Multi-task generative topographic mapping in virtual screening

Arkadii Lin, Dragos Horvath, Gilles Marcou, Bernd Beck, Alexandre Varnek

Journal of Computer-Aided Molecular Design
https://doi.org/10.1007/s10822-019-00188-x

Multi-task generative topographic mapping in virtual screening

Arkadii Lin[1,2] · Dragos Horvath[1] · Gilles Marcou[1] · Bernd Beck[2] · Alexandre Varnek[1]

Received: 15 September 2018 / Accepted: 2 February 2019

Abstract

The previously reported procedure to generate "universal" Generative Topographic Maps (GTMs) of the drug-like chemical space is in practice a multi-task learning process, in which both operational GTM parameters (example: map grid size) and hyperparameters (key example: the molecular descriptor space to be used) are being chosen by an evolutionary process in order to fit/select "universal" GTM manifolds. After selection (a one-time task aimed at optimizing the compromise in terms of neighborhood behavior compliance, over a large pool of various biological targets), for any further use the manifolds are ready to provide "fit-free" predictive models. Using any structure–activity set—irrespectively whether the associated target served at map fitting stage or not—the generation or "coloring" a property landscape enables predicting the property for any external molecule, with zero additional fitable parameters involved. While previous works have signaled the excellent behavior of such models in aggressive three-fold cross-validation assessments of their predictive power, the present work wished to explore their behavior in Virtual Screening (VS), here simulated on hand of external DUD ligand and decoy series that are fully disjoint from the ChEMBL-extracted landscape coloring sets. Beyond the rather robust results of the universal GTM manifolds in this challenge, it could be shown that the descriptor spaces selected by the evolutionary multi-task learner were intrinsically able to serve as an excellent support for many other VS procedures, starting from parameter-free similarity searching, to local (target-specific) GTM models, to parameter-rich, nonlinear Random Forest and Neural Network approaches.

Keywords Generative topographic mapping · Multi-task learning · Ligand-based virtual screening · Big data · Universal maps · ChEMBL · DUD · Neural networks

Abbreviations

GTM	Generative topographic mapping
UGTM	Universal generative topographic mapping
GA	Genetic algorithm
CV	Cross-validation
DUD	Directory of Useful Decoys
NN	Neural network
RF	Random forest

Introduction

Generative Topographic Mapping (GTM) [1] is a dimensionality reduction method corresponding to a probabilistic extension of Self-Organizing Maps (SOM) [2]. In order to project the data onto a 2D latent space, the method injects a 2D hyperplane, called manifold, into the descriptor space, in which each item of the "Frame Set" (FS) spanning this space corresponds to a point defined by its high-dimensional descriptor vector. The manifold is mathematically described by a square grid of reference points (nodes) and a set of Radial Basis Functions (RBF, Gaussian functions). The FS items serve to "bend" the manifold in order to make it visit a maximum of their descriptor space positions. Using a gradient descent, the method tries to fit positions of the RBF centers, in order to maximize Gaussian function levels at all the FS data points. In other words, it tries to fit the data maximizing a LogLikelihood (LLh) value, which is a logarithm of a cumulated probability of a compound to be related to each node of the manifold [3]. When the manifold

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10822-019-00188-x) contains supplementary material, which is available to authorized users.

✉ Alexandre Varnek
varnek@unistra.fr

[1] Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, 4, Blaise Pascal Str., 67081 Strasbourg, France

[2] Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorferstrasse 65, 88397 Biberach an der Riss, Germany

Journal : Large 10822	Article No : 188	Pages : 13	MS Code : JCAM-D-18-00287	Dispatch : 7-2-2019

is built, each compound is characterized by its LLh value and is described by the vector of its probabilities to "reside" in each node. This vector, R_{nk}, representing the probability of compound k to reside in node n is called the responsibility vector. Since any compound is certain to reside somewhere on the map, $\sum_n R_{nk} = 1, \forall k$. A library of several compounds can be described by the vector of cumulated responsibilities CR of its members k, $CR_n = \sum_k R_{nk}$. Given compounds of known property or bioactivity values, an activity/property Landscape can be created and visualized. This is useful not only for data visualization and analysis but also as a QSAR/QSPR model. After projecting a new compound on it, the class/property value can be easily predicted from the landscape.

Initially, GTM was tested as a tool for Quantitative Structure–Activity Relation (QSAR) tasks on typical structure–property sets [4, 5], where the known actives and inactives of the set were used both as FS and as property set for coloring of the herewith fitted manifold. From this perspective, the initial descriptor space yielding the top predictive manifold could be freely tuned, together with the manifold parameters (number of nodes, number of Gaussians, Gaussian width and Regularization term). The resulting GTM thus represents a predictive model fully dedicated to a specific QSPR problem, and exclusively trained on specific QSPR data. It is the results of a typical single-task learning process, like many other in Ligand-Based Virtual Screening: Decision Trees, Artificial Neural Networks (ANN), Support Vector Machine, Similarity search on binary fingerprints, etc. [6, 7] In addition to this list, SOM method was also tried as a VS technique in many studies [8–10]. For instance, it was used to identify several purinergic receptor agonists [10]. Later, SOM was compared with a Similarity search with data fusion, and, despite a poor predictive performance, the results of such comparison show that in principle SOM can be used as a tool for the VS tasks [8].

However, GTM was also tested successfully as a tool for large public chemical database (PubChem-17, ChEMBL-17 and FDB-17) visualization and analysis [3]. In 2015, Sidorov et al. [11] used GTM in order to create a compound set-independent "universal" map of Chemical Space (CS). The manifold and its underlying descriptor space were not selected with respect to any peculiar property but were aimed at representing the best possible consensus, ensuring a broad "polypharmacological competence", i.e. ability to host predictive property landscapes for a maximum of diverse properties. Conceptually, this is a form of Multi-Task Learning (MTL): based on a generic FS randomly picked to cover the entire ChEMBL CS, structure–activity data from about 100 unrelated target-specific series of ligands of known $\mathrm{pK_i}$ values were used to challenge each manifold in terms of its ability to "host" predictive activity landscapes for each of these series. Selection with respect to the mean predictive performance over all series produced not an optimal manifold dedicated to a given QSPR problem, but a best-compromise manifold of optimal robustness and ability to host any arbitrary property landscape, all while maintaining a certain predictivity level. This ability was eventually validated in showing that it can easily distinguish active from inactive compounds for more than 400 ChEMBL targets (others than the ~ 100 used for selection). Results report an averaged Balanced Accuracy (BA) higher than 0.6 for all the targets (none of which served for map parameter selection).

The above approach is thus related to MTL [12, 13], consisting in learning the choices (descriptors, GTM grid size, etc.) leading to a "consensual" manifold, i.e. learning the choices that are generally relevant to QSPR in drug design, all targets confounded.

MTL is a wide-spread strategy in chemoinformatics and is embodied by numerous distinct strategies, from the use of calculated properties by a previously fitted model as input descriptor to a higher-order model (feature nets [14], FN), to multiple-output multilayered ANNs [13] to strategies in which both ligands and targets are descriptor-encoded (computational chemogenomics [15–19]). Conceptually, the "universal" map approach is different from all the above and is closest related to the multiple-output multilayered ANNs. Manifold building conceptually matches the fitting of parameters of the common layers of the ANN, crystallizing the knowledge of the common features that are important to all the learning tasks. Landscape creation by coloring with specific data sets, followed by prediction, matches the task-specific output neurons of the ANNs—with the notable difference that the latter may still be fine-tuned to improve task-specific predictability. By contrast, at given manifold, coloring of a landscape by projection of a property set and thereupon-based prediction is deterministic and parameter-free. Thus, there is no perfect analogy between the "universal" GTM style of MTL and above-mentioned classical MTL methods. Unlike chemogenomics approaches, "universal" manifolds do not require at all any injection of information about the considered targets, which can be of arbitrary diversity. While chemogenomics focusses on groups of related activities (i.e. for biologically related targets) "universal" manifolds were successfully hosting landscapes for completely unrelated chemical and biological properties, ranging from target-specific activities to cell- or organism-based screen results. Learning features that are "universally" important in structure–activity relationships ensures, on one hand, the generality of "universal" GTMs. On the other, generality will unsurprisingly result in lesser predictive propensity for some targets, as the inductive transfer of knowledge operating at manifold construction step basically resumes to a generic ability to span drug-relevant CS.

So far, no comparison of GTMs and—in particular—of Universal GTMs (UGTM) to other VS methods was undertaken. In order to evaluate the quantitative benefits of building "universal" manifolds, their performance in VS was compared to—firstly—single-task "local" GTMs, dedicated to each biological properties, and also to state-of-art single-task machine learning methods, namely Similarity search and Similarity search with data fusion, Neural Networks (NN), and Random Forest (RF).

Methods

Data

For this project two public databases are used: ChEMBL (version 23) [20] and Directory of Useful Decoys (DUD) [21]. To extract the data, the previously described [11] target-specific structure–activity series extraction protocol has been reenacted on the later release 23 of the ChEMBL database. A total of 618 human single proteins were retained, after "categorization" of ChEMBL-reported activity scores into "actives" and "inactives", respectively. To this purpose, a set of activity classification rules embodied in scripts (available in Supplementary Material of the cited paper) were applied. Compounds with reported percentage of inhibition were considered inactive if values were below 50%, otherwise they were ignored. If dose–response activity measures were available, various cutoffs ranging from low nanomolar to micromolar range were tried out. Compounds better than the threshold were labeled "active" (a minimum of 15 required), the ones of activity weaker that the ten-fold threshold value were "inactives" (at minimum 50), with in-between molecules being ignored (in order to facilitate the separation problem). The actual target-specific cutoff eventually retained was the one ensuring a reasonable balance, closest to one active (or more) for four inactives (but never exceeding parity one active: one inactive—series having, at all considered cutoffs, more reported actives than inactives were discarded). Files (labeled Target-ChEMBLID.smi_ID_class) reporting, for each target, the standardized SMILES string, compound ChEMBL ID and assigned class are now provided as Supplementary Material for the nine targets of the VS simulation, together with their corresponding DUD files. Equivalent data for the remaining 609 targets used in internal validation are available upon request.

Next, DUD data were used to extract independent, external compound series, by focusing on the subset of ChEMBL targets that are also present in DUD and pruning all DUD compounds already encountered in the ChEMBL series. This often meant elimination of virtually all the actives from the DUD series, thus failure to obtain an external data set. However, in nine cases (Table 1) the DUD target-specific series contained sufficiently numerous original actives and were retained for external validation of ChEMBL-trained models (Table 2).

Table 1 A list of nine DUD targets taken for the external validation

Target ID	Target name
CHEMBL1827	Phosphodiesterase 5A
CHEMBL1952	Thymidylate synthase
CHEMBL251	Adenosine A2a receptor
CHEMBL260	MAP kinase p38 alpha
CHEMBL279	Vascular endothelial growth factor receptor 2
CHEMBL301	Cyclin-dependent kinase 2
CHEMBL4282	Serine/threonine-protein kinase AKT
CHEMBL4338	Purine nucleoside phosphorylase
CHEMBL4439	TGF-beta receptor type I

Structure standardization, assignment of activity classes (active vs. inactive) for structures associated to human targets, and rejection of targets with too small or too imbalanced structure–activity series were employed as already described. DUD compounds were likewise standardized, and their given activity class labels (active vs. inactive = decoy) were adopted as such. The set of data passed the data curation procedure contained 1.5M and 914K compounds from ChEMBL and DUD databases, respectively.

Molecular descriptors

One hundred different fragmentation schemes supported by the ISIDA Fragmentor software, [23, 24] and gathered according to the experience of previous works [3, 11] were used as a starting pool for the search of suitable descriptor space. Recall that descriptor space selection is a key meta-parameter of the evolutionary map sampling tool.

Universal (multi-task) GTM manifolds

For technical reasons (the release of a major, faster version of the GTM software), the already published "universal" map selection protocol has been rerun, with another important change with respect to the previously published version; the use of structure–activity class series as selection sets instead of the originally employed (less data-rich) structure-pK_i (continuous) affinity data. Out of the 618 ChEMBL structure–activity series, 236 were randomly designed as selection sets (see file "selection.targets" in the zipped dataset repository in Supplementary Material) for UGTM training (attached "external.targets" enumerates the remaining 382 targets not involved in selection). The FSs were constructed as sets of random ChEMBL samples of different sizes (between 8.5K and 26K compounds). Here, a GA was used to optimize GTM parameters, such as the number of nodes, the number of Gaussian functions (RBF), the regularization coefficient and the width of an RBF. In addition to the best descriptors set and the best GTM parameters, GA also has chosen the most suitable descriptors normalization scheme. At a given GTM parameter set, the manifold training procedure is run in incremental mode [25]. The size of each block was 10,000 compounds. Then, for each selection set, a threefold cross-validation of the current manifold was performed, where landscapes are iteratively built based only on 2/3 of the ChEMBL set, while the remaining tier will be projected into the landscape and ranked by a probability to be active, representing the "color" (relative population of actives vs. inactives) in their target area. For technical details about the rigorous formalism to construct and predict with class and activity landscapes, please refer to our previous GTM publications. According to this selection criterion of mean threefold cross-validated BA of prediction, four best universal maps, each based on a different descriptor space, with the mean BA ranging within 0.7–0.75 have been selected (Table 3). Corresponding GTM parameters and FS sizes are presented in Table 4.

Table 2 The datasets used for the screening procedure

Target ID	DUD data sets			ChEMBL data sets			Thresholds[a]
	Actives	Inactives	Total	Actives	Inactives	Total	K_i/IC/EC_{50} (nM)
CHEMBL1827	170	25,334	25,504	691	824	1515	50
CHEMBL1952	63	6113	6176	124	455	579	1000
CHEMBL251	79	28,001	28,080	1303	3618	4921	100
CHEMBL260	100	32,925	33,025	1453	2567	4020	100
CHEMBL279	94	22,595	22,689	2047	4663	6710	100
CHEMBL301	189	25,675	25,864	638	2305	2943	500
CHEMBL4282	52	14,228	14,280	725	2619	3344	500
CHEMBL4338	102	6334	6436	100	111	211	50
CHEMBL4439	82	8013	8095	282	385	667	50

[a] Compounds with dose–response affinity value below or equal to threshold (in nM) are considered active, while those with values exceeding the 10-fold threshold value are inactives. At intermediate activities, compounds are discarded from the ChEMBL set. Note that the DUD definition of "actives" does not comply to the same rules—they routinely include co-crystallized ligands, irrespective of their affinities

Table 3 The best selected descriptors sets [23]

Map	Abbreviation	Definition	Descriptor set size
1	IA-FF-FC-AP-2-3	Sequences of atoms with a length of two to three atoms labeled by force field type and formal charge flag, using all paths	987
2	IIRAB-FF-1-2	Atom-centered fragments of restricted atom and bonds of a length one to two atoms labeled by force field types	1029
3	IAB-PH-FC-AP-2-4	Sequences of atoms and bonds of a length two to four atoms labeled by pharmacophoric atom types and formal charges using all paths	779
4	IA-2-7	Sequences of atoms of a length two to seven atoms	728

Table 4 Selected GTM metaparameters for the four best chromosomes chosen by the genetic algorithm

Map	FS size	Number of nodes per line	Number of RBF per line	Regularization coefficient	RBF width	Normalization scheme[a]
1	17,000	41	23	1.122	1.1	2
2	17,000	47	29	0.018	1.6	1
3	25,500	37	19	0.017	2.1	2
4	25,500	38	19	3.55	1.9	2

[a] The standardization schemes: 1—centering on the mean value; 2—Z-normalization (centering on the mean value and division by the standard deviation)

Monitored success scores

In this benchmarking study, the mean area under the Receiver Operating Characteristic (ROC AUC) when predicting half of the compound series based on landscapes colored (or models learned, for other methods—*vide infra*) on the other half is used in the internal validation procedure. This further on named <AUC>$_{1/2}$ criterion will be consistently used to compare models (except for single-query similarity searching, where it cannot be defined—see following subsection). The mean is taken over ten independent repeats of the above procedures, where splitting into training and kept-out compounds is fully randomized. No specific care is taken to ensure that each compound is strictly kept out once and only once per iteration.

Internal validation results were alternatively depicted as density distribution plots of the ROC AUC values over the training subsets (Figs. 1, 2, *vide infra*). For each method each ChEMBL target-specific set returns the ten distinct ROC AUC values from the randomized internal validation experiments described in the "Methods" section. Plotting the density (number of targets) in counting each target 10 times, into the specific bins matching each of its ROC AUC values achieved on the random splits (and followed by a normalization of the density to compensate for multiple counts)—would however produce one "global" histogram, with no information on the expected fluctuation of density bar heights. Estimating those error bars is however of paramount importance, in order to ensure that the histogram shape is not an artefact of the peculiar randomized choice of training/test splits. For this specific purpose, this work proceeds to first generate "splitting accident-prone" histograms, considering each target-specific compound set to be represented by one randomly picked ROC AUC out of the 10. Depending on the pick, the set will be counted in a lower or higher bin, i.e. its localization on the X axis will reflect the intrinsic uncertainly induced by the train/test splitting. Every set is counted exactly once—only its X-axis bin may fluctuate. Therefore, every such "splitting accident-prone" histogram will differ in shape. One thousand of these are generated, which allows a thorough monitoring of the expected fluctuation of bar heights as a consequence of splitting artefacts. Eventually, the plot shows the mean bar heights (which converge to the above-mentioned "global" histogram) with associated error bars (if readable—occasionally, fluctuations are too small).

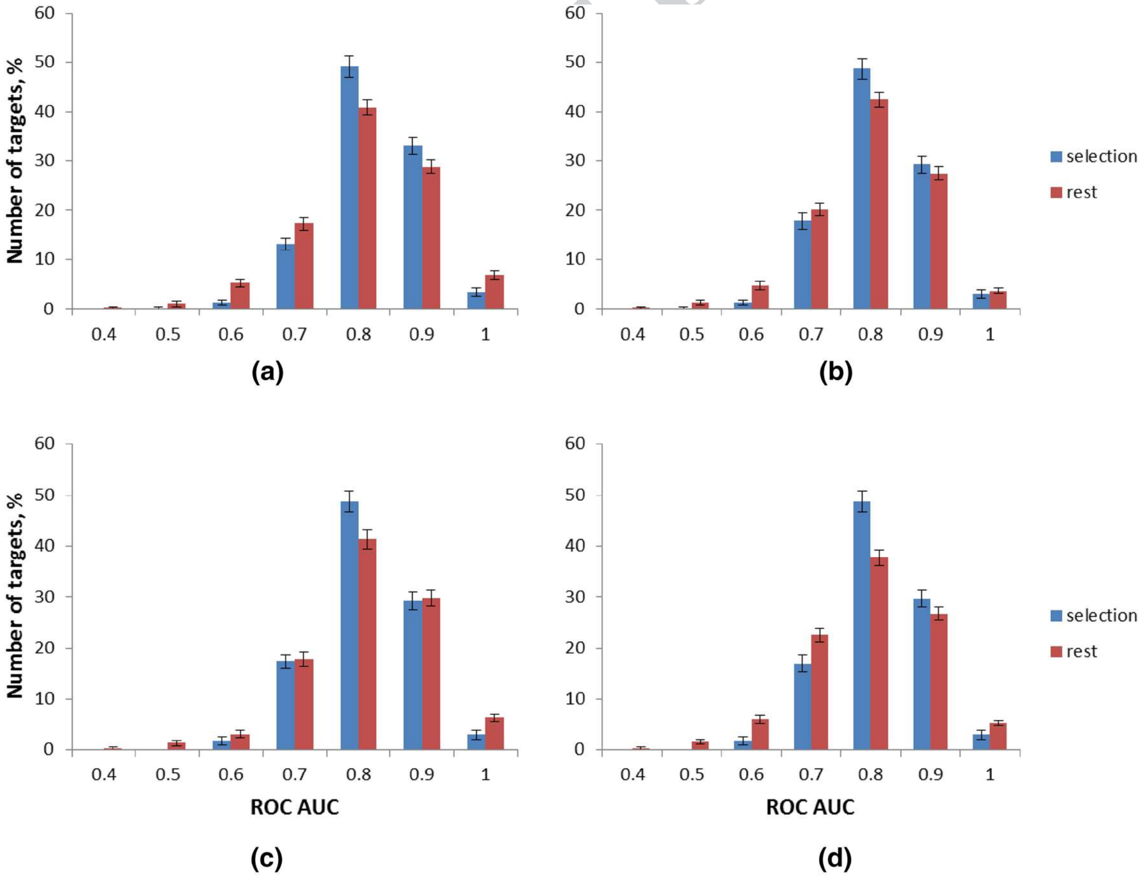

Fig. 1 ROC AUC values for the selection set and rest targets: **a** map 1, **b** map 2, **c** map 3, **d** map 4

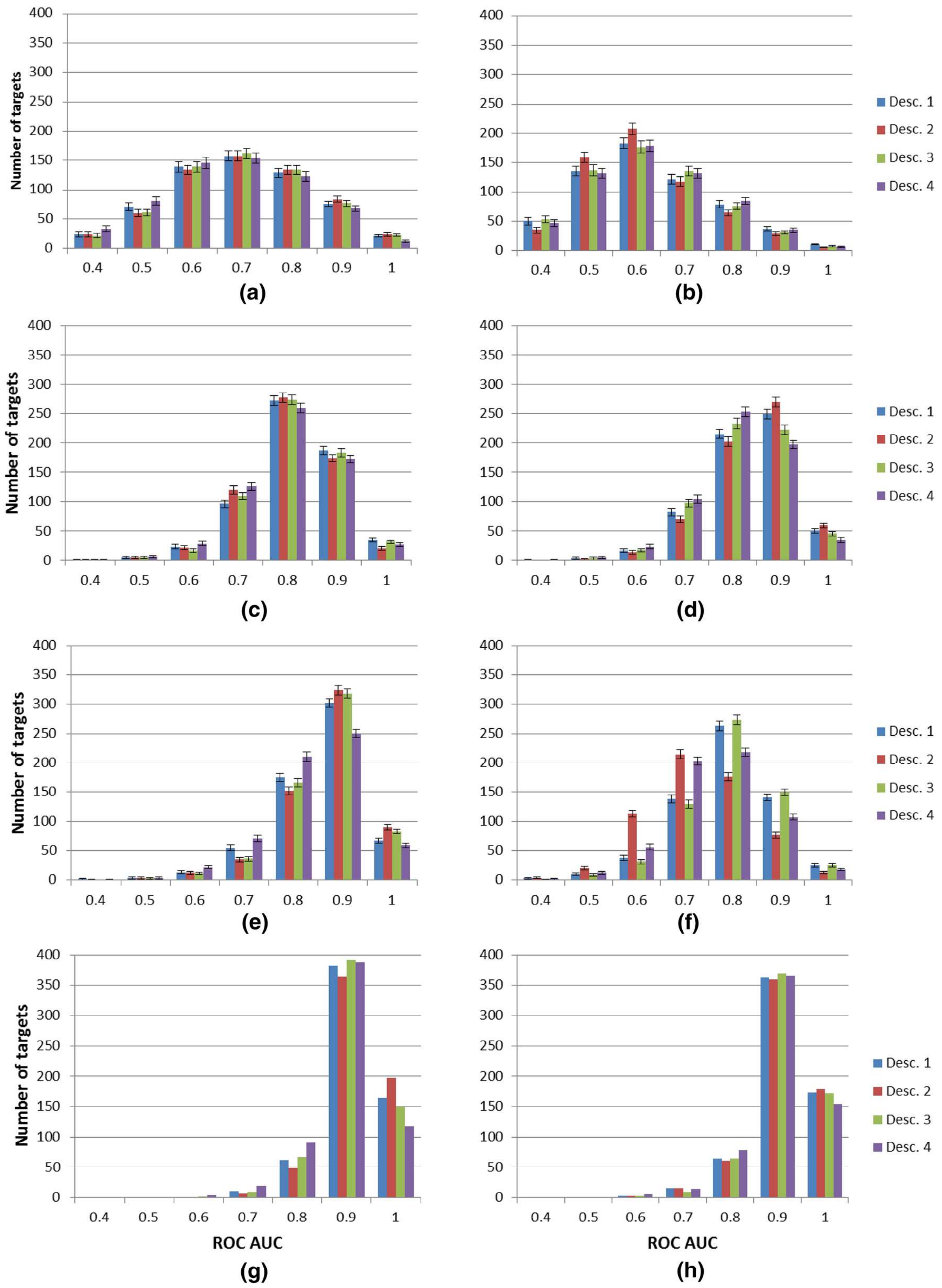

Fig. 2 Internal validation results on 618 ChEMBL targets: single-query Similarity search in **a** descriptors and **b** latent spaces, **c** UGTM, **d** local GTM, Similarity search with data fusion in **e** descriptors and **f** latent spaces, **g** NN, and **h** RF. Here, Desc. 1–4 correspond to the descriptors sets shown in the Table 3

In actual virtual screening, the DUD series is projected onto the "complete" landscape generated from the entire ChEMBL set. To estimate the predictive performance of a particular map, ROC AUC (further on referred to as VSAUC) is computed, after ranking DUD compounds as above-mentioned [26].

Benchmarked models

For each of the 618 targets, single-task (local) models were set up in each out of four descriptors spaces chosen in Table 3 using the following methods:

- Regular (local) GTM
- Similarity search
- Similarity search with data fusion
- RF
- NN

Depending on the nature of the model, setting it up requires distinct protocols, involving parameter selection or fitting (local GTM, PF, NN) or decisions on used similarity scoring, etc. These aspects will be detailed in the dedicated paragraphs below, while the same success score monitoring procedure outlined above was applied to all models. The descriptors normalization scheme was not changed and corresponds to the one that is shown in Table 4.

The parameters of local GTM were not optimized, but were taken by default: the number of nodes is 625 (25×25), the number of Gaussian functions is 144 (12×12), the width of a Gaussian function is 2.82, the regularization coefficient is 1.0. To perform the experiments with NN and RF, SciKit Learn implementations of Multi-Layer Perceptron (MLP) (https://scikit-learn.org/stable/modules/neural_networks_supervised.html) and RandomForestClassifier (https://scikit-learn.org/stable/modules/ensemble.html#forest) were employed [26–29]. Here, the MLP parameters are taken by default: the number of hidden layers is 1, the number of the nodes in a layer is 100, the rectified linear unit function (relu) is used as an activation function [30], and the "adam" solver is used for the weights optimization [31]. Backpropagation approach is applied to train the net [26–28]. In case of RF, an ensemble of trees is built on a random half of compounds where the original ratio actives/inactives is kept. All the parameters are taken by the default, mentioned in SciKit Learn (https://scikit-learn.org/stable/modules/ensemble.html#forest), where the number of trees in a forest is 10.

As a gold standard for the VS tasks, Similarity search and Similarity search with data fusion were chosen. Both these methods are based on a simple similarity principle: similar compounds should share similar activity. Therefore, the idea of similarity searching is to find compounds out of a screening pool which are similar to the reference point with a known label (i.e. active). While there are better suited criteria [32, 33] to specifically monitor neighborhood behavior compliancy, herein the generally applicable ROC AUC criterion is used to score the potential predictive performance of the method, after ranking candidates in decreasing similarity order (Tanimoto scores) to the used query. Also, as an alternative to a simple similarity searching, similarity searching with data fusion is taken. Within this approach the screening pool is compared not to one but to *N* reference compounds (in this project the pool of reference compounds was chosen to embody a randomly picked 50% of all ChEMBL actives available for a target). To rank a candidate, the highest Tanimoto score is taken out of the *N* computed values. As it was done earlier, in order to ensure reproducible results, averaging out the dependence on the randomly picked query compound(s), all similarity-based calculations were repeated 10 times, and the mean ROC AUC was computed for each target. In single-query searches, the $<AUC>_s$ value resulted from 10 individual similarity ranking simulations using 10 randomly picked active queries. With data fusion, 10-fold repeats of searches employing one half of the pool of actives generate the corresponding $<AUC>_{1/2}$ criterion that will be directly compared with equivalent $<AUC>_{1/2}$ criteria of the other VS methods, and the single-search $<AUC>_s$.

Eventually, the DUD pool was screened to obtain a VSAUC score using only the data fusion-based strategy, i.e. ranked according to their Tanimoto score with respect to their nearest neighbor of the entire corresponding ChEMBL series.

In order to measure the impact of dimensionality reduction/information loss by the GTM transformation of initial descriptors into responsibility vectors, similarity searching was performed in both descriptor and GTM responsibility vector spaces.

Results and discussion

Internal validation of the new UGTM versions

For above-cited technical reasons, this article introduces new, refitted "universal" GTM manifolds using a new GTM software release and extended selection sets of 236 (randomly picked) ChEMBL structure–activity class series associated to as many single protein targets. This undertaking is completely independent of the herein presented VS benchmark, as it focuses on the "multi-task" learning of the optimal compromise in terms of neighborhood behavior compliance over a large panel of targets, and even though this by no means a preparation step of the actual VS, UGTM performance analysis must be briefly discussed here. First, it must not be forgotten that, out of the 618 ChEMBL target-specific series exploited by this study, 236 have a special status with respect to UGTMs: they served as selection sets for the optimal UGTM manifolds. This concerns two of the nine targets used in the VS simulation are included here (CHEMBL4439 and CHEMBL1952). By contrast, the remaining 382 external sets (including the other seven VS targets) were never used in UGTM tuning. It is thus legitimate to verify whether these 236 targets are favored—better predicted—by UGTMs, with respect to the latter. Figure 1 reports the distribution of "selection" versus "external" target-specific sets with respect to the internal validation ROC AUC values (see density distributions plots, in the Scoring section of methods). While the histograms show the expectable shift in favor of better results for the selection sets, this trend is very limited. Therefore, in the following analysis, no further distinction between selection and external ChEMBL sets will be done—statistics will indiscriminately refer to the set of 618 target-specific series. Furthermore, this observation is interesting, as it proves that MTL over ~200 structure–activity sets associated to fully non-related biological properties allows to cartograph the drug-relevant CS with a precision that is sufficient to ensure a same level of prediction accuracy for a large number of distinct biologically relevant targets to date.

Last but not least, let it be noted that even for the two targets CHEMBL4439 and CHEMBL1952 which served at map selection stage, the external validation by VS is no less rigorous than for any other of the herein benchmarked models. Any predictive model issued from supervised learning uses target-related information for calibration, and then is challenged to predict an independent compound set—as is the case here (DUD molecules filtered in order to ensure that they do not include any ChEMBL members). For all the nine targets, "coloring" of UGTM manifolds with ChEMBL data is the prerequisite to predict the likelihood to be active for the external DUD compounds—this is the equivalent of aforementioned model "calibration", except that it occurs in a deterministic and non-supervised manner—the manifold being already given. To resume, for two targets the injection of training information into UGTM models implies both manifold fitting and coloring, whilst for the seven others it implies only non-supervised manifold coloring. In either case, external validation concerns independent, never encountered compounds.

Internal validation benchmark

Comparative internal validation results for the various methods in terms of the above-defined $<AUC>_{½}$ ($<AUC>_S$ for single-query similarity screening) are given in Fig. 2. The poorest results come from single-query similarity, which is normal because the quantity of injected knowledge (one active reference) is minimal. Things are even worse after dimensionality reduction: moving to responsibilities decreases performances even more. Nevertheless, with 50% of the mass of known actives used to color GTM fuzzy class landscapes, predictivity increases dramatically over single-query searches, and in spite of moving into the responsibility vector space.

Local maps are, as expected, better than universal maps. To begin with, they are already based on molecular descriptors known—thanks to the MTL of UGTM hyperparameters—to be generally pertinent choices, for a large pool of targets, Even though their control parameters were set to default values (likewise, the parameters of UGTMS being locked to the ones defining the best compromise neighborhood behavior), the degrees of freedom controlling the "bending" of their manifolds are now free to adjust specifically in response to the dedicated structure–activity series. Local maps might presumably be improved even more if their hyperparameters would be optimized.

Yet, similarity with data fusion, which is comparable to the GTM-based approach in terms of input SAR knowledge—50% of the actives—outperforms the former when driven in the original descriptor spaces: projection on a map inexorably costs in terms of information loss.

Eventually, NNs and RFs, are machine-learning approaches featuring a wealth of tunable parameters—unlike the fixed Universal and local GTM manifolds. Therefore, they are clearly the better performers.

In view of virtual screening of the DUD series, the best map for each target has been selected basing on its $<AUC>_{1/2}$ score. The number of targets for which the best map/descriptors space achieves a $<AUC>_{1/2} > 0.8$ have been counted for each method (Fig. 3).

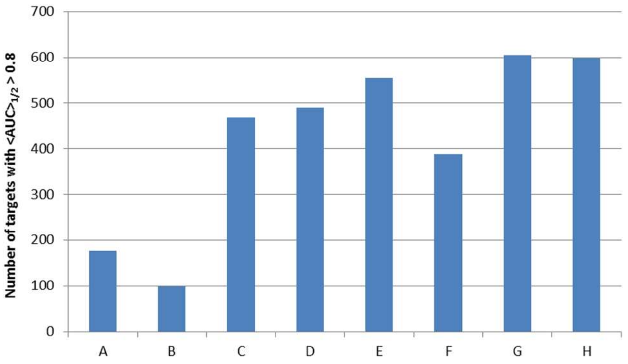

Fig. 3 The number of targets for which the best model over the four descriptor spaces returns $<AUC>_{1/2} > 0.8$. If, for a target, at least one of the four models of given type, based on the four descriptor spaces reaches this threshold, then the target will be added to the type bin: A—similarity search in initial space, B—similarity search in responsibility space, C—UGTM, D—local GTM, E—similarity search with data fusion in initial space, F—similarity search with data fusion in responsibility space, G—NN, H—RF

The bar chart in Fig. 3 keeps the trend seen in Fig. 1 and demonstrates that RF and NN outperform the GTM approach. At the same time, local GTM demonstrates the ability to be used successfully for 490 targets which makes it comparable with Similarity search with data fusion, which successfully handles 555 of the targets.

Virtual screening simulation using DUD compounds

The last part of the project is devoted to the retrieval, by VS, of actives among DUD compounds, with the ChEMBL-data-driven models. As it was described earlier, nine targets were found in common for DUD and ChEMBL (Tables 1, 2), where the smallest series includes more than 6000 compounds from DUD and more than 200 compounds from ChEMBL. The most data-rich target contains more than 33,000 compounds from DUD and more than 6000 compounds from ChEMBL.

Note that the DUD classification into actives and (presumably) inactive decoys is conceptually different from the classifications employed in the training sets. DUD actives may, for example, include co-crystallized ligands of high micromolar to millimolar potency, which are far from qualifying as "actives" by ChEMBL standards. This fact is potentially harmful for the external "prediction" performance monitored here—yet, this class of artefacts generally applies to classification models, which are the last recourse in response to highly heterogenous affinity measures that cannot be directly compared unless they are converted to "classes" according to more or less rigorous criteria. However, relative comparison of method performances should still be possible—if extrapolation from ChEMBL data to the DUD set is successfully accomplished by at least some methods, failure to do so by others cannot be ascribed due to classification artefacts. This is the case in the present work.

To screen the DUD pool, the best maps were chosen based on their mean ROC AUC value obtained in internal-validation (Table 5).

In this VS simulation, the QSAR-based approaches were used, with the hypothesis (colored landscape, learned model) being based on the entire ChEMBL series of the nine above-mentioned targets. Single-query similarity searching was not considered here, as its intrinsic limitations due to the poverty of injected knowledge (a single active) were clear from internal validation results. In addition to ROC AUC, an Enrichment Factor (EF) within the 10% of top ranked compounds was added as a second criterion to estimate the quality of the predictions. The results of the external validation are shown in the Figs. 4 and 5.

Here, the predictive performance for the UGTM approach varies within 0.55 ÷ 0.9 in terms of ROC AUC and within 0.2–6.2 in terms of the EF. Local GTMs show much better performance (ROC AUC ranges within 0.75–0.9, EF ranges within 2.2–8.2). While NNs were on par with RF and outperformed GTM models in terms of internal validation results, it appears that they are no longer systematically among top performers in VS, where similarity searching, RF and local GTM models are often much more robust. The activity landscapes and the DUD projections done for the target CHEMBL4282 and presented in Fig. 6 show that most of the DUD compounds are within the occupied zones (in other words, within the GTM applicability domain).

It is also seen from the DUD and ChEMBL activity landscapes that active DUD compounds are projected onto active zones of ChEMBL, which makes the ROC AUC and EF very high.

Discussion

The construction procedure of "universal" maps supporting multiple predictive landscapes on a same GTM manifold is a novel strategy in MTL. It is atypical in several aspects:

- First, it includes both operational parameters of the GTM model and hyperparameters. The key hyperparameter here is the choice of the molecular descriptor space,

Table 5 ROC AUC values and corresponding descriptors space for the best models computed within the internal validation

Target ID	UGTM	Local GTM	Similarity search in initial space	Similarity search in latent space	NN	RF
CHEMBL1827	0.89/4[a]	0.88/2	0.92/2	0.82/4	0.97/1	0.97/1
CHEMBL1952	0.88/4	0.84/4	0.85/4	0.76/4	0.92/1	0.92/3
CHEMBL251	0.84/3	0.84/2	0.91/2	0.81/3	0.95/2	0.96/3
CHEMBL260	0.76/2	0.77/2	0.9/3	0.81/3	0.95/3	0.95/1
CHEMBL279	0.74/2	0.71/3	0.89/3	0.76/3	0.93/3	0.93/4
CHEMBL301	0.82/4	0.83/4	0.91/2	0.8/3	0.94/2	0.95/3
CHEMBL4282	0.83/3	0.88/2	0.94/2	0.83/3	0.96/2	0.96/2
CHEMBL4338	0.83/1	0.86/3	0.85/3	0.78/3	0.94/2	0.93/2
CHEMBL4439	0.88/2	0.9/2	0.89/2	0.87/3	0.94/2	0.94/3

[a] Mean ROC AUC/No. of a map/descriptors space corresponded to Table 3

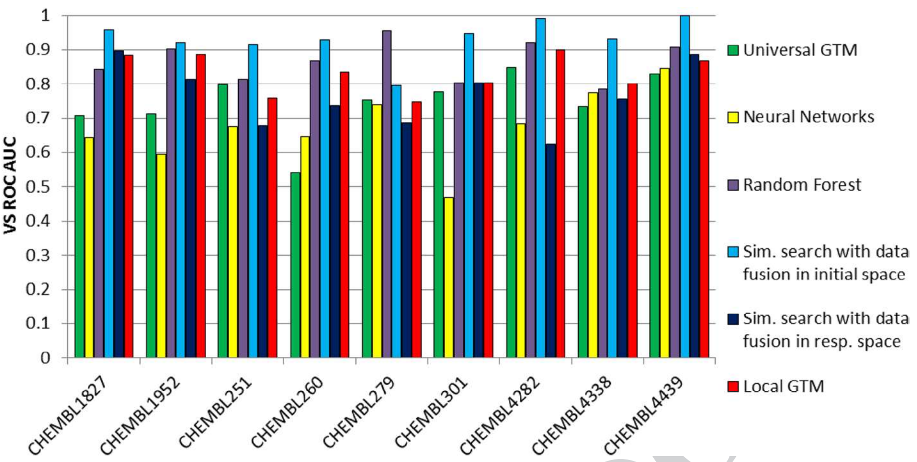

Fig. 4 The comparison of the VS methods, where each column corresponds to the best map in terms of its ROC AUC value computed in the internal validation (see Table 5)

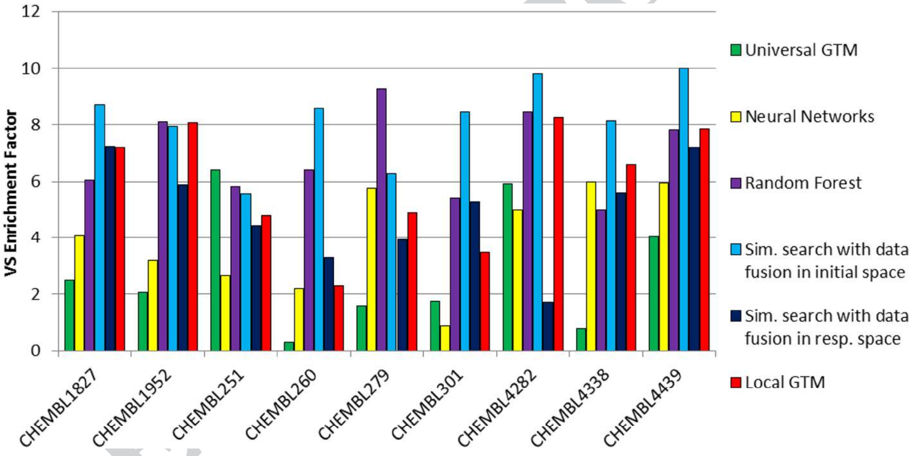

Fig. 5 The EF for different VS approaches where the EF value is given for the map with the highest ROC AUC value computed in the internal validation (see Table 5)

allowing the procedure to select those descriptor spaces which remain neighborhood behavior-compliant after GTM-driven dimensionality reduction

- Second, its multi-task nature is given by the construction of a common manifold, which is, per se, an unsupervised learning process aimed at maximizing the coverage of FS compounds by this manifold. This common manifold is challenged to host fuzzy classification landscapes for many different biological targets. Each of them is a classical single-task model for the property associated to the ligands that were used to color the specific landscape. However, since these landscape-based predictive models do not feature any specific fitable parameters, their quality can be regarded as an intrinsic property of the underlying common manifold. Creation of the manifold implicitly provides access to as many landscape-driven predictive models as available property-annotated ligand series. The MTL—here primarily consisting in selecting optimally suited descriptor spaces and optimally associated GTM grid size, manifold flexibility parameters, etc.—was directed by the goal of discovering (hyper) parameter combinations maximizing the mean quality of 236 distinct "selection" series of target-specific activity-annotated ligands
- Third, it does not focus on specific transfer of knowledge within biologically related targets, such as is the case in computational chemogenomics. This MTL simultaneously addressed the rather exhaustive set of all human protein targets with sufficient activity annotations in ChEMBL, all protein families confounded. Neither the 236 "selection" series of target-specific activity-annotated ligands, nor the remaining 382 series used for external validation (with comparable success rate to the former 236) include any intended family-specific bias in terms of biological targets. Here, MTL would not target typical questions like "What are the common features of kinase binders?", but more general "What are the common features of bioactive molecules, all targets confounded?"

Uncovering the few ISIDA fragmentation schemes that are optimally suited for this endeavor is a first key result of this atypical multitask learning setup. Since descriptor spaces cannot host predictive GTM models unless they are,

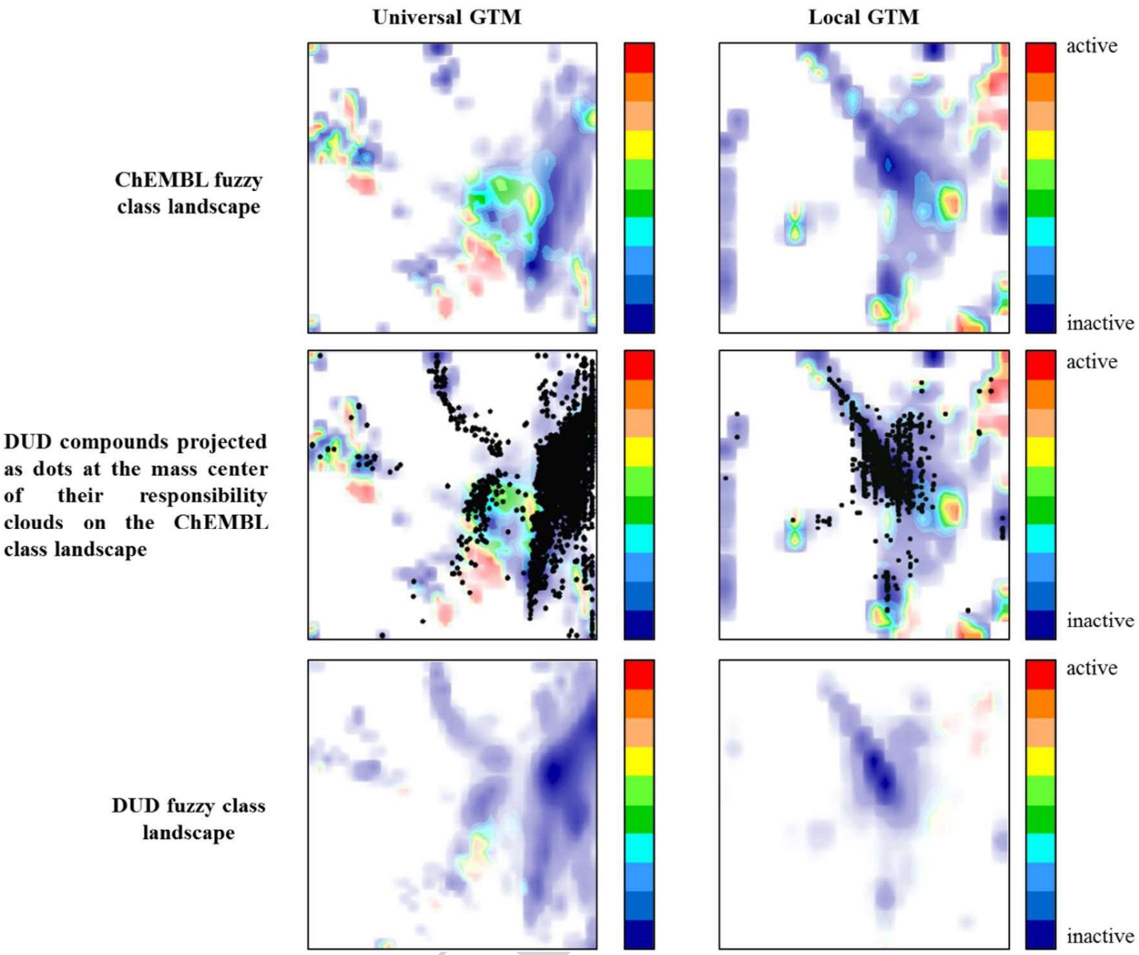

Fig. 6 Fuzzy class landscape representations of the (ChEMBL and respectively DUD) sets associated to target CHEMBL4282 on universal map 3 (left) versus the local GTM (right)

per se, neighborhood behavior-compliant, it is unsurprising to observe that all the alternative approaches—from data-fusion-driven similarity searching to target-dedicated local GTM, RF and NN models—were rather successful, both in terms of internal validation and external VS. There was no need to rescan, for each predictive method, the entire set of available molecular descriptor spaces—the choices of the evolutionary UGTM builder were appropriate. Note that the 100 different descriptor spaces out of which the four herein used were selected have themselves emerged as a historical accumulation of descriptor spaces that were used in the past [3, 11], on rather unrelated problems such as library comparison, and were seen to be successful. In this sense, if we declare all the cases in which knowledge from previous experiences is actively used to restrain the scope of effectively considered working hypotheses as some form of "multi task" learning, then MTL is rather the rule than the exception in chemoinformatics.

UGTM models are remarkably robust in VS—for models with zero adjustable parameters, albeit they are systematically outperformed—in particular with respect to enrichment of the top selection—by the equally parameter-free data-fusion similarity searching, not affected by information loss upon dimensionality reduction. However, UGTM models are specifically failing to rank a significant number of actives among the top 100 candidates—they are not effective in ensuring high EF values in VS. By contrast, their global ROC AUC scores show that they do, overall, manage to eventually rank actives ahead of most of the inactives, only slightly less effective than the other methods—without systematically placing actives at the top of the list.

Responsibility vectors are still maintaining some degree of neighborhood behavior-compliance, but their use in similarity searching is not recommended, as landscape-driven prediction on UGTM manifolds is the more powerful method. Note that data fusion-based similarity screening with Q actives being used as queries would scale like QxN in terms of computational effort required to virtually screen a database on N candidates. By contrast, landscape-based prediction effort is simply proportional to N and does not depend on the training set size used to create the predictive landscape. Thus, the latter would become computationally more interesting after a given Q value—not to mention all the benefits stemming from intuitive visualization provided by the GTM approach.

Conclusions

The previously reported strategy to generate "universal" maps, able to support predictive models for a broad spectrum of biological activities represents a generic MTL approach, where optimal molecular descriptors are selected alongside with optimal operational parameters of the GTM algorithm. A first important outcome of the approach is uncovering "multicompetent" molecular descriptor spaces that remain neighborhood behavior-compliant even after the dimensionality reduction process—leading to GTM responsibility vectors and ultimately to a (x, y) point in 2D GTM latent space. These tend to correspond to ISIDA fragmentation schemes restricted to rather small fragment sizes but incorporating information-rich atom labels such as pH-dependent pharmacophore types or CVFF force field types.

It could be shown that descriptors herewith selected are not only an excellent support for GTMs, but also for many other predictive models—starting with plain similarity screening. In this sense, all models here implicitly benefitted from the initial MTL, which provided a pool of four descriptor spaces that turned out to be highly relevant for all the envisaged QSAR model building procedures for more than 600 completely independent targets.

Tanimoto-score-based similarity screening (using a data fusion scenario, thus ensuring that the amount of information injected into it—active examples—matches the sizes of the training sets used by other approaches) is actually more successful than UGTM-driven predictions, as information loss upon dimensionality reduction is unavoidable.

Local GTMs, where manifolds are allowed to focus on the chemical subspace populated by a single target-specific ligand series, are unsurprisingly better performers than their universal, consensus-oriented counterparts. Note, however, that the latter would always represent a better choice whenever the activity-annotated data set pertaining to a target of interest is not sufficient to support the fitting of local maps. The same holds true for parameter-rich non-linear RF and NN models.

Author contributions The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding The project leading to this article has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant agreement No 676434, "Big Data in Chemistry" ("BIGCHEM", http://bigchem.eu).

References

1. Bishop CM, Svensén M, Williams CK (1998) GTM: the generative topographic mapping. Neural Comput 10(1):215–234
2. Kohonen T (1990) The self-organizing map. Proc IEEE 78(9):1464–1480
3. Lin A, Horvath D, Afonina V, Marcou G, Jean-Louis R, Varnek A (2018) Mapping of the available chemical space versus the chemical universe of lead-like compounds. ChemMedChem 13:540–554. https://doi.org/10.1002/cmdc.201700561
4. Kireeva N, Baskin I, Gaspar H, Horvath D, Marcou G, Varnek A (2012) Generative topographic mapping (GTM): universal tool for data visualization, structure—activity modeling and dataset comparison. Mol Inform 31(3–4):301–312
5. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A (2015) GTM-based QSAR models and their applicability domains. Mol Inform 34(6–7):348–356. https://doi.org/10.1002/minf.201400153
6. Muegge I, Oloff S (2006) Advances in virtual screening. Drug Discov Today 3(4):405–411. https://doi.org/10.1016/j.ddtec.2006.12.002
7. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today 20(3):318–331. https://doi.org/10.1016/j.drudis.2014.10.012
8. Hristozov D, Oprea TI, Gasteiger J (2007) Ligand-based virtual screening by novelty detection with self-organizing maps. J Chem Inf Model 47(6):2044–2062. https://doi.org/10.1021/ci700040r
9. Kaiser D, Terfloth L, Kopp S, Schulz J, de Laet R, Chiba P, Ecker GF, Gasteiger J (2007) Self-organizing maps for identification of new inhibitors of P-glycoprotein. J Med Chem 50(7):1698–1702. https://doi.org/10.1021/jm060604z
10. Schneider G, Nettekoven M (2003) Ligand-based combinatorial design of selective purinergic receptor (A2A) antagonists using self-organizing maps. J Comb Chem 5(3):233–237
11. Sidorov P, Gaspar H, Marcou G, Varnek A, Horvath D (2015) Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. J Comput Aided Mol Des 29(12):1087–1108. https://doi.org/10.1007/s10822-015-9882-z
12. Rosenbaum L, Dörr A, Bauer MR, Boeckler FM, Zell A (2013) Inferring multi-target QSAR models with taxonomy-based multi-task learning. J Cheminform 5(1):33
13. Varnek A, Gaudin C, Marcou G, Baskin I, Pandey AK, Tetko IV (2009) Inductive transfer of knowledge: application of multi-task learning and feature net approaches to model tissue-air partition coefficients. J Chem Inf Model 49(1):133–144. https://doi.org/10.1021/ci8002914
14. Xu Y, Ma J, Liaw A, Sheridan RP, Svetnik V (2017) Demystifying multitask deep neural networks for quantitative structure–activity relationships. J Chem Inf Model 57(10):2490–2504
15. Brown JB, Okuno Y, Marcou G, Varnek A, Horvath D (2014) Computational chemogenomics: is it more than inductive transfer? J Comput Aided Mol Des 28(6):597–618. https://doi.org/10.1007/s10822-014-9743-1
16. Heikamp K, Bajorath J (2013) Prediction of compounds with closely related activity profiles using weighted support vector machine linear combinations. J Chem Inf Model 53(4):791–801. https://doi.org/10.1021/ci400090t
17. Medina-Franco JL, Giulianotti MA, Welmaker GS, Houghten RA (2013) Shifting from the single to the multitarget paradigm in drug discovery. Drug Discovery Today 18(9–10):495–501. https://doi.org/10.1016/j.drudis.2013.01.008

18. Bieler M, Heilker R, Koeppen H, Schneider G (2011) Assay related target similarity (ARTS)—chemogenomics approach for quantitative comparison of biological targets. J Chem Inf Model 51(8):1897–1905. https://doi.org/10.1021/ci200105t
19. Jacob L, Hoffmann B, Stoven V, Vert J-P (2008) Virtual screening of GPCRs: an in silico chemogenomics approach. BMC Bioinform 9(1):363
20. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B (2011) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40(D1):D1100–D1107
21. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular Docking. J Med Chem 49(23):6789–6801. https://doi.org/10.1021/jm0608356
22. Horvath D, Brown J, Marcou G, Varnek A (2014) An evolutionary optimizer of libsvm models. Challenges 5(2):450–472
23. Ruggiu F, Marcou G, Varnek A, Horvath D (2010) ISIDA property-labelled fragment descriptors. Mol Inform 29(12):855–868. https://doi.org/10.1002/minf.201000099
24. Ruggiu F, Marcou G, Solov'ev V, Horvath D, Varnek A (2017) ISIDA fragmentor 2017-user manual
25. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A (2015) Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. J Chem Inf Model 55(1):84–94. https://doi.org/10.1021/ci500575y
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12(Oct):2825–2830
27. Ruck DW, Rogers SK, Kabrisky M, Oxley ME, Suter BW (1990) The multilayer perceptron as an approximation to a Bayes optimal discriminant function. IEEE Trans Neural Netw 1(4):296–298. https://doi.org/10.1109/72.80266
28. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533
29. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
30. Dahl GE, Sainath TN, Hinton GE (2013) Improving deep neural networks for LVCSR using rectified linear units and dropout. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), 2013, IEEE, Vancouver, pp 8609–8613
31. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980
32. Horvath D, Koch C, Schneider G, Marcou G, Varnek A (2011) Local neighborhood behavior in a combinatorial library context. J Comput Aided Mol Des 25(3):237–252. https://doi.org/10.1007/s10822-011-9416-2
33. Papadatos G, Cooper AWJ, Kadirkamanathan V, Macdonald SJF, McLay IM, Pickett SD, Pritchard JM, Willett P, Gillet VJ (2009) Analysis of neighborhood behavior in lead optimization and array design. J Chem Inf Model 49(2):195–208. https://doi.org/10.1021/ci800302g

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.