JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Stargate GTM: Bridging Descriptor and Activity Spaces

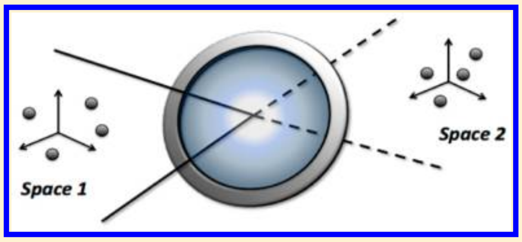
Héléna A. Gaspar,[†] Igor I. Baskin,[‡,§] Gilles Marcou,[†] Dragos Horvath,[†] and Alexandre Varnek*[,†,§]

[†] Laboratoire de Chemoinformatique, UMR 7140, Université de Strasbourg, 1 rue Blaise Pascal, Strasbourg 67000, France

[‡] Faculty of Physics, M.V. Lomonosov Moscow State University, Leninskie Gory, Moscow 119991, Russia

[§] Laboratory of Chemoinformatics, Butlerov Institute of Chemistry, Kazan Federal University, Kazan 420008, Russia

Supporting Information



ABSTRACT: Predicting the activity profile of a molecule or discovering structures possessing a specific activity profile are two important goals in chemoinformatics, which could be achieved by bridging activity and molecular descriptor spaces. In this paper, we introduce the "Stargate" version of the Generative Topographic Mapping approach (S-GTM) in which two different multidimensional spaces (e.g., structural descriptor space and activity space) are linked through a common 2D latent space. In the S-GTM algorithm, the manifolds are trained simultaneously in two initial spaces using the probabilities in the 2D latent space calculated as a weighted geometric mean of probability distributions in both spaces. S-GTM has the following interesting features: (1) activities are involved during the training procedure; therefore, the method is supervised, unlike conventional GTM; (2) using molecular descriptors of a given compound as input, the model predicts a whole activity profile, and (3) using an activity profile as input, areas populated by relevant chemical structures can be detected. To assess the performance of S-GTM prediction models, a descriptor space (ISIDA descriptors) of a set of 1325 GPCR ligands was related to a *B*-dimensional (B = 1 or 8) activity space corresponding to pK_i values for eight different targets. S-GTM outperforms conventional GTM for individual activities and performs similarly to the Lasso multitask learning algorithm, although it is still slightly less accurate than the Random Forest method.

1. INTRODUCTION

Generative Topographic Mapping (GTM)[1] is a nonlinear dimensionality reduction method and a probabilistic extension of Self-Organizing Maps (SOM).[2] Similarly to some other dimensionality reduction methods,[3] GTM can be used to visualize multidimensional data as a 2D (two-dimensional) plot. The main advantages of GTM over other methods stem from its fully probabilistic nature, which enables us to estimate the probability distribution functions (PDF) both in the initial D-dimensional descriptor space and in the "latent" or "hidden" two-dimensional space. This greatly extends the capabilities of the method to obtain classification[4−7] and regression models[8] as well as to define and visualize model applicability domains.[7]

In this article, we introduce the new approach Stargate GTM (S-GTM), through which two different spaces, *e.g.*, the space of chemical descriptors and the space of activities, are reduced to a 2D latent space. The 2D latent space plays the role of a "gate" between the descriptors and activity spaces, which can map data from one space into the other. S-GTM can assess the activity profile of a given compound using molecular descriptors as input, or, starting with an activity profile, it can detect areas in the descriptor space containing relevant chemical structures.

Predicting an activity profile is a hot topic in computer-aided drug design.[9,10] By assessing the selectivity and promiscuity of molecules, activity profiles can be used to estimate undesired side effects. In most of the cases, profiling is achieved with the help of an ensemble of individual models, each describing one activity.[15] There are, however, some multitask learning (MTL) algorithms that can predict all activities simultaneously, which include the multilayer perceptron with several output units, each corresponding to a particular activity,[11] or linear methods such as the $l_{2,1}$-norm Lasso[12−15] algorithm. In this context, S-GTM also belongs to MTL methods and provides additional benefits such as data visualization and inverse QSAR analysis.

Inverse QSAR is often defined as the process of predicting descriptors corresponding to a specific activity (or activity profile) followed by generation of new structures.[16,17] In this study, we demonstrate how S-GTM reveals some regions in the chemical space populated by molecules possessing the desired activity profiles; the corresponding structures can be retrieved in a screening procedure.

Below, we describe the S-GTM algorithm and present S-GTM models for ligand affinities (pK_i) for eight different targets built on a data set of 1325 molecules extracted from the ChEMBL database.

2. CONVENTIONAL VS S-GTM

2.1. Conventional GTM. The conventional GTM introduced by Bishop et al.[1] is a Bayesian nonlinear dimensionality reduction method. GTM models a data distribution by means of a manifold, which could be seen as a two-dimensional

Received: June 21, 2015
Published: October 12, 2015

"rubber sheet" injected into the D-dimensional descriptor space. This injection could be described as a mapping of points $\{\mathbf{x}_k\}$ from the 2D latent space to points $\{\mathbf{y}_k\}$ on the manifold in the D-dimensional space. The algorithm defines a regular grid of K nodes $\{\mathbf{x}_k\}$. Each related node on the manifold $\mathbf{y}_k$ is considered as the center of a normal distribution with an inverse variance β, which is used for data density approximation. Each data point $\mathbf{t}_n$ has a nonzero probability to be generated from this distribution, i.e., to be associated with a given grid node. The set of normal probability distributions centered on grid nodes $\mathbf{y}_k$ forms a $K \times D$ matrix characterizing the shape of the manifold; $\mathbf{Y}$ is computed using M RBF functions and a parameter matrix $\mathbf{W}$:

$$\mathbf{y}_d(\mathbf{x}) = \sum_{m}^{M} \phi_m(\mathbf{x}) W_{md} \tag{1}$$

GTM can be used not only as a tool for nonlinear dimensionality reduction and data visualization, as it has originally been suggested, but also for building classification[4−7] and regression models.[8] To build regression models, we used a method described in a previous article:[8] the activity landscape prediction (GTM activity landscape). The averaged activity value $\bar{a}_k$ associated with the node $\mathbf{x}_k$ is calculated using activities and responsibilities of all molecules in the training set

$$\bar{a}_k = \frac{\sum_{n=1}^{N} a_n R_{kn}}{\sum_{n=1}^{N} R_{kn}} \tag{2}$$

where N is the number of molecules, a_n is the experimental activity of the n-th molecule, and R_{kn} is the corresponding responsibility of node $\boldsymbol{x}_k$. The predicted activity $\hat{a}_j$ of a test compound $\mathbf{t}_j$ projected onto the map with responsibilities $\{R_{kj}\}$ is computed using the averaged activity values $\{\bar{a}_k\}$ of the training set molecules at each node $\mathbf{x}_k$:

$$\hat{a}_j = \sum_{k} \bar{a}_k R_{kj} \tag{3}$$

2.2. Stargate GTM.

Unlike its conventional analogue, Stargate GTM builds a model for *two* initial spaces (*e.g.*, the space of structural descriptors and the space of experimental activities). It represents a sort of "gate" between two spaces (see Figure 1), offering a way to map objects from one space into the other.

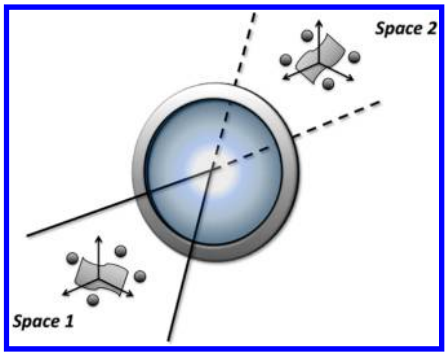


Figure 1. Schematic representation of the Stargate GTM approach: the S-GTM model links different spaces (Space 1 and Space 2) and acts as a "gate" to map molecules from one space into the other, using two different manifolds fitting the data in each space.

Instead of building a common manifold for both spaces, S-GTM creates an individual manifold in each of the two spaces (descriptors and activities) and combines the individual probability distributions estimated for each of them to obtain a joint probability distribution for both spaces. These two manifolds are built so that one grid node in the 2D latent space is associated with the centers of the corresponding normal probability distributions on both manifolds.

The S-GTM workflow consists of two stages:

i. **Training stage:** building manifolds in Space 1 and Space 2 trained via joint responsibilities.

ii. **Test stage:** data projection from Space 1 into the S-GTM 2D latent space followed by mapping into Space 2.

S-GTM Training Stage. At the training stage (Figure 2), two different probability distributions are considered for Space 1 and Space 2: $p(\mathbf{t}_n^{\text{Space1}}|\mathbf{x}_k,\mathbf{W}^{\text{Space1}},\beta^{\text{Space1}})$ and $p(\mathbf{t}_n^{\text{Space2}}|\mathbf{x}_k,\mathbf{W}^{\text{Space2}},\beta^{\text{Space2}})$, respectively. They can be computed using two different manifolds $\mathbf{Y}^{\text{Space1}}$ and $\mathbf{Y}^{\text{Space2}}$ as well as two different inverse variances β^{Space1} and β^{Space2}. Posterior probabilities (responsibilities) for Space 1 and Space 2, respectively $\mathbf{R}^{\text{Space1}}$ and $\mathbf{R}^{\text{Space2}}$, are computed during the expectation step:

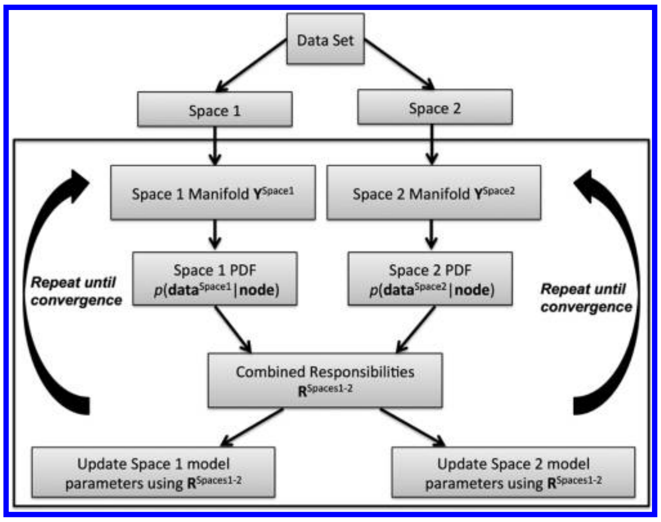

Data Set
Space 1
Space 2
Space 1 Manifold Y^Space1
Space 2 Manifold Y^Space2
Space 1 PDF p(data^Space1|node)
Space 2 PDF p(data^Space2|node)
Repeat until convergence
Repeat until convergence
Combined Responsibilities R^Spaces1-2
Update Space 1 model parameters using R^Spaces1-2
Update Space 2 model parameters using R^Spaces1-2

Figure 2. S-GTM training stage workflow.

$$\begin{aligned} R_{kn}^{\text{Space1}} &= p(\mathbf{x}_k|\mathbf{t}_n^{\text{Space1}}, \mathbf{W}^{\text{Space1}}, \beta^{\text{Space1}}) \\ &= \frac{p(\mathbf{t}_n^{\text{Space1}}|\mathbf{x}_k, \mathbf{W}^{\text{Space1}}, \beta^{\text{Space1}})}{\sum_{k'} p(\mathbf{t}_n^{\text{Space1}}|\mathbf{x}_{k'}, \mathbf{W}^{\text{Space1}}, \beta^{\text{Space1}})} \end{aligned} \tag{4a}$$

$$\begin{aligned} R_{kn}^{\text{Space2}} &= p(\mathbf{x}_k|t_n^{\text{Space2}}, \mathbf{W}^{\text{Space2}}, \beta^{\text{Space2}}) \\ &= \frac{p(\mathbf{t}_n^{\text{Space2}}|\mathbf{x}_k, \mathbf{W}^{\text{Space2}}, \beta^{\text{Space2}})}{\sum_{k'} p(\mathbf{t}_n^{\text{Space2}}|\mathbf{x}_{k'}, \mathbf{W}^{\text{Space2}}, \beta^{\text{Space2}})} \end{aligned} \tag{4b}$$

Combined responsibilities R_{kn} are assessed assuming a conditional independence of data distributions in Space 1 and Space 2

$$R_{kn} = \frac{p(\mathbf{t}_n^{\text{Space1}}|\mathbf{x}_k, \mathbf{W}^{\text{Space1}}, \beta^{\text{Space1}})^{w^{\text{Space1}}} \cdot p(\mathbf{t}_n^{\text{Space2}}|\mathbf{x}_k, \mathbf{W}^{\text{Space2}}, \beta^{\text{Space2}})^{w^{\text{Space2}}}}{\sum_{k'} p(\mathbf{t}_n^{\text{Space1}}|\mathbf{x}_{k'}, \mathbf{W}^{\text{Space1}}, \beta^{\text{Space1}})^{w^{\text{Space1}}} \cdot p(\mathbf{t}_n^{\text{Space2}}|\mathbf{x}_{k'}, \mathbf{W}^{\text{Space2}}, \beta^{\text{Space2}})^{w^{\text{Space2}}}} \tag{5}$$

DOI: 10.1021/acs.jcim.5b00398

where w^{Space1} and w^{Space2} are weight factors attributed to each space, ranging from 0 to 1: $[0 \leq w^{\text{Space1}} \leq 1, w^{\text{Space2}} = 1 - w^{\text{Space1}}]$. These combined responsibilities **R** are used to adjust the shape of manifolds $\mathbf{Y}^{\text{Space1}}$ and $\mathbf{Y}^{\text{Space2}}$ during the maximization step. This procedure is repeated until convergence.

S-GTM Test Stage. As shown in Figure 3, at the test stage, data in Space 1 (descriptor space) $\mathbf{T}^{\text{Space1}}$ is mapped into Space 2 (activity space) to obtain $\hat{\mathbf{T}}^{\text{Space2}}$. The coordinates of a data point in Space 2 correspond to its activities. The mapping from Space 1 into Space 2 ($\mathbf{T}^{\text{Space2}} \rightarrow \hat{\mathbf{T}}^{\text{Space2}}$) proceeds in two steps. First, data points from Space 1 are projected into the 2D latent space: $\mathbf{T}^{\text{Space1}} \rightarrow \mathbf{X}(\mathbf{T}^{\text{Space1}})$; then, they are mapped from the 2D latent space into Space 2: $\mathbf{X}(\mathbf{T}^{\text{Space1}}) \rightarrow \hat{\mathbf{T}}^{\text{Space2}}$.

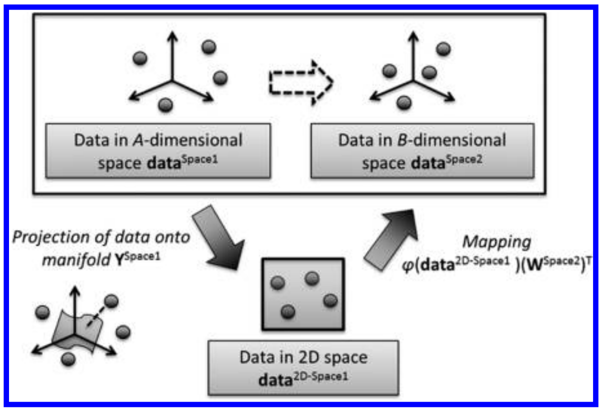

Data in *A*-dimensional space data^Space1
Data in *B*-dimensional space data^Space2
Projection of data onto manifold Y^Space1
Mapping φ(data^2D-Space1)(W^Space2)^T
Data in 2D space data^2D-Space1

Figure 3. S-GTM test stage workflow.

The projection $\mathbf{T}^{\text{Space1}} \rightarrow \mathbf{X}(\mathbf{T}^{\text{Space1}})$ is performed by computing the distances in Space 1 between the data points $\mathbf{T}^{\text{Space1}}$ and manifold points $\mathbf{Y}^{\text{Space1}}$. Then, these distances are used to compute the associated probability distribution function and responsibilities, from which the 2D coordinates of the *n*-th molecule $\mathbf{x}(\mathbf{t}_n^{\text{Space1}})$ can be obtained:

$$\mathbf{x}(\mathbf{t}_n^{\text{Space1}}) = \sum_{k}^{K} \mathbf{x}_k R_{kn}^{\text{Space1}} \tag{6}$$

The "reverse" mapping of the *n*-th molecule starts by applying radial basis functions to $\mathbf{x}(\mathbf{t}_n^{\text{Space1}})$

$$\phi_m(\mathbf{x}(\mathbf{t}_n^{\text{Space1}})) = \exp\left(-\frac{\| \mathrm{x}(\mathbf{t}_n^{\text{Space1}}) - \boldsymbol{\mu}_m \|^2}{2r^2}\right) \tag{7}$$

where $\boldsymbol{\mu}_m$ is the center of the *m*-th RBF and a parameter tunable by the user; these functions are used together with a parameter matrix $\mathbf{W}^{\text{Space2}}$ to map the 2D data points into Space 2:

$$\hat{t}_{nd}^{\text{Space2}} = \sum_{m}^{M} \phi_m(\mathbf{x}(\mathbf{t}_n^{\text{Space1}}))W_{md}^{\text{Space2}} \tag{8}$$

2.3. Prediction with S-GTM. The activity profile of a given molecule corresponding to its coordinates in the activity space ($\hat{\mathbf{t}}_n^{\text{Space2}}$) can be directly assessed by eq 8, using its coordinates in the descriptor space ($\mathbf{t}_n^{\text{Space1}}$). Notice that this method leads to reasonable results only if the responsibility distribution of the *n*-th molecule $\mathbf{r}_n^{\text{Space1}}$ in the 2D latent space is unimodal. Typically, this is the case for the mapping of molecules from descriptor space into activity space. On the other hand, an activity profile $\mathbf{t}_n^{\text{Space2}}$ mapped from activity space into descriptor space has a multimodal responsibilities distribution $\mathbf{r}_n^{\text{Space2}}$ in the 2D latent space (see section 4), and, hence, an equation analogous to eq 8 cannot be directly applied. In that case, several scenarios might be envisaged to retrieve the corresponding structures. One of these scenarios consists in selecting the nodes $\{\mathbf{x}_k\}$ with the highest responsibility values $\{R_{kn}^{\text{Space2}}\}$ for the *n*-th activity profile and retrieving compounds projected from descriptor space onto areas in the vicinity of these nodes.

3. METHODS

3.1. Data Preparation. For S-GTM modeling, both structural (molecular descriptors) and activity data for each molecule are required. However, it is sometimes difficult to find enough experimental data to completely fill the structure−activity matrix. In this work, we used a data set containing 1325 ligands of eight different rhodopsin-like GPCR receptors[18] extracted from the ChEMBL database.[19] The set was chosen so that all molecules would have an experimental pK_i value for one of these receptors (Dopamine D_2). For the other receptors, missing experimental affinities were completed by values theoretically predicted by SVM models reported by Brown et al. (2014).[18] Although predicted data might be quite different from experimental measurements (if available), we believe that the data set used in this work is suitable for comparing performances of S-GTM and conventional machine learning methods. The target names and number of experimental and predicted affinities available are given in Table 1.

The data set was randomly split into training and test sets containing 883 and 442 compounds, respectively.

Table 1. Some Information Concerning the Affinity Data Set Containing 1325 Molecules with Affinities for Eight Targets[a]

			experimental		predicted	
ChEMBL ID	short name	target name	#cpds	pK_i range	#cpds	pK_i range
1867	A2a	α_{2A}-adrenergic receptor	22	5.54−8.77	1303	6.16−7.6
217	D2	dopamine D_2 receptor	1325	2.85−10.24	0	
234	D3	dopamine D_3 receptor	684	4.17−10.13	641	5.71−8.77
219	D4	dopamine D_4 receptor	355	4.85−9.6	970	5.34−8.79
214	S1a	serotonin 1a receptor (5-HT$_{1a}$)	229	2.54−10.85	1096	3.14−9.72
224	S2a	serotonin 2a receptor (5-HT$_{2a}$)	187	5.25−11.0	1138	6.14−9.45
3155	S7	serotonin 7 receptor (5-HT$_7$)	20	5.79−8.9	1305	6.36−8.49
228	ST	serotonin transporter	58	5.21−10.0	1267	5.38−8.61

[a] D_2 was the only target for which measured pK_i values were available for all compounds; for the others affinities of some molecules were predicted theoretically (see Section 3.1). For each target the table reports its ChEMBL ID, full and short names, the number (#cpds) of experimental and predicted affinities and their ranges (min-max).

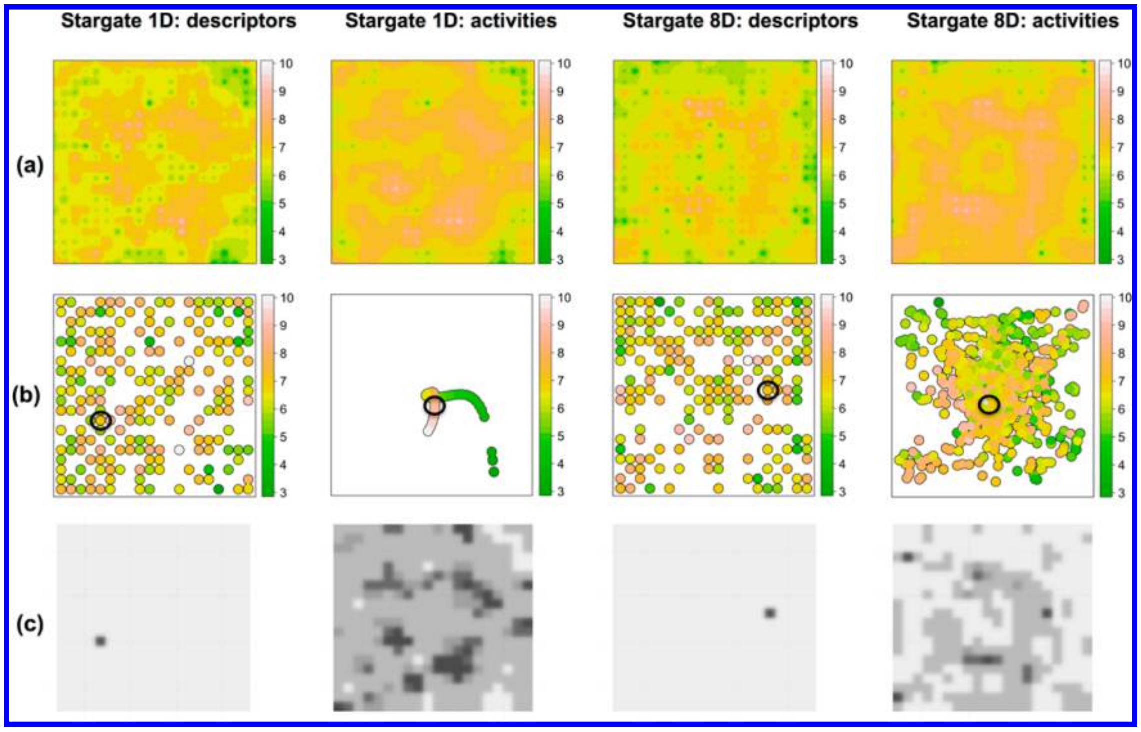


Figure 4. S-GTM maps obtained with the S-GTM/8D (eight affinities predicted simultaneously) and S-GTM/1D (activities predicted one by one) protocols. Three different representations for each protocol: (*top*) activity landscapes of D_2 affinity, (*middle*) average position of data points where one is selected by a black circle, and (*bottom*) responsibility distribution of the data point selected by a black circle. The color code corresponds to D_2 affinities in pK_i units.

As in any modeling study, results of S-GTM calculations depend on the chosen descriptors. Selection of optimal descriptors for a given data set is, however, out of the scope of this article. Therefore, in order to characterize the descriptor space we simply used ISIDA atom-centered fragments IIA(1−5)_P[20,21] providing the best SVM models for Dopamine D_2.[18] They were generated by ISIDA Fragmentor v. 2013.[21] Descriptors with more than 90% null values were discarded; thus, only 144 descriptors were used in the modeling.

3.2. Modeling Workflow. The performance of models was assessed both in 3-fold cross-validation (3-CV) and on the external test set using determination coefficients (Q^2 or R^2) and root mean squared error *RMSE*

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (a_{\exp_i} - a_{\text{pred}_i})^2}{N}} \tag{9}$$

$$R^2 \text{ (or } Q^2) = 1 - \frac{\sum_{i=1}^{N} (a_{\exp_i} - a_{\text{pred}_i})^2}{\sum_{i=1}^{N} (a_{\exp_i} - \bar{a}_{\exp})^2} \tag{10}$$

where $\mathbf{a}_{\exp}$ are experimental and $\mathbf{a}_{\text{pred}}$ are predicted values, and $\bar{a}_{\exp}$ is the average of experimental values. For clarity purposes, we will use notations R^2 and Q^2 for determination coefficients assessed on the external test set and in cross-validation, respectively.

3.2.1. Models Preparation. For conventional GTM, four parameters need to be selected: σ (radial basis functions width factor), λ (regularization coefficient), M (number of RBF functions), and K (number of grid nodes). In our calculations, we fixed M = 400 and K = 400, whereas the two other parameters were systematically varied: σ = [0.25, 0.5, 1, 1.5, 2], λ = [0.01, 0.1, 1, 10, 100]. Therefore, 25 models corresponding to different combinations of σ and λ were built. The prediction method for conventional GTM was the GTM activity landscape method (eqs 2 and 3). The parameters providing the best 3-CV performance were selected to build a model, which was then used to predict an external test set.

For S-GTM built with an eight-dimensional activity space and the ISIDA descriptor space, 25 models were also built, corresponding to different combinations of σ and λ parameters, whereas the weight factors were fixed at $w^{\text{Space1}} = w^{\text{Space2}} = 0.5$. Generally, the algorithm achieves a rapid convergence of log-likelihood $\mathcal{L}$ from both initial spaces. Unlike conventional GTM, the Stargate version with a multidimensional activity space predicts the whole activity profile simultaneously, but prediction quality still needs to be assessed for each individual target. Therefore, the values of parameters that achieved the highest number of targets (N_Q) predicted with $Q^2 > 0.50$ in 3-CV calculations were selected as optimal. If two models had the same N_Q value, the one with the highest Q^2 averaged across all properties was selected. The σ and λ parameters corresponding to the best models were then used to predict the activities of compounds in the external test set.

Performances of GTM and S-GTM models were compared to those obtained with popular Random Forest (RF)[22] and $l_{2,1}$−norm Lasso[12−15] machine learning methods. We used the RF algorithm implemented in the WEKA program[23] v. 3.6.9 with 500 trees, 12 attributes for random selection, and otherwise default parameters. The performances were measured by internal 3-CV on the training set and validation on the external test set. Lasso multitask modeling was performed using the MALSAR[24] package of the MATLAB[25] software. The regularization parameter of the Lasso model achieving the best average 3-CV RMSE in the training set was selected and used to predict compounds in the external test set.

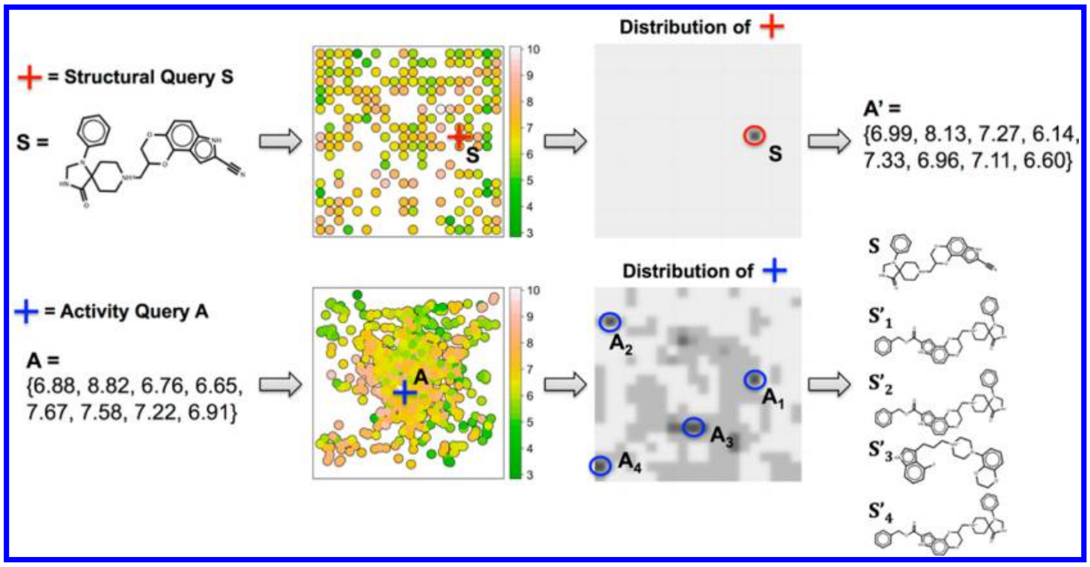


Figure 5. Examples of activity profile predictions (*top*) and structure retrieval (*bottom*) performed by S-GTM. Activity profile assessment proceeds in three steps: (1) mapping the query structure **S** onto the 2D latent space, (2) identification of its position in the 2D latent space, and (3) mapping this data point from the latent to the activity space. In an inverse QSAR task (*bottom*) S-GTM retrieves several structures with activities similar to the query **A**. This also proceeds in three steps: (1) mapping the activity query **A** onto the 2D latent space, (2) identifying some areas with high responsibility values, and (3) recovery of structures within the delimited areas. The numbers between curly brackets correspond to ligand activities (pK_i values) with respect to eight considered receptors (in the same order as in Table 1). Notice that the structural **S** and activity **A** queries correspond to one same molecule. Conventional QSAR calculations predict for the structure **S** an activity profile **A**′ which is close to **A**, whereas inverse QSAR retrieves several structures for the query **A**, including **S**.

3.2.2. GTM, S-GTM/1D, and S-GTM/8D Protocols. GTM models were generated for three protocols: conventional GTM, S-GTM/1D, and S-GTM/8D. In the S-GTM/8D protocol, the descriptor space and an eight-dimensional activity space were used to build a model able to predict all eight affinity types simultaneously. Instead of the eight-dimensional activity space, a one-dimensional activity space was used for the S-GTM/1D protocol, where models for each affinity type were built separately. Only S-GTM/8D parameters were optimized (section 3.2.1) and used to build both S-GTM/8D and S-GTM/1D models.

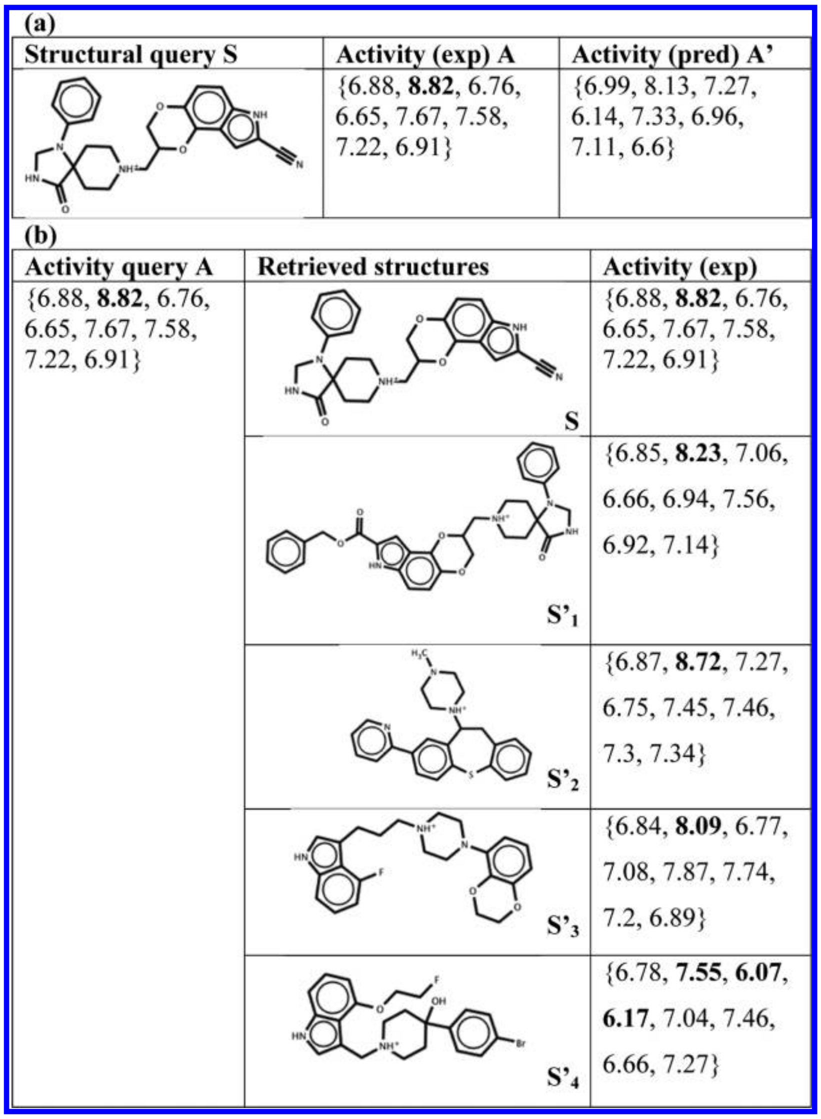
4. RESULTS AND DISCUSSION

As mentioned in section 3.2.2, for the S-GTM modeling of the Affinity data set, we used an eight-dimensional (entire profile, with prediction-filled missing values) and a one-dimensional affinity space (eight models for eight separate affinities). For each protocol, Figure 4 presents an average position for data points in 2D latent space projected from either descriptor space or activity space (*middle*) and examples of activity landscape (*top*) and responsibility distribution of one selected data point (*bottom*). Activity landscapes were calculated according to eq 2 for D_2 affinity, for which experimentally measured values were available for all data points.

The two data projections - from descriptor space (with responsibilities $\mathbf{R}^{\text{Space1}}$) and from activity space (with responsibilities $\mathbf{R}^{\text{Space2}}$) - look very different for S-GTM/8D and S-GTM/1D protocols (Figure 4 *middle*). Indeed, data projection from descriptor space into 2D latent space results in a set of points scattered on the map for both protocols. This significantly differs from the data projection from activity space which looks like a one-dimensional curve (S-GTM/1D) or a compact data cloud (S-GTM/8D). This can be explained by significant differences between $\mathbf{R}^{\text{Space1}}$ and $\mathbf{R}^{\text{Space2}}$ distributions, which define positions of data points according to eq 6. In fact, the $\mathbf{r}_n^{\text{Space1}}$ distribution for the *n*-th molecule is likely to be unimodal (Figure 4 *bottom*), whereas the $\mathbf{r}_n^{\text{Space2}}$ distribution for the *n*-th activity profile is usually multimodal. The former indicates that a molecule is well-defined by specific values of descriptors, whereas the latter can be explained by the fact that structurally different molecules sometimes have similar activity values. In particular, "inactives" are not subject to any structural constraints − they might be scattered anywhere in the descriptor space. This is a consequence of the intrinsic asymmetry of QSAR modeling and chemical similarity concept: similar molecules should have similar activities, but similar activities might correspond to quite different chemical structures. Thus, passing a molecule through the "gate" (S-GTM model) from the descriptor to the activity space, S-GTM predicts its activities. On the other hand, passing an activity query (a set of specified activity values) from the activity to the descriptor space, S-GTM performs an inverse QSAR analysis: it finds areas in the descriptor space in which other molecules might have the specified activity values.

Notice that in this paper we use the term "inverse QSAR" to denote only the first step (from activities to molecular descriptors) of what is actually known as "inverse QSAR" in the literature[16,17] (moving from desired activities to descriptors and then to novel chemical structures). S-GTM does not directly generate new structures from scratch; it only delineates areas where known structures can exhibit specific activities. These areas correspond to different modes of the multimodal distribution depicted by the most intense dark gray spots in Figure 4c. Populated areas in Figure 4 *middle*, along with the shape of the corresponding manifold in the activity space, define an applicability domain for the inverse QSAR model, exactly in the same manner as the applicability domain for direct QSAR models is defined on conventional 2D GTM maps.[7]

Table 2. Conventional and Inverse QSAR Procedures Using the S-GTM/8D Protocol: (*a*) Structural Query S and Corresponding Experimental and Predicted Activity Profiles A and A′ and (*b*) Activity Profile Query A and the Structures {S′$_i$} Found Using S-GTM, with Their Associated Experimental Activity Profiles[a]



(a)

Structural query S	Activity (exp) A	Activity (pred) A'
	{6.88, **8.82**, 6.76, 6.65, 7.67, 7.58, 7.22, 6.91}	{6.99, 8.13, 7.27, 6.14, 7.33, 6.96, 7.11, 6.6}

(b)

Activity query A	Retrieved structures	Activity (exp)
{6.88, **8.82**, 6.76, 6.65, 7.67, 7.58, 7.22, 6.91}	S	{6.88, **8.82**, 6.76, 6.65, 7.67, 7.58, 7.22, 6.91}
	S'$_1$	{6.85, **8.23**, 7.06, 6.66, 6.94, 7.56, 6.92, 7.14}
	S'$_2$	{6.87, **8.72**, 7.27, 6.75, 7.45, 7.46, 7.3, 7.34}
	S'$_3$	{6.84, **8.09**, 6.77, 7.08, 7.87, 7.74, 7.2, 6.89}
	S'$_4$	{6.78, **7.55**, **6.07**, **6.17**, 7.04, 7.46, 6.66, 7.27}

[a] The affinities (pK_i) are given in the same order as in Table 1. Experimental activities are highlighted in bold.

Several important observations can also be made from close inspection of Figure 4. First, the mode position on the unimodal distribution of the data coming from the descriptor space (first and third columns on Figure 4 for S-GTM/8D and S-GTM/1D, respectively) coincides exactly with one of the modes of the multimodal distribution of the data from the activity space (second and fourth columns on Figure 4). This means that a molecule, traveling through the gate from descriptor space to the activity space and then returning back, is reconstructed. The second observation results from the analysis of activity landscapes (Figure 4 *top*): the areas corresponding to different modes of this distribution have similar affinity values. This means that the molecule returns from the activity space along with several other molecules with similar activity values or similar activity profiles. The third observation is that the number of these "companion molecules" is smaller for an 8-dimensional affinity space than for a one-dimensional affinity space. This could be explained by a much lower probability to find a molecule with similar affinities for several targets than that for a single target.

It is interesting to note that D_2 activity landscapes obtained by projecting data (descriptor values) from the descriptor space into the 2D latent space and by projecting data (activity values) from the affinity space into the 2D latent space are very similar (Figure 4 *top*). The Pearson correlation coefficient between descriptor and activity landscapes is 75% for both S-GTM/8D and S-GTM/1D protocols.

Figure 5 demonstrates how a single S-GTM model can be used for both conventional and inverse QSAR. For this purpose, we used the maps built with the S-GTM/8D protocol (see Figure 4). In a conventional QSAR procedure, an activity profile **A**′ is predicted for a structural query **S** characterized by a unimodal responsibility distribution in the 2D latent space (Figure 5 *top*). On the other hand, an activity profile query **A**, characterized by a multimodal responsibility distribution on the 2D map (Figure 5 *bottom*), retrieves several structures {**S**′$_i$}, each of which corresponds to a particular mode (Table 2). This

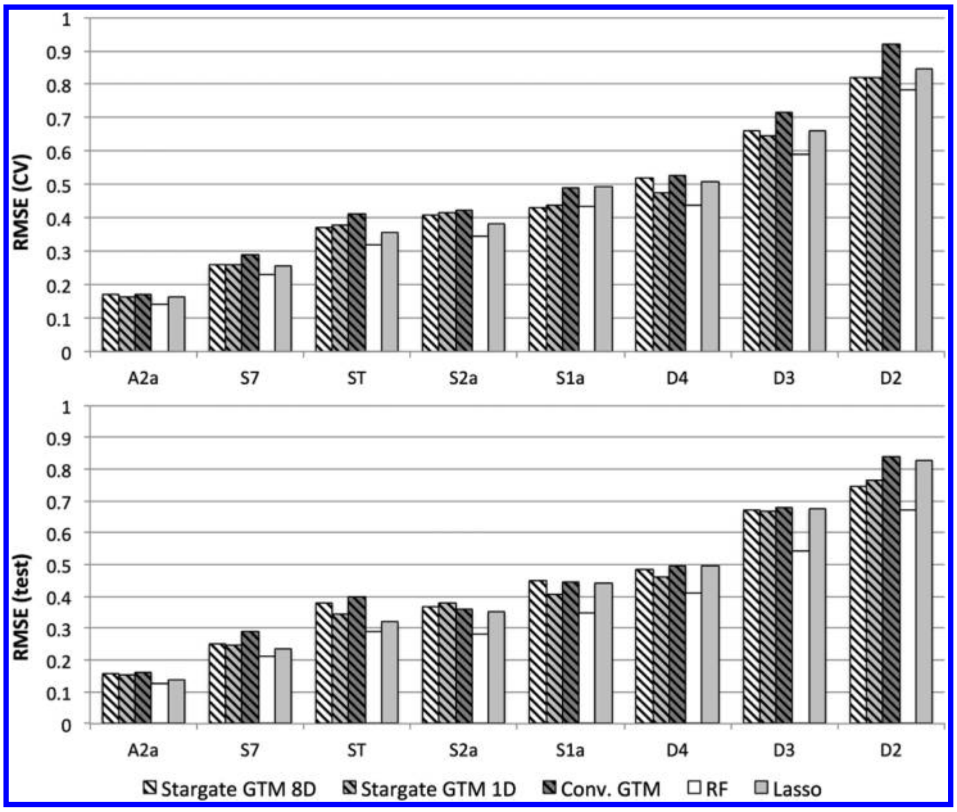


Figure 6. Predictive performance (RMSE) of Random Forest (RF), Lasso, conventional GTM, and S-GTM models for eight different activities in the Affinity data set (see Table 1) assessed in 3-fold cross-validation on the training set (*top*) and on the external test set (*bottom*). The data points are sorted in ascending order of cross-validated RMSE for S-GTM. Notice that Lasso and S-GTM/8D models predict all eight affinities simultaneously, whereas RF, conventional, and S-GTM/1D models predict activities one by one.

confirms that a data set contains more compounds sharing a single activity value than compounds sharing an entire activity profile. Notice that the diversity of retrieved structures increases with the number of modes of **A** which, in turn, decreases with the dimensionality of the activity space. This can be clearly demonstrated by comparing distributions of the query **A** responsibilities in S-GTM/8D and S-GTM/1D protocols (see Figure 4 *bottom*, second and fourth columns, respectively).

The model performances in terms of RMSE either assessed in 3-CV on the training set or on the external test set for conventional GTM, S-GTM/1D, S-GTM/8D, Lasso, and Random Forest models are given in Figure 6 for each affinity listed in Table 1. Internal (CV) and external (test) predictive performances are very close. S-GTM models outperform conventional GTM models and are close to Lasso models, although they are still less efficient than Random Forest models. The fact that S-GTM and Lasso multitask learning (MTL) methods underperform the Random Forest single-task method might be related to the fact that affinities for the different targets are unrelated and predicting them simultaneously does not provide a significant advantage. On the other hand, S-GTM has additional benefits: it can be used both for inverse QSAR and as a visualization tool, which is not the case for conventional Lasso or Random Forest methods.

■ CONCLUSION

In this paper, we described Stargate GTM (S-GTM), a new visualization and multitarget prediction approach. Unlike conventional GTM in which data points in a *D*-dimensional descriptor space are approximated by a 2D manifold, S-GTM connects two different spaces (*e.g.,* molecular descriptor space and activity space) via a 2D latent space acting as a "gate" between them. S-GTM can be used for both conventional and inverse QSAR. An activity profile can be predicted by mapping an object (chemical structure) from the descriptor space into the activity space; on the other hand, mapping objects from the activity space into the descriptor space delineates descriptor space areas populated by molecules possessing specific activity values.

As a supervised learning method, S-GTM regression outperforms conventional GTM models. Compared to conventional machine learning methods, the accuracy of S-GTM predictions is similar to that of Lasso models, although it is still slightly inferior to that of Random Forest models. We believe that the S-GTM performance could be improved by choosing a more appropriate set of descriptors and optimizing the weight factors characterizing the relative contributions of each space.

It should be noted that unlike classical machine-learning techniques which explicitly focus on minimizing RMSE, S-GTM is not *directly* bound to achieve minimal error, but rather to optimize the likelihood of matching data points to the 'rubber band' manifold. GTM in general is bound to achieve the 'most meaningful map' of compounds as a totally unsupervised method, while the novelty in S-GTM is that bringing in a second, property space, meaningful mapping with respect to the latter will implicitly improve predictive power.

Since S-GTM builds simultaneously two manifolds, the two different data projections from the two initial spaces (descriptors and activities) can be visualized. Typically, the

DOI: 10.1021/acs.jcim.5b00398

probability distribution of a data point mapped from the descriptor space into the 2D latent space is unimodal, whereas the probability distribution for a mapping from the activity space into the 2D latent space is multimodal. This illustrates that a chemical structure encoded by a set of molecular descriptors has a particular activity profile, whereas a given activity profile might correspond to several different chemical structures.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00398.

> Results of methodological tests on S-GTM performed on four data sets in which descriptor and property spaces have been described by ISIDA and MOE descriptors, respectively (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: varnek@unistra.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Prof. Jürgen Bajorath (University of Bonn), for the curated data from ChEMBL. A.V. and I.I.B. thank the Russian Scientific Foundation (Agreement No. 14-43-00024 from October 1, 2014) for support.

■ REFERENCES

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

(2) Kohonen, T. *Self-Organizing Maps*; Springer; 2001.

(3) *Nonlinear Dimensionality Reduction*; Lee, J. A., Verleysen, M., Eds.; Information Science and Statistics; Springer New York: New York, NY, 2007.

(4) 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.

(5) Kireeva, N.; Kuznetsov, S. L.; Bykov, A. A.; Tsivadze, A. Y. Towards in Silico Identification of the Human Ether-A-Go-Go-Related Gene Channel Blockers: Discriminative vs. Generative Classification Models. *SAR QSAR Environ. Res.* **2013**, *24* (2), 103–117.

(6) Kireeva, N.; Kuznetsov, S. L.; Tsivadze, A. Y. Toward Navigating Chemical Space of Ionic Liquids: Prediction of Melting Points Using Generative Topographic Maps. *Ind. Eng. Chem. Res.* **2012**, *51* (44), 14337–14343.

(7) Gaspar, H. A.; Marcou, G.; Horvath, D.; Arault, A.; Lozano, S.; Vayer, P.; Varnek, A. Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS). *J. Chem. Inf. Model.* **2013**, *53* (12), 3318–3325.

(8) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inf.* **2015**, *34* (6–7), 348–356.

(9) Morphy, J. R.; Harris, C. J. *Designing Multi-Target Drugs*; Royal Society of Chemistry: 2012.

(10) Bao, L.; Sun, Z. Identifying Genes Related to Drug Anticancer Mechanisms Using Support Vector Machine. *FEBS Lett.* **2002**, *521* (1–3), 109–114.

(11) Varnek, A.; Gaudin, C.; Marcou, G.; Baskin, I.; Pandey, A. K.; Tetko, I. V. Inductive Transfer of Knowledge: Application of Multi-Task Learning and Feature Net Approaches to Model Tissue-Air Partition Coefficients. *J. Chem. Inf. Model.* **2009**, *49* (1), 133–144.

(12) Evgeniou, A.; Pontil, M. Multi-Task Feature Learning. *Advances in Neural Information Processing Systems* **2007**, *19*, 41.

(13) Argyriou, A.; Evgeniou, T.; Pontil, M. Convex Multi-Task Feature Learning. *Mach. Learn.* **2008**, *73* (3), 243–272.

(14) Liu, J.; Ji, S.; Ye, J. Multi-Task Feature Learning via Efficient L 2, 1-Norm Minimization. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*; AUAI Press: 2009; pp 339–348.

(15) Nie, F.; Huang, H.; Cai, X.; Ding, C. H. Efficient and Robust Feature Selection via Joint *l*2, 1-Norms Minimization. *Advances in neural information processing systems* **2010**, 1813–1821.

(16) Skvortsova, M. I.; Baskin, I. I.; Slovokhotova, O. L.; Palyulin, V. A.; Zefirov, N. S. Inverse Problem in QSAR/QSPR Studies for the Case of Topological Indexes Characterizing Molecular Shape (Kier Indices). *J. Chem. Inf. Model.* **1993**, *33* (4), 630–634.

(17) Visco, D. P., Jr.; Pophale, R. S.; Rintoul, M. D.; Faulon, J.-L. Developing a Methodology for an Inverse Quantitative Structure-Activity Relationship Using the Signature Molecular Descriptor. *J. Mol. Graphics Modell.* **2002**, *20* (6), 429–438.

(18) Brown, J. B.; Okuno, Y.; Marcou, G.; Varnek, A.; Horvath, D. Computational Chemogenomics: Is It More than Inductive Transfer? *J. Comput.-Aided Mol. Des.* **2014**, *28* (6), 597–618.

(19) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100–D1107.

(20) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: A Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput.-Aided Mol. Des.* **2005**, *19* (9–10), 693–703.

(21) *ISIDA Fragmentor 2013*; Laboratoire de Chémoinformatique, UMR 7140, Université de Strasbourg: France, 2013.

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

(23) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explor. Newsl.* **2009**, *11* (1), 10–18.

(24) Zhou, J.; Chen, J.; Ye, J. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*; Arizona State University: 2011.

(25) *MATLAB and Statistics Toolbox*, Release 2014a; The MathWorks, Inc.: Natick, Massachusetts, United States, 2014.