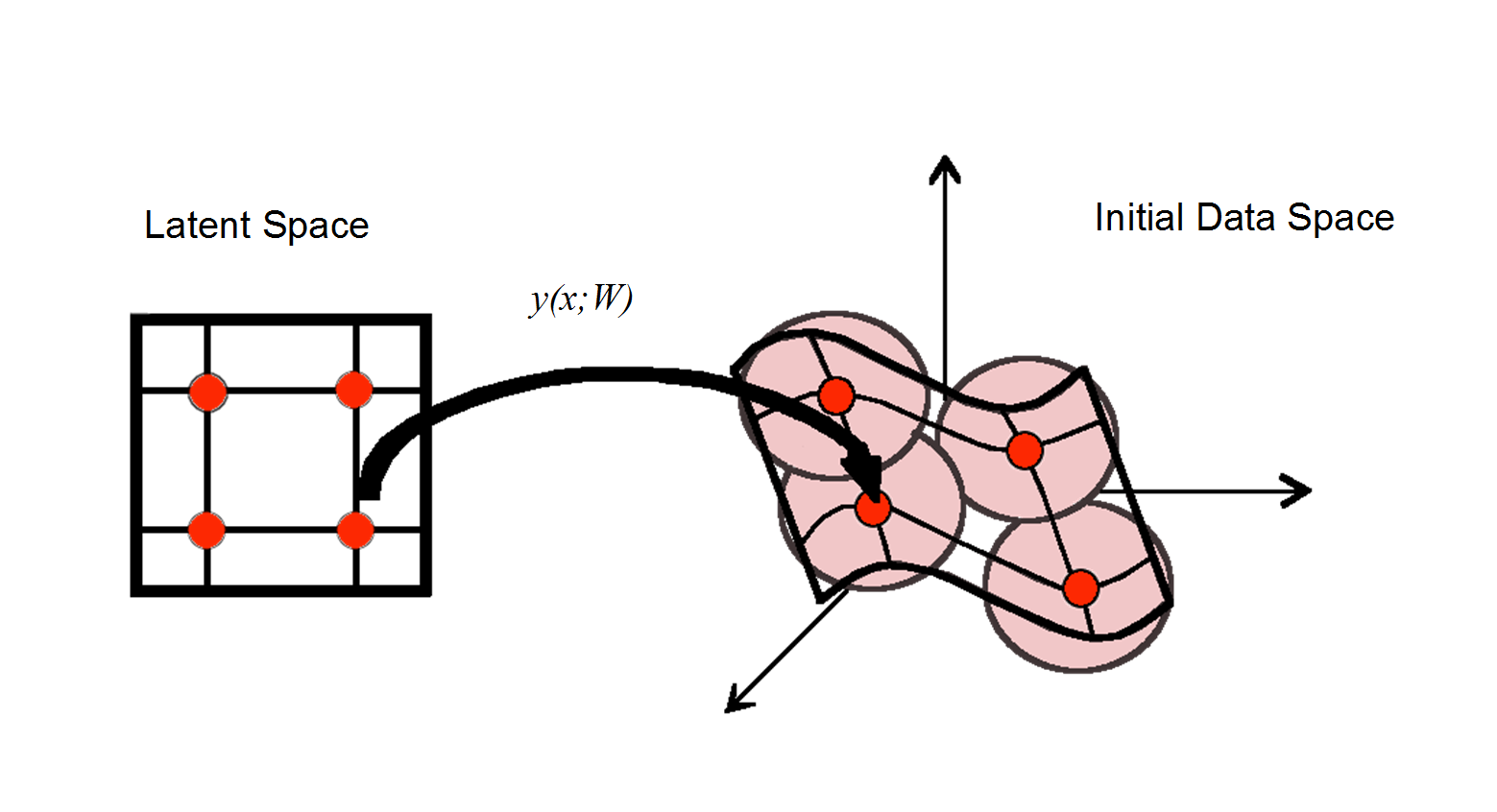
Generative Topographic Mapping

Generative Topographic Mapping (GTM), introduced by Bishop et al is a dimensionality reduction technique which transforms the initial, multi-dimensional dataspace into 2D dimensional latent space (the “map”) by fitting a 2-dimensional non-linear manifold into the data space (Figure 1).



**Figure 1.** Dimensionality reduction concept. Each node *xk* in the latent space (red point on the grid) is mapped to the corresponding manifold point *yk*in the initial data space by the non-linear mapping function *y (x;W)*

The GTM algorithm starts with generation of 2D latent space in the form of a square matrix containing *k* number of nodes. Each node is mapped to a manifold point *yk* embedded in the D-dimensional data space using the non-linear mapping function *y (x;W)*. The manifold points (*yk*) are the centers of normal probability distributions (NPDs) of *t (eq. 1)*.

|  |  |
| --- | --- |
|  | (1) |

where *tn* is a data instance and β is the common inverse variance of these distributions. The ensemble of *N* data instances (in cheminformatics, *N* molecules) spans the relevant zone of the problem space to be mapped. Molecules are represented by their molecular descriptor vectors *tn*, (1…N), which define a “frame” within which the map is positioned, and will therefore be termed “the frame set”.

In Kohonen maps a compound is unambiguously assigned to a node, making compounds within a node indistinguishable. On the contrary, in GTM for every compound projected on the manifold there is a certain probability to “reside” in every node of the grid. The responsibility, or posterior probability, that a point *tn* in the data space is generated from the *k*th node is computed based on current β and **W** using Bayes’ theorem:

|  |  |
| --- | --- |
|  | (2) |

The responsibilities *Rkn* are used to compute the mean (real value) position of a molecule on the map, **s (*tn***) by averaging over all nodes with responsibilities as weighting factors:

|  |  |
| --- | --- |
|  | (3) |

Thus, each point on the GTM corresponds to the average position of one molecule. This step completes the mapping by reducing the responsibility vector to a plain set of 2D coordinates, defining the position of the projection point of the initial *D*-dimensional vector on the map plane. The responsibility vector has the property of being bound to a square grid, a common reference system that may be visually rendered in spite of its still high dimensionality *k*. A molecule characterized by its *rn* vector can be visualized by the pattern of grid nodes that it “highlights”, *i.e*., with respect to which its responsibility values are significant.

GTMs can be used both as classification and as regression tools. Given a training set of *m* molecules assigned, on the basis of experimental input, to different and non-overlapping categories *ci* (typically, actives ϵ *c1*, inactives ϵ *c2*), then the responsibility vector of each molecule can be used to transfer class information onto its associated nodes Intuitively, if the class assignment is visualized as a color, then each molecule will “transfer” some of its color to the nodes, proportionally to responsibilities. Transferred colors accumulate in the nodes, eventually defining nodes where one specific color dominates over the others (Figure 2).



**Figure 2.** Example of GTM latent space classification model with applicability domain (AD) AchE inhibitors (red) and decoys (blue). Lighter regions have a lower probability of association to the winning class P (xk|cbest) and may therefore be discarded from the applicability domain of the model. The points on the map represent individual compounds colored by class

Note, again, that training molecules *m* are not necessarily related to the frame set compounds *n* that served to generate the manifold equations: frame and training sets may be identical, partially overlapping or even fully disjoined. Mathematically, the (normalized) amount of color on each node represents the probability of association of the node *k* to class *ci*:

|  |  |
| --- | --- |
|  | (5) |

where is computed as follows:

|  |  |
| --- | --- |
|  | (6) |

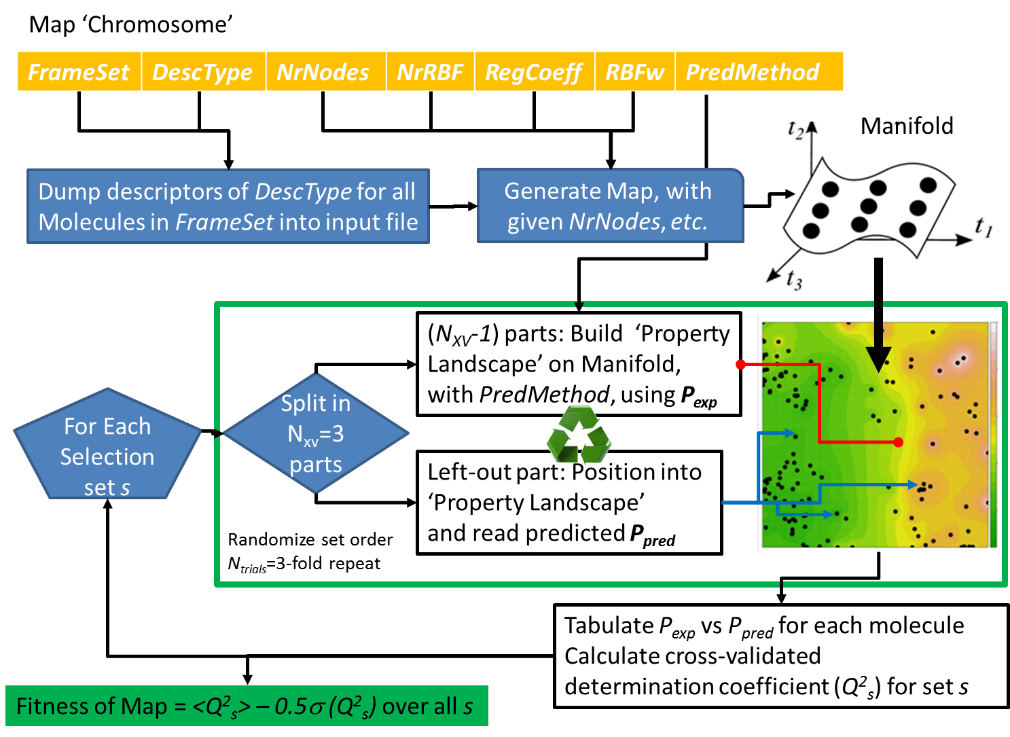
whereresponsibility of node *k* for a molecule belonging to class *ci*, *ni* enumerates training set compounds belonging to class *ci*, *Nci* is the number of training set compounds belonging to class *ci*, and P (*ci*) = , represents the prior probability of class *i, i.e*., the fraction of class members within the training set.

If P (*c1|xk*) > P (*c2|xk*), node *k* will be formally assigned to class 1, and visually rendered in the associated color (Figure 12), with an intensity modulated by P (*xk|ci*). This allows checking whether the local dominance of class 1 corresponds, indeed, to a significant local accumulation of members of that class, or whether the prevalence is the result of unreliable extrapolations of distribution tails to nodes far off the actual regions of interest.

Now, “colored” nodes represent a repository of the knowledge extracted from the training set compounds, and can be subsequently used for predictions, by transferring the acquired “color” back to query compounds *q* to be classified. As a first step, a query compound *q* defined by its descriptor vector tq will be located on the GTM, *i.e.*, associated to responsibilities {Rkq}, and optionally mapped to its 2D residence point s. In this study, the so-called local method was chosen for definition of projected compounds class. The local methodbased on the 2D representation only uses the conditional probability of the node closest to the molecule in 2D, :

Using some continuous property value instead of class labels, GTM manifolds can be “colored” according to a very similar formalism, in order to create property landscapes. In the process, GTM nodes acquire property values equal to the responsibility-weighted average of the properties of (partially) residing training items. Herewith, the nodes become “repositories” of the structure-activity knowledge present in the training set. Projection of a new item on this landscape allows the prediction of its property, by reversely taking the responsibility-weighted averages of the node(s) in which it is found to reside.

Studies dedicated to GTM modeling highlighted the fundamental distinction between actual unsupervised map (manifold) construction, based on a frame set, and subsequent (supervised) learning or “coloring” of this map, based on a potentially different training set. Some options or parameters only concern the unsupervised manifold fitting step and include the four GTM setup parameters: the grid size *k¸* the number of RBFs *M,* the RBF width factor (*w*). and the weight regularization coefficient (l), in addition to the frame set choice, which can be formally regarded as an additional degree of freedom. Eventually, one meta-parameter of paramount importance affects both manifold construction and learning process: the choice of the initial descriptor space, the primary conveyor of numerically encoded structural information. All these parameters have an impact on the quality of the final predictive model supported by the manifold.If a map is designed to describe the chemical space of compounds possessing a certain property, map quality must be evaluated by its classification capacity**.** Thus, an evolutionary algorithm needed to choose the best among models based on the same frame set but different parameters and descriptors can be used. Choices of parameters and descriptors can be synthetically represented as a “chromosome”, with loci dedicated to each mentioned degree of freedom. Some loci represent categorical variables, denominating the choice of frame set, descriptor type or prediction method; some are integers (size, RBF number), and others are real numbers. Evolutionary computing readily supports browsing such heterogeneous search spaces, which makes it a method of choice for the quest of optimally tuned GTM models. The chromosome (“genotype”) unambiguously encodes the “recipe” to build a GTM model (the associated “phenotype”). This phenotype is defined by the ability to “survive” in the competitive environment of a fixed-size chromosome population (under steady evolution through crossover and mutation events involving current members), *e.g.*, its “fitness” score. The nature of this fitness score has already been hinted at: some mean of cross-validated predictive power scores, over selection sets. This might be refined by introducing a penalty related to the spread (standard deviation) of individual scores per set: at equal mean predictive power, the map performing roughly equally well for each selection model is to be preferred to a map doing very well on few models but failing for others (Figure 3).



**Figure 3.** Scheme of the detailed process of estimating the fitness score for a multiproperty-competent GTM model operating in regression mode, and employing repeated, randomized leave-1/3-out cross-validation for a robust assessment of individual quality criteria *Q2* for each selection set