Trustworthiness, the Key to Grid-Based Map-Driven Predictive Model Enhancement and Applicability Domain Control

Dragos Horvath, Gilles Marcou, Alexandre Varnek

Trustworthiness, the Key to Grid-Based-Map driven Predictive Model Enhancement and Applicability Domain Control.

Dragos Horvath [a*], Gilles Marcou [a], Alexandre Varnek [a]

[a] Laboratory of Chemoinformatics, UMR 7140 University of Strasbourg/CNRS, 4 rue Blaise Pascal, 67000 Strasbourg, France

* Corresponding Authors (dhorvath@unistra.fr)

Abstract.

In chemography, grid-based maps sample molecular descriptor space by injecting a set of nodes, then linking them to some regular 2D grid representing the map. They include Self-Organizing Maps (SOM), and Generative Topographic Maps (GTM). Grid-based maps are predictive because any compound thereupon projected can "inherit" the properties of its residence node(s) – node properties themselves "inherited" from node-neighboring training set compounds. This article proposes a formalism to define the trustworthiness of these nodes as "providers" of structure-activity information captured from training compounds. An empirical four-parameter Node Trustworthiness (*NT*) function of density (sparsely populated nodes are less trustworthy) and coherence (nodes with training set residents of divergent properties are less trustworthy) is proposed. Based upon it, a trustworthiness score *T* is used to delimit the Applicability Domain (AD) by means of a trustworthiness threshold *TT*. For each parameter setup, success of ensuing inside-AD predictions is monitored. It is seen that setup-specific success levels (averaged over large pools of prediction challenges) are highly covariant, irrespectively of the targets of prediction challenges, of the (classification or regression) type of problems, of the specific parameterization and even the nature (GTM or SOM) of underlying maps. Thus, success levels determined on the basis of regression problems (445 target-specific affinity QSAR sets) on GTMs and levels returned by completely unrelated classification problems (319 target-specific active/inactive-labeled sets) on SOMs were seen to correlate to a degree of 70%. Therefore, a common, general-purpose setup of the herein proposed parametric AD definition was shown to generally apply to grid-based map-driven property prediction problems.

Abbreviations: AD – Applicability Domain, GTM - Generative Topographic Mapping, LLh-LogLikelihood, NB – Neighborhood Behavior, [Q]SAR – [Quantitative] Structure-Activity Relationships, R – Responsibility vector, RBF – Radial Basis Function, SOM – Self-Organizing Maps

1 Introduction.

Computer-aided management of chemical information exploits the "chemical space" (CS) defined by a molecular descriptor vector, which positions molecules in this framework [1]. It assumes "Neighborhood Behavior" (NB) compliance [2-4] – molecules with similar descriptors

should have similar properties. Chemography is a domain of chemoinformatics dedicated to "flattening out" the CS, to be rendered as a human-readable 2D map. In virtue of the NB principle, close analogues within a CS sphere centered on a reference compound of property *P* will likely have property values close to *P*. Or, one may conceive *P* as a local characteristic of the CS – like a physical field filling the entire space, not only points where its "sources" (here – the reference compounds) are located. Structure-activity (SA) information is herewith "disembodied" from its original providers (the training molecules) and transferred to the CS. If so, then this high-dimensional property "field" should be mappable as intuitive 2D property landscapes – so that the position of a compound on the 2D map may be predictive of its property.

One powerful approach in chemography, Generative Topographic Mapping (GTM) was introduced by Bishop, Svensen & Williams [5, 6]. Essentially a probabilistic, fuzzy generalization of Kohonen Self-Organizing Maps (SOMs) [7, 8], GTM draws its multivalence [9] specifically from its fuzzy-logics approach. Both approaches are ***grid-based***, relying on a 2D grid of nodes laid out according to a regular pattern in the map plane (the "latent space"). These nodes are linked to the initial descriptor space and are associated to items from their descriptor space neighborhoods – albeit the mathematical formalism used to achieve this strongly differs in the two approaches (SOMs employ "code vectors" while in GTMs nodes are bound to a flexible "manifold" inserted in CS and fitted against the "frame set" of compounds). In both approaches, nodes may serve as "probes" reporting local property values at corresponding CS coordinates, and hence become instrumental in generating 2D property landscapes.

By contrast to SOMs, where an item (a molecule) is assigned to one and only one node, GTM however interprets the statement "molecule *M* resides in node *N*" as a fuzzy truth, of real value $0<R_N(M)<1$. The sum of all $R_N(M)$ – further on referred to as the Responsibility vector *R* – over all nodes equals one. In practice, the concept of responsibility is associated to the GTM algorithm, never to SOMs. Formally, one may nevertheless think about SOM "responsibilities" as a binary vector with $R_W(M)=1$ for the "winning" node *W*, and $R_N(M)=0$ for all others $N \neq W$. With this important specification, GTM and SOM-based property prediction are in this work described by the same *R*-based formalism, irrespective of the real (GTM) or binary (SOM) nature of this vector. Note that this – landscape-based, *vide infra* – property prediction procedure proposed here is an extension of the standard GTM prediction mechanism, that can be seamlessly applied to GTMs and SOMs – and other "grid-based" maps – alike. Many other

property prediction mechanisms based on SOMs and GTMs could be envisaged – yet, this one is generally applicable to grid-based maps and therefore would also allow for an unified approach defining Applicability Domains.

Above-mentioned landscape construction is nothing but responsibility-mediated transfer of structure-activity information from the training set onto the nodes. Prediction represents the inverse – first, the compound is projected, and the predicted property is taken as the *R*-weighed mean of node properties. In GTM, any small structural change impacts *R* levels, and smoothly modifies predicted properties. Therefore, GTMs support full-blown predictive regression [10] and classification [11] models (Quantitative Structure-Property/Activity Relationships, QS[P/A]Rs). On the contrary, SOMs are limited to postulating that the predicted property of any node resident would equal the mean of training set resident properties, a fixed value for each node – a direct consequence of the binary nature of their *R* vector. Thus, any change of structure not impacting on the SOM node assignment will have no consequence on SOM-predicted properties. SOM-based property landscapes are intrinsically "granular" with GTM-based landscapes are smooth. This being said, above differences are in practice not as clear-cut. On one hand, *R* vectors on GTM often are *de facto* binary, as compounds (routinely) happen to be associated to a single node, at *R* levels above 0.99999. On the other, SOMs could be enhanced by a fuzzy-logics formalism, defining real-value *R* scores as some decreasing function of the distance of item and code vector nodes. In this paper, however, the goal is not an in-depth benchmarking of GTMs against SOMs, but the use of two different paradigms of grid-based maps, GTMs and SOMs in order to search an Applicability Domain formalism which may apply to both, in spite of their differences.

Moreover, grid-based maps offer straightforward means to assess the trustworthiness of its landscape-based predictions. Two criteria of node trustworthiness can be envisaged. The first is the cumulated responsibility of node residents, *i.e.* the node *density*. Nodes with low cumulated training set responsibilities are basically *terra incognita* – marginal levels of association to training set compounds makes the assignment of a node property value technically possible, but not trustworthy. The second is the *coherence* of the property data contributing to a node, *i.e.* the *spread* (R-weighed standard deviations) of the resident properties – showing that some map zones may be more NB-compliant than others. High coherence is mandatory for high trustworthiness.

The prediction of a compound property is, in GTM, however tributary to *all* the nodes to which this compound is associated with tangible responsibility values. Each node has its own "Node Density" (*ND*) and "Node Coherence" (*NC*) – how would these impact on the global trustworthiness of prediction? Furthermore – if the predicted compound is equally strongly associated to a node characterized by a high activity value and a node of low activity value, is it senseful to accept the *R*-weighted mean of these diverging values as predicted value? This aspect – quantitatively measured by the standard deviation associated to the *R*-weighted mean value – will be further on termed "Prediction Coherence" (*PC*) to be distinguished from above-mentioned "Node Coherence" (*NC*). Eventually – how to best combine all the cited aspects into one clear-cut decision-making trustworthiness score?

Paradoxically, even though GTM-based property prediction offers an extremely versatile control of its Applicability Domain (AD), and even though the potential power of such AD control has been understood and advertised in previous publications [10, 12], this versatility makes it impossible to easily formulate "the" obviously best mode to define trustworthiness, and its threshold value delimiting the AD. So far, in our hands GTM-driven predictions were used as such, with at best some empirical, ill-defined minimum density requirements to be satisfied, awaiting for a systematic study to explore the relative merits of envisageable trustworthiness scoring schemes – hence, this contribution.

Actually, this study goes one step further, and first investigates whether landscape nodes involved in prediction should be no longer contribute proportionally to the *R* values of the compound to predict, but have their effective impact modulated by their density and coherence factors. Intuitively, such a strategy makes sense – nodes are the "sources of knowledge" which transfer to the candidate to be predicted the knowledge learned about the CS distribution of property *P* on the basis of training set compound. Sources most relevant to the compound to predict (nodes of higher level of association *R*) should impact most on prediction (*R*-weighted averaging is paramount). Yet, it might be senseful to dampen the relative impact of less trustworthy sources (empty or non-coherent nodes). A mathematical formalism in this sense is proposed here, involving a few tunable parameters.

Eventually, trustworthiness of node sources (*NC*, *ND*) and coherence of prediction (*PC*) are combined into a final trustworthiness score *T*. Predictions at *T* above user-specified thresholds were carried out and evaluated. AD-dependent predictions need be assessed in terms of two

(often) conflicting criteria: prediction accuracy and AD coverage of the external set. A good *T* criterion would typically allow increasing accuracy at the cost of lowering coverage, as more constraining thresholds are applied. In order to avoid complex Pareto front analysis, (coverage, accuracy) pairs were here characterized by three complementary quality scores *QC*, *QA*, *QU*. All the three are geometric means of coverage and accuracy, with one being biased to be more sensitive on Coverage (*QC*), another to respond more to Accuracy (*QA*), while the default third is Unbiased (*QU*) and equally sensitive to both.

A large number of pK_i (-log of the thermodynamic instability constant of protein-ligand complexes) quantitative (regression) prediction problems were run over a pool of 445 different biological targets endowed with enough (>100) associated ligands of reported K_i in ChEMBL v.26. Systematically, a randomized 30% of each set was taken out as external set, while the remaining 70% served to generate pK_i landscapes on the seven general-purpose "universal" maps previously developed by our team. Prediction of external pK_i values was then performed, for all envisaged AD-defining parameters. Coverage (fraction of external set within herein defined AD) and accuracy (here, R^{2*} values, *vide infra*) were monitored. For each target, randomized external set extraction was repeated 25 times, as prediction scores may significantly change in response to training *versus* external set composition. It is thus possible to count how many of these 25 trials returned (coverage, accuracy) pairs of high *QC*, *QA* or *QU* – and implicitly, to monitor the percentage of "successful" predictions throughout the pool of prediction challenges featuring the 445 QSAR sets on the seven GTMs. The question whether preferred AD-defining strategies would depend on the nature of the used GTM (based on significantly different molecular description schemes) was also addressed.

Furthermore, it is important to verify whether the findings of the regression-based AD definition quest are of general validity – irrespective of both target nature and QSAR problem nature. 319 additional biological targets for which ChEMBL does not offer sufficient pK_i data for large enough quantitative QSAR sets but reports enough activity data to generate a classification QSAR series (discriminating between empirically defined "actives" versus "inactives") were employed to this purpose. As GTM-driven "fuzzy binary classification landscapes" treat the probability to belong to a class rather than the other as a real-value score, they technically behave like regression landscapes. The mathematical formalism provided here applies irrespectively of whether the molecular property *P(M)* is a real value or a class number (1=inactive/2=active).

However, Balanced Accuracy (*BA*) is used as accuracy criterion in the latter case. As a side remark, multi-class classification problems require a different mathematical formalism and are not covered here. However, this is not a real limitation, because any *C*-class classification problem can be reformulated as *C* independent binary classification challenges – each focused on segregating members of a class *c=1..C* from respective non-members.

The final key point addressed here is checking whether GTM-based trustworthiness definitions may as well apply to SOMs, herewith showing that the herein developed formalism may apply to several grid-based mapping algorithms. To this purpose, four "universal" SOMs US_1-US_4 were constructed in following the evolutionary universal GTM selection procedure [13], using strictly the same 236 active/inactive binary classification QSAR problems for map quality assessment – but adapting the map-encoding "chromosome" to accommodate SOM-specific instead of GTM-specific parameters. The 319 above-mentioned classification problems were eventually enacted on US_1-US_4, following the established protocol. Collaterally, this allows to quantitatively assess if – and in how far – GTMs are, as expected but so-far never formally proven – more effective predictors than SOMs.

As an outline, this article focuses on strategies to optimize predictivity of grid-based map landscapes, in proposing a means to quantify node trustworthiness as "providers" of neighborhood information learnt at training stage. To this purpose, a mathematical formalism featuring a few tunable parameters and a series of prediction success criteria is introduced.

Predictive landscapes rely on "universal GTMs" or on-purpose built "universal SOMs". They are built on hand of ChEMBL-extracted training sets associated with either continuous regression (pK_i values) or binary (activity class) data. Since universal SOMs were not described before, short Methods and Results chapters are needed to properly introduce them.

The relative performance of the various setups is impacted by the chosen "points of view" embodied by the complementary success criteria: setups guaranteeing high coverage are not the same as ones providing high prediction accuracy. With this is mind, the following key questions were addressed here:

- Are preferred AD-defining strategies dependent on the nature of the used GTM (based on significantly different molecular description schemes)?
- Are they problem category specific, or are there consensus setups which maximize success in both regression and classification problems?

- Are they map type specific, or are there consensus setups which maximize success of both GTM and in SOM-driven models?

Based on this study, and dependent on the coverage *versus* accuracy-oriented point of view of the user, the tunable parameters were assigned values guaranteeing a general compliance of the AD control schemes with all the prediction scenarios covered here.
Eventually, a visual illustration of the concepts involved in the AD definition is shown, as an aid to highlight and understand prediction errors.

2 Methods

GTM construction has been already extensively described in literature. Likewise, the philosophy and technical details at the basis of the herein used Universal maps were also largely described [13]. Therefore, this article will only focus on the methodology of predictive landscapes, based on the already introduced responsibility vector $R_N(M)$, featuring the level of association of an item (compound) *M* to each of the nodes *N* of the GTM.

2.1 Universal SOMs

Universal SOMs were generated by an evolutionary algorithm exploring the parameter space associated to the herein used SOM_PAK software [14], while the fitness function used to select the maps was the same mean cross-validated *BA* score over the 236 active/inactive binary classification QSAR problems used to power the universal GTM search [13, 15]. Each of the four US was based on the same descriptors used by the corresponding universal GTM, *i.e.* descriptor choice as a degree of freedom of the evolutionary process was disabled. The six SOM-specific degrees of freedom encoded by the chromosome are given in the Appendix document available as Supplementary Information.

2.2 Predictive Landscape Construction

As mentioned in Introduction, the methodology below applies to both GTM and SOM-driven property prediction, with the provision that SOM "responsibility" vectors are binary, with $R_W(M)=1$ for the "winning" node *W*, and $R_N(M)=0$ for all others $N \neq W$. A property landscape is defined by transferring properties *P* from training set (TS) molecules (*M*) onto the strongly associated nodes. *P* may be any continuous molecular property, or a binary class label in fuzzy classification landscapes. Many-class classification problems are not considered here. This

paragraph outlines how to calculate a predictive landscape (*i.e.* the set of node-associated property values NP_N) together with associated AD-relevant criteria (density, coherence).

The cumulated responsibility on a node, CR_N, is the sum of training molecule responsibilities:

$$CR_N = \sum_{M \in TS} R_N(M) \quad (1)$$

Note that theoretically CR_N is always positive, since all $R_N(M)$ are positive values rendered by radial basis functions. Practically, R values are stored on file with a precision of 10^{-5}. Thus, nodes with no "tangible" responsibility – $R_N(M) < 10^{-5}$ for all molecules – have $CR_N = 0$.

The magnitude CR_N is dependent on training set size and map resolution (number of nodes), thus requires some normalization before being used as a node density (*ND*) criterion in trustworthiness estimation. Here, it was empirically decided to assign the most populated node *ND=1.0* and empty nodes *ND=0.0*, hence

$$ND_N = \frac{CR_N}{\max\limits_N CR_N} \quad (2)$$

The responsibility-weighed mean property value of residents is computed on each node N as the node property NP_N:

$$NP_N = \frac{1}{CR_N} \sum_{M \in TS} R_N(M) P(M) \quad (3)$$

The associated responsibility-weighed standard deviation value of the property, $\sigma(NP_N)$ is:

$$\sigma(NP_N) = \sqrt{\frac{1}{CR_N} \sum_{M \in TS} R_N(M) P^2(M) - NP_N^2} \quad (4)$$

This standard deviation reflects the degree of consistency of the property on the map. It is small if most compounds located in the same region of the map have similar property values. It is large if there is no relation between the property value and this location on the map.

This observation can be translated into a *coherence* score, NC_N, defined as follows:

$$NC_N = 1 - \frac{\sigma(NP_N)}{\max\limits_{M \in TS} P(M) - \min\limits_{M \in TS} P(M)} = 1 - \frac{\sigma(NP_N)}{\Delta P} \quad (5)$$

The node coherence is a dimensionless value <1, but always positive (the standard deviation of a distribution cannot exceed its range width). The score is maximal if the compounds located near the considered node share similar property values (NB compliance); it is maximal if there is no relation between this node and the property value. The symbol *ΔP* denotes property range width. At null CR_N, NC_N is undefined and set to zero: empty nodes are completely incoherent, by definition. Otherwise, $NC_N > 0$, as the standard deviation of a property cannot exceed its range width.

2.2.1 The case of empty nodes and the Min-Mean Toggle (MMT).

For nodes void of any tangible responsibilities, $CR_N = 0$ and hence equations (3) to (5) are not applicable. Default *NP* values are assigned to empty nodes, depending on a "Min-Mean Toggle" *MMT*:

- With the toggle set to "mean", empty nodes are assigned to the average property of the training set: $NP_N|_{CR_N=0} = \langle P(M) \rangle_{M \in TS}$. It is the reasonable expectation if nothing is known about a chemical space zone.
- Toggle at "min" assigns a chosen expectation of the activity level. In the present case it is the lowest property value observed in the training set: $NP_N|_{CR_N=0} = \min_{M \in TS} P(M)$.

The minimum value was chosen in the present case, because with all herein predicted bioactivity scores (pK_i or activity classes) "minimum" is synonymous to low activity. Thus, the *MMT* degree of freedom chooses between two empirical postulates about the behavior of compounds in chemical space zones not covered by the training set: "mean" assumes those activities to be "average" (as predicted by a null model), "min" assumes those molecules to be inactive.

2.3 Node Trustworthiness

The below proposed Node Trustworthiness (*NT*) score serves to modulate the participation of each node in the prediction process – with highly populated and homogeneous nodes expected to contribute more. *NT* is postulated to increase with node density *ND*, equation (2), and coherence *NC*, equation (5), according to the simple working hypothesis below:

$$NT_N = \frac{\tau + ND_N^{\alpha} \times NC_N^{\beta}}{\max_N \left(\tau + ND_N^{\alpha} \times NC_N^{\beta} \right)} \tag{6}$$

where τ, α and β are tunable parameters, which will be subject to an exhaustive scan in this work. The relative importance of ND_N and NC_N is controlled by the values of the exponents α and β that were allowed to take values of (0.0, 0.01, 0.1, 0.5, 1.0, 2.0). Trustworthiness becomes independent of the density and the coherence if α=β=0. The parameter τ plays the role of a default trustworthiness level and is set to take the values (0.0, 0.01, 0.1, 0.5); if non-null, it prevents the score to be undefined in the case both density and coherence are together null. The denominator in equation (6) is a normalization factor, ensuring that the most trustworthy of all nodes of the trained landscape is assigned $NT_N = 1$. At the opposite end, $NT_N = 0$ signals that such nodes will be completely ignored in predictions. The entire grid of (τ, α, β, *MMT*) combinations was explored, excluding redundancies (if α=β=0, all nodes will have $NT_N = 1$ irrespective of τ). Note that NT_N may become zero only for empty nodes and only if τ=0 and α>0. Otherwise, their trust level remains positive and thus the *MMT*-chosen NP_N value becomes relevant, while at τ=0 and α>0 prediction results are *MMT*-independent.

2.4 The NT-sensitive prediction protocol.

The rule to interpolate the property (below, "^" stands for "predicted") of a molecule *M* accounting for node trustworthiness is:

$$\hat{P}(M) = \frac{\sum_N R_N(M) NT_N \times NP_N}{\sum_N R_N(M) NT_N} \tag{7}$$

which resumes to $\hat{P}(M) = \sum_N R_N(M) \times NP_N$ if all nodes are equally trustworthy. The normalizing factor at denominator is the mean node trustworthiness of residence nodes concerning *M*:

$$\overline{NT}(M) = \sum_N R_N(M) NT_N \tag{8}$$

The weighted mean $\hat{P}(M)$ serving as predicted value is associated to a standard deviation:

$$\widehat{\sigma P}(M) = \sqrt{\frac{1}{\overline{NT}(M)} \sum_N R_N(M) NT_N \times NP_N^2 - \hat{P}(M)^2} \tag{9}$$

The standard deviation expresses the divergence of the node properties based on which the molecular property is extrapolated. It is, in addition to the mean node trustworthiness $\overline{NT}(M)$, the other key contributor to the *trustworthiness* $T(M)$ of the prediction of *M*:

$$T(M) = \overline{NT}(M) \times \left(1 - \frac{\widehat{\sigma P}(M)}{\Delta P}\right) \tag{10}$$

$T(M)$ is large if (1) *M* predominantly resides in trustworthy nodes and (2) the standard deviation of the prediction is small in comparison to the activity range ΔP, *e.g.* nodes of residence have nearly equal node property values; it is small otherwise.

2.5 Applicability domain definition and performance criteria.

The score is finally used to take an applicability domain decision. For an external set, at given (τ, α, β, *MMT*), all molecules *M* reaching a user-chosen Trustworthiness Threshold *TT* are considered as inside the AD. The AD coverage fraction is:

$$f_{AD} = [\text{number of molecules with } T(M) > TT]/[\text{external set size}] \tag{11}$$

Eventually, $\hat{P}(M)$ is compared to actual $P(M)$ for all compounds within the AD, in order to establish the prediction quality criterion. For categorical problems, *BA* is classically defined as the mean of proportions of well classified actives and well classified inactives, respectively – ranging from zero to one, with random classifier performance at 0.5. For regression problems, the root-mean-squared error of prediction $RMSE\left[\hat{P}(M), P(M)\right]|_{T(M)>TT}$ has been reported to the standard deviation $\sigma_0(P)$ of $P(M)$ over the entire QSAR series, prior to its randomized split into training and external sets. It is translated into a determination coefficient of a given prediction simulation:

$$R^{2*}(\tau, \alpha, \beta, MMT; TT) = 1 - \frac{RMSE^2\left[\hat{P}(M), P(M)\right]|_{T(M)>TT}}{\sigma_0^2(P)} \tag{12}$$

This measure is independent on fluctuations of the dynamic range of the property within the randomly picked external set. Note that the quality criteria above (f_{AD} and R^{2*}, respectively *BA*) are all tributary to the five parameters $(\tau, \alpha, \beta, MMT; TT)$. The first four control the prediction

mechanism, whereas the latter controls the subset of predictions considered inside the AD. For TT, considered values were (0.0, 0.5, 0.7, 0.8, 0.9, 0.95). In view of above-mentioned redundancies, 1062 distinct $(\tau, \alpha, \beta, MMT; TT)$ combinations were systematically scanned.

2.6 Data sets

All the ligand structures used were imported from ChEMBL [16] and standardized according to the default procedure of our web server http://infochim.u-strasbg.fr/webserv/VSEngine.html. GTM landscape-based regression models were benchmarked against a series of 445 QSAR sets extracted from ChEMBL v.26. Each such set consists of ligands binding with known thermodynamic inhibition constant to a given biological target. The 445 considered targets are all the ones featuring ≥100 distinct ligands of known K_i, excluding imprecise entries (K_i larger or smaller than indicated value). Since employed molecular descriptors are stereochemistry-insensitive, strict uniqueness of standardized, stereochemistry-void canonical SMILES is required. Unique SMILES associated to multiple K_i values diverging by more than one order of magnitude were discarded. The property used in predictions was pK_i.

Classification models were acted on binary active/inactive QSAR sets previously extracted for Universal map construction and validation [13]. Please refer to that article for a detailed discussion of their preparation and the assignment protocol of active *versus* inactive status. After excluding targets already covered by the 445 regression problems, 319 distinct classification QSAR sets remained.

2.7 Benchmarking protocol

Each regression QSAR set was projected on each of the seven Universal maps, further on referred as $U_1 \ldots U_7$. As prediction was run independently for each map, this amounts to 445×7=3115 series of predictions. For each such series, a systematic scan over combinations of the four tunable parameters $(\tau, \alpha, \beta, MMT)$ is started.

A "prediction challenge" is run at given $(\tau, \alpha, \beta, MMT)$ for each set, on every map. Therefore, the number of regression prediction challenges performed here equals 3,308,130 = 1062 (scanned parameter quintuplets) times 3115 (QSAR set, Universal Map) combinations. A prediction challenge consists of the following key steps:

1. First, a random split of associated QSAR data into "training" (70%) and "external test" (30%) is proposed.

2. Training data are used to build the pK_i landscape on the given map and to define the *NT* scores of nodes therein in accordance to set $(\tau, \alpha, \beta, MMT)$ values.
3. Eventually, external compounds are projected on the above landscape, with output of $\hat{P}(M)$ and respectively $T(M)$.
4. Looping over considered *TT* values, entries satisfying the AD threshold are selected and quality criteria f_{AD} and R^{2*} are reported as associated to the given setup π.
5. The cycle of (randomized data splitting – landscape construction – prediction – evaluation) is repeated, until having recorded 25 (f_{AD}, R^{2*}) entries for each *TT* value.

Thus, a prediction challenge returns 25 (f_{AD}, R^{2*}) entries per *TT* value, concerning a given QSAR data set on a given map.

The same procedure was then applied to the 319 binary classification problems, employing U_1 as a representative GTM and monitoring problem-specific (f_{AD}, BA) pairs. Eventually, these classification problems were also processed on the four universal SOMs US_1-US_4.

At any given setup, f_{AD} and R^{2*} and respectively *BA* may significantly fluctuate in response of the randomized composition of training *versus* external set. Are there any setups providing *systematic* advantages in terms of (f_{AD}, R^{2*}), respectively (f_{AD}, BA) ?

As mentioned in Introduction, the quality of (coverage, accuracy) pairs cannot be captured by a single number but can be characterized by three "view-point-specific" coverage-focused (*QC*), accuracy-focused (*QA*) and unbiased (*QU*) quality criteria. To this purpose, the accuracy criteria R^{2*} and *BA* must first be normalized to a [0,1] range, by mapping the lowest relevant value to 0. Here, $\min(R^2) = 0.3$ and $\min(BA) = 0.6$ while $\max(R^2) = \max(BA) = 1$ Let the normalized values be designed as *Q*:

$$Q = \begin{cases} max[0, (R^{2*} - \min(R^2))/(\max(R^2) - \min(R^2)] & for\ regression \\ max[0, (BA - \min(BA))/(\max(BA) - \min(BA)] & for\ classification \end{cases} \quad (13)$$

Herewith the three quality criteria are defined as:

$$QU = \sqrt{Q \times f_{AD}}$$
$$QA = \sqrt[4]{Q^3 \times f_{AD}} \quad (14)$$
$$QC = \sqrt[4]{Q \times f_{AD}^3}$$

Three quality thresholds (low, *L=0.4*; medium, *M=0.6*; high, *H=0.8*) were envisaged, making it straightforward to count the percentage of situations in which a given parameter vector π managed to reach a prediction quality score exceeding a given threshold. Formally, the success rate $S(\pi)|_{\Pi,QX,Y}$ of a setup π over a given pool Π of predictions according to criterion *QX* at threshold *Y* represents the percentage of situations in which setup π delivered the expectation *QX>Y*, out of the total number of times π has been used for prediction within the pool Π. The latter consists of conveniently regrouped prediction challenges – by problem type, by maps or map families, as listed in Table 1 below.

Table 1: Pools of prediction challenges.

Pool Designation Π	Description
reg@U	Prediction challenges of the 445 regression QSAR sets on all the seven universal GTMs (445×7×25=77875 challenges)
reg@U*i*	Prediction challenges of the 445 regression QSAR sets on a specific universal GTM #*i*=1..7 (subsets of reg@U)
class@U1	Prediction challenges of the 319 classification QSAR sets on universal GTM #1 (for direct comparison to reg@U1)
class@US	Prediction challenges of the 319 classification QSAR sets on all the four universal SOMs (319×4×25=31900 challenges)
class@US*i*	Prediction challenges of the 319 classification QSAR sets on a specific universal SOM #*i*=1..4 (subsets of class@US)

The analysis of the success rate aims at answering the following questions:

1. How do success rates depend on the employed quality criteria, QX, and success rate threshold, Y?
2. Are there π setups that can be considered globally better, irrespective of challenge pools, and the nature of the problem (classification or regression)?
3. Are there π setups that can be considered globally better, irrespective of map parameters, and even the nature (GTM or SOM) of the mapping algorithm?

Positive answers to points 2 and 3 above would mean that a context- and method-independent consensus on how to define trustworthiness in grid-based chemical space maps can be found. For each challenge pool Π, quality criteria, QX, and success rate threshold, Y, a set of 1062 success rate $S(\pi)|_{\Pi,QX,Y}$ values was collected – one for each investigated setup π. These vectors of success rate values are subjected to covariance analysis. If $S(\pi)|_{\Pi,QX,Y}$ is positively correlated to $S(\pi)|_{\Pi',QX',Y'}$ over all the π, this means that roughly the same setups maximizing the success rate $QX>Y$ of prediction challenges over pool Π are also the ones maximizing success rates $QX'>Y'$ of prediction challenges over pool Π'. Reciprocally, setups causing weak success rates in the context Π, QX, Y will not work in the context Π', QX', Y' either. The degree of covariance is reported as the Pearson correlation coefficient of the linear regression line $S(\pi)|_{\Pi,QX,Y} = aS(\pi)|_{\Pi',QX',Y'} + b$. The higher the Pearson score, the stronger the assumption that some common parameter set π can be found to perform well in both the contexts Π, QX, Y and Π', QX', Y'. Low correlation means, on the contrary, that any specific setup π yielding satisfactory results for problems in pool Π according to the criterion $QX>Y$ would be a bad choice when applied to the context Π', QX', Y'.

Eventually, the last key aspect is to understand whether top performing setups are characterized by specific values or value ranges of the individual parameters $(\tau, \alpha, \beta, MMT; TT)$, herewith establishing practical recipes on how exactly to best harness the trustworthiness issues in grid-based map predictors in order to maximize their predictive quality.

3 Results

3.1 Universal SOMs

Details about the corresponding SOM-specific setups can be found in the Table A1 of the Appendix document. Evolutionary "growth" of universal SOMs was a relatively easy exercise compared to the previously achieved GTM [15, 17] construction, both because (a) the descriptor sets of the four top GTMs were used, without further considering descriptor choice as a degree of freedom and (b) SOM configuration has less tunable parameters. Therefore, if these SOMs would be shown to be less proficient active/inactive separators (in repeated three-fold cross-validated simulations, following the same protocol used for GTM construction) over the 236 selection QSAR sets, this cannot be a consequence of insufficient sampling of SOM parameter

space by the evolutionary procedure. Indeed, the distribution histograms of the shifts in cross-validated Balanced Accuracies *XVBA* of QSAR sets on the GTM, with respect to the equivalent performance of the same set on the equivalent SOM, $XVBA(set@U_k) - XVBA(set@US_k)$ are clearly biased towards positive values, for all the *k=1..4* four pairs of corresponding maps.

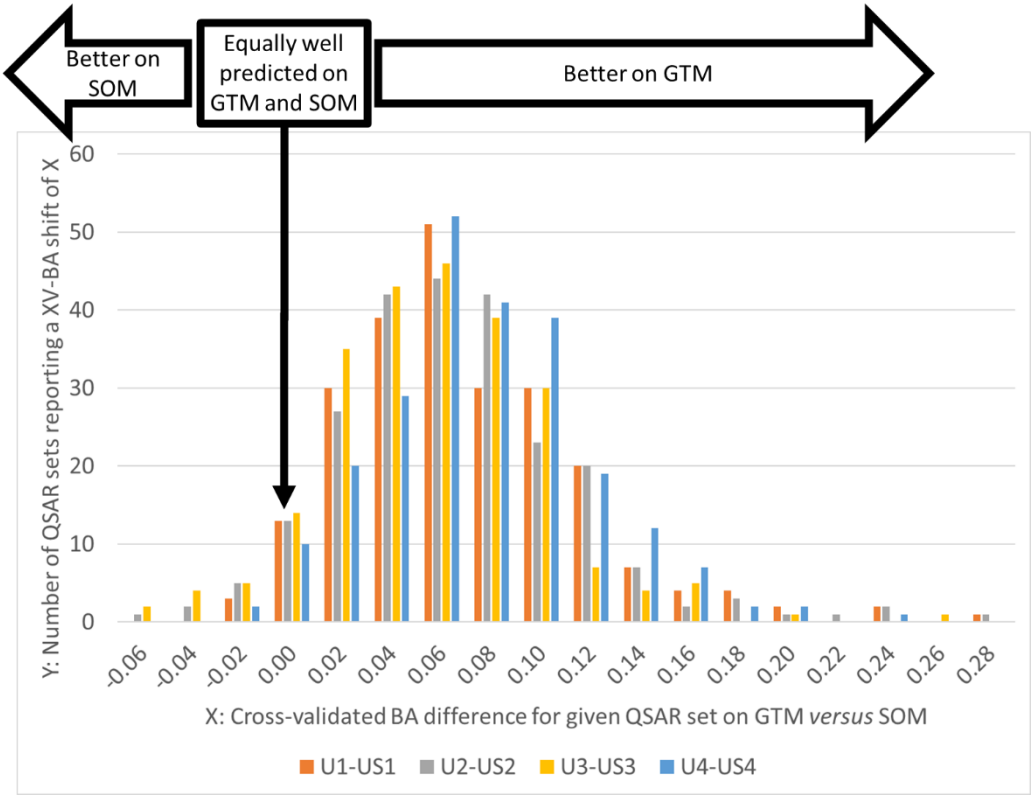


Figure 1: Distribution of the XVBA shift $XVBA(set@U_k) - XVBA(set@US_k)$ for each of the 236 QSAR sets achieved on universal GTMs *versus* equivalent (same descriptor space-based) universal SOMs. U_i are universal GTMs , US_i are universal SOMs; the *i* index refers to a given descriptor space common to both U_i and $\underline{US_i}$.

Targets seen to better cross-validate on GTMs are clearly a majority, whereas QSAR sets better discriminated on SOMs are rare, and are only marginally enhanced. Fuzzy logics-based grid-based mapping is clearly a winning strategy.

3.2 The double putative impact of trustworthiness: general discussion.

The key novelty – and also putative source of confusion – of this work is that node trustworthiness is being assigned two distinct roles: it (conditionally) influences upon predicted values, according to equation (7), and then contribute to estimate the global trustworthiness of predictions, according to (10). Property prediction using grid-based maps is practically a "data fusion" exercise: nodes to which the compound is assigned are "data sources" on hand of which

a final "decision" concerning the compound predicted properties must be made. By default, this "decision" consists in taking the *R*-weighed mean of node properties. The present work however argues that trustworthiness of the "data sources" should also be considered here – a common sense strategy in data fusion. Here, importance of robust data sources (trustworthy nodes) is enhanced proportionally to *NT*. If a compound is shared between a coherent and a low coherence, activity cliff-ridden node, the *NT* factor will enhance the contribution of the former "safer source" in mean prediction (relative to the default, *R*-weighted contribution). Turning this *NT*-bias off is implicitly achieved by setting the exponents α and β in equation (6) to zero, as covered by the present benchmarking – the current formalism is fully "backwards compatible" with the standard grid-based map prediction process.

Of course, the above only concerns cases in which there *are* several data sources to ponder upon. In SOMs – and, very often, on GTMs, whenever *R* vectors are *de facto* binary (~100% of residence in a single node) – the only node in question becomes, by default, the most trustworthy one. If its *NT* equals zero, no property prediction can be performed and the molecule is forcibly out of the AD. Else, the molecule property will be assigned to the mean property of the node – the best envisageable estimate under given circumstances. However, even if *NT* did not directly impact on predicted values, it will nevertheless serve in the user's decision-making on whether to trust or to discard this prediction, its second key role. Thus, a predicted value based on a single low-density or uncoherent node cannot be "corrected" – but will be labeled as untrustworthy according to its low *T(M)* value from equation (10).

3.3 Quality Criteria and Thresholds

A setup is successful according to a given quality criterion at a given threshold if its coverage and accuracy values are fulfilling the specified constraints. The third paragraph of the Appendix in Supplementary Information intuitively illustrates what it practically means to achieve "success" according to a quality criterion *QC*, *QU* and *QA*. Covariance analysis of success scores (details in Appendix) showed that success counts at considered thresholds remain largely proportional – $S(\pi)|_{P,QX,L} \sim S(\pi)|_{P,QX,M} \sim S(\pi)|_{P,QX,H}$ – meaning the relative merit of parameter sets is actually independent on the threshold. Therefore, further discussion will focus on the medium and high levels of performance only. Concerning the nature of the criteria, the analysis (details in Appendix) showed that success scores based on compromise *QU* are fairly correlated

to both the *QA*- and the *QC*-based ones – while the accuracy- and coverage-based criteria are indeed complementary (uncorrelated). Therefore, *QU* can be further on ignored.

3.4 Is setup success map-dependent?

Success scores for regression problems were monitored according to GTMs U_1 to U_4, then compared to each other and to the global success scores obtained over the full set of seven universal GTMs. Likewise, success rates of classification problems monitored on US_1 to US_3 were compared to each other and to the global success scores obtained over the full set of four universal SOMs. These results are reported in Figure 2. The correlation is high for all pairwise map comparisons. The proficiency of a setup π is therefore independent of the map.

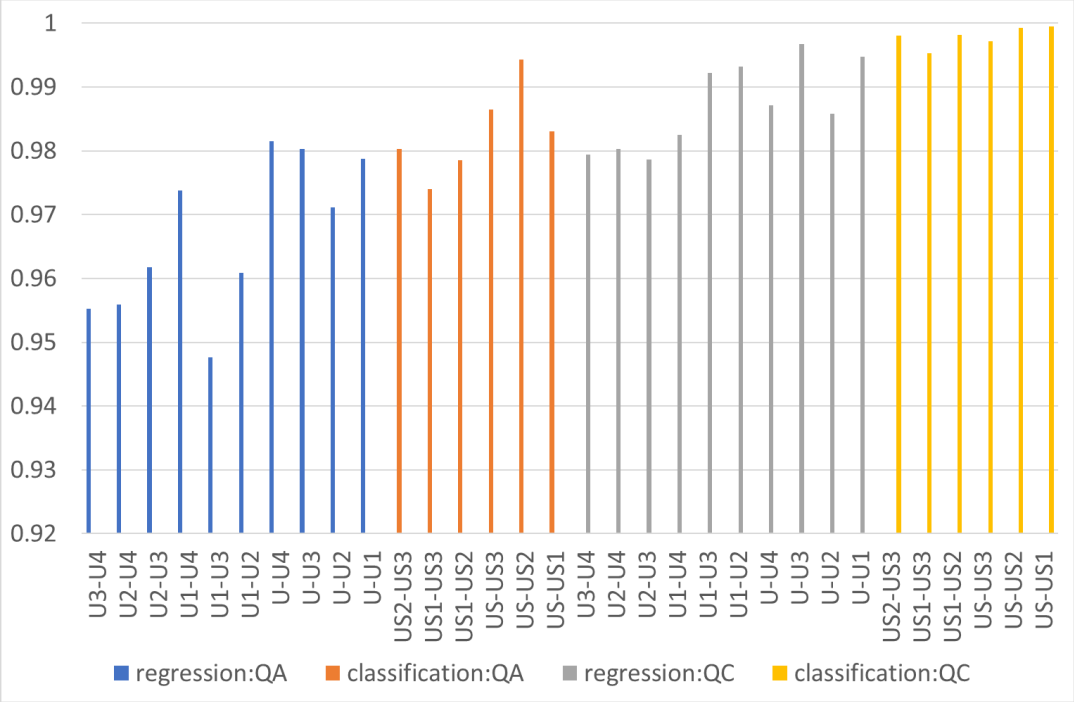


Figure 2: Level of correlation (Pearson R^2) between setup success scores advocated by specific maps (U*i* – universal GTM #*i*, US*i* – universal SOM #*i*) or by the consensus of the entire set of maps of a given category (U – all 7 GTMs, US – all 4 SOMs), at M threshold. Color coding refers to the prediction pool-quality criterion combination. The left-most blue U3-U4 bar reflects the degree of correlation between $S(\pi)|_{reg@U4,QA,M}$ and $S(\pi)|_{reg@U3,QA,M}$, *etc*

Within a large pool of QSAR problems of a same type and within a given category of grid-based maps (GTM or SOM), the setups optimizing coverage-focused and respectively accuracy-focused predictive performance appear to be independent on the map parameters, including the molecular descriptors it is based upon. Of course, these descriptors are the ultimate reason for which a molecule is (fuzzily or not) assigned to a given node and not to another. However, the

same (well-tuned) trustworthiness analysis schemes tends to give the good results, irrespective of the map. There are no π setups to work well with one specific map and fail with all others, in spite of the widely different nature of the chemical information represented on these maps. So far, results suggest that node trustworthiness may indeed be coherently tuned as a function of node density and coherence, independently of the nature of the map.

3.5 Is setup success dependent on the nature of the QSAR problem and QSAR sets?

The universal GTM U_1 represents the common ground on which both the 445 regression problems and the 319 classification problems were processed. Setup success scores obtained for the pool of regression problems, $S(\pi)|_{reg@U1,QA,M}$ and $S(\pi)|_{reg@U1,QC,M}$ respectively can be directly compared to their classification pool analogues $S(\pi)|_{class@U1,QA,M}$ and $S(\pi)|_{class@U1,QC,M}$. Note that this comparison encompassed two key changes: the nature of the prediction problem (regression *versus* classification) and the QSAR sets *per se* – the two series of 445 and 319 sets do not have a single biological target in common. Nevertheless, the Pearson correlation scores are of 0.675 for regression-focused and 0.775 for classification-focused success rates, respectively. Trends in setup success scores are thus resilient with respect to simultaneous and radical changes in terms of QSAR sets, and the nature of the prediction problem. The pertinence of a setup scheme therefore does not significantly depend on the particular problem it is being applied to. This is a key result, which completes the previous observation that well-tuned trustworthiness analysis schemes are also independent of the map nature. Please refer to the fourth paragraph of the Appendix document for a detailed analysis, with examples, of this result.

3.6 Is setup success dependent on the SOM vs GTM nature of the map?

The Pearson correlation score of $S(\pi)|_{class@U1,QA,H}$ *versus* $S(\pi)|_{class@US1,QA,H}$ reaches 0.744, witnessing that the impact of setup on the success scores largely follows a same trend, irrespective of the fundamental difference between the SOM and the GTM (all other things being, in as far as possible, equal – notably molecular descriptors). Since it was already established that the various SOMs tend to be highly covariant in terms of success score rankings, $S(\pi)|_{class@U1,QA,H}$ also strongly correlates to the generic $S(\pi)|_{class@US,QA,H}$, all SOMs confounded (Pearson score 0.75). *QC*-based scores typically tend to correlate even better than their *QA*-based counterparts and exceed 0.8. However, in terms of absolute scores, success rates

continue to be significantly higher on the GTM, in line with the cross-validation results from Figure 1. The most successful setup on US_1 achieved *QA>H* in 18.5% of class@US1 attempts, while being successful in 23.7% of equivalent GTM-base challenges class@U1 – with respect to which it is ranked only 10th. The absolute best result achievable within class@U1 amounts to 25.6%.

Ultimately, a triple "jump" in problem configuration space – from GTMs to SOMs, from regression to classification, from one pool of biological targets to another – is characterized by the correlation level of $S(\pi)|_{reg@U,QA,H}$ *versus* $S(\pi)|_{class@US,QA,H}$. With a robust Pearson score of 0.69 (0.76 for *QC*), this finally underlines that node trustworthiness may indeed be quantified independently of node construction history and underlying mathematics. A few representative π setups can be seen to define node trustworthiness in a way that systematically enhances prediction success and intelligently delimits AD throughout the spectrum of grid-based maps.

3.7 What are the good setups?

Note that a full-blown correlation of success score values throughout the series of monitored setup parameter combinations is not even necessary to define one or more parameter combinations of general use with grid-based maps. It is sufficient to find some combinations that are systematically ranked amongst the best in each of the prediction challenges on the GTMs and SOMs. Sorting $S(\pi)|_{reg@U,QA,H}$ and respectively $S(\pi)|_{class@US,QA,H}$ in order to eventually pick parameter combinations with the best mean rank in both lists returns the following Table 2.

Table 2: Setups being consensually top-ranked in terms of accuracy- and respectively coverage-focused success scores (*QA*, *QC*) at H thresholds, for both regression problems on GTMs and classification problems on SOMs. For each setup, its rank with respect to regression problems is listed in column U, while US reports its rank within the classification problem pool. "< >" stands for the mean of the two ranks – the criterion by which these setups were selected. On yellow background – the "default" setup corresponding to no trustworthiness considerations and no AD control.

Setup	Rank (QA>H)			Setup	Rank (QC>H)		
$(\tau, \alpha, \beta, MMT; TT)$	U	US	< >	$(\tau, \alpha, \beta, MMT; TT)$	U	US	< >
(0.01,0.10,1.00,mean;0.70)	9	1	5	(0.50,0.01,0.01,mean;0.00)	9	14	11.5
(0.10,0.10,1.00,mean;0.70)	7	5	6	(0.50,0.01,1.00,mean;0.00)	18	10	14
(0.00,0.10,1.00,mean;0.70)	8	6	7	(1.00,0.00,0.00,mean;0.00)	1	31	16
(0.01,0.10,1.00,min;0.70)	11	4	7.5	(1.00,0.00,0.00,mean;0.50)	1	31	16
(0.50,0.10,1.00,min;0.80)	5	10	7.5	(1.00,0.00,0.00,mean;0.70)	2	31	16.5

(0.10,0.10,1.00,min;0.70)	15	2	8.5	(1.00,0.00,0.00,mean;0.80)	4	31	17.5
(0.50,0.10,1.00,mean;0.80)	17	7	12	(0.10,0.10,1.00,mean;0.00)	14	28	21
(0.50,0.10,2.00,min;0.70)	13	26	19.5	(0.01,0.50,2.00,mean;0.00)	45	5	25
(0.50,0.10,2.00,mean;0.70)	22	23	22.5	(0.10,0.50,0.01,mean;0.00)	13	38	25.5
(0.01,0.10,2.00,min;0.50)	27	19	23	(0.10,0.50,0.10,mean;0.00)	22	33	27.5

As expected, optimal parameterization of the trustworthiness-driven AD delimiter will differ respective to whether the focus is set on accuracy or on coverage. Of course, there is no simple relationship between the trustworthiness threshold *TT* and coverage – there is only a local rule stating that at given method, training and test sets, coverage will decrease with increasing *TT*. Otherwise, the fraction of test set predictable at trustworthiness > *TT* first of all depends on the test set, and its degree of overlap with training molecules. It is expected to see rather large *TT* values selected when the focus is on accuracy in left-hand Table 2 (0.7 or 0.8 in 9 out of the 10 setups), and zero (in 7 out of the 10; right-hand Table 2) when exhaustive coverage is preferred. This however does not prevent *QA*-selected setups to occasionally support very high, or even total coverage of test sets.

Higher τ values are also associated to intrinsically enhanced coverage, $\tau > 0$ ensuring that all nodes may be technically used for prediction, including empty nodes. Unless $\tau > 0$, if the dependency of trustworthiness on node density has a non-zero exponent α, empty nodes will not contribute at all to prediction and therefore test compounds having tangible responsibilities only on empty nodes are non-predictable, even at *TT=0*. If $\tau > 0$ or $\alpha = 0$, this "hard" AD exclusion is deactivated: full coverage of any arbitrary test set can be guaranteed, at least at *TT=0*. Notably, the default "AD-less" setup (1.00,0.00,0.00, mean;0.00) implemented by the current GTM predictor unsurprisingly qualifies amongst the top coverage-focused setup strategies but it is not the top one. This is just one of many setups guaranteeing 100% coverage, and it is not the most accurate one. It is closely followed in terms of ranking by analogues at *TT>0*. In those cases, prediction trustworthiness is controlled only by the coherence of prediction $\widehat{\sigma P}(M)$ as in equation (9), since at $\tau = 1, \alpha = \beta = 0$ all nodes are equally trustworthy. Yet, setting *TT>0* does not improve ranking in terms of coverage. In accuracy-focused ranking *QA>H*, this "AD-less" setup is ranked #153 out of 1062, with an absolute success score of 0.14% (roughly twice as small as top accuracy-focused setups).

In terms of Mean-Min-Toggle *MMT*, under coverage-focus, a clear preference is observed for "coloring" empty nodes by the mean of training set compound property values. The situation is less clear when accuracy focus is applied – which is expected, as the latter setups actively downweigh the impact of empty nodes on prediction, hence making the choice of the assigned property largely irrelevant. Yet, if empty nodes must be used for prediction in the name of complete test set coverage, the most rational strategy is *MMT*="mean". This choice has no negative impact on accuracy-focused setups, thus *MMT*="mean" is the universally best option.

With focus on accuracy, the α exponent controlling the impact of node density on its trustworthiness is best set to 0.1 – low, but never zero. Essentially, such a low exponent specifically penalizes empty and nearly empty nodes but has a limited impact for reasonably populated nodes. By contrast, results prone a rather strong dependence (linear, or even quadratic) $\beta = 1 \ldots 2$ of node trustworthiness on node coherence.

Thus, in an accuracy-focused strategy – and independently of the nature of the prediction problems and the underlying grid-based maps – we herein propose to modulate node trustworthiness – see equation (6) – as proportional to $0.01 + ND^{0.1} \times NC$, with a trustworthiness threshold of the order of 0.7 to delimit the AD.

As already mentioned, co-opting empty nodes into the prediction process is a key "strategy" to increase coverage, so coverage-based focus typically allows for even lower $\alpha = 0.0..0.01$, albeit values of 0.5 also appear towards the bottom of the preference list. There is no clear trend in terms of β in the right-hand columns of Table 2 – meaning that there is no unique recipe to scale down contributions from nodes lacking coherence and herewith achieve an improvement throughout the (entire) test set. Node coherence is best used as an out-of-AD trigger for test set compounds with significant residence rates in low coherence nodes. Still, if focus is on coverage then the result quality will be little impacted by the way of modulating node weights in terms of coherence. The first two ranked setups of the coverage-focused scenario are practically not significanlty more performant than the default setup in the third position – thus, the latter can be safely used for predictions in situations when a predicted value should be mandatorily returned for the whole training set ($1.00 + ND^{0.} \times NC^{0.}$). This is notably the case in cross-validation, for failure to do so would result in incomplete set of experimental-predicted data.

3.8 Tracking Prediction Errors on the Trustworthiness Landscape

The strength of grid-based map predictions is that the prediction process can be visually followed on the map, and better understood. Below, the prediction challenges of the affinity (pK_i) of trypsin (CHEMBL209) inhibitors on six universal map landscapes are traced in Figure 3. Five of the landscapes represent the training set (70%, 647 compounds), and surround the central bottom landscape E, representing the external test molecules (30%, 275 molecules) colored by their prediction error (see spectrum bar below), on universal GTM #2 (U_2). Shown training set information is as follows:

(A) represents the binary occupancy of U_2 nodes by training set compounds: all nodes having some tangible cumulated responsibility ($>10^{-5}$) are rendered in black. These are nodes for which a mean property value and its standard deviation are technically calculable based on training data, *i.e.* nodes in which there is at least some neighborhood information "trickling down" into them.

(B) renders the actual density distribution (*ND*) of the training set – from the dark red most populous node (containing > 20 of the 647 training compounds) to the barely visible marginally populated ones.

(C) represents the coherence (*NC*) landscape of the training nodes: darker blue means higher coherence. Coherence improves at the borders to the empty zones and degrades in the denser areas. Marginally populated nodes with cumulated responsibilities between 10^{-5} and 10^{-2} frequently acquire these values as a single contribution of one remote compound. The mean property value results from such single contributing compound, and the associated standard deviation is null. In dense nodes, by contrast, several chemical species with diverging property values are clustered. It is impossible to expect all residents to have strictly the same activity

(D) eventually renders the node trustworthiness score according to the chosen top accuracy-focused setup $0.01 + ND^{0.1} \times NC$. As τ*=0.01*, even empty nodes have non-null trustworthiness, hence the homogeneous background. Thanks to the density modulation, perfectly coherent but marginally populated nodes are not among the most trustworthy. Dense nodes harboring residents with strongly diverging affinity values are being penalized as well.

(E) represents the property landscape used for prediction. It is not density-modulated, thus displaying empty nodes at property value equal to the mean training set pK_i value (which is a remarkably high 6.7, over this QSAR set very rich in potent actives).

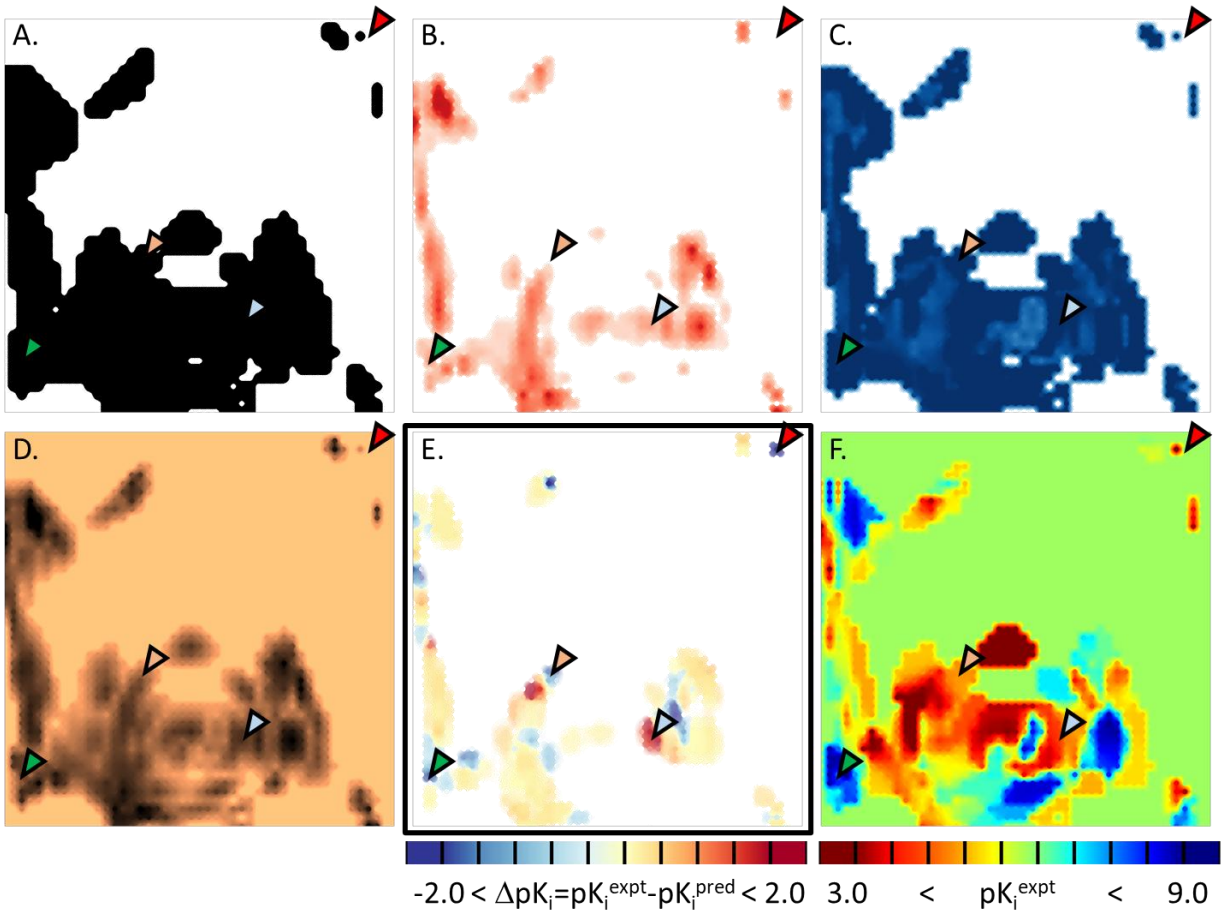


Figure 3: Key landscapes for tracing of prediction errors of trypsin affinity on universal GTM#2. They represent (A) binary occupancy by the training set, (B) training set density distribution, (C) node coherence *NC*, (D) node trustworthiness *NT*, (E) prediction error of test set compounds and (F) the property landscape used for prediction. The four colored pointers correspond to the specific compounds highlighted in the text. Their structures are shown in the following Figures.

The north-western red marker represents a test compound (Figure 4) falling outside the training set-covered zone (Figure 3). Its affinity is set to the training set mean, which is nevertheless two orders of magnitude above its actual affinity. Similarly, the central orange marker pinpoints towards an only slightly more populated zone of average trustworthiness: the in there projected training set compound (Figure 4) is also predicted to be ~100 times more potent than it actually is. There are several more examples of misprediction that may be associated to low local trustworthiness: at *TT*=0, the root-mean-squared pK_i prediction error over the entire test set is of 0.95 log units (R^2=0.67). AD restrictions at increased *TT* translate in better prediction accuracy (at decreased coverage): RMSE=0.91 for 258/275 compounds within the AD at *TT*=0.7, whereas

at *TT*=0.8 only 78 test compounds remain within AD but are accurately predicted within RMSE=0.52.

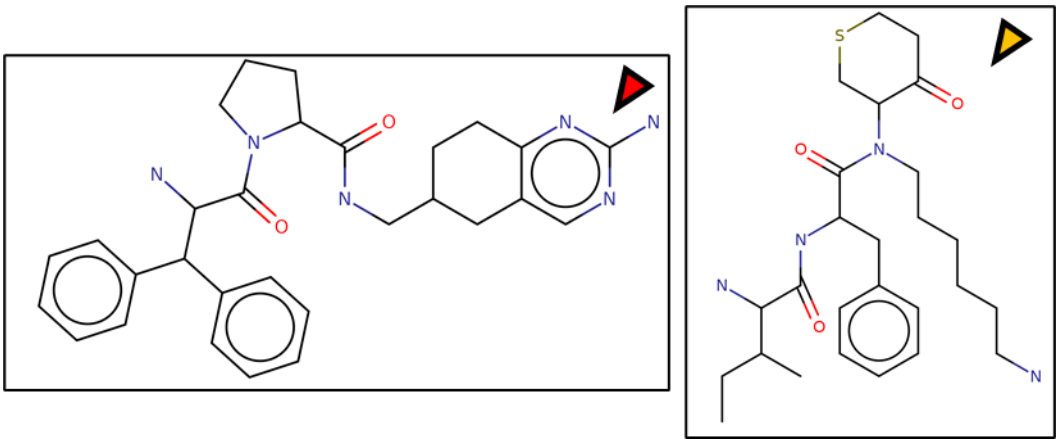


Figure 4: Test set compounds located in zones void of, or sparsely populated by training molecules

However (Figure 5), not all prediction failures can be traced back to lacking node trustworthiness – the light-blue and green markers point towards rather trustworthy map areas. This is particularly true for the former case, representing a zone populated by rather weak training inhibitors of overall linear shape (the two training set inhibitors, gray background in FFF, have pK_i values of 4.6 and 6, respectively). There are no obvious NB violations concerning training compounds – yet, the mispredicted test compound may be the one less well fitting into the area. Its estimated affinity of 5.2 is much lower than its experimental value, of 7.9. Last but not least, the green marker corresponds to a genuine activity cliff – both training and test compounds are clearly similar, but the two former are highly potent (9.1 and 7.4, respectively) whereas the latter (predicted at 8.6) is not (experimental 6.7).

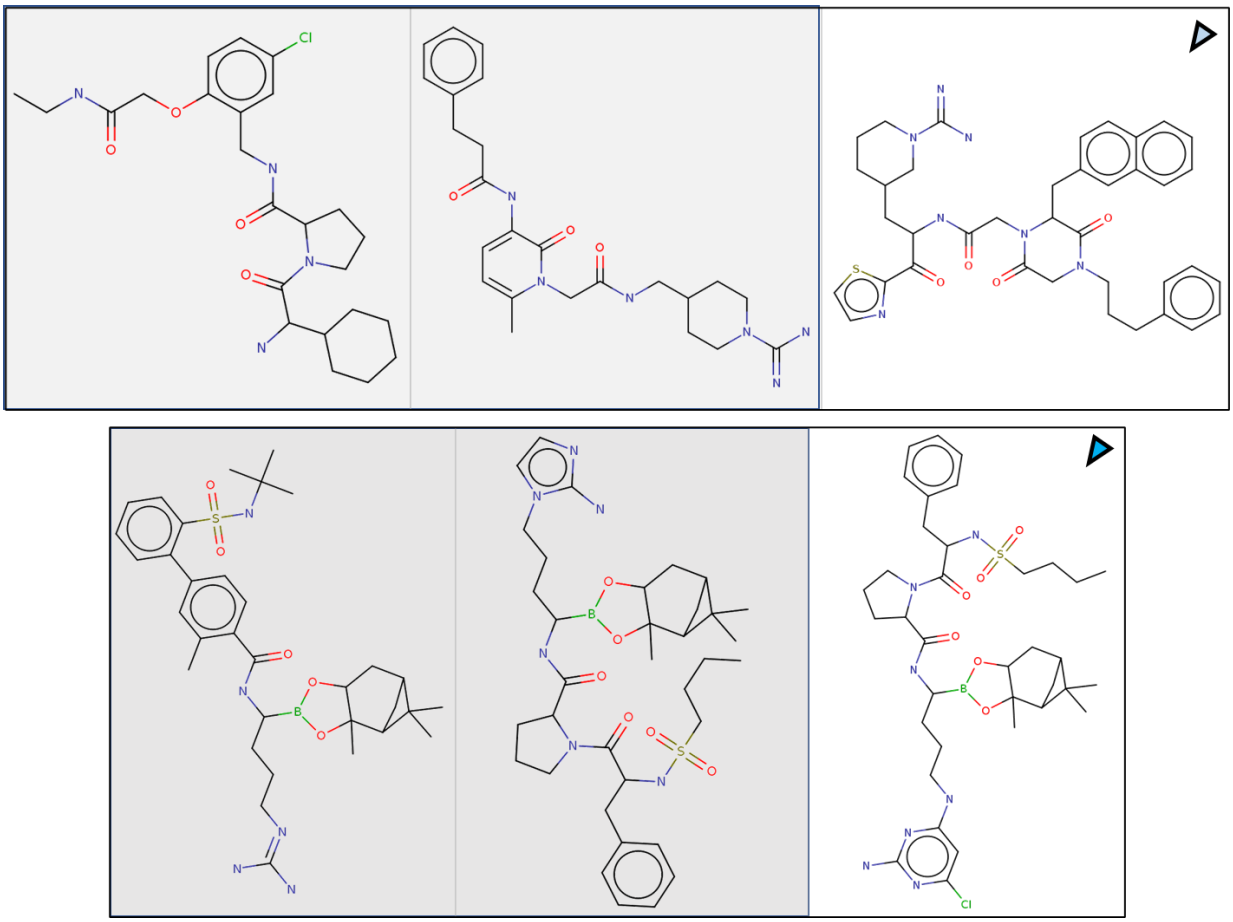


Figure 5: Examples of large prediction errors in relatively dense and coherent map areas (training set compounds are against a grey background, by contrast to the mispredicted test molecule).

4 Conclusions

In chemography, grid-based maps such as SOM and GTM sample molecular descriptor space by injecting a set of nodes, then linking them to some regular 2D grid representing the map. They support property prediction models, because any compound thereupon projected can "inherit" the properties of its residence node(s) – node properties themselves "inherited" from node-neighboring training set compounds. In previous publications, the transparent control of the Applicability Domain of such approaches was often mentioned as one of their inherent strengths. This contribution illustrates how to practically implement such control.

This work is however not an exhaustive approach to all possible AD definitions, but it is one of general applicability to GTM and SOM. The composite AD criterion introduced here integrates two key aspects of applicability: closeness to dense training space zones [18] and data/prediction coherence [19]. Other – not necessarily node-based – AD criteria exist. The GTM-specific log

likelihood criterion[10] may serve to reject species that are too "far" from the manifold. Eventually, nothing would prevent the usage classical "bounding box"-type ADs [20] in initial descriptor space.

This article is a systematic study focusing on the trustworthiness of map nodes as "providers" of structure-activity information captured from training compounds, controlling prediction trustworthiness score as the key to delimit the AD of a predictive landscape. An empirical four-parameter Node Trustworthiness (*NT*) function of density *ND* (sparsely populated nodes are less trustworthy) and coherence *NC* (nodes with training set residents of divergent properties are less trustworthy) is proposed. *NT* is postulated to depend on *ND* and *NC* as $\tau + ND_N^{\alpha} \times NC_N^{\beta}$, where τ, α, β are three of the four above-mentioned tunable parameters. The fourth, the Min-Mean Toggle *MMT* encodes the empirical choice of how to "color" the empty nodes, which are not tangibly populated by any of the training set compounds.

Based upon the *NT* function, a trustworthiness score *T* is defined as the product between the mean *NT* of nodes participating to the prediction and the coherence *PC* of the predicted mean. The role of *T* is to define the Applicability Domain (AD) within a trustworthiness threshold *TT*.

Prediction simulations were run on a large scale, co-opting a significant part of to-date publicly available structure-activity data sets (ChEMBL v. 26). Regression problems were represented by 445 target-specific sets of ligands with reported K_i values, biological targets being as diverse as possible (all with ligand series exceeding 100 members were featured). A series of classification problems was selected to include, out of the in-house curated active/inactive-labeled ChEMBL ligand sets, 319 targets that were distinct from the 445 above.

The previously constructed universal GTMs served on one hand as supports for predictive challenges. On the other, in order to expand the scope of this study to other grid-based mapping techniques, analogous "universal SOMs" were calibrated and entered in the study. This collaterally represented an opportunity to eventually provide quantitative proof to the – presumably true, but previously never explicitly checked – claim that GTM fuzziness is paramount to improve their predictive power over SOMs.

For each parameter setup over all considered "pools" of challenges (combinations of QSAR sets and various maps), success of ensuing inside-AD predictions was monitored. This success is tributary to the end user's needs – in some circumstances, accurate predictions at the cost of discarding large parts of the external set as out of the AD are paramount. By contrast,

compulsory return of a prediction for the entire external set is mandatory. Accordingly, "accuracy-focused" and "coverage focused" success criteria were designed.

It is seen that setup-specific success levels (averaged over large pools of prediction challenges) are highly covariant, irrespectively of the targets of prediction challenges, of the (classification or regression) type of problems, of the specific parameterization and even the nature (GTM or SOM) of underlying maps. Thus, success levels determined on the basis of regression problems on GTMs and levels returned by completely unrelated classification problems on SOMs were seen to correlate to a degree of 70%. Therefore, a common, general-purpose setup of the herein proposed parametric AD definition was shown to generally apply to grid-based map-driven property prediction problems.

It appears that node trustworthiness can be intrinsically defined to characterize any node as a "supplier" of learnt structure-activity information – and this irrespectively of the training compounds, the nature of the learnt variable or the mathematics behind the grid-based mapping mechanism. There are two key distinctions between GTM and SOM. The first is the algorithm defining the coordinates of the nodes: "manifold-based" for GTM and "code vector-based" for SOMs. The other is the nature of the *R* vector: continuous for GTM (albeit in practice often binary within employed numerical precision), binary for SOM. Thus, in a "metaheuristics" space of possible grid-based maps, encoding node localization as 0="manifold-based" *versus* 1="code vector based", respectively the nature of *R* (0=continuous, 1=binary), GTM is metaheuristics (0,0) and SOM its diagonally opposed (1,1). Alternative options (0,1): GTM with forced binary *R* vectors and (1,0): SOM with fuzzy-logics sharing of an item over several near nodes, are technically valid possibilities, albeit not customary ones. Or, the same AD-defining strategy, with the same parameters, applies to both "extreme" metaheuristics (0,0) and (1,1). It thus may be safely assumed to apply to other grid-based mapping techniques, such as the hypothesized (0,1) and (1,0), approaches.

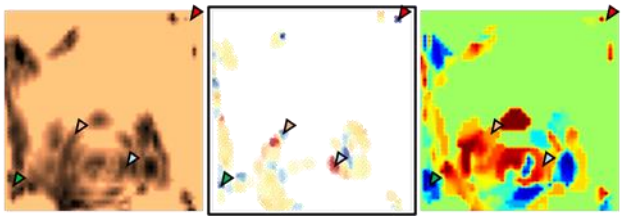
For an accuracy-focused strategy – and independently of the nature of the prediction problems and the underlying grid-based maps – we herein propose to modulate node trustworthiness – see equation (6) – as proportional to $0.01 + ND^{0.1} \times NC$, with a trustworthiness threshold of the order of 0.7 to delimit the AD. If focus is on coverage, then the default approach (not modulating node weights by their trustworthiness and setting *MMT* to "mean") is still one of the best strategies.

Last but not least, the herein introduced trustworthiness criteria represent *per se* mappable properties that characterize chemical space and may help to track down prediction errors and activity cliffs in daily QSAR practice.

5 Supporting Information

Two tar archives provide the curated compound sets and their activity values: RegSets.tar.gz contains <target>.smi_chid_pki three-column files, where <target> are the CHEMBL IDs of the 445 biological targets with sufficient pK_i data in ChEMBL version 26 (a dictionary file reporting the biological names of the targets associated to the CHEMBL IDs is also included in the archive). The three columns contain, as the extension .smi_chid_pki suggests, the (stereochemistry-depleted) standardized compound SMILES, the compound CHEMBL IDs (multiple "+"-concatenated entries if there are several ChEMBL compounds converging to this same stereochemistry-depleted structure) and associated pK_i value (for <target>). In ClassSets.tar.gz the 319 compound series used in classification challenges (these stem from ChEMBL version 23) are reported likewise, except that the .smi_chid_class files report the activity class (1=inactive, 2=active) in their third and last column. A MS Word document, Appendix.docx, provides the more detailed analysis of certain issues mentioned in the main text.

6 Table of Contents Graphic



7 References

1. Oprea, T. I.; Gottfries, J., Chemography: the art of navigating in chemical space. *J Combin Chem* **2001**, 3, 157-166.
2. Papadatos, G.; Cooper, A. W. J.; Kadirkamanathan, V.; Macdonald, S. J. F.; McLay, I. M.; Pickett, S. D.; Pritchard, J. M.; Willett, P.; Gillet, V. J., Analysis of Neighborhood Behavior in Lead Optimization and Array Design. *J. Chem. Inf. Model.* **2009**, 49, 195-208.
3. Horvath, D.; Barbosa, F., Neighborhood Behavior – the Relation Between Chemical Similarity and Property Similarity. *Curr. Trends Med. Chem.* **2004**, 4, 589-600.
4. Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E., Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, 39, 3049-3059.

5. Bishop, C. M.; Svensén, M.; Williams, C. K. I., GTM: The Generative Topographic Mapping. *Neural Computation* **1998**, 10, 215-234.
6. Bishop, C. M.; Svensén, M.; Williams, C. K. I., Developments of the generative topographic mapping. *Neurocomputing* **1998**, 21, 203-224.
7. Kohonen, T., *Self-Organizing Maps*. Springer: Heidelberg, Belin, Germany, 2001.
8. Kohonen, T., *Self-Organization and Associative Memory*. Springer: Heidelberg, 1984.
9. Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping Approach to Chemical Space Analysis. In *Advances in QSAR Modeling. Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences*, Roy, K., Ed.; Springer Verlag: Frankfurt, 2017, pp 167-199.
10. Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A., GTM-Based QSAR Models and Their Applicability Domains. *Molecular Informatics* **2015**, 34, 348-356.
11. 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, 3318-25.
12. Kaneko, H., Data Visualization, Regression, Applicability Domains and Inverse Analysis Based on Generative Topographic Mapping. *Molecular Informatics* **2019**, 38.
13. Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D., Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *Journal of Computer-Aided Molecular Design* **2015**, 29, 1087-1108.
14. Kohonen, T.; Hynninen, J.; Kangas, J.; Laaksonen, J. Helsinki University of Technology, 1996.
15. Lin, A.; Horvath, D.; Marcou, G.; Beck, B.; Varnek, A., Multi-task generative topographic mapping in virtual screening. *Journal of Computer-Aided Molecular Design* **2019**, 33, 331-343.
16. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2011**, 40, D1100-D1107.
17. Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A., Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *Journal of Chemical Information and Modeling* **2019**, 59, 564-572.
18. Tropsha, A., Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, 29, 476-488.
19. Horvath, D.; Marcou, G.; Varnek, A., Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J. Chem Inf. Model.* **2009**, 49, 1762-1776.
20. Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A., Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, 48, 1733–1746.