation and Validation of Models. The GTM algorithm is based on likelihood (L) maximization. In this respect, the parameters of the method—the number of RBF (m), the map resolution (k), the RBF width (w), and the weight regularization parameter (l)—providing the highest L value should be selected. The situation becomes more complicated if one applies GTM in structure−activity modeling. Indeed, the selected parameters should allow one to achieve maximal L for both the training set (data description and visualization performance) and the test set (model prediction performance). In this study, we systematically varied m from 16 to 64, k = 64, 100, 144, 196, 256, 625, and 1296; w = 0.5, 1, 1.5, and 2; l from 10^{-2} to 10^2. For each set of parameters, GTM models were 2-fold cross-validated. In these calculations, the likelihood of the training set (L(train)) varied from 105 to 166, whereas a relative training/test set's likelihood $\delta L = (L(\text{train}) - L(\text{test}))/L(\text{train})$ varied from 3% to 33%. A reasonable compromise between L(train) and δL was found for $m = 36$, $k = 1296$, $w = 1.5$, and $l = 1.0$, with $\delta L = 17\%$ and L(train) = 150.

For comparison purposes, random forest (RF)[20] and naïve Bayes[21] models have been obtained using the same descriptors as for GTM models. RF implementation in Weka version 3.7.1[22] has been used in this work. RF is a bagging algorithm using an ensemble of random decision trees. The number of trees in the forest was not optimized and taken equal to 50. Each random tree was developed using a random subset of descriptors whose size was set to the logarithm of the number of descriptors plus one (default setting of the method in Weka). The naïve Bayes classification is a probabilistic approach based on Bayes' theorem, which estimates conditional probabilities and assumes that descriptors are independent given one class (conditional independence assumption).

Predictive performance of GTM, NB, and RF models were assessed in 10-fold cross-validation. Confidence intervals for each statistical parameter (balanced accuracy, precision, recall, etc.) were estimated using t-statistics with a confidence of 95%. In order to tackle a chance correlation issue, the modeling was performed for 20 y-scrambled data sets using 10-fold cross-validation. The upper bound performances of the scrambled models (lower bound for FP rate) were estimated using t-statistics with a confidence of 95%. Two statistical parameters were used to evaluate the performance of the model: balanced accuracy[23] (BA) for the overall success of the model (also called macroaverage arithmetic[24]) and *F-measure* for the success of predictions for a given class:

$$\text{BA} = \frac{1}{N}\sum_i \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{14}$$

$$\text{F-measure} = 2\frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{15}$$

Here, N is the number of classes, TP is the number of true positives, FN is the number of false negatives, recall is a ratio of true positives to all actual positives, and precision is a ratio of true positives to all instances predicted as positive.

3.4. Software. GTM maps were constructed using our in-house ISIDA/GTM program (version 2012), implemented in Free Pascal. It includes a command-line tool and a graphical user interface (GUI) with a visualizer that links the chemical structure to its position on the 2D map. This program processes descriptor files in the *svm* format, and operates in

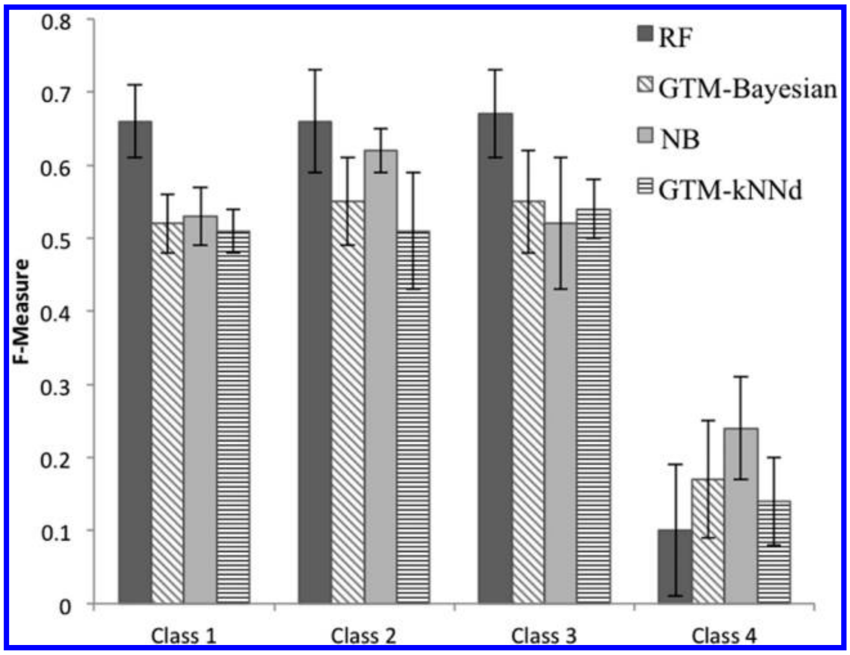


Figure 4. Performances of random forest (RF), naïve Bayes (NB), GTM−Bayesian, and GTM−*k*NNd classification models assessed in 10-fold cross-validation.

normal, projection, and cross-validation modes. In the normal mode, a GTM map is computed using the four user-defined parameters *m*, *k*, *w*, and *l* (cf. section 2.1). Already developed models can be saved and further used in projection mode to map new molecules onto the 2D map. A GTM model may be cross-validated by estimating the likelihood of training and test sets. The program is also able to build, cross-validate and visualize GTM-based classification and regression models, by using classes or continuous activities from an SDF file or an external file and computes basic statistical measures such as F-measure and determination coefficient for these models.

4. RESULTS AND DISCUSSION

4.1. Data Visualization and Modeling. On the 2D GTM map of the entire BDDCS data set, one can see a number of clusters representing classes 1−3 that slightly overlap in numerous frontier areas (Figure 3a). The molecules of class 4 do not form a distinct cluster, which may explain bad performances of models for this class. For comparison purposes, we also built maps for solubility and metabolism (Figure 3b and c). One can see that the classes (high/low) form much bigger clusters with less frontier regions than on the GTM map for BDDCS.

Figure 4 shows that the RF classification model performs better than GTM-based models. For classes 1 and 3, performance of the NB (naïve Bayes) model is similar to that of GTM-based models, whereas for class 2 NB outperforms GTM (see the Supporting Information for details). The average value of balanced accuracy (BA) is 0.52 for RF and NB, 0.45 for GTM−Bayesian, and 0.43 for GTM−*k*NNd model. Notice that RF statistics are very close to those reported by Broccatelli et al.[2] Since the two GTM-based models (eqs 8 and 9) perform similarly, only the Bayesian one has been used for further computational experiments.

In order to explain the moderate performance of classification models for the four BDDCS classes, binary GTM Bayesian models for solubility (high/low) and metabolism (high/low) have been obtained and validated in 10-fold cross-validation. The latter models perform much better than those for BDDCS: the averaged BA = 0.70 for both properties. Better performance of binary over four classes BDDCS models could be explained by more distinct classes clustering on GTM maps for metabolism and solubility (Figure 3b and c). This reveals that GTM does not actually distinguish four but rather two classes.

4.2. Impact of AD on the Model Performance. In this section, we discuss the influence of applicability domain definitions according to eqs 10−13 on the performance of RF and GTM Bayesian models assessed in 10-fold cross-validation. The LF, CPF, or CLF parameters were systematically varied as it is mentioned in section 2.3. At a given value of a parameter, the molecules of the test set outside the AD were discarded and the models (built on the training set) were applied to remaining molecules. Increase of the parameters lead to the increase of the number of discarded molecules and, hence, the reduction of the test set coverage. The procedure was repeated ten times, once for each fold, each test set including one tenth of the initial data set. The results of this analysis are represented by the balanced accuracy averaged across the 10 test sets as a function of coverage (Figure 5). The faster the raise of BA with decreasing coverage, the more pertinent is the associated AD definition. Unfortunately, this is not the case for the curves presented in Figure 5: the model performance significantly increases only after discarding a large amount of molecules. In this context, CLF behaves better than LF and CPF: at 80% coverage, for the GTM−Bayesian model BA increases by 0.05 for CLF, 0.01 for LF, and practically does not change for CPF.

It should be noted that CPF, CLF, and LF do not discard molecules in the same order. The first molecules discarded with LF are the least well fitted by the GTM manifold. On the other hand, the molecules discarded in class-dependent approaches can be well fitted to the map but situated in the areas of overlapping clusters representing different BDDCS classes.

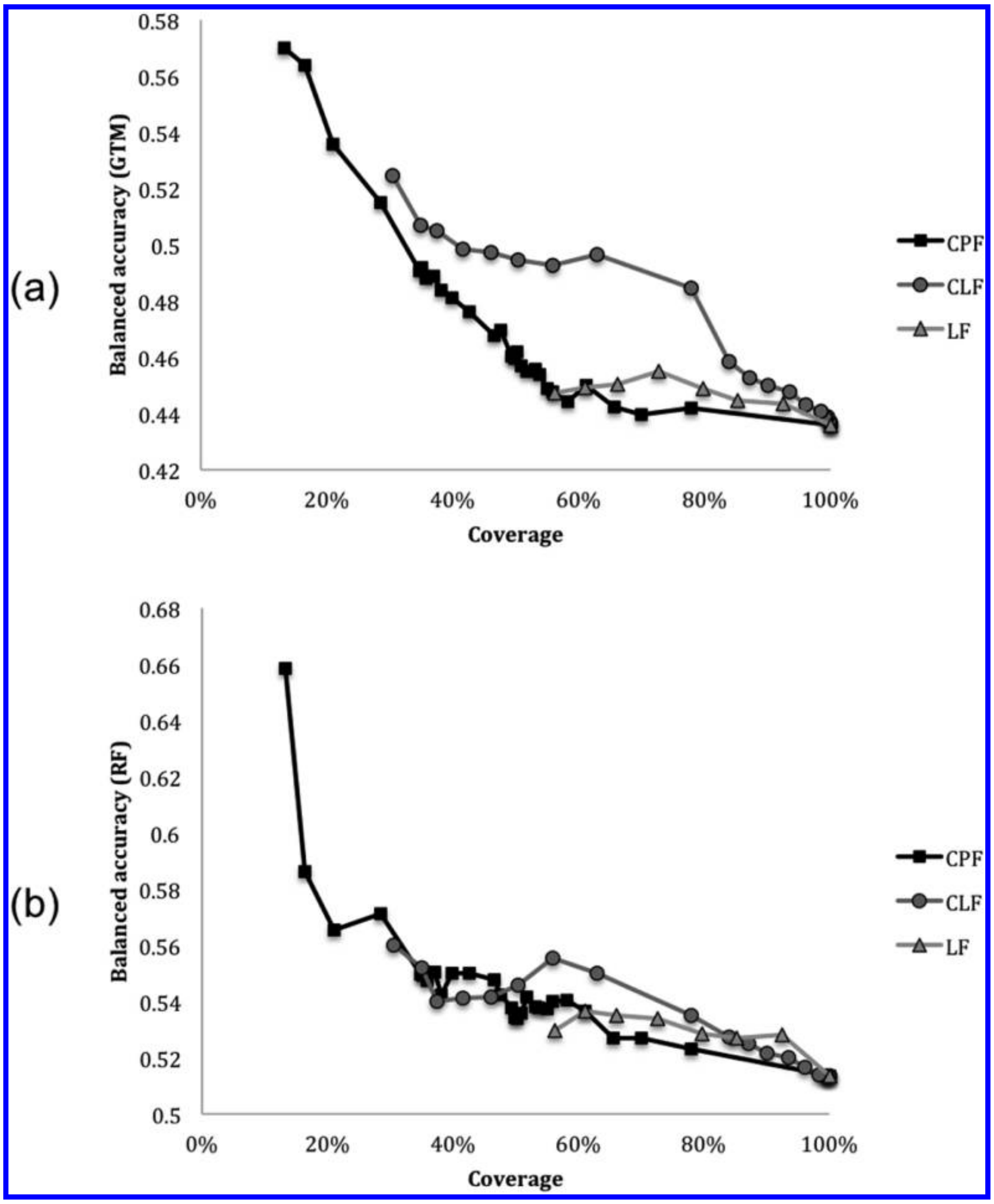


Figure 5. Performance (balanced accuracy) of (a) GTM−Bayesian and (b) RF models as a function of coverage (% of test molecules in AD). The balanced accuracy and the coverage were averaged across the 10 test sets in 10-CV. Different GTM-based AD definitions were used: class prevalence factor (CPF), class-likelihood factor (CLF), and likelihood factor (LF).

Conventional bounding box (BB) applicability domain in combination with RF model has also been tested for the purpose of comparison. With BB, a test molecule is considered to be inside the AD if each of its descriptors lies within an interval $[v_{\min}, v_{\max}]$. The $v_{\min}$ and $v_{\max}$ values are set from the Qth and 100-Qth percentile of the training set. Here, Q ranges from 1 (almost all compounds in) to 10 (almost all compounds out). Performance of the RF model as a function of coverage for both LF and BB approaches is shown in Figure 6. One can see that LF performs slightly better than BB up to 50% coverage.

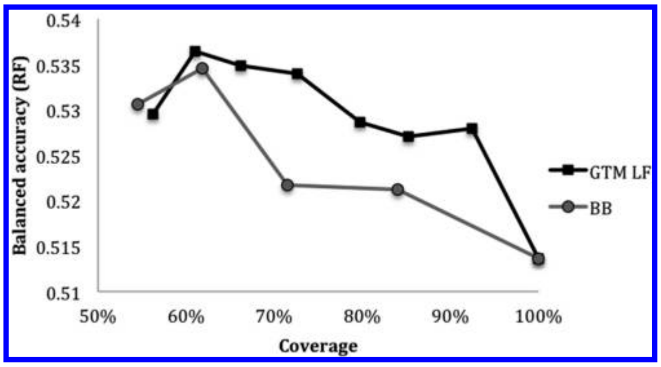


Figure 6. Balanced accuracy of RF models as a function of coverage for GTM-based likelihood factor AD (LF) and classical bounding box AD (BB).

An interesting feature of the CPF approach is the possibility to visualize the applicability domain on a GTM map. Zones colored according to the predominant class probability for CPF = 1 and 5 are visualized in Figure 7. For CPF = 1, all molecules are projected onto colored regions defining the applicability

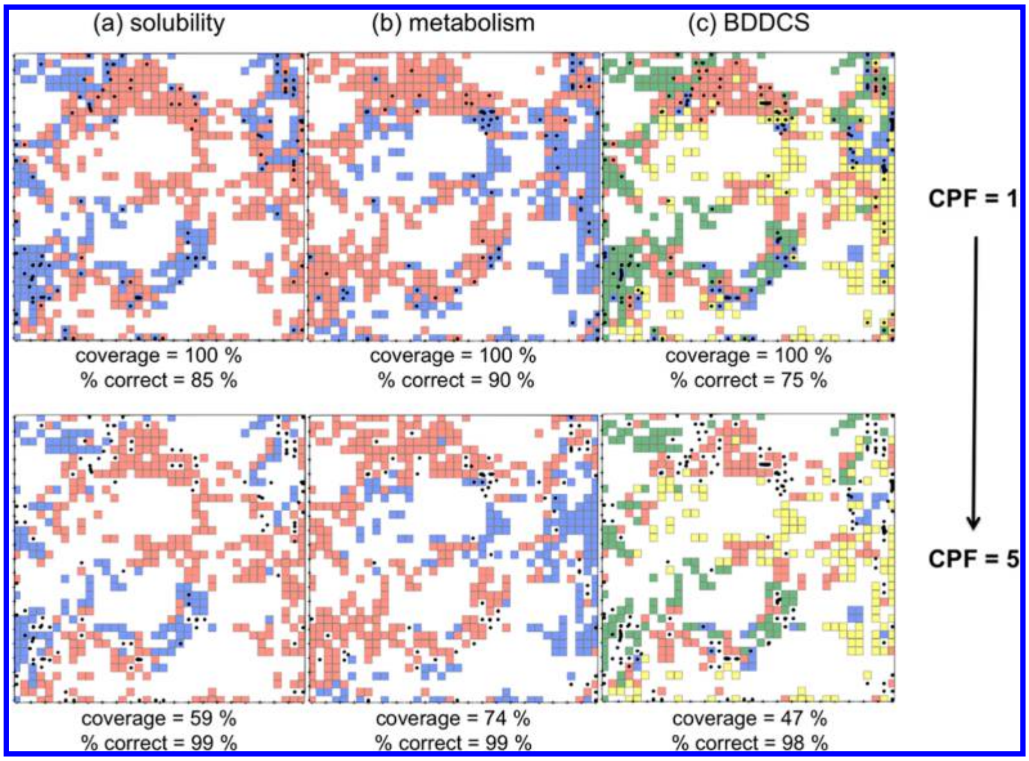


Figure 7. Graphical interpretation of the applicability domain (AD) for GTM Bayesian classification models for (a) high (red) and low (blue) solubility, (b) high (red) and low (blue) metabolism, and (c) the BDDCS classes (see Figure 1 for the color code). On the map prepared for the entire set of 893 molecules, the color of each node stands for the class having the highest probability compared to the other classes. Black points correspond to incorrectly classified molecules. Increasing the class prevalence factor (CPF) from 1 (up) to 5 (down) leads to shrinking the AD and to the increase of the proportion of correctly classified molecules inside the AD (% correct). The maps have been built with the ISIDA/GTM program.

domain of the model. Increasing the CPF leads to shrinking the colored area: at CPF = 5 only half of the data set is projected there. Another half is found in "empty" regions corresponding to uncertainty zones. For solubility and metabolism maps, most of these empty regions are situated at the frontier between the two classes. As it follows from Figure 5, discarding molecules from these zones results in better performance of models.

5. CONCLUSION

Here, we have demonstrated that data probability density distribution in GTM latent space can be used to build classification models. Although in the BDDCS case these models do not perform as well as random forest models, GTM maps give an insight into the interpretation of the model by means of the data visualization. Thus, better performance of binary models for solubility and metabolism compared to 4-classes BDDCS models is explained by the difference in classes clustering and the number of frontier zones on the maps. Three GTM-based definitions of the applicability domain (AD) of models have been suggested. They consider, respectively, GTM likelihood, probabilities of different classes in a given node of the map, and informational entropy. It has been shown that all ADs improve the performance of models by discarding the molecules outside the AD, although in some cases this effect is moderate. An important feature of probability-based AD is that it can directly be visualized on GTM maps and, in such a way, is very useful for the interpretation of the model.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

(i) Information about performance of BDDCS classification models for random forest, naïve Bayes, and GTM and (ii) the list of the compounds excluded from the training set. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: varnek@chimie.u-strasbg.fr.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Wu, C. Y.; Benet, L. Z. Predicting drug disposition via application of BCS: Transport/absorption/elimination interplay and development of a biopharmaceutics drug disposition classification system. *Pharm. Res.* **2005**, *22* (1), 11−23.

(2) Broccatelli, F.; Cruciani, G.; Benet, L. Z.; Oprea, T. I. BDDCS Class Prediction for New Molecular Entities. *Mol. Pharmaceutics* **2012**, *9* (3), 570−580.

(3) Benet, L. Z.; Amidon, G. L.; Barends, D. M.; Lennernas, H.; Polli, J. E.; Shah, V. P.; Stavchansky, S. A.; Yu, L. X. The use of BDDCS in classifying the permeability of marketed drugs. *Pharm. Res.* **2008**, *25* (3), 483−488.

(4) Custodio, J. M.; Wu, C. Y.; Benet, L. Z. Predicting drug disposition, absorption/elimination/transporter interplay and the role of food on drug absorption. *Adv. Drug Delivery Rev.* **2008**, *60* (6), 717−33.

(5) Benet, L. Z. Predicting Drug Disposition via Application of a Biopharmaceutics Drug Disposition Classification System. *Basic Clin. Pharmacol. Toxicol.* **2010**, *106* (3), 162−167.

(6) Benet, L. Z. The Drug Transporter-Metabolism Alliance: Uncovering and Defining the Interplay. *Mol. Pharmaceutics* **2009**, *6* (6), 1631−1643.

(7) Shugarts, S.; Benet, L. Z. The Role of Transporters in the Pharmacokinetics of Orally Administered Drugs. *Pharm. Res.* **2009**, *26* (9), 2039−2054.

(8) Chen, M. L.; Amidon, G. L.; Benet, L. Z.; Lennernas, H.; Yu, L. X. The BCS, BDDCS, and Regulatory Guidances. *Pharm. Res.* **2011**, *28* (7), 1774−1778.

(9) Broccatelli, F.; Larregieu, C. A.; Cruciani, G.; Oprea, T. I.; Benet, L. Z. Improving the prediction of the brain disposition for orally administered drugs using BDDCS. *Adv. Drug Delivery Rev.* **2012**, *64* (1), 95−109.

(10) Broccatelli, F.; Mannhold, R.; Moriconi, A.; Giuli, S.; Carosati, E. QSAR Modeling and Data Mining Link Torsades de Pointes Risk to the Interplay of Extent of Metabolism, Active Transport, and hERG Liability. *Mol. Pharmaceutics* **2012**, *9* (8), 2290−2301.

(11) Rowland, M. Microdosing: a critical assessment of human data. *J. Pharm. Sci.* **2012**, *101* (11), 4067−74.

(12) Benet, L. Z.; Broccatelli, F.; Oprea, T. I. BDDCS Applied to Over 900 Drugs. *AAPS J.* **2011**, *13* (4), 519−547.

(13) Khandelwal, A.; Bahadduri, P.; Chang, C.; Polli, J.; Swaan, P.; Ekins, S. Computational Models to Assign Biopharmaceutics Drug Disposition Classification from Molecular Structure. *Pharm. Res.* **2007**, *24* (12), 2249−2262.

(14) Bishop, C.; Svensén, M.; Williams, C. GTM: A principled alternative to the Self-Organizing Map. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN 96)*; Springer-Verlag: Berlin, Heidelberg: 1996; pp 165−170.

(15) Bishop, C.; Svensén, M.; Williams, C. Developments of the generative topographic mapping. *Neurocomputing* **1998**, *21* (1−3), 203−224.

(16) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10* (1), 215−234.

(17) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inf.* **2012**, *31* (3−4), 301−312.

(18) Kohonen, T. *Self-Organizing Maps*; Springer-Verlag: Berlin, Heidelberg, 1995.

(19) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *J. Mol. Struct.: THEOCHEM* **2000**, *503* (1), 17−30.

(20) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.

(21) John, G. H.; Langley, P. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*; Morgan Kaufmann Publishers Inc.: Montréal, Québec, Canada, 1995; pp 338−345.

(22) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **2009**, *11* (1), 10−18.

(23) Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, Aug 23−26, 2010; IEEE Computer Society, 2010; pp 3121−3124.

(24) Ferri, C.; Hernández-Orallo, J.; Modroiu, R. An experimental comparison of performance measures for classification. *Pattern Recogn. Lett.* **2009**, *30* (1), 27−38.