Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge

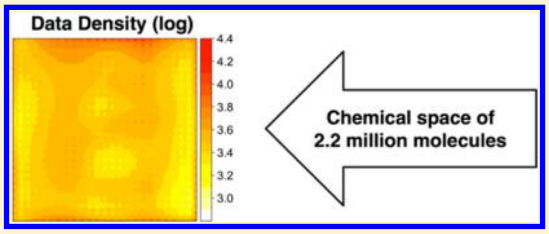
Héléna A. Gaspar,[†] Igor I. Baskin,[†,‡,§] Gilles Marcou,[†] Dragos Horvath,[†] and Alexandre Varnek*,[†,§]

[†]Laboratory of Chemoinformatics, University of Strasbourg, 67081 Strasbourg, France

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

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

S *Supporting Information*



ABSTRACT: This paper is devoted to the analysis and visualization in 2-dimensional space of large data sets of millions of compounds using the incremental version of generative topographic mapping (iGTM). The iGTM algorithm implemented in the in-house ISIDA-GTM program was applied to a database of more than 2 million compounds combining data sets of 36 chemicals suppliers and the NCI collection, encoded either by MOE descriptors or by MACCS keys. Taking advantage of the probabilistic nature of GTM, several approaches to data analysis were proposed. The chemical space coverage was evaluated using the normalized Shannon entropy. Different views of the data (property landscapes) were obtained by mapping various physical and chemical properties (molecular weight, aqueous solubility, LogP, etc.) onto the iGTM map. The superposition of these views helped to identify the regions in the chemical space populated by compounds with desirable physicochemical profiles and the suppliers providing them. The data sets similarity in the latent space was assessed by applying several metrics (Euclidean distance, Tanimoto and Bhattacharyya coefficients) to data probability distributions based on cumulated responsibility vectors. As a complementary approach, data sets were compared by considering them as individual objects on a meta-GTM map, built on cumulated responsibility vectors or property landscapes produced with iGTM. We believe that the iGTM methodology described in this article represents a fast and reliable way to analyze and visualize large chemical databases.

1. INTRODUCTION

In chemoinformatics, molecules are considered as objects in chemical space defined either by molecular graphs or by vectors of descriptors.[1] The exploration of this chemical space is a challenge for chemists seeking to understand its structure, to discover its unexplored regions, and to analyze the structural relationship between the compounds that it encompasses. Nowadays, millions of chemical structures are stored in public and proprietary databases and this number exponentially increases because of the implementation of parallel and combinatorial synthesis approaches, as well as new experimental techniques like flow or microwave reactors. A way to meet this Big Data Challenge[2,3] is to visualize large arrays of chemical data within a human-understandable, yet information-rich framework.

Approaches to data visualization in graph-based and in descriptor-based chemical spaces differ. In graph-based chemical space, the scaffold tree[4] or scaffold network[5] approaches characterize data sets by a hierarchy network connecting nodes representing individual molecules, scaffolds, and subscaffolds. In the descriptor-based chemical space, where a *D*-dimensional vector represents each molecule, two popular approaches are used: similarity network graphs and dimensionality reduction. In similarity network graphs, the nodes representing molecules are connected if their similarity exceeds a user-defined threshold (see review paper by Magiorra and Bajorath[6] and references therein). On the other hand, dimensionality reduction techniques[7] are used to transfer the objects from the *D*-dimensional chemical space into a latent space of 2 or 3 dimensions.

Among various popular methods used for chemical data visualization, only a few are suitable for visualizing large databases. For instance, Sammon mapping[8] and multidimensional scaling[9] cannot be used for processing large databases because of their high memory requirements. The linear scaling mapping based on space-filling Hilbert curves[10] preserves neither distance nor topology. On the other hand, principal component analysis (PCA) and self-organizing Kohonen maps (SOM) are more suitable for exploration of large chemical spaces. Thus, Horvath et al.[11] used SOM built on the basis of a set of some 11 000 reference drugs, bioactive compounds, and commercial molecules[12] to analyze a large database of about 200 000 molecules. PCA is the most popular method of chemical data visualization. In the context of Big Data, we could mention publications by Singh et al.,[13] Le Guilloux et al.,[14] and Reymond,[15] where this approach was used for chemical space analysis.

However, both PCA and SOM have some clear drawbacks. PCA, as a linear method of dimensionality reduction, may

Received: September 22, 2014
Published: November 25, 2014

process poorly nonlinear data. In some cases, a small number of principal components explains only a small part of data variance. As noted by Bengio et al.[16] "the expressive power of linear features is very limited: they cannot be stacked to form deeper, more abstract representations since the composition of linear operations yields another linear operation". This hampers drastically the ability of PCA to reveal disentangled factors responsible for data variation, especially in the case of Big Data. Another problem comes from the low information richness of PCA plots, resulting from the tendency to concentrate most of the data points in a certain region in the form of a Gaussian cloud, while leaving the rest of the plot poorly populated.[17] This behavior could be explained with the help of the probabilistic interpretation of PCA, which casts it as a factor analysis based on a single multivariate normal distribution function.[18]

SOM is a nonlinear dimensionality reduction method. Due to its topology-preserving character, SOM provides more information-rich plots than PCA. However, SOM suffers of its purely empirical nature and lacks solid statistical foundations.[19] As a result, the output information is truncated to the assignment of a molecule into its residence node, and the indication of how well it fits into this node. SOM tools, by default, would not report whether other nodes might have hosted a molecule as well, at only slightly higher quantization errors (mean dissimilarity between each molecule and the code vectors of its residence neuron). Since SOM does not define any probability distribution function, any powerful tool of statistical analysis and inference cannot be applied. The training algorithm for SOM does not optimize an objective function[20] and, therefore, does not guarantee convergence. The choice of SOM parameters (learning rate and width of neighborhood functions) proceeds essentially in an empirical manner, without any statistical justification.

The above is the key issue prompting Bishop et al.[21] to suggest generative topographic mapping (GTM) as a probabilistic extension of SOM. GTM overcomes most of the limitations of SOMs without introducing disadvantages. GTM is a probabilistic topology-preserving dimensionality reduction method,[21] which projects the D-dimensional chemical space onto a two-dimensional space. It has been shown that GTM could be used not only as a chemical data visualization tool[17,22,23] but also to build classification[17,22] and regression[24] structure–property models.

In their paper, Bishop et al.[25] described an incremental version of GTM (here referred to as iGTM), which could reduce computational costs for large data sets. In this work, an iGTM algorithm was implemented in the in-house ISIDA-GTM program and applied to the analysis of a database containing more than 2 million compounds.[26] This database consists of 36 libraries of commercially available compounds and the NCI database; the size of the individual libraries varies from several hundred to several hundred thousands. Therefore, we could hardly use small samples to analyze the chemical space occupied by the libraries. Indeed, equal size samples would underestimate the contribution and diversity of large libraries, whereas samples proportional to the library size would not represent small libraries with enough compounds. Thus, all molecules should be accounted for when performing a comprehensive analysis of this database.

Apart from data visualization, we suggested here some parameters useful for the analysis of both the whole database (GTM property landscapes) and individual libraries (Relative Landscape Elevation index and normalized Shannon entropy). We also introduced GTM-based measures of libraries similarity, accounting for their position in the chemical space as well as the properties of their compounds.

This paper is organized as following: first, the iGTM algorithm and its implementation are described; then, we introduce some parameters useful for the analysis of data distribution and subset similarity in GTM latent space. The robustness of the iGTM algorithm and its comparison with conventional GTM were studied on the model data set of cox2 ligands and decoys containing about 5000 molecules. Finally, we demonstrated the application of the suggested techniques to the large database of commercial and NCI compounds. Most of the calculations were performed using MOE 2D descriptors.[27] Some of them were repeated with MACCS keys for comparison purposes.

2. METHOD

2.1. Incremental GTM: The Algorithm and Its Implementation.

GTM constructs a two-dimensional manifold in the D-dimensional data space, which fits the shape of the data cloud using the function $\mathbf{y}(\mathbf{x};\mathbf{W})$ that maps from the two-dimensional latent space to the D-dimensional data space:

$$\mathbf{y}(\mathbf{x};\ \mathbf{W}) = \mathbf{W}\boldsymbol{\Phi}(\mathbf{x}) \tag{1a}$$

$$y_d(\mathbf{x};\ \mathbf{W}) = \sum_{m=1}^{M} W_{md}\phi_m(\mathbf{x}) = \sum_{m=1}^{M} W_{md}\exp\left(-\frac{\|\mathbf{x} - \mathbf{x_m}\|^2}{2\sigma}\right) \tag{1b}$$

where d runs from 1 to D, M is the number of radial basis functions (RBFs) $\boldsymbol{\Phi}$ used to approximate the mapping function, $\mathbf{W}$ is the weight matrix, $\mathbf{x}_m$ is the center of the mth RBF function in 2D latent space. The RBFs are Gaussian basis functions depending on the distance between the latent space points $\mathbf{x}$ and their center $\mathbf{x}_m$. The RBF centers form a square grid in the latent space; their number M and variance σ are method parameters. The parameter σ controls the smoothness of the manifold and is set as the minimum distance between RBF centers multiplied by a factor w. The optimal weight matrix $\mathbf{W}$ can be found using the EM (expectation-maximization) optimization algorithm, using a regularization coefficient l, also set by the user, which has an impact on the flexibility of the manifold.

GTM operates with K nodes, which form a square grid in the latent space. The images obtained by mapping these nodes to the D-dimensional data space form the centers of the Gaussian functions used to approximate the data distribution.

The E-step of the EM algorithm updates the posterior probabilities (the so-called responsibilities) R_{kn} of a node $\mathbf{x}_k$ for a data point $\mathbf{t}_n$:

$$R_{kn} = \frac{p(\mathbf{t}_n|\mathbf{x}_k,\ \mathbf{W},\ \beta)}{\sum_k p(\mathbf{t}_n|\mathbf{x}_k,\ \mathbf{W},\ \beta)} \tag{2}$$

where $\mathbf{x}_k$ is the position of the kth grid node on the GTM 2D map, and β^{-1} is the variance of the Gaussian functions approximating the distribution of data $\mathbf{t}$ in the initial D-dimensional data space:

$$p(\mathbf{t}|\mathbf{x}_k,\ \mathbf{W},\ \beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\frac{\beta}{2}\ \|\mathbf{y}(\mathbf{x}_k;\ \mathbf{W}) - \mathbf{t}\|^2\right\} \tag{3}$$

The M-step of the EM algorithm updates the weight matrix **W** and the variance β^{-1}. The optimization runs until the convergence of the GTM objective function, the *log likelihood*.

$$\mathcal{L}(W, \beta) = \sum_{n=1}^{N} \ln\left\{\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{t}_n|\mathbf{x}_k, \mathbf{W}, \beta)\right\} \tag{4}$$

In the conventional GTM algorithm,[21] the input data is considered as a single matrix of dimensions $[N, D]$, where N is the number of molecules and D is the dimensionality of the input space, i.e., the number of descriptors. When N is too large, it becomes impossible to keep this matrix in a computer's RAM. In order to tackle this problem, the incremental version of the GTM algorithm (iGTM) was suggested by Bishop et al.[25] Unlike its conventional analogue, iGTM performs learning by small data increments (blocks) instead of using the whole data matrix at once. In the E-step, responsibilities are updated incrementally. During the M-step, the parameters **W** and β are computed using both new ($\mathbf{R}'^{\text{new}}$) and old ($\mathbf{R}'^{\text{old}}$) responsibilities of the molecules in a given block:

$$\mathbf{W} = (\mathbf{\Phi}^{\text{T}}\mathbf{G}\mathbf{\Phi} + l\mathbf{I})^{-1}\{(\mathbf{RT})^{\text{old}} + (\mathbf{R}'^{\text{new}} - \mathbf{R}'^{\text{old}})\mathbf{T}'\} \tag{5}$$

$$\beta^{-1} = \beta^{-1} + \frac{1}{DN'}\sum_{i}^{N'}\sum_{k}^{K}(R'^{\text{new}}_{ik} - R'^{\text{old}}_{ik})\, \|y(x_k;\, W) - T'_i\|^2 \tag{6}$$

Here l is the regularization coefficient, **T** is the data matrix, **T**′ is the current data block (a submatrix of **T**), N' is the number of data points in **T**′, **I** is the identity matrix, and **G** is a diagonal matrix $G_{kk} = \sum_i^N R_{ik}$. The responsibilities that are not currently updated are kept in memory.

The following workflow was used in this work (Figure 1):

1. The GTM manifold is initialized on a random data subset, from which the GTM parameters W and β are computed.
2. Initialization E-step: the first estimation of responsibilities of all molecules is performed using the W and β parameters computed in step 1.
3. The first optimization of parameters is performed (initialization M-step), using the responsibilities estimated in step 2.
4. The manifold and the parameters **W** and β are updated for each data block according to eqs 5 and 6. The ensemble of all blocks forms a data scan (Figure 1). Reaching convergence may require several data set scans.

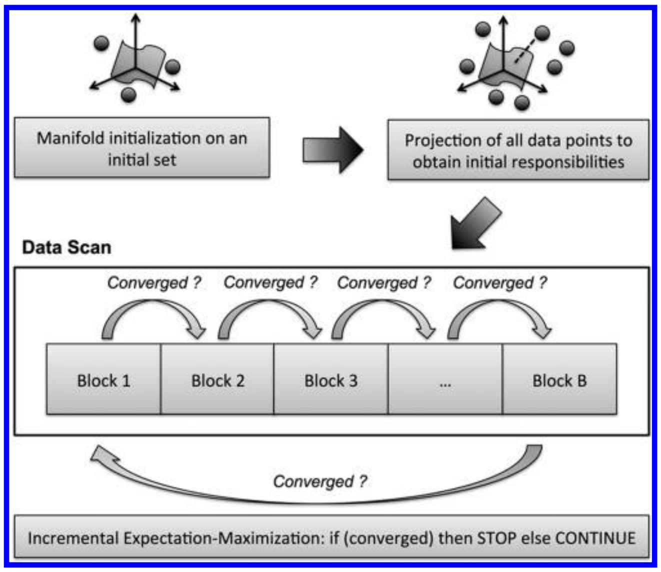

Manifold initialization on an initial set
Projection of all data points to obtain initial responsibilities
Data Scan
Converged ?
Converged ?
Converged ?
Converged ?
Block 1
Block 2
Block 3
...
Block B
Converged ?
Incremental Expectation-Maximization: if (converged) then STOP else CONTINUE

Figure 1. Incremental GTM algorithm.

This algorithm was implemented in the in-house ISIDA/GTM program.[22] The number of blocks is a parameter to be tuned by the user. If the number of blocks is equal to the number of instances, then the model is optimized instance by instance; if it is equal to 1, the algorithm becomes equivalent to the conventional GTM. There are two possibilities for checking convergence of incremental EM: block convergence and scan convergence. For block convergence, the algorithm checks the likelihood variation between two data blocks; it does not wait until it scanned the entire database. If the update is small enough (usually 10^{-4} likelihood units), the algorithm stops. For a scan convergence, the algorithm checks the likelihood gain between two entire data scans. The less time-consuming block convergence was chosen: if convergence is reached after a certain number of data blocks, the algorithm stops without going through the rest of the data set. Ng and McLachlan[28] suggested the optimal number of blocks to be close to $N^{3/8}$, where N is the size of the entire data set. This number is a trade-off between minimizing the number of M-steps and maximizing the speed of M-step computations.

2.2. GTM-Based Data Analysis.

In this section we introduce some parameters to be used for the analysis of chemical data represented on a GTM map. These are GTM property landscape, relative landscape elevation (RLE) index, normalized Shannon entropy, and γ-score.

2.2.1. GTM Property Landscape.

Molecular properties can be visualized on a GTM map using properties' distribution functions in the latent space (*GTM property landscapes*). In practice, the property landscape function is calculated as an ensemble of property values $\hat{A}_k$ assessed at each node $\mathbf{x}_k$ of the latent space:

$$\hat{A}_k = \frac{\sum_{n=1}^{N} A_n R_{kn}}{\sum_{n=1}^{N} R_{kn}} \tag{7}$$

Where A_n is the property value of the nth molecule, R_{kn} are responsibilities defined by eq 2, and N is the number of molecules in the data set. Thus, a property landscape is represented by a K-dimensional vector, where K is the total number of nodes. Here, the gstat package[29] in R[30] was used to perform a kriging interpolation[31] between these values and obtain smoothed maps. Below, we use property landscapes both to analyze the entire data set and to assess the similarity between its different subsets.

2.2.2. Relative Landscape Elevation Index.

Another useful measure is the relative landscape elevation (RLE) index, which characterizes a difference between property distributions of two libraries. It is calculated as a relative number of nodes where a given property landscape value for the ith library is higher than that for the reference landscape.

$$\text{RLE}_i = \frac{1}{K}\sum_{k=1}^{K} H(\hat{A}_k(\text{library}_i) - \hat{A}_k(\text{reference})) \times 100\% \tag{8}$$

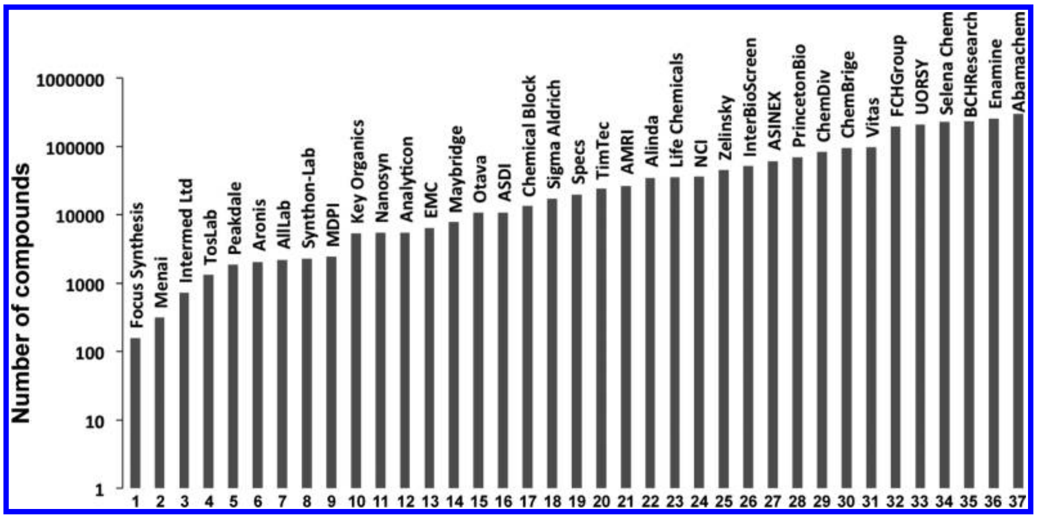


Figure 2. Number of compounds within each of the 37 filtered libraries (in log scale). The complete data set of 2 193 550 compounds was used to build a GTM map.

where H is the Heaviside function or unit step function, which returns 1 for positive arguments and 0 for negative arguments: $H(x) = 1$ if $x > 0$ and $H(x) = 0$ if $x \leq 0$. Therefore, if the property landscape value of the library is higher than the reference, the function returns 1; otherwise it returns 0. For example, a library with $\mathrm{RLE}_i = 100\%$ has higher property values than the reference set in all the nodes of the GTM latent space. As a reference, we can use the landscape of the whole database or that of any individual library j $(j \neq i)$.

2.2.3. Normalized Shannon Entropy. The cumulated responsibilities $\mathrm{Cum}R_{ki}$ are probabilities of finding a new instance close to a node x_k, obtained by averaging the responsibilities $R_{kn}(C_i)$ over N_{C_i} compounds belonging to the ith library in the latent space:

$$\mathrm{Cum}R_{ki} = \frac{\sum_n R_{kn}(C_i)}{N_{C_i}} \tag{9}$$

The Shannon entropy[32−35] measures how well a given library covers the chemical space. The higher the entropy is, the more the library is dispersed on the map:

$$E(C_i) = -\sum_k \mathrm{Cum}R_{ki} \log(\mathrm{Cum}R_{ki})$$

The entropy was normalized in the range [0, 1], by dividing by $\log(K)$:

$$E_{\mathrm{Norm}}(C_i) = \frac{E(C_i)}{\log(K)} \tag{10}$$

At $E_{\mathrm{Norm}}(C_i) = 0$ all molecules are mapped onto the same GTM node, and at $E_{\mathrm{Norm}}(C_i) = 1$, the molecules cover the chemical space uniformly.

2.3. Similarity of Libraries. *2.3.1. Libraries Overlap in the Latent and Initial Spaces.* In order to compare the overlap of libraries in the latent space several similarity or distance measures[36] were used: the Tanimoto coefficient, the Bhattacharyya coefficient, and the Euclidean distance.

$$S_{\mathrm{Tanimoto}}(C_i, C_j) = \frac{\sum_k \mathrm{Cum}R_{ki}\mathrm{Cum}R_{kj}}{\sum_k \mathrm{Cum}R_{ki}{}^2 + \sum_k \mathrm{Cum}R_{kj}{}^2 - \sum_k \mathrm{Cum}R_{ki}\mathrm{Cum}R_{kj}} \tag{12}$$

$$S_{\mathrm{Bhattacharyya}}(C_i, C_j) = \sum_k \sqrt{\mathrm{Cum}R_{ki}\mathrm{Cum}R_{kj}} \tag{13}$$

$$S_{\mathrm{Euclidean}}(C_i, C_j) = \sqrt{\sum_k |\mathrm{Cum}R_{ki} - \mathrm{Cum}R_{kj}|^2} \tag{14}$$

where the cumulated responsibilities $\mathrm{Cum}R_{ki}$ were computed using eq 9. In the initial data space, Tanimoto coefficients T_{IJ} for two individual libraries I and J containing, respectively, N_i and N_j molecules, were calculated as

$$T_{IJ} = \frac{1}{N_i N_j} \sum_{i \in I} \sum_{j \in J} T(i, j) \tag{15}$$

Here, Tanimoto coefficients $T(i, j)$ assessing similarity of two molecules i and j were calculated using molecular descriptors.

The Spearman rank correlation coefficient was used to compare the similarity matrices.

2.3.2. Mapping a Library As a Single Object on a GTM. The initial GTMs trained on all molecules may serve to encode any individual library by a vector of length equal to the total number of nodes K in the latent space, based either on cumulated responsibilities $\mathrm{Cum}R_{ki}$, or on property landscape values $\hat{A}_k$ at nodes $\mathbf{x}_k$. These vectors can be used as inputs for a second generation of library-driven GTMs that we coined meta-GTM (μGTM), built on the output of primary GTMs, and in which each library is considered as a unique object. The parameters used to build the μGTMs ($K = 625$, $M = 25$, $w = 2$, $l = 1$) were the same as for the original GTM, see section 4.3.

3. DATA AND DESCRIPTORS

The data set, composed of compounds from 36 commercial libraries and the NCI database, was taken from the article by Petrova et al.[26] It contains 2 193 550 druglike molecules filtered out from a set of 15.6 million compounds. The data set used in this work contains several duplicates because some molecules are present in different libraries. The size of the libraries ranges

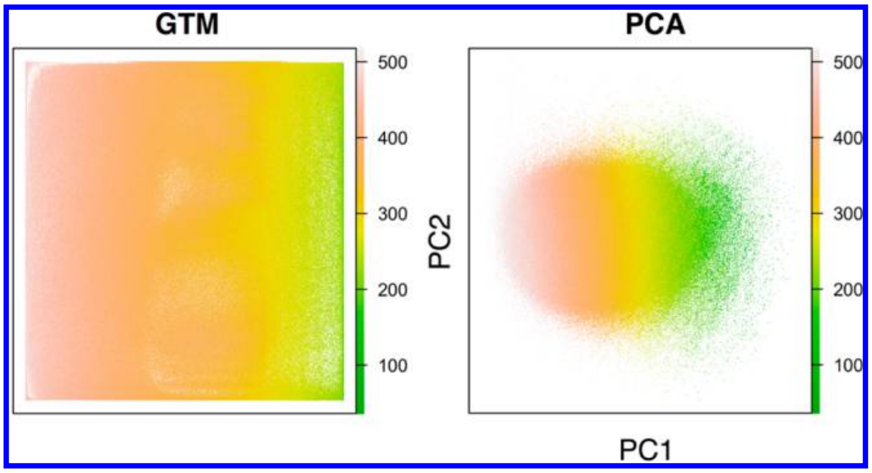


Figure 3. GTM and PCA maps of the entire database (∼2.2 million compounds), colored by molecular weight (in Dalton). The two first principal components (PC1 and PC2) explain only 40% of the variance. The colors vary from green (low molecular weight) to pink (high molecular weight).

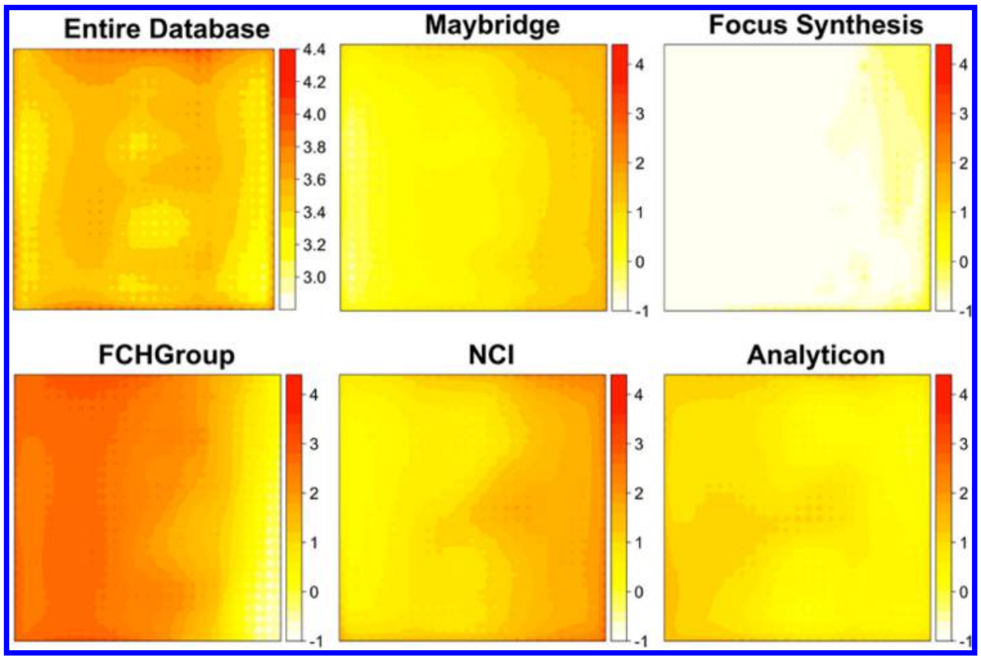


Figure 4. Density maps of the entire database of ∼2.2 million molecules and several individual libraries in GTM latent space. For a better visualization, two different scales were used for the maps of the entire database (from 3 to 4.4 log units) and for the five individual libraries (from −1 to 4 log units) All maps are based on the common manifold calculated for the entire database.

from 158 (Focus Synthesis) to 299 898 (Abamachem); see Figure 2.

The data set was standardized using the Instant JChem (ChemAxon) program (http://www.chemaxon.com). In most of cases, 186 MOE 2D descriptors (referred to as MOE) were used. They include some physicochemical parameters, structural characteristics, and connectivity indices. Some of these descriptors were used as “properties” to calculate property landscapes.

For the purpose of comparison, some calculations were performed with 166 MACCS keys (referred to as MACCS) representing an array of 166 bits encoding the absence or presence of certain structural features in a molecular graph. The number of MACCS descriptors was reduced to 162 after removing descriptors with all values equal to 0. Both MOE and MACCS descriptors were computed using the MOE v.2011.10 software.[27]

For the methodological tests (section 4.1), we used a data set of 4499 compounds containing 409 ligands and 4090 decoys of cox2 (cyclooxygenase 2) from the DUD (Directory of Useful Decoys).[37] This data set has a reasonable size both for running conventional GTM and to study different options of iGTM.

4. RESULTS AND DISCUSSION

4.1. Methodological Tests and Choice of the iGTM Parameters. The methodological tests of iGTM focused on two questions: (1) does iGTM produce a similar map as a conventional GTM, when initialized on all compounds and (2) more generally, does the resulting map depend on the iGTM parameters—the size of the initial subset and the number of batches? The calculations were performed on a data set of 4499 compounds including 409 ligands and 4090 decoys of cox2 (cyclooxygenase 2) from the DUD (Directory of Useful Decoys).[37] Results given in the Supporting Information show

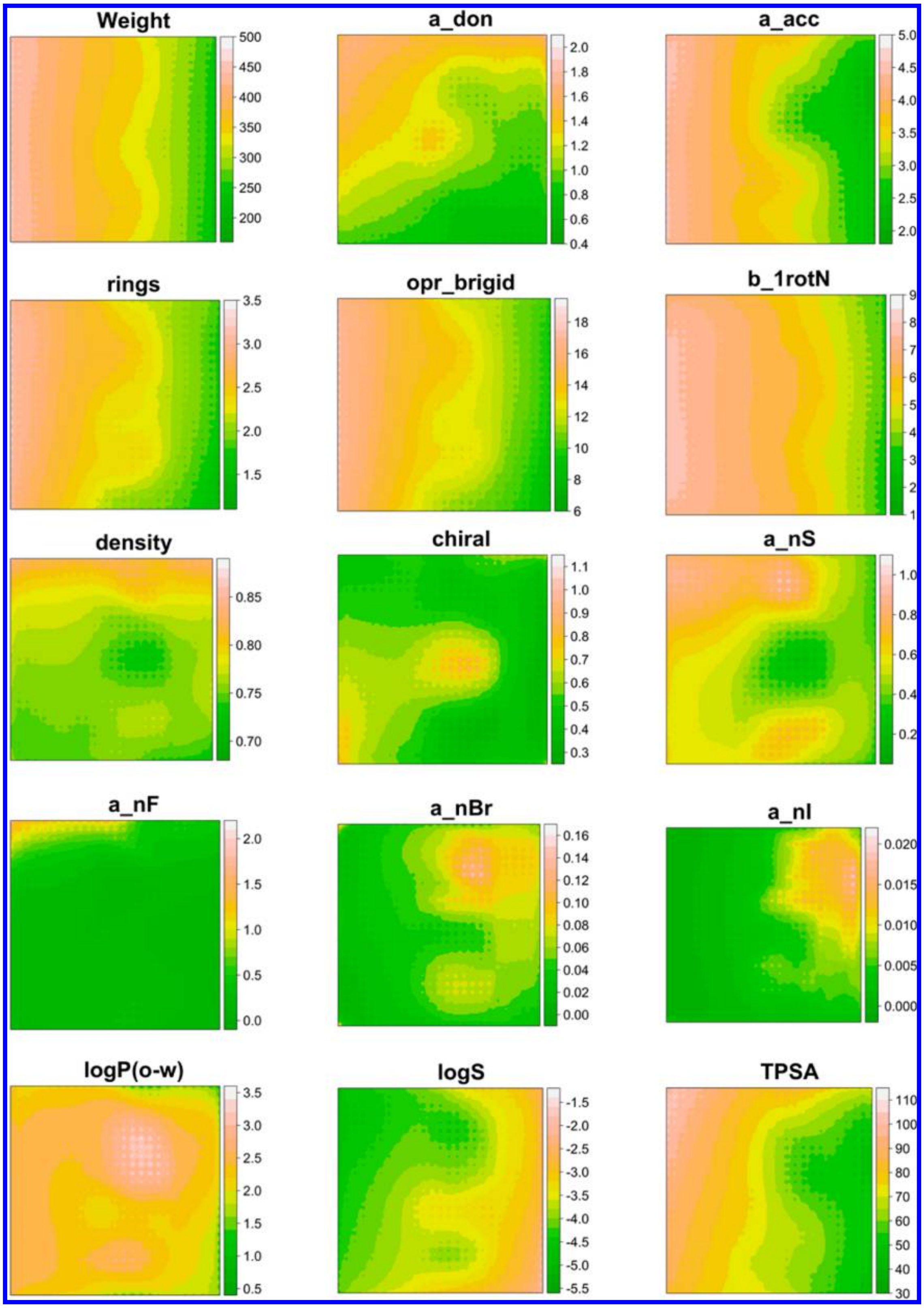


Figure 5. GTM property landscapes for molecular weight (Weight), the number of H-bond donor (a_don) and acceptor (a_acc) atoms, rings, rigid (opr_brigid) and rotatable (b-1rotN) bonds, chiral centers, sulfur (a_nS), fluor (a_nF), bromine (a_nBr), iodine (a_nI) atoms, molecular density (density), logP, logS, and TPSA.

that (i) iGTM initialized on the entire data set produces a similar map as conventional GTM, which proves the convergence of these two algorithms, (ii) the variation within a certain range of the number of blocks and the size and composition of the initial set does not affect the clustering ability nor the dispersion of the points on the map.

Four preliminary calculations have been performed on the entire database of $N = 2\,193\,550$ compounds in order to check

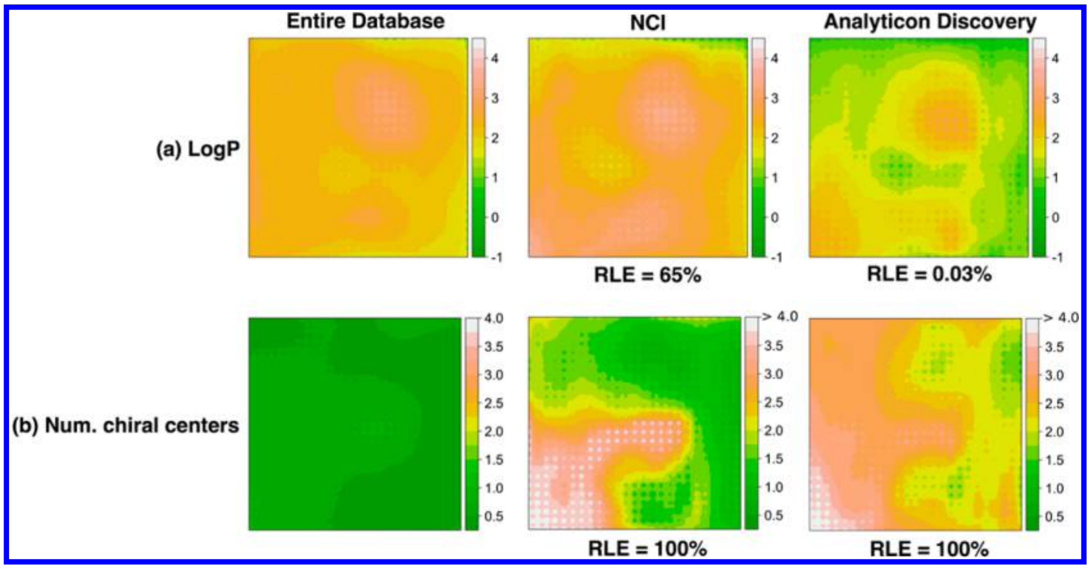


Figure 6. Property landscapes (logP and the number of chiral centers) of the entire database and two libraries—NCI and Analyticon Discovery. RLE indices are calculated taking as a reference the property landscape of the entire database.

the variation of the iGTM output as a function of the initial subset and operational parameters: the number M of radial basis functions, the RBF width multiplication factor w, and the regularization coefficient l (see section 2.1). The grid resolution K was set to 625. The calculations were performed using two random initialization subsets of 5000 compounds and two different parameters sets: $[K = 625, M = 25, w = 2, l = 1]$ and $[K = 625, M = 16, w = 1, l = 10]$. Following the recommendations by Ng and McLachlan,[28] the number of blocks was set to 190. In each run, the normalized Shannon entropy values were calculated for each library. Thus, each map could be characterized by a vector of 37 entropy values. The four resulting vectors were highly correlated, with a Spearman rank correlation $\rho = 0.997 \pm 2 \times 10^{-3}$. We also compared the four Euclidean distance matrices of the 37 libraries using Mantel's test with Spearman's ρ and the correlation was always very high ($\rho = 0.996 \pm 2 \times 10^{-3}$). This demonstrates that iGTM is pretty stable with respect to the variation of the method parameters. We selected the parameter set $[K = 625, M = 25, w = 2, l = 1]$ for further calculations. With these parameters, iGTM requires approximately 19 CPU h on a 3.4 GHz Intel Core i7 processor, and 8 Gb RAM.

4.2. Chemical Space Layout of the entire database and individual libraries. The iGTM map obtained for the entire database is shown in Figure 3. The data points, each corresponding to a molecule, cover all the latent space. This differs from a plot made in the space of the two first principal components, where the data occupy only a small part of the latent space (Figure 3). Another way to represent the database is, instead of plotting the position of each compound on the map, to plot the data density distribution calculated from the sum of all responsibilities in each node of the latent space. Since each compound in the database is annotated with respect to its affiliation to a particular library, the data density distributions of individual libraries could be easily extracted from that of the entire database. On the maps presented in Figure 4, significant data density variations may be observed. Some of the individual libraries cover more or less homogeneously the whole latent space, whereas others cover only a part of it. In order to better characterize the latent space zones underlying these density distributions, some property landscapes were prepared (Figure 5). Molecular weight gradually increases from the left to the right of the map—unsurprisingly, since size is a key parameter of chemical space. The number of H-bond acceptor atoms, rings, rigid and rotatable bonds, and TPSA logically follow this trend. The density of H-bond donor atoms is particularly high in the upper left corner, whereas the density of chiral centers reaches its maxima in the center and in the bottom left corner of the map. Molecules containing halogens or sulfur are localized mostly in the upper area of the map; this explains high molecular density values (molecular weight divided by van der Waals volume) there. As may be seen from the distributions of logP and aqueous solubility logS, hydrophobicity is larger for the molecules occupying the left-hand side of the map.

Looking at the individual libraries distributions (Figure 5) in the property landscape context gives an idea about their content. For instance, NCI is mostly located on the right-hand side of the map, which is a low molecular weight area.

Property landscapes of individual libraries may also be analyzed. As an example, library-specific distributions of logP and number of chiral centers, respectively, are shown in Figure 6. Two selected libraries may have similar landscapes for a property A, and different landscapes for another property B—which is the case for the properties plotted in Figure 6. For some libraries, the logP property distribution is similar to that of the whole database (e.g., NCI), whereas for others it is quite different (e.g., Analyticon Discovery). The chirality landscape values (an averaged number of chiral centers per molecule) for NCI and Analyticon Discovery varies in the range 0−8 (Figure 6) whereas for the entire database it varies in the range 0−1.1 (Figure 5). This explains why RLE = 100% for these two libraries, when the chirality landscape for the whole database is taken as a reference. On the other hand, taking NCI chirality landscape as a reference, the RLE of Analyticon Discovery is 75%. This shows that Analyticon has, on average, a higher number of chiral centers than NCI, over all the chemical space area it covers. This is not surprising taking into account that Analyticon Discovery specializes in natural compounds rich in chiral centers.

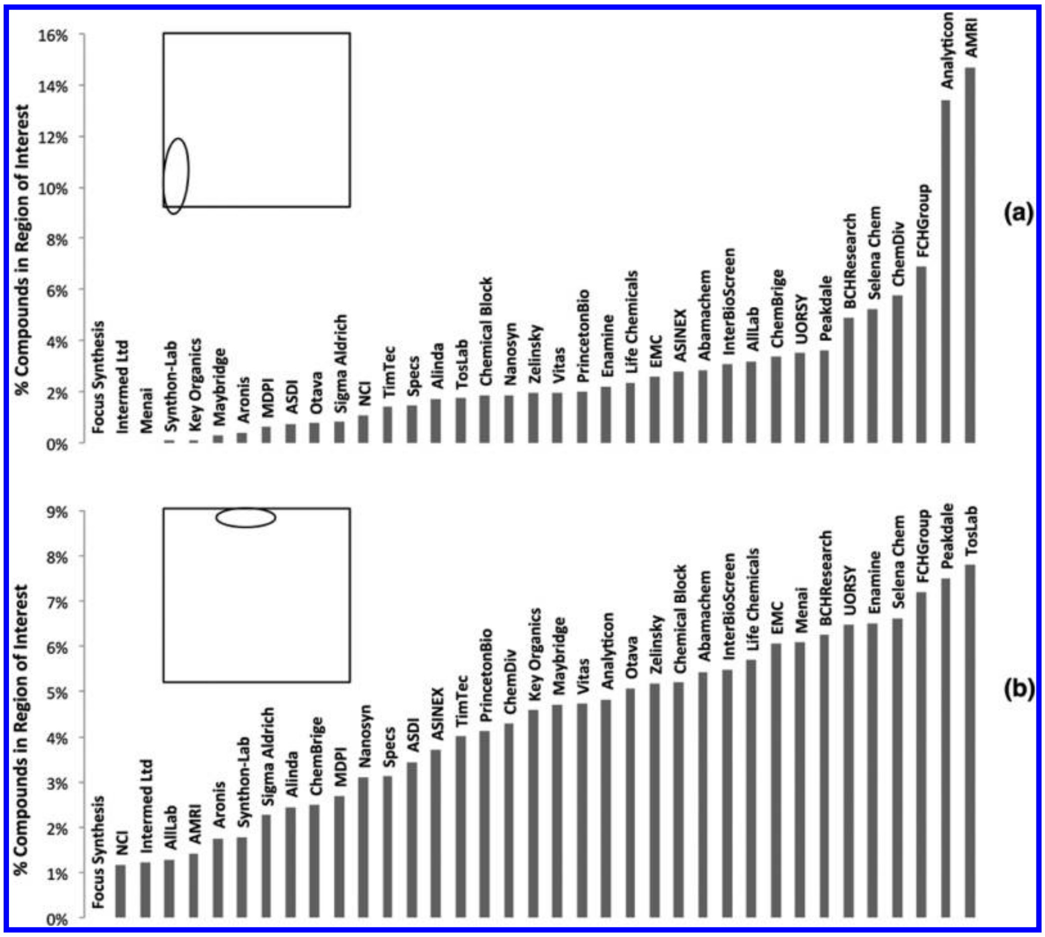


Figure 7. Regions of interest in GTM latent space populated by compounds possessing a selected profile: (a) high average number of chiral centers (>0.7), high molecular weight (>400 Da), and low logS (<−3.5 log units); (b) high average number of F atoms (>0.95) and relatively high logS (>−4.3). Each region of interest, delineated by an ellipse, was defined by superimposing the corresponding activity landscapes shown in Figure 5. The histograms show the relative populations (number of compounds divided by the total size of each library, in percent) of the individual libraries in these zones.

Property landscapes could be particularly useful to select "regions of interest" populated by compounds possessing desirable property profiles. For example, in order to choose compounds with low solubility, a large number of chiral centers, and high molecular weight, we can superimpose the corresponding property landscapes shown in Figure 5. The delineated region in the bottom left-hand corner of the map has the required profile (Figure 7a). The information about the contribution of individual libraries in this zone can be easily extracted: Analyticon Discovery and AMRI, both rich in natual compounds, have a large proportion of compounds in this region (Figure 7a). The relative contribution of individual libraries varies from one region of the map to another. Thus, another region of interest, rich in fluorinated and relatively highly soluble compounds (resulting from the superimposition of both solubility and number of fluorine atoms landscapes), is well populated by Peakdale and TosLab (Figure 7b).

The question arises: how well do individual libraries cover the GTM latent space? The normalized Shannon entropy ranges from 0.75 (Focus Synthesis) to 0.98 (ASINEX), showing that all libraries have a high coverage of the chemical space. Figure 8a compares entropy values and sizes of libraries. Generally, no correlation between these two parameters is observed. In most cases, relatively small libraries, e.g. Life Chemicals, EMC, ASINEX, cover a large portion of the chemical space, and therefore have large entropy values.

To figure out the impact of descriptors on the chemical space construction, a GTM map was also computed with MACCS keys. Surprisingly, a correlation between the MOE and MACCS entropies was observed (Figure 8b). However, some libraries do not follow this trend. For instance, comparing EMC and FCHGroup libraries, it may be observed that the former covers better the MOE-based space, whereas the latter covers better the MACCS-based space (see the Supporting Information for the details).

4.3. Similarity of Libraries.

In this section, individual libraries are compared using two different but complementary approaches: (i) similarity indices and distance measures (see section 2.2) and (ii) clustering in GTM library space (section 2.3).

The similarity of libraries was evaluated using $S_{\text{Bhattacharryya}}$, S_{Tanimoto}, and $S_{\text{Euclidean}}$ (eqs 12−14) for GTM distributions, resulting, respectively, in three 37 by 37 matrices. These matrices are intercorrelated: the Spearman rank correlation coefficient varies from 0.90 to 0.96. In order to compare similarity measures in the latent and initial data spaces, "conventional" Euclidean distances were also calculated for pairs of molecules. Since these calculations for the entire database of ~2.2 million compounds would have been be very

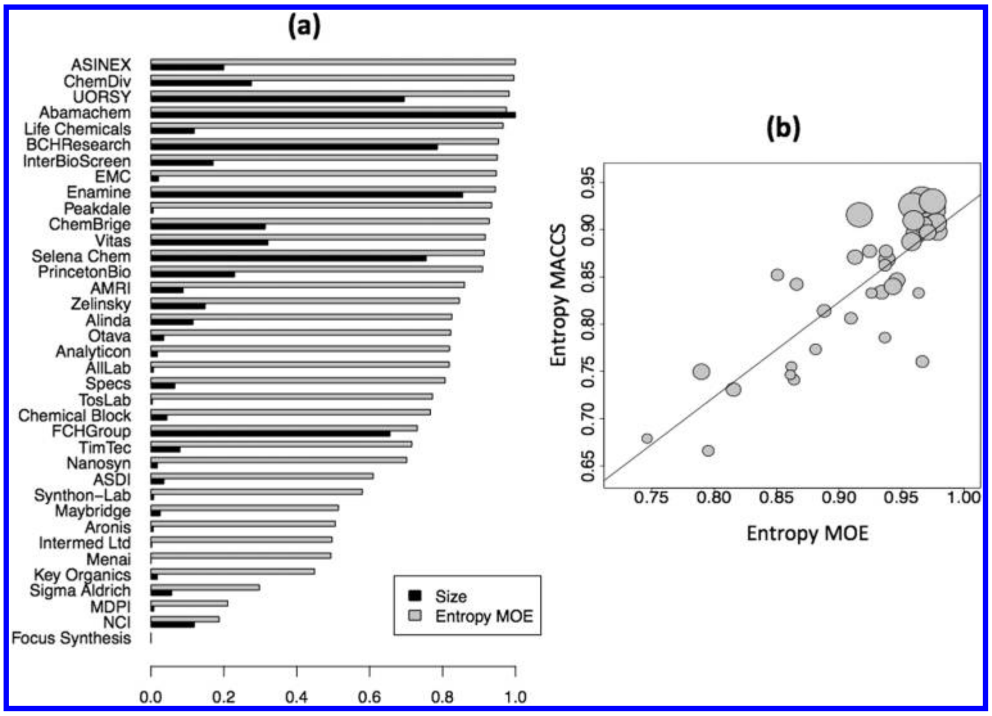


Figure 8. (a) Size (black) of individual libraries and their normalized Shannon entropy (gray). The libraries are sorted by decreasing entropy. (b) Correlation between GTM-based entropies calculated with MACCS and MOE descriptors. Each library is represented by a point with a size proportional to the number of compounds.

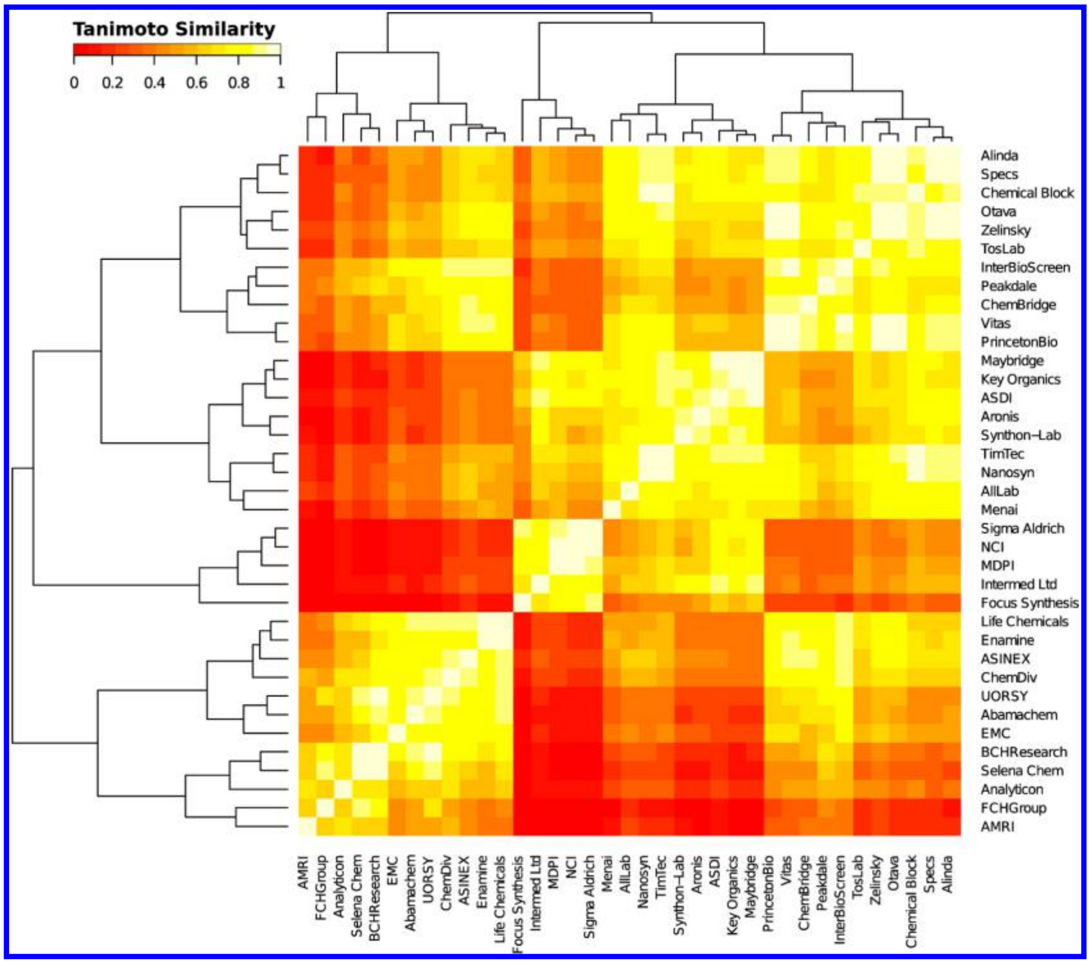


Figure 9. Heatmap representing similarities between 37 libraries on 2D GTM map using the GTM-based Tanimoto coefficient (eq 12) as a metric.

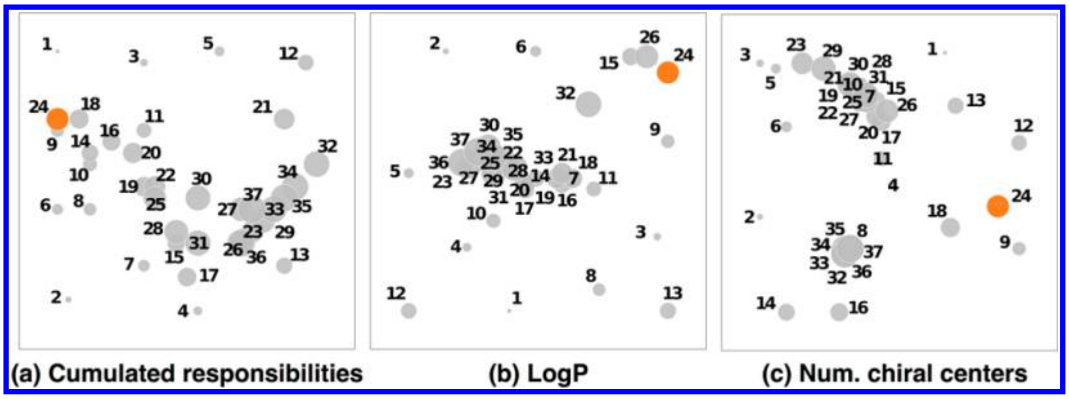


Figure 10. Examples of μGTM maps built on (a) cumulated responsibilities, (b) logP landscape, and (c) chirality landscape from iGTM. On the maps, each library is represented by a circle, of which the diameter is proportional the library's size. The correspondence between the objects' numbers and library names is given in Figure 2. The NCI collection (number 24) is shown in orange.

time-consuming, we selected the 14 smallest libraries with a number of molecules <10 000. Euclidean similarity matrices (14 × 14) corresponding to libraries' similarities in the latent and initial spaces correlate nicely: the Spearman rank correlation coefficient was equal to 0.78. The latter is an important result showing that some conclusions drawn for the objects' similarity in the latent space can be extended to the initial space.

Similar calculations repeated with MACCS descriptors resulted in similar correlations between different metrics in the latent space as well as between similarities measured in the initial and latent spaces (see the Supporting Information for the details).

The heatmap in Figure 9 represents the similarity matrix in GTM latent space for 37 libraries using Tanimoto coefficients S_{Tanimoto} calculated according to eq 12. It clearly shows several clusters formed by similar libraries ($S_{\text{Tanimoto}} > 0.7$). Each cluster corresponds to libraries with similar data density distributions. Thus, looking at Figure 9 we may easily explain why Focus Synthesis and NCI on the one hand and Analyticon Discovery and FCHGroup on the other hand are in the same clusters.

This analysis could be particularly useful to select database(s) for virtual or real screening experiments. Thus, we could impose a Tanimoto threshold to avoid the selection of databases that overlap too much.

Another way to analyze libraries similarities is to build μGTM maps on the vectors based on cumulated responsibilities or property landscapes derived from the original GTM (see section 2.3.2). Thus, each library is characterized by a vector of dimension equal to the number of nodes in the 2D map. Since in our calculations, the grid resolution parameter K was always equal to 625, the vectors used for μGTM construction had a length of 625. Three examples of μGTM maps where each library is represented by a data point are given in Figure 10.

The μGTM built on cumulated responsibilities (Figure 10a) displays neighborhood behavior similar to the Tanimoto coefficients heatmap in Figure 9. For example, NCI (number 24 on μGTM) in both cases displays high similarity to Sigma-Aldrich (18), MDPI (9), Intermed Ltd. (3), Focus Synthesis (1), Maybridge (14), Key Organics (10), Nanosyn (11), and ASDI (16). This shows that in the chemical space of MOE descriptors and in the corresponding GTM latent space (Figure 3), all these libraries overlap. It should be noted that the similarity between libraries depends on the overlap of their data density functions in the chemical space. This differs from the approach by Le Guilloux et al.[14] who suggested considering "delimited reference chemical subspaces" delimited by convex hulls on a plot of two first principal components without accounting for data density functions.

μGTM maps in Figure 10b and c characterize the similarity of two different property landscapes. In the μGTM built on logP landscapes (Figure 10b), NCI is in the vicinity of FCHGroup (32) and Otava (15). This looks surprising because according to Figures 9 and 10a, neither FCHGroup nor Otava significantly overlap with NCI in the chemical space. Similar observations could be made for the μGTM built on chirality landscapes (Figure 10c), where Analyticon Discovery (12) is a neighbor of NCI because their chirality landscapes are similar (see Figure 6). Thus, despite the small overlap of the density distributions of the two libraries, their property landscapes (i.e., property distribution functions) may nicely overlap in the GTM latent space.

5. CONCLUSION

In this work, we have demonstrated that incremental GTM is a valuable solution for visualization and analysis of large arrays of chemical data. The iGTM algorithm implemented in our in-house program is rather fast: it processes more than 2 million molecules for 19 h only on a desktop monoprocessor machine.

One of the significant advantages of GTM over other popular visualization approaches is the possibility to assess various probability distribution functions (data density and property landscapes). These functions, together with related GTM-based parameters introduced in this work (normalized Shannon entropy and relative landscape elevation index), can efficiently be used for data analysis. In particular, they may help chemists to analyze the content of a given database in terms of chemical structure and molecular properties, to select compounds possessing a desirable properties profile, to compare different databases according to their position in chemical space, and to perform a virtual screening by identifying the neighborhood of the query.

In our opinion, the use of the data probability distribution functions instead of ensembles of data points for chemical space analysis is inevitable in the context of Big Data challenge. We believe that the methodology described in this article represents a fast and reliable way to analyze and visualize large chemical databases.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

(i) iGTM methodological tests, (ii) data density of 37 libraries in the GTM latent space calculated with MOE descriptors, (iii) landscapes of all 186 MOE properties, (iv) normalized Shannon entropy values using MACCS keys, (v) comparison of libraries' similarity using different metrics. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: varnek@unistra.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.V. and I.I.B. thank Russian Scientific Foundation (Agreement No 14-43-00024 from October 1, 2014) for support. Dr. A. Chuprina is acknowledged for data providing.

■ REFERENCES

(1) Varnek, A.; Baskin, I. I. Chemoinformatics as a Theoretical Chemistry Discipline. *Mol. Inform.* **2011**, *30*, 20−32.

(2) Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-driven medicinal chemistry in the era of big data. *Drug Discovery Today* **2014**, *19*, 859−868.

(3) Szlezák, N.; Evers, M.; Wang, J.; Pérez, L. The Role of Big Data and Advanced Analytics in Drug Discovery, Development, and Commercialization. *Clin. Pharmacol. Ther.* **2014**, *95*, 492−495.

(4) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47−58.

(5) Varin, T.; Schuffenhauer, A.; Ertl, P.; Renner, S. Mining for Bioactive Scaffolds with Scaffold Networks: Improved Compound Set Enrichment from Primary Screening Data. *J. Chem. Inf. Model.* **2011**, *51*, 1528−1538.

(6) Maggiora, G. M.; Bajorath, J. Chemical space networks: a powerful new paradigm for the description of chemical space. *J. Comput. Aided Mol. Des.* **2014**, *28*, 795−802.

(7) Pletnev, I. V.; Ivanenkov, Y. A.; Tarasov, A. V. Dimensionality Reduction Techniques for Pharmaceutical Data Mining. In *Pharmaceutical Data Mining*;Balakin, K. V., Ed.; John Wiley & Sons, Inc., 2009; pp 423−455.

(8) Sammon, J. W. A Nonlinear Mapping for Data Structure Analysis. *IEEE Trans. Comput.* **1969**, *18*, 401−409.

(9) Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **1964**, *29*, 1−27.

(10) Delaney, J. Linear scaling mapping of chemical space. *239th ACS National Meeting*, San Francisco, CA, United States, March 21, 2010.

(11) Horvath, D.; Lisurek, M.; Rupp, B.; Kühne, R.; Specker, E.; von Kries, J.; Rognan, D.; Andersson, C. D.; Almqvist, F.; Elofsson, M.; Enqvist, P.-A.; Gustavsson, A.-L.; Remez, N.; Mestres, J.; Marcou, G.; Varnek, A.; Hibert, M.; Quintana, J.; Frank, R. Design of a General-Purpose European Compound Screening Library for EU-OPEN-SCREEN. *ChemMedChem.* **2014**, n/a−n/a.

(12) Bonachera, F.; Marcou, G.; Kireeva, N.; Varnek, A.; Horvath, D. Using self-organizing maps to accelerate similarity search. *Bioorg. Med. Chem.* **2012**, *20*, 5396−5409.

(13) Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L. Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository. *J. Chem. Inf. Model.* **2009**, *49*, 1010−1024.

(14) Le Guilloux, V.; Colliandre, L.; Bourg, S.; Guénegou, G.; Dubois-Chevalier, J.; Morin-Allory, L. Visual Characterization and Diversity Quantification of Chemical Libraries: 1. Creation of Delimited Reference Chemical Subspaces. *J. Chem. Inf. Model.* **2011**, *51*, 1762−1774.

(15) Ruddigkeit, L.; Awale, M.; Reymond, J.-L. Expanding the fragrance chemical space for virtual screening. *J. Cheminformatics* **2014**, *6*, 27.

(16) Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798−1828.

(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. Inform.* **2012**, *31*, 301−312.

(18) Tipping, M. E.; Bishop, C. M. Probabilistic Principal Component Analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 611−622.

(19) Svensen, J. F. M. GTM: The Generative Topographic Mapping. Ph.D. Thesis, University of Aston in Birmingham, 1998.

(20) Erwin, E.; Obermayer, K.; Schulten, K. Self-organizing maps: ordering, convergence properties and energy functions. *Biol. Cybern.* **1992**, *67*, 47−55.

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

(22) 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*, 3318−3325.

(23) Chupakhin, V.; Marcou, G.; Gaspar, H.; Varnek, A. Simple Ligand−Receptor Interaction Descriptor (SILIRID) for alignment-free binding site comparison. *Comput. Struct. Biotechnol. J.* **2014**, *10*, 33−37.

(24) 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*, 14337−14343.

(25) Bishop, C. M.; Svensén, M.; Williams, C. K. I. Developments of the generative topographic mapping. *Neurocomputing* **1998**, *21*, 203−224.

(26) Petrova, T.; Chuprina, A.; Parkesh, R.; Pushechnikov, A. Structural enrichment of HTS compounds from available commercial libraries. *MedChemComm* **2012**, *3*, 571.

(27) *Molecular Operating Environment (MOE)*, 2011.10; Chemical Computing Group Inc., 2011.

(28) Ng, S. K.; McLachlan, G. J. On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. *Stat. Comput.* **2003**, *13*, 45−55.

(29) Pebesma, E. J. Multivariable geostatistics in S: the gstat package. *Comput. Geosci.* **2004**, *30*, 683−691.

(30) R Development Core Team. *R: A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013; ISBN 3-900051-07-0, http://www.R-project.org.

(31) Williams, C. K. I. Prediction With Gaussian Processes: From Linear Regression To Linear Prediction And Beyond. In *Learning and Inference in Graphical Models*; Kluwer, 1997, pp 599−621.

(32) Shannon, C. A Mathematical Theory of Communication. *Bell Syst. Technol. J.* **1948**, *27*, 379−423.

(33) Jost, L. Entropy and diversity. *Oikos* **2006**, *113*, 363−375.

(34) Wang, Y.; Geppert, H.; Bajorath, J. Shannon Entropy-Based Fingerprint Similarity Search Strategy. *J. Chem. Inf. Model.* **2009**, *49*, 1687−1691.

(35) Godden, J. W.; Stahura, F. L.; Bajorath, J. Variability of Molecular Descriptors in Compound Databases Revealed by Shannon Entropy Calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 796−800.

(36) Cha, S.-H. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *Int. J. Math. Models Methods Appl. Sci.* **2007**, *1*, 300−307.

(37) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.