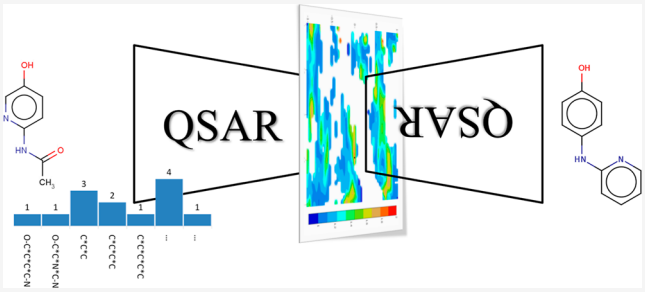
# Inverse QSAR: Reversing Descriptor-Driven Prediction Pipeline Using Attention-Based Conditional Variational Autoencoder

William Bort, Daniyar Mazitov, Dragos Horvath, Fanny Bonachera, Arkadii Lin, Gilles Marcou, Igor Baskin, Timur Madzhidov, Alexandre Varnek

# Inverse QSAR: Reversing Descriptor-Driven Prediction Pipeline Using Attention-Based Conditional Variational Autoencoder

William Bort, Daniyar Mazitov, Dragos Horvath, Fanny Bonachera, Arkadii Lin, Gilles Marcou, Igor Baskin, Timur Madzhidov, and Alexandre Varnek*

**ABSTRACT:** To better formalize the notorious inverse-QSAR problem (finding structures of given QSAR-predicted properties) is considered in this paper as a two-step process including (i) finding "seed" descriptor vectors corresponding to user-constrained QSAR model output values and (ii) identifying the chemical structures best matching the "seed" vectors. The main development effort here was focused on the latter stage, proposing a new attention-based conditional variational autoencoder neural-network architecture based on recent developments in attention-based methods. The obtained results show that this workflow was capable of generating compounds predicted to display desired activity while being completely novel compared to the training database (ChEMBL). Moreover, the generated compounds show acceptable druglikeness and synthetic accessibility. Both pharmacophore and docking studies were carried out as "orthogonal" *in silico* validation methods, proving that some of *de novo* structures are, beyond being predicted active by 2D-QSAR models, clearly able to match binding 3D pharmacophores and bind the protein pocket.




## 1. INTRODUCTION

Predictive quantitative structure−activity/property relations (QSAR/QSPR)[1] are regression or classification models that are able to compute, upon input of a molecular structure, an estimate of the activity/property value the compound is expected to display. One may formulate the above as activity = $f$(structure), where function $f$ needs first to be calibrated in order to have $f$(structure) returning accurate approximations of known activity values. If the above holds, then *inverse mapping* would allow to retrieve the "optimal" chemical structure(s), maximizing the expectancy of having an activity matching the input argument, that is, the desired activity level needed to achieve success in the current research project.

Since the first pioneering linear regression model by Hansch and Leo,[2] procedures to "fit," for example, machine learn $f$(structure), have progressed to the point of routine calibration of nonlinear models based on a plethora of machine learning methods (support vector machines, partition trees, neural networks—to cite only the most popular[3−7]).

Typically, the *structure* argument in $f$(structure) is the molecular graph with vertices colored by chemical elements and edges colored by bond types. Since $f$(structure) returns a real number, it is obvious that the information content of the input molecular graph could first be translated in this process into some purely numerical representation—a vector of $N$ real numbers $\vec{D}$ known as the "molecular descriptor vector." In classical QSAR, the two formal steps, descriptor calculation $\vec{D}$ = $\theta$(structure) and model fitting, activity = $\mu(\vec{D})$ are clearly separated into successive steps, and hence activity = $\mu(\theta$-(structure)) = $f$(structure). Hence, the inverse QSAR problem may be conceptualized as a succession of two formal steps:[8−10]

1. finding descriptor vectors ("seed vectors") matching the desired activity level: $\vec{D} = \mu^{-1}$(activity)
2. finding the structures that correspond to the $\vec{D}$ above: structure = $\theta^{-1}(\vec{D})$

Since $\mu$: $\mathbb{R}^N \rightarrow \mathbb{R}$, searching extremal points of $\mu(\vec{D})$ is a standard optimization problem, and albeit solving may prove challenging when $\mu$ is highly nonlinear or if $N$ is large, this step of inverse QSAR is conceptually an easy one.

By contrast, step 2 is both technically and conceptually hard—to the point that, until recently, the typical way to discover molecules with activity values matching a desired activity level is to enumerate candidate structures and apply, to each, the QSAR model until all input candidates were herewith "virtually screened[11,12]" or until enough events $f$(structure) ≈ desired activity occurred, for example, "virtual hits" were

**Received:** August 27, 2022

found. Virtual screening (VS), however, is limited by the choice of candidate structures either from public/commercial databases or from user-designed virtual libraries. In contrast to systematic VS, sampling techniques of chemical structures consider molecular structure as evolvable.[13−15] This is *de novo* design,[16−23] which fundamentally differs from VS by the fact that structures are not a predefined library but are generated and/or modified "on the fly" by some automated molecular structure editor.

The recent advent of deep neural networks (DNNs), able to extract information from arbitrary "brute" data and herewith learn to recognize patterns, had a major impact in the field of QSAR.[24−28] The idea of DNNs is mimicking a human brain in which neurons communicate by generating and passing signals. Along with many applications of DNNs, Rana *et al.*[29] reviewed the application of the simplest example of DNN models—multilayer perceptron (MLP)—to disease diagnostics. MLP was also shown as a method to build successive QSAR models.[30] Later, parsing a chemical structure given in the form of a SMILES string by DNNs using the natural language processing technique was proposed as a new approach for QSAR model training.[31] This success was not the end, and soon graph convolutional networks were proposed as a replacement of recurrent neural networks (RNNs) in QSAR modeling.[32] As the research domain is in full effervescence, an exhaustive overview of already envisaged DNN architectures is beyond the scope of this article. The reader is encouraged to access the most recent reviews.[33]

Some DNN architectures, namely, autoencoders, relate input structure (simply rendered as SMILES[34] ) to activity within a unique computational framework, apparently bypassing the need for molecular descriptors in QSAR. *De facto*, SMILES string encoder architectures first translate structure to a "latent" real vector $\vec{L}$, which the associated decoder would use to regenerate the SMILES. Thus, $\vec{L}$ is nothing but a machine-generated molecular descriptor vector. Therefore, the decoder is a deep-learning-based model based on latent space descriptors $\vec{L}$ implicitly allowing for a solution to the inverse problem.

So far, the majority of QSAR models are still based on classical, human expert-designed descriptors. This is first due to historical reasons, latent space descriptors $\vec{L}$ being very new. However, expert-designed descriptors $\vec{D}$ may still have a key advantage over the former (such as atom order invariance, which may be an issue in $\vec{L}$ spaces—and their support of relatively small training sets in contrast to "big data"-dependent DNN approaches). So far, only a few attempts to convert arbitrary descriptor space $\vec{D}$ back to structure have been described. One work[35] reports two distinct RNN-driven approaches labeled PCB (physchem-based) and FPB (fingerprint-based). The former inputs a vector of predicted physicochemical properties (including a QSAR-predicted bioactivity value) to generate SMILES strings of compounds matching these properties. The latter uses Morgan fingerprints for input. Similarly, a transformer architecture has been implied to "translate" various classical chemoinformatics fingerprints back to structure.[36] Both works can be considered as examples of "hard" inverse QSAR approaches and were successfully used to generate structures in the neighborhood of known actives. However, they stopped short of coupling "easy" and "hard" QSAR problems in order to investigate how their approaches would cope with input vectors corresponding to optima of the QSAR landscape, not to already known molecules.

For the above reasons, the current contribution wishes to explore the feasibility of a genuine solution for the inverse QSAR problem for models based on classical, expert-defined molecular descriptors. The core of this work consists in the development of an attention-based conditional variational autoencoder (ACoVAE) based on transformer architecture. Given the seed vectors of ISIDA fragment descriptors, the ACoVAE generates corresponding molecules.

We have used two types of in-house generated QSAR models of ABL tyrosine kinase 1 (CHEMBL1862) activity:

1. Support vector regression (SVR) models for the inhibition constant ($pK_i$) using $\vec{D}$ = ISIDA[37,38] circular fragment counts. Seed vectors prepared with the help of a genetic algorithm used to sample $\vec{D}$ space with predicted $pK_i$ value as fitness.

Additionally, the descriptor vector of the molecule possessing the highest affinity ("lead molecule" LM) from the CHEMBL1862 set was also used as a seed vector.

2. Generative topographic mapping (GTM)-based predictive activity class landscapes using the "universal" map[39] based on $\vec{D}$ = force field-type colored[40] ISIDA atom sequence counts. Sampling of $\vec{D}$ was performed around the coordinates of active-enriched nodes of the landscape.

The inverse QSAR problem is considered solved if (i) the obtained structures are valid and chemically feasible and (ii) the obtained structures are submitted to classical forward QSAR model prediction and return conveniently high activity values.

Here, the ultimate goal was to obtain *de novo* structures that are perceived by a QSAR model to be highly active—whether they really are active or not is a question of underlying model quality, not of the quality of the inverse QSAR approach. Nevertheless, an alternative orthogonal *in silico* validation of these structures as ligands of the considered targets has been performed by pharmacophore analysis with the LigandScout[41] program and by docking using both LeadIT[42] and S4MPLE[43] approaches.

## 2. METHODS

**2.1. ACoVAE.** The proposed ACoVAE transformer model is shown in Figure 1. It consists of three main parts:

(1) During the training procedure, a GRU-based encoder parametrizes a random latent vector distribution based on the training set SMILES. Hyperspherical distribution with zero mean and variance equal to 1 is used as target latent vector distribution;

(2) A condition vector encoder uses a grouped linear transformation (GLT) layer[44] to transform initial descriptor vectors to a conditional latent vector;

(3) A standard autoregressive multihead attention decoder[45] translates from condition and random latent vectors to SMILES. A more detailed architecture of the network is given in Supporting Information, Figures S1 (training stage) and S2 (inference stage). During the training, a SMILES strings and their corresponding descriptor vectors are used to train the ACoVAE. A reparameterization trick for latent vector sampling is used to train the network end-to-end. In the inference stage, the latent vector is sampled from a prior (0, 1) hyperspherical distribution, and a desired descriptor vector is used as

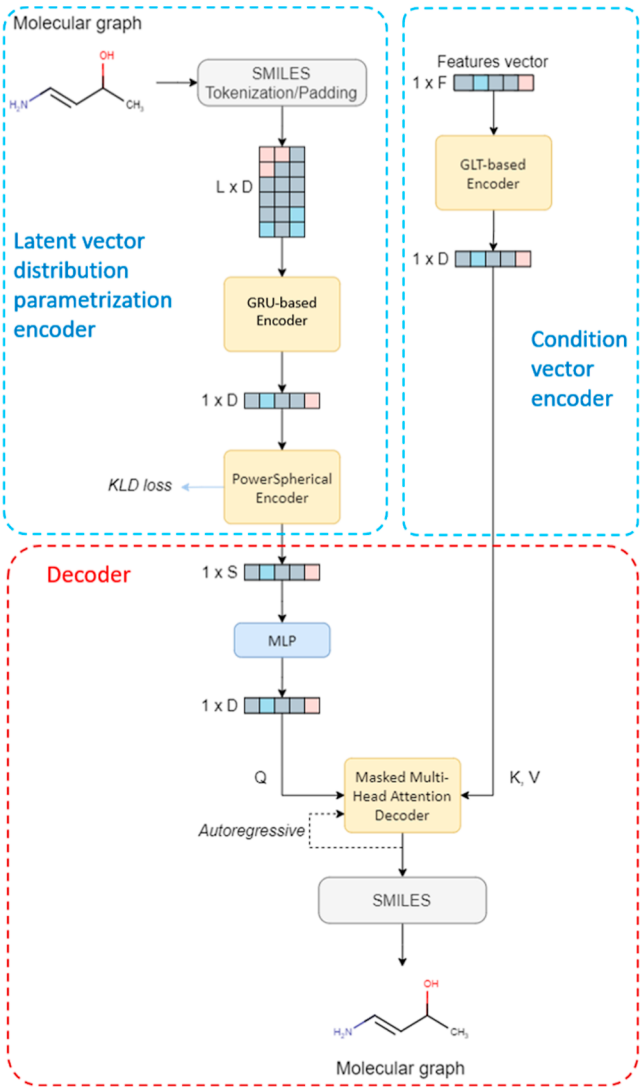



**Figure 1.** General scheme of the ACoVAE architecture used in this study. The GRU-based encoder (top left) parametrizes SMILES into latent vectors following a hyperspherical distribution, which is used upon inference for random sampling. The descriptor vector which is used as a condition in the generation is embedded by a GLT layer (top right). Autoregressive transformer is used to decode random latent vectors and combined conditions into SMILES strings. A detailed representation of all three networks is given in the Supporting Information.

condition. Based on the random and condition vector, the decoder generates a wanted SMILES. Notice, that alternative SMILES for a given condition descriptor vector can be generated both (i) by running inference stage with different random vectors sampled from a prior distribution and (ii) by sampling different text strings using categorical sampling from token probabilities predicted by the transformer for a given random and condition vector.

The proposed architecture of the ACoVAE transformer was inspired by the one proposed by Lin *et al.*[46] In a similar way, a random latent vector is fed as a START token. However, substantial changes were introduced which helped us to achieve better performance. In our architecture, a random latent vector is encoded directly using a GRU, while Lin *et al.* used a trick with a priori undefined random distribution parameterized by a separate network. Additionally, a hyperspherical uniform distribution was preferred to a standard Gaussian one because during the tuning stage, the former performed better. A von Mises−Fisher distribution is commonly used for sampling from hyperspherical uniform distribution[47] with the reparameterization trick. However, we found that the power spherical distribution[48] used instead of von Mises−Fisher one allows a speeding up of the learning process without loss of the performance. Application of a GLT transformation layer[49] better translates the descriptor vector into the internal representation used by the decoder network than MLP. Finally, inspired by the GELU approximation,[50] new activation function FTSwishG resulted from some modifications of the previously reported FTSwish[51] was used throughout the ACoVAE network

$$\mathrm{FTSwishG} = \mathrm{RELU}(x) \times \mathrm{sigmoid}(1.702x) - 0.2 \qquad (1)$$

According to our tests, it gives better results compared to the ReLU, GeLU, and FTSwish activation functions. In such a way, our ACoVAE transformer architecture is a novel one, having only a few in common with the one proposed by Lin *et al.*[46] The designed architecture is implemented using the TensorFlow framework and can be readily retrained for other descriptor types. It is available on our GitHub storage https://github.com/Laboratoire-de-Chemoinformatique/ACoVAE.

**2.2. SVR Models.** A series of ligands for ABL tyrosine kinase (CHEMBL1862) from the ChEMBL v.23 database was standardized using a protocol reported by Sidorov *et al.*[39] SVR models for thermodynamic instability constants of protein−ligand complexes ($pK_i$) were generated using the evolutionary *libsvm* model tuner,[52] which supports selection of the best suited descriptor space yielding to best performance models as a key hyperparameter. The best-suited ISIDA fragmentation schemes were defined together with the SVR-specific parameters (kernel type, cost, $\gamma$, *etc.*) optimizing model quality. The models were built on a training set containing 739 molecules and validated on a test set of 82 molecules. The test set data were collected from recent publications posterior to model training. The best model relies on IIRAB-1-3 ISIDA fragment count descriptors (7372 atom-centered fragments with a radius of 1 to 3 atoms with restricted fragmentation) and the Gaussian kernel option. It displayed a reasonable performance in cross-validation ($R^2$ = 0.79 and RMSE = 0.70) and on the test set ($R^2$ = 0.80 and RMSE = 0.67).

Computation of the "optimal" seed vectors has been confided to an evolutionary heuristic browsing through the $\vec{D}$ space in search of vectors maximizing computed $pK_i$ values. The "chromosome" of the approach is a 20-dimensional integer vector in which loci may contain either zero or a number denoting a training set compound. The vector encoded by such a chromosome is taken as the mean $\langle\vec{D}\rangle$ of descriptor vectors of the training set compounds mentioned in the chromosome (a compound may be mentioned several times in different loci, which amounts to increasing its weight in the computed average). The fitness score of the chromosome is nothing but the corresponding $pK_i$ = SVR($\langle\vec{D}\rangle$) to be maximized. Hence, the evolutionary algorithm is bound to find, by applying cross-over and mutation operators, chromosomes enumerating optimal sets of training set compounds, with the property that the centroid of the descriptor vector of the set is predicted to correspond to high affinity values. The procedure was applied for each SVR model for 150,000 generations. Sampled "high-affinity" $\langle\vec{D}\rangle$ values

were used as the condition vector for the ACoVAE decoder. Details about evolutionary model building can be found in our publication,[52] which also provides instruction on how to obtain and download that tool. Here, it was used with default setup, meaning 12-fold-repeated three-fold cross-validation (with steadily reshuffled cross-validation tiers at every iteration). The model fitness score was the mean cross-validated determination coefficient $\langle Q^2 \rangle$ penalized by 1 standard deviation, fitness = $\langle Q^2 \rangle - \sigma(Q^2)$.

**2.3. GTM Landscape-Driven Models.** GTM is a dimensionality reduction technique developed by Bishop *et al.*[53,54] The method performs a nonlinear projection of an *N*-dimensional space onto a 2D latent space. The former corresponds to the descriptor space, where each molecule is defined by an *N*-dimensional molecular descriptor vector. The 2D latent space corresponds to a manifold which is defined by a set of radial basis functions and evaluated on sample points called "nodes." Simply put, the manifold can be seen as a rubber band that can be folded in *N*-dimensions during training to fit the data distribution in a way maximizing its coverage of the space zones populated by relevant items (the "frame set"). Any compound can subsequently be projected on the manifold. For visualization purposes, the manifold is "unfolded" into a 2D plane, organizing the nodes into a square grid. GTM is a probabilistic method, meaning that compounds are fuzzily projected on all nodes of the manifold. As such, an item is associated with ("resident in") each node with different probabilities. The sum of the probabilities—technically named responsibilities—over all nodes of the manifold equals 1. In practice, this means that one compound will be defined by a responsibility "pattern" potentially involving several nodes instead of being confined to one node only. When projecting compounds of experimentally known properties, neighborhood behavior[55] (NB) compliance implies that residents of the same node should have related property values, so that the node may be seen to "represent" that local average property, and "colored" accordingly. Resulting property "landscapes" are nothing but NB-driven QSAR models: the property of any external item can be predicted from the "local color" of the landscape zone onto which it is projected. In this work, the fuzzy class landscapes (monitoring the likelihood to classify as "active" with respect to a target) were employed. They were based on the previously published[56] universal map #1 (UM1)—the first of a series of GTMs parameterized (using ChEMBL data), such as to maximize their "polypharmacological competence," that is, their ability to host a large battery of highly predictive fuzzy class landscapes associated with diverse biological targets. Note that landscape-based QSAR models are parameter-free (the landscapes are built by projection of existing structure−activity data on the given manifold in an unsupervised manner). Therefore, landscape-based QSAR models are implicitly available as soon as the supporting structure−activity data are available.

The structure−activity data set associated with the CHEMBL1862 target was projected on the manifold of the first universal map UM1[56] and was seen to "spontaneously" segregate into zones populated predominantly by "actives" and "inactives," respectively. This map was built based on ISIDA[40] atom sequence counts with a length of two to three atoms labeled by CVFF force field types and formal charge status (IA-FF-2-3-FC). Recall that construction of activity landscapes on a given GTM manifold is not supervised but a purely deterministic procedure. The separation proficiency of the considered manifold was obtained by repeated leave-1/3-out cross-validation, in which iteratively two-third of the items are projected on the map in order to "color" the activity class landscape, whereas the remaining one-third of compounds *a posteriori* projected onto that landscape and have their activity classes assigned on basis of their residential zones in the landscape. Cross-validated balanced accuracy was 0.78, significantly above the randomness threshold of 0.5. The structure−activity dataset is herewith proven to be robust and modelable by both machine-learning (SVR) and neighborhood analysis-based mapping.

Activity class landscape for CHEMBL1862 was used to identify zones in the chemical space in which "active" compounds tend to cluster preferentially. Note that the label "active" was assigned to compounds with the ∼25% highest affinity values according to the initial automated data curation procedure used for universal map fitting. The GTM nodes *n* in which active compounds were seen to preferentially reside were identified as key points if

$$\frac{\sum_{c \in \text{Actives}} R_{cn}}{\sum_{\text{all } c} R_{cn}} \gg \frac{N_{\text{Actives}}}{N_{\text{all}}} \tag{2}$$

$R_{cn}$ represents the responsibility of compound *c* with respect to node *n*, summed over actives (numerator) and over all training compounds (denominator), with the ratio representing the fuzzy-logic propensity to expect an active "resident" in node *n*. This propensity should be much higher than the baseline propensity to encounter an active throughout the training set (top nodes were selected according to the ratio of summed responsibilities). Coordinates of these key nodes correspond to vectors in ISIDA descriptor chemical space zones expected to harbor active compounds. The Gaussian neighborhoods of key node vectors were sampled by generating a multidimensional Gaussian distribution with a width of $w = 0.05$. Several vectors were generated from the initial node vector using this method.

**2.4. Solution of Inverse QSAR Problem: The ACoVAE Algorithm.** Sampling with the ACoVAE transformer is accomplished by giving a descriptor vector to the trained decoder part of the model. Each descriptor vector, which corresponds to the "condition" part of the ACoVAE, is combined with a batch of random vectors from a power spherical distribution, which serves as the basis for the latent space. Each descriptor vector/random latent vector combination returns a sample of generated SMILES. Categorical sampling is the preferred method of generation since it allows, for the same input, to explore different possibilities, thus maximizing the generative "coverage." Therefore, the batch of latent vectors returns a batch of generated SMILES. For example, for one descriptor vector concatenated with 200 different sampled random vectors with a batch size of 512, the algorithm returns 200 × 512 = 102,400 generated SMILES. In such a way, a given descriptor vector can be used several times leading to different SMILES. In-house CGRtools[57] software is used to verify the validity of the generated text string, directly removing any incoherent or incorrect SMILES.

The following parameters were analyzed when monitoring the pertinence of the inverse QSAR approach:

1. *Validity = #valid SMILES/#all generated text strings*, which measures success to generate a syntactically valid SMILES string (assessed by CGRtools), starting from the input "high-affinity" $\langle \vec{D} \rangle$ vectors.

2. *Feasibility* assessing chemical feasibility and drug-likeness according to Ertl[58] and QED[59] indices.

3. *Novelty*. A compound generated with ACoVAE is considered "novel" if it is not contained in the training database.

A coherence between the ISIDA descriptor vector recalculated for the generated SMILES string and the input vector at the source of that SMILES was assessed using the Tanimoto similarity score.

**2.5. Filtering of Nonvalid SMILES Strings.** During the sampling procedure, output SMILES were parsed and standardized using CGRtools. Then, they were transformed into Kekulé form followed by verification of valences. If no error detected, the SMILES strings were rearomatized and then written to the output. Failure of any step in this workflow leads to discarding the given text string as invalid SMILES.

## 3. RESULTS AND DISCUSSION

**3.1. Finding Candidate Descriptor Vectors Associated with High Affinity.** For the SVR model, the evolutionary sampler of the ISIDA descriptor space outlined in Section 2.2 is very fast to visit "high-affinity" $\langle\vec{D}\rangle$ values. Points in the ISIDA descriptor space corresponding to predicted p$K_i$ values close to the ones of the most active compounds included in the training set can be discovered in matter of tens of minutes on Linux workstations with the following specification: Intel Xeon Silver 4214 2.20 GHz, 48 cores, 64 GB RAM, Ubuntu 18.04.6 LTS. However, the discovery of points with activities predicted to be *better* than the one of the best training compounds was never achieved despite of the total run times of the order of 48 h, resulting in >150 K visited $\langle\vec{D}\rangle$ values. On the one hand, it is not clear whether such points may actually exist—SVR may suffer (in particular when based on the Gaussian kernel) from the "regression towards the mean" effect, consisting of systematic underestimation of high and overestimation of low property values. Moreover, it is even less likely that points where the SVR model nevertheless predicts a value beyond the largest observed p$K_i$ would actually be located within the "fragment control bounding box" defining the applicability domain[54] (AD) of the model. Given the fact that herein visited $\langle\vec{D}\rangle$ values are generated as means of descriptor vectors of randomly selected subsets of compounds, these points are guaranteed within the bonding box AD (each vector element $D_i$ will be larger or equal than the minimal and, respectively, smaller or equal than the maximal $D_i$ value ever encountered within the training set). Third, the top affinities for all these targets are already within the 0.1 nM range—discovery of significantly more potent molecules is extremely unlikely in this context. Therefore, the five visited $\langle\vec{D}\rangle$ values corresponding to the highest predicted p$K_i$ scores (comparable but not better than the affinity of the most active compound) were used to tackle the inverse QSAR problem (see Figure 2).

As a complementary study to the inverse-SVR descriptor selection, the most active ChEMBL compound shown in Table 2 (compound **A**) was selected as a seed to show the difference between the generation from optimized vectors and a real active molecule.

For the GTM-based activity class predictors, two nodes that were most highly enriched in "active" residents were selected, as represented in Figure 3. Candidate descriptor vectors were obtained by augmenting the $D$ space coordinates of these nodes with Gaussian noise as described in the Methods section

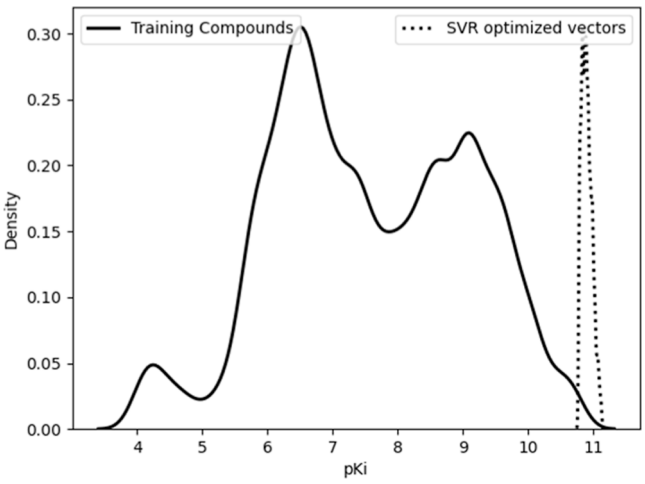



**Figure 2.** Distribution of p$K_i$ for the compounds used to train the model. The dotted line renders the distribution of predicted p$K_i$ for the vectors of the final population emerging from the evolutionary sampling approach.

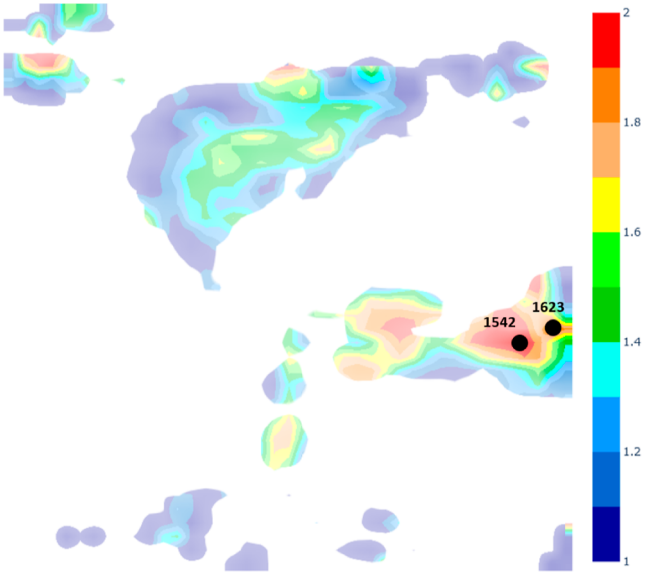



**Figure 3.** Selected nodes for target ChEMBL1862 on the fuzzy activity class landscape where color encodes the relative populations of actives (class 2, red when pure) *vs* inactives (class 1, blue when pure). Intermediate color design nodes with residents of both classes in various proportions. Numbers of the node are represented.

(see 2.3). Projection of these seed vectors on the landscapes below unsurprisingly assigns quasi-unitary responsibility values to their "source" nodes, implicitly qualifying them as "probable actives."

**3.2. ACoVAE Calibration Results.** Two distinct ACoVAEs were trained—one for each relevant ISIDA descriptor space:[40] IIRAB-1-3 for the inverse-SVR problem and IA-FF-2-3-FC for the inverse-GTM challenge. Each training set contained the same 1,540,615 compounds from ChEMBL-23, standardized using ChemAxon[60] standardizer, following the procedure implemented on the VS server of the Laboratory of Chemoinformatics in the University of Strasbourg (http://infochim.u-strasbg.fr/webserv/VSEngine.html). The following standardization steps were applied: (i) dearomatization and final aromatization according to the "basic" setup of the

ChemAxon procedure (heterocycles like pyridone are not aromatized), (ii) dealkalization, (iii) conversion to canonical SMILES, (iv) removal of salts and mixtures, (v) neutralization of all species, except nitrogen(IV), and (vi) generation of the major tautomer with ChemAxon. This resulted in 1,540,615 unique, stereochemistry-depleted SMILES strings used for training (stereochemical information was removed because the herein used molecular descriptors do not capture it).

Model training was done for 100 epochs and lasted for about 30 h on a QUADRO RTX 6000 graphic card. The loss function tends to stabilize early during training as shown in Figure 4; however, the model continues to learn as character-specific reconstruction rates and pure reconstruction rates continue to grow. Arguably, the model could be trained for somewhat longer since the reconstruction rate (*val_rec_rate*) has seemingly not reached a plateau at 100 epochs. However, we believed that the achieved accuracy—some 50% reconstruction rate and 98% character-specific reconstruction rate, was sufficient for the model acceptance. Notice that variational autoencoders have a tendency for lower reconstruction rates than their deterministic counterparts because of the element of randomness introduced by sampling latent vectors from a given distribution instead of having deterministic latent vectors.

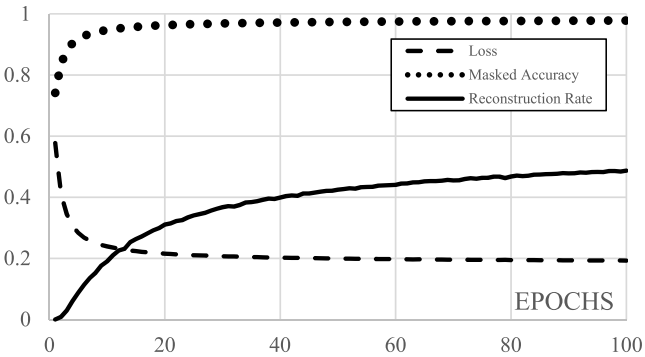



**Figure 4.** Training metrics for the ACoVAE transformer model based on ISIDA descriptors. "Loss" is the loss function of the model. "Masked accuracy" corresponds to the character-specific reconstruction rate. "Reconstruction rate" corresponds to the full SMILES string reconstruction rate.

### 3.3. Inverse QSAR Results.

*3.3.1. Inverse-SVR and Inverse-Lead Compounds.* According to Table 1 displaying various quality criteria of inverse-SVR compounds, the low success rate in the sampling procedure can be mitigated if we consider the time factor. Sampling of 512.000 SMILES strings (using 5 conditional vectors corresponding to the 5 vectors of highest activity predicted by the SVR model) resulting in 6899 valid, unique candidates takes only about 4 to 5 h on a QUADRO RTX 6000 GPU. Comparing lead molecule sampling to inverse-SVR sampling shows that both perform similarly in terms of unique valid compounds and activity prediction, although lead molecule sampling scores a bit lower on the latter metric.

**Table 1. Performance of the ACoVAE Transformer Model for the CHEMBL1862 Target When Sampling from Seed Descriptor Vectors from Different Sources**

| seed vector source | number (percentage) of valid compounds | number (percentage) of unique compounds | novelty compared to ChEMBL (%) | predicted active[a] (%) |
|---|---|---|---|---|
| SVR | 12,432 (2.43%) | 6,899 (55.49%) | 100 | 48.6 |
| GTM | 70,684 (13.8%) | 61,342 (86.78%) | 99.98 | 6.9 |
| lead molecule | 23,559 (4.60%) | 7,600 (32.26%) | 99.95 | 41.6 |

[a]"Predicted active" implies predicted $pK_i > 7$ by the SVR model. This latter is more stringent than GTM landscape-based predictions, which positions a vast majority of inverse-GTM compounds close to their "source" nodes and herewith classifies them as "actives."

A descriptor vector marking a position in the chemical space may or may not translate to a chemically meaningful structure, knowing that the initial vector is typically not a slightly perturbed position vector of a real molecule but merely a chemical space point associated with high predicted activity according to a machine-learned, action mechanism-agnostic model. However, the ACoVAE decoder process injecting randomized latent vectors (see Section 2.1) may produce an arbitrary number of SMILES strings based on a given chemical space point. For each of the five considered chemical space points of high predicted affinity, chemically meaningful molecules were obtained (at a low success rate of 1.34%—but this is merely an order of magnitude of the likelihood to draw a random latent vector *i.e.,* "compatible" with the current chemical space position). The complexity of the molecule that the model is trying to generate is implicitly affecting the chance to retrieve a valid structure. Since the model generates SMILES strings, it must conform to a very specific grammar which is intolerant to errors. Any misplaced character in the SMILES sequence can render it incorrect and bring up an error—a well-known problem in chemoinformatics. Without extensive understanding of the chemical meaning behind a SMILES string, it can be very difficult to correctly open and close multiple rings to recreate valid structures with correct aromaticity and stable behavior. This, in part, explains why the model may be very successful in some parts of chemical space and struggle more in other parts. A possible solution to that problem would be the use of DeepSMILES[61,62] or SELFIES[63] which use a simpler syntax eliminating the risks of incorrect ring closures and parenthesis errors.

GTM landscapes identify zones enriched in actives, nevertheless containing some inactives. The sampling is performed using an ensemble of seeds generated from a given GTM node. These seeds can occasionally be located in the vicinity of inactives. In contrast, sampling from the most active compound generates structures similar to this seed. This explains the difference in the proportion active/inactive for different seeds in Table 1.

Generated compounds were filtered to remove both chemically inconsistent species (by CGRtools) and duplicates and were compared to the initial training database (ChEMBL) to compute the "novelty" rate which corresponds to the percentage of valid unique generated compounds not appearing in the training set of the model. Table 1 shows that all generated compounds are novel. The trained SVR model was used to estimate the $pK_i$ values of the generated compounds, which were then classified as actives or inactives by using a threshold 7. As such, about half of the generated compounds were predicted to be active.

Compounds predicted as inactives by the model were filtered out. Generated compounds were compared to the GA-optimized vectors used as input to the model. Results in Figure

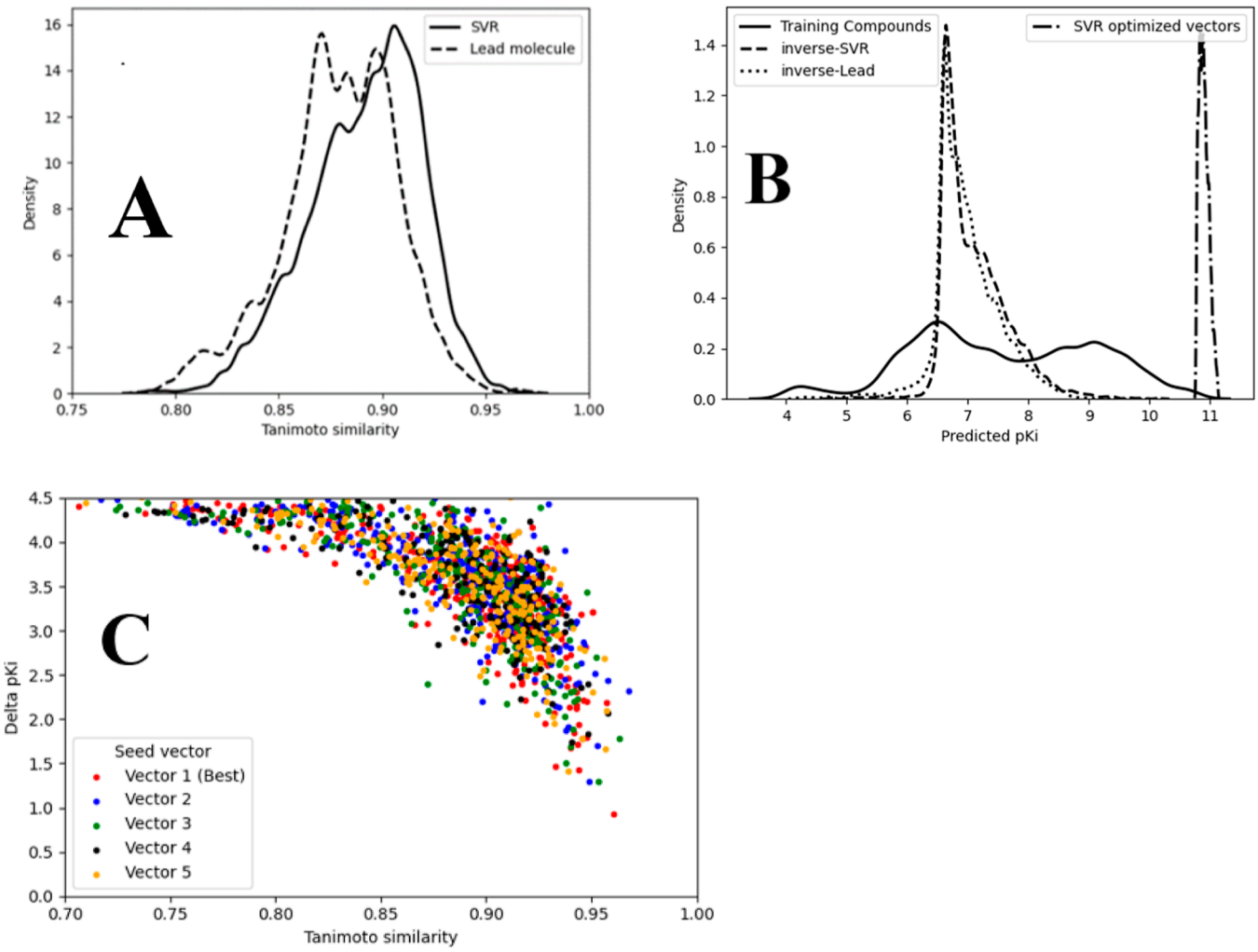



**Figure 5.** (A) Distribution of Tanimoto similarity calculated between sampled compounds and the ISIDA descriptors used for their sampling (obtained *via* SVR GA and lead molecule). (B) Distribution of predicted activities for inverse-SVR compounds, lead molecule sampled compounds, training compounds, and vectors optimized by GA. (C) Scatter plot with the *x*-axis being the Tanimoto similarity between the sampled compound and the GA vector and the *y*-axis, the difference in (calculated) p$K_i$ between the inverse-SVR compounds and the original GA vector. The different colors correspond to the five different "seed" vectors used for the sampling procedure.

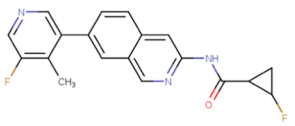
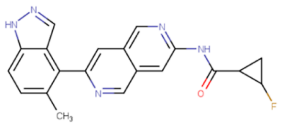
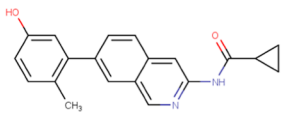
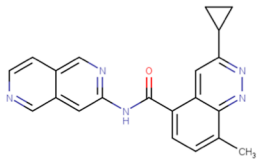
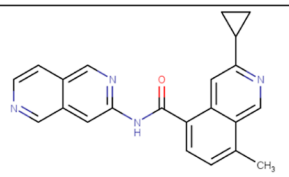
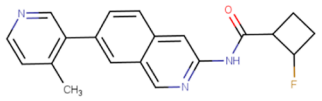
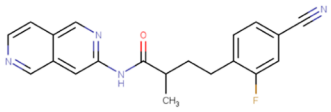
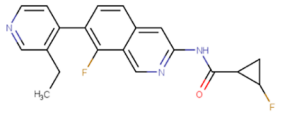
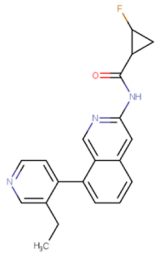
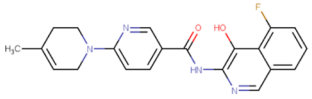
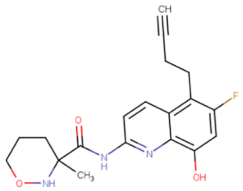
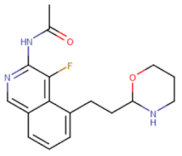
5A show that most compounds are very similar ($T_c$ > 0.85/0.90) to their "seed," meaning the model was able to understand the information contained in the descriptor vector and translate it in terms of SMILES. Given that the value contained in the vectors may not be integers or that some of the descriptor values may be incompatible, an average of $T_c$ = 0.9 is a sign that the model was able to extract hidden knowledge from the ISIDA descriptor and adapt it to a chemically feasible structure. Some generated compounds approach the activity values of the GA-optimized vectors as shown in Figure 5B, although all active compounds have lower p$K_i$. Figure 5C shows the difference in predicted p$K_i$ between the generated compounds (based on their actual $\vec{D}$ vectors) and the "source" GA-optimized vectors $\overrightarrow{\langle D \rangle}$, plotted against the Tanimoto coefficient $T_c(\vec{D}, \langle \vec{D} \rangle)$. Unsurprisingly, the SVR QSAR models are neighborhood-behavior compliant: the closer the source vector $\langle \vec{D} \rangle$ remains to the actual compound descriptor, the higher the likelihood to have the latter predicted at high affinity levels—(virtual) activity cliffs notwithstanding (p$K_i$ shifts of 2 orders of magnitude may occasionally happen for 90% similar descriptor vector pairs).

The three most active compounds from ChEMBL, the three inverse-SVR and three inverse-lead molecules predicted that the most active were extracted and compared in terms of structural similarity and p$K_i$ values. The most active inverse-SVR and inverse-lead compounds are structurally very similar in terms of substructure counts but not necessarily in terms of overall topology to the most active ChEMBL compounds, as shown in Table 2. Similar substructures or features like quinoline, cyclopropane, peptide bonds, and fluoride atoms appear in both ChEMBL and generated compounds—but they may be interconnected in a different way. Sampling the neighborhood of a given compound is likely to witness the neural network return typical "building blocks" seen in those compounds, all while recombining them and placing them in original contexts.

*3.3.2. "Inverse-GTM" Compounds.* Inverse-GTM sampling, in this case, gives better results in terms of validity and uniqueness than inverse-SVR compounds.

Compounds generated from a GTM node vector consistently tend to be projected into the same area they were sampled from. This is not true of all compounds, a minority

**Table 2. Most Active ChEMBL-Reported Compounds (A, B, C) against the ChEMBL1862 Target as Well as the Most Potent Structures Generated from the Different Seed Vectors**[a]

| ChEMBL compounds | | |
|---|---|---|
| **A** | **B** | **C** |
|  10.73 |  10.70 |  10.70 |
| inverse-SVM compounds | | |
|  10.20 |  9.84 |  9.82 |
| inverse-Lead compounds[b] | | |
|  10.08 |  9.45 |  9.35 |
| inverse-GTM compounds | | |
|  7.88 |  7.84 |  7.83 |

[a]The numbers correspond to experimentally measured (for ChEMBL compounds) or predicted with SVR models p$K_i$ values. [b]Compounds generated for the descriptor vector generated for molecule **A**, which is the highest affinity molecule (inverse-LEAD) with p$K_i$ = 10.73.

being projected in different areas of chemical space—in inactive-dominated zones (see Figure 6).

In inverse-GTM, random noise is also used to perturb the input descriptor (GTM node vector), whereas inverse-SVR compounds were strictly sampled on hand of the five optimized descriptor vectors. Accordingly, the resulting compounds are more diverse but less prone to score very high predicted p$K_i$ values as shown in Table 2. Rather than focusing on recombination of fragments maximally contributing to SVR-predicted p$K_i$ values, the model incorporates fragments of all training compounds occupying the vicinity of the chosen "seed" vector.

*3.3.3. "Inverse-SVR" and "Inverse-Lead" Versus "Inverse-GTM".* Sampling with inverse-SVR and inverse-lead has a chance to return molecules predicted highly active, which is not the case for compounds generated with inverse-GTM. This can be explained by the fact that inverse-SVR (inverse-lead) vectors served as the generation seed correspond to high activity values, which is not the case for the GTM node vectors. Inverse-GTM molecules have lower SVR-predicted p$K_i$ values comparatively because "active" GTM landscape areas were defined to harbor "actives" of p$K_i \geq 7$, and the categorical nature of the landscape makes no further distinction between submicromolars and subnanomolars. The two methods produce active compounds, but molecules

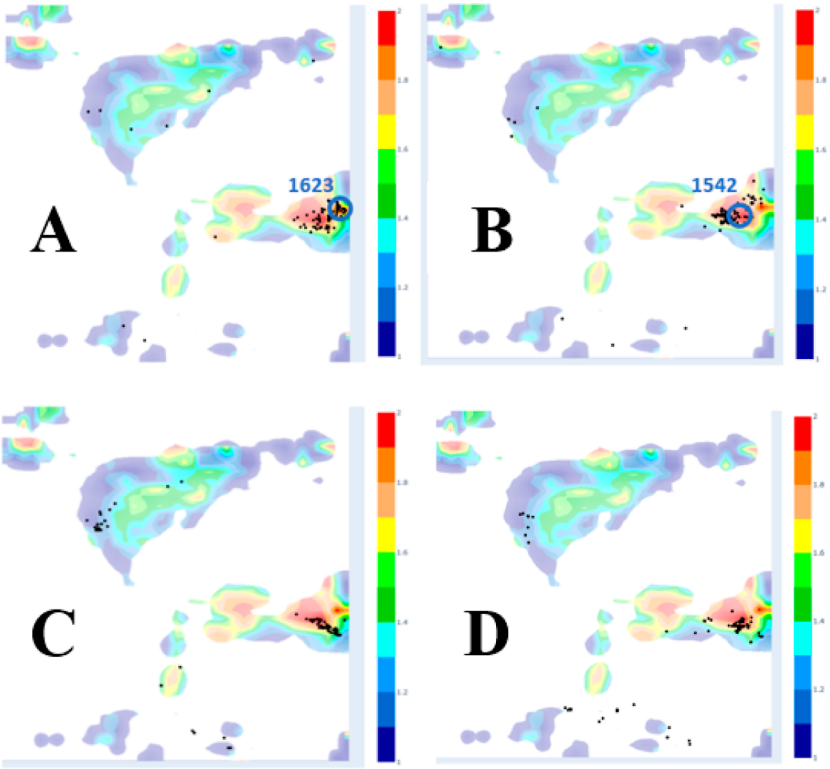



**Figure 6.** Projection of the 100 most active compounds predicted by the SVR models, generated in different fashions. See caption of Figure 3 for landscape color coding. (A) Compounds were generated from "node" vectors obtained from node 1623. (B) Compounds were generated from "node" vectors obtained from node 1542. (C) Inverse-SVR compounds. (D) Inverse-lead compounds.

generated from inverse-SVR tend to be more focused on specific chemical space zones predicted to stand for very high affinity. Therefore, they reproduce structural features typical to the few top actives—the "originality" mostly consisting in the way in which these features (scaffolds, linkers) are reorganized in the final structures. Inverse-GTM seeds tend by contrast to stem from structurally less specific neighborhoods, generating a more diverse set.

Figure 7 confirms this trend as we see that the distribution of activities of inverse-SVR and inverse-lead compounds has a tail in the very active regions, while the distribution of p$K_i$ for GTM-based compounds has a lower mean and is centered.

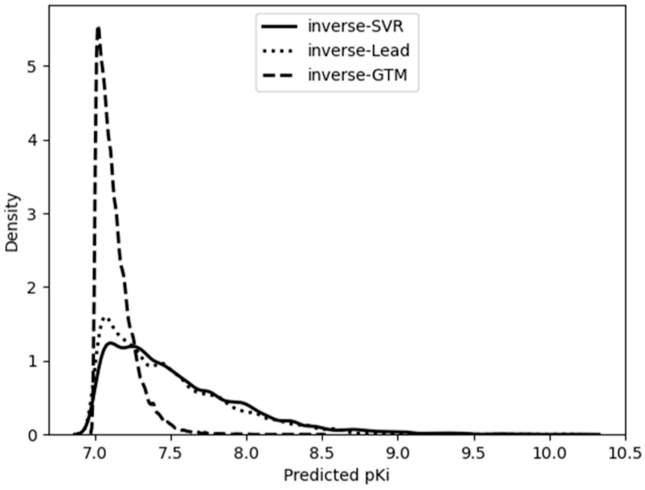



**Figure 7.** Comparison between the distribution of (SVR-predicted) activities between inverse-SVR, inverse-lead, and inverse-GTM compounds.

Interestingly, most of inverse-SVR compounds are projected in the large active zone where inverse-GTM compounds were sampled—even though the GTM-driven categorical QSAR is based on other descriptors than the SVR approach. This is additional proof that SVR-based and GTM-based models are not fundamentally divergent in terms of prediction but merely conflicting in terms of the specific definition of "actives" as continuous *versus* categorical magnitudes.

As it follows from Figure 8, synthetic accessibility score for the generated compounds (inverse-SVR, inverse-lead, and inverse-GTM) have on average a higher SA score than ChEMBL compounds. According to this score, generated structures are more difficult to synthetize than real ChEMBL molecules. On the other hand, they are still in the range of ChEMBL distribution (which goes up to 4.5−5) meaning that generated structures are not synthetically unreachable and therefore viable. The quantitative estimate druglikeness index shows that on average, inverse-SVR and inverse-lead compounds are of more interest for medicinal chemists than inverse-GTM compounds.

*3.3.4. Validation of Inverse-SVR and Inverse-Lead Compounds Using Pharmacophore Modeling.* Pharmacophore models were trained using LigandScout[41] (4.4) to check whether the generated compounds would also comply to the ligand- and structure-based hypothetic binding patterns that can be inferred on hand of current structure−activity data. Both structure-based and ligand-based approaches were applied in an effort to be as comprehensive as possible. The compounds present in the training set of the SVR model (821 compounds) were used for ligand-based model training.

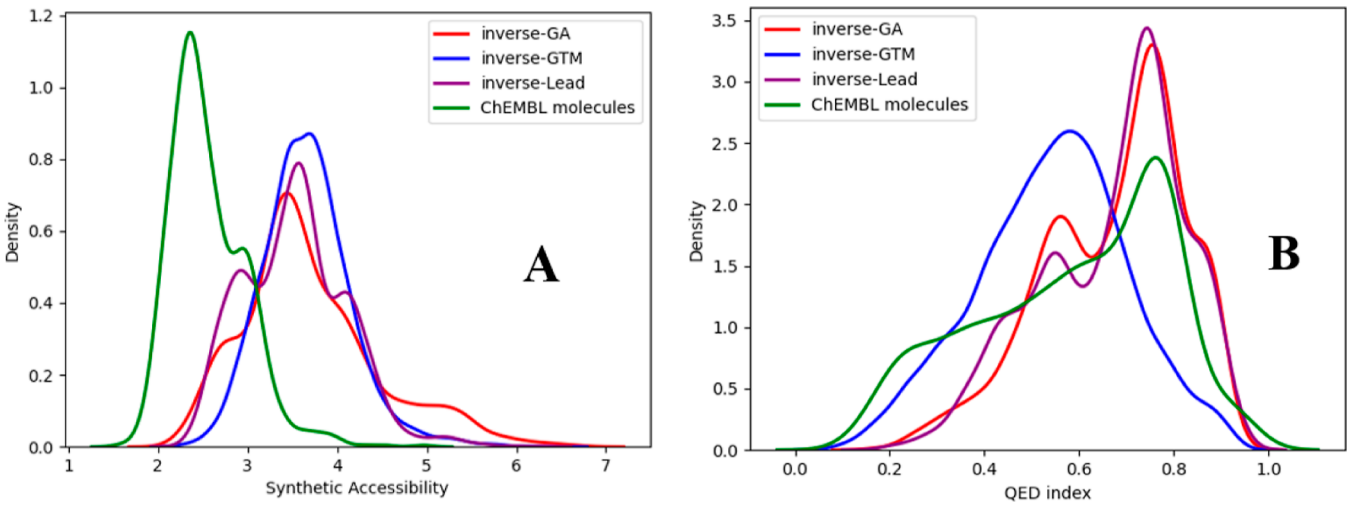
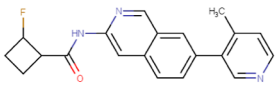
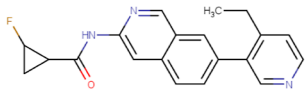
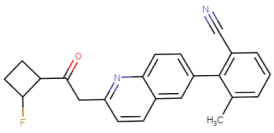
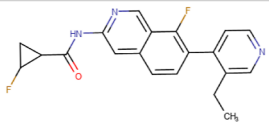



**Figure 8.** (A) Synthetic accessibility score for the four datasets calculated. (B) Quantitative estimate druglikeness index distribution for the three different datasets.

**Table 3. Hits Found with Pharmacophore Models and Their Validation with Docking for Inverse-SVM (I–III) and Inverse-Lead (IV) Compounds**

| | Hits | calculated pKi / activity rank | Pharmacophore | Docking score (LeadIT) |
|---|---|---|---|---|
| I |  | 9.82<br>3rd most active | Model 1 | -33.2 |
| II |  | 9.34 /<br>16th most active | Model 1 | -31.4 |
| III |  | 9.18<br>25th most active | Model 1 | -23.27 |
| IV |  | 9.45<br>2nd most active | Model 2 | -31.8 |

Ligand-based pharmacophores should reflect consensus features in highly active binders. Therefore, a threshold of $pK_i \geq 9$ was considered here to define "actives," in contrast to the default $pK_i \geq 7$ defining "actives" in other contexts of this work (GTM landscape, docking studies—*vide infra*). In addition, only the inverse-SVR and inverse-lead compounds with predicted $pK_i \geq 9$ were screened. This subset of the initial generated compounds contains 39 inverse-SVR molecules and 8 inverse-lead compounds which makes 47 generated compounds in total.

For ligand-based pharmacophores, conformations for the training set compounds were calculated using the pre-loaded FAST parameters of the software. These settings returned a maximum of 25 conformations by compound. Ligand-based pharmacophores were built and clustered by LigandScout.[41] Pharmacophore models were calculated for two clusters containing 78 and 5% (163 and 9 molecules, respectively) of all training set actives (model 1 and model 2, respectively). Different pharmacophore models were generated for each cluster using sets of 5 to 10 molecules.

Structure-based pharmacophores were built based on PDB crystal structures of human proto-oncogene tyrosine-protein kinase ABL1. 2HZI and 2CQG crystal structures were used to generate the shared pharmacophore model which was screened against the 47 generated compounds for which $pK_i > 9$ was predicted.

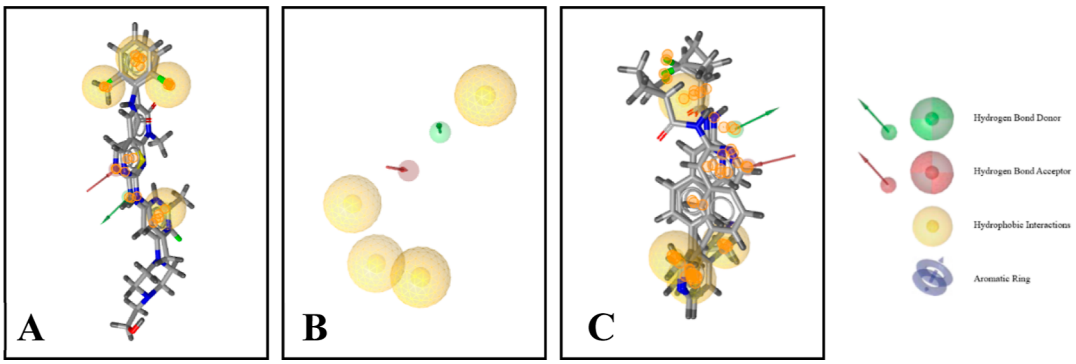



**Figure 9.** (A) Pharmacophore aligned with both PDB crystal structure ligands. (B) Shared pharmacophore model. (C) Selected inverse-SVR hits aligned with the pharmacophore model.

*3.3.4.1. Ligand-Based Pharmacophore.* The screening of 47 inverse-SVR and inverse-lead molecules "hidden" in a set of 328 inactive decoys selected from the training set inactives allowed to understand if the two ligand-based pharmacophore models were selective enough to primarily focus on putative actives. If the considered pharmacophore models were observed to be as likely to match inactive decoys, it may be inferred that "matching" the pharmacophore model is no reliable indicator of putative activity against CHEMBL1862 but merely that the ligand-based pharmacophore models are too generic (easily matched by random compounds).

Model 1 and model 2 returned, respectively, three and one hits. The hits align well with the pharmacophore model, and most features match as shown in[40] Figures S4 and S5 in the Supporting Information. Table 3 shows that the four hits have relatively high ranking among the most actives, one of them being the third predicted most active inverse-SVR compound and another the second most active inverse-lead compound.

*3.3.4.2. Structure-Based Pharmacophore Screening.* The shared pharmacophore model computed for two PDB structures (2HZI and 2GQG) is mostly based on hydrophobic interactions with one hydrogen bond donor and one hydrogen bond acceptor as shown in Figure 9B. The ligands contained in the PDB crystal structures are typically larger than inverse-SVR molecules. However, Figure 9A shows that crystalized ligands may include specific moieties not directly involved in binding. VS with the shared pharmacophore returned eight hits (see Table S2 in Supporting Information), four of which correspond to those found with ligand-based pharmacophores (Table 3). Notice that inverse-SVR compounds nicely match the pharmacophore, all while being smaller than the PDB ligands (see Figure 9C). These results show that the generated compounds are not only predicted active by the SVR models because they were optimized to do so but also fit the activity criteria of external validation methods like pharmacophore models. The fact that these three compounds were found by both methods and predicted highly active by the SVR model indicates that these compounds may be good candidates for further testing.

*3.3.5. Validation of Inverse-SVR Compounds Using Ligand-To-Protein Docking.* In the docking challenge, both LeadIT and S4MPLE were able to predict the correct binding geometry of the native ligand of 2E2B (in protein-rigid redocking mode), and both were seen to significantly prioritize "actives" ($pK_i > 7$), for LeadIT, the area under the ROC curve obtained after redocking the 821 training set compounds (out of which only 816 could be docked) was of 0.77. S4MPLE also performed reasonably well (ROC AUC = 0.69 after the docking of 550 of the training set compounds, in random order). At that point, a quantitative correlation of $R^2 = 0.51$ between LeadIT and S4MPLE scores could be observed. Unfortunately, neither the LeadIT score ($R^2 = 0.21$, over 816 redocked compounds) nor S4MPLE ($R^2 = 0.16$ over the 550 ligands) can return docking scores that quantitatively correlate with the experimental $pK_i$ values. We refer the reader to the Supporting Information section for a detailed analysis of the relationships between docking scores and actual, respective predicted $pK_i$ values. It was observed that 76% of the experimentally confirmed training set actives ($pK_i > 7$) dock with LeadIT scores below or equal to −30, whereas LeadIT score ≤−25 would retrieve 92% of them. Therefore, the percentage of a library achieving LeadIT scores better (more negative) than this order of magnitude is a first rough estimate of how strongly CHEMBL-1862-focused that library is. Indeed, these percentages are significantly higher within the mixed collection of inverse-GTM and inverse-SVR leads (blue in Figure 10) than within the random subset of ZINC random decoys (orange bars). It should be noticed that only two out of three hits selected by pharmacophore models (molecules I, II, and IV, Table 2) were validated in docking calculations as actives: the LeadIT score for molecule III was larger than the threshold of −25. The fact that the molecules I, II, and IV were

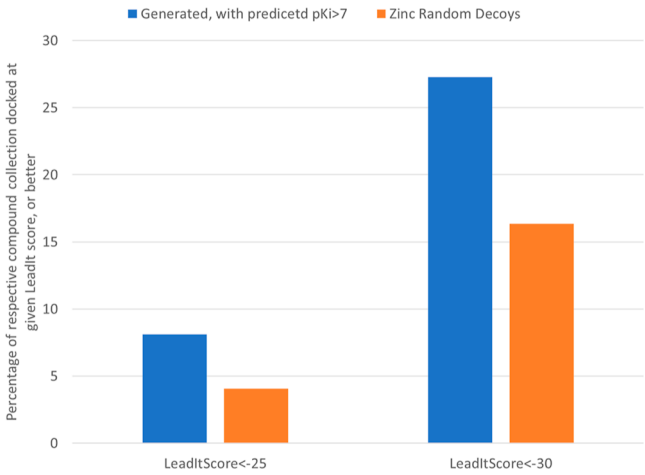



**Figure 10.** Percentages within the collection of inverse-GTM and inverse-SVR leads (blue) and the set or random ZINC decoys (orange) achieving LeadIt docking scores typical of experimentally validated actives of $pK_i > 7$.

found by both pharmacophore and docking methods as well as predicted highly active by the SVR model indicates that these compounds may be good candidates for further testing. We do not exclude that application of a docking score correlating with studied activity (*e.g.*, that reported by Ahmed *et al.*[64]) may better validate generated molecules.

## 4. CONCLUSIONS AND PERSPECTIVES

This article introduced a new type of architecture based on state-of-the-art deep learning method which is capable, given a descriptor type and successful training, to generate compounds possessing wanted activity and structural features from "seed" descriptor vectors—where the descriptor vectors are not "latent" vectors themselves produced by some encoder architecture but standard, state-of-the-art descriptors typically used in QSAR (here, ISIDA fragment counts). This provides an elegant solution for the inverse QSAR problem—the inference of novel molecular structures matching model-predicted high activity zones of the descriptor space. Finding descriptor "seeds" corresponding to aforementioned interesting zones has been herein addressed in two model-specific ways: evolutionary search for $D$ vectors corresponding to high predicted affinity values ($pK_i$) according to SVR models or $D$ vectors within the immediate neighborhood of GTM nodes preferentially populated by active compounds. Additionally, the descriptor vector generated for the highest affinity ligand from the training set was also used as a seed. Selecting only descriptor vectors associated with very high predicted affinity values ($pK_i$) equal or close to the best ever values reported in ChEMBL lead to inverse-SVR and inverse-lead molecules being structurally related to already existing top-active ChEMBL compounds—in the sense that they share significant common substructures, all while preserving their global originality. An external pharmacophore study performed on inverse-SVR compounds shows that several molecules with high predicted activity show good matches with existing active molecules in terms of pharmacophores. Selecting the vectors based on generative topographic mapping is focused on a binary, class-based definition of activity, and inverse-GTM molecules appear more diverse, all while predicted to have remarkable $pK_i$ values by the SVR models (better than 100 nM, but not yet close to the top-active ChEMBL compounds). Original compounds of acceptable synthetic feasibility index could be readily obtained. Therefore, the inverse QSAR problem—fast discovery of original feasible compounds specifically selected for being predicted active by a given QSAR model—can be considered as conveniently solved, at least for the (rather widely used) class of fragment-based molecular descriptor-based QSAR models. Of course, the ultimate promise of prospective discovery of experimentally validated actives may only be kept if the "inversed" model lives up to its promises in terms of prediction—but this is an altogether different problem, which is not covered by the present, purely methodological work. It is clearly not expected to necessarily see inverse-QSAR *de novo* compounds automatically score well in docking if docking scores are decorrelated from the QSAR-predicted affinity estimator. In particular, fragment-count-based QSARs may overrate the importance of given molecular fragments if the latter happen to appear *by chance* only within the structures of actives, thus establishing the mechanistically wrong shortcut "presence of key fragments → activity" simply because inactive counterexamples containing the same fragments in a different mutual configuration were not found at the training stage. ACoVAE-based approaches may, as seen in this work, readily suggest structures issued by recombining such key fragments—guaranteed to achieve high ratings by the parent QSAR model but not sure to still feature a global pharmacophore compatible with the target. The goal of this work was to present genuine solutions for the QSAR inversion problem based on "classical" fragment descriptors rather than on DNN-specific latent space vectors. Technically, this was a success, but it also clearly reveals that QSAR inversion *alone* is too risky a path to take in drug design: the actual pursuit of the synthesis efforts of sometimes challenging (but—granted—novel) structures may or may not pay, given the intrinsically incomplete and error-prone nature of QSAR models. However, if inverse QSAR is coupled with orthogonal activity prediction techniques, as done here, it can be observed that many of compounds alleged to be active by the initial QSAR models fail to pass the additional, independent activity assessment tests (pharmacophore matching, docking). This is no surprise because the consensus rate of chemoinformatics predictors based on premises as radically different as 2D-QSAR, pharmacophore screening and docking are typically very low. Nevertheless, we were successful in discovering some *de novo* structures which did pass the latter tests. This shows that the exploration of the initial inverse-QSAR-relevant chemical space is sufficient to visit areas in which not only the original QSAR model but also the alternative approaches indicate that biological activity is likely, pending experimental validation.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c01086.

Detailed description of neural network architecture and some complementary results of QSAR and pharmacophore modeling (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Alexandre Varnek** − *Laboratory of Chemoinformatics, UMR 7140 University of Strasbourg/CNRS, 67000 Strasbourg, France*; orcid.org/0000-0003-1886-925X; Email: varnek@unistra.fr

### Authors

**William Bort** − *Laboratory of Chemoinformatics, UMR 7140 University of Strasbourg/CNRS, 67000 Strasbourg, France*

**Daniyar Mazitov** − *Laboratory of Chemoinformatics and Molecular Modeling, A. M. Butlerov Institute of Chemistry, Kazan Federal University, 420008 Kazan, Russia*

**Dragos Horvath** − *Laboratory of Chemoinformatics, UMR 7140 University of Strasbourg/CNRS, 67000 Strasbourg, France*; orcid.org/0000-0003-0173-5714

**Fanny Bonachera** − *Laboratory of Chemoinformatics, UMR 7140 University of Strasbourg/CNRS, 67000 Strasbourg, France*

**Arkadii Lin** − *Laboratory of Chemoinformatics, UMR 7140 University of Strasbourg/CNRS, 67000 Strasbourg, France*

**Gilles Marcou** − *Laboratory of Chemoinformatics, UMR 7140 University of Strasbourg/CNRS, 67000 Strasbourg, France*; orcid.org/0000-0003-1676-6708

**Igor Baskin** − *Department of Material Science and Engineering, Technion—Israel Institute of Technology, 3200003 Haifa, Israel*

**Timur Madzhidov** − *Laboratory of Chemoinformatics and Molecular Modeling, A. M. Butlerov Institute of Chemistry, Kazan Federal University, 420008 Kazan, Russia;* orcid.org/0000-0002-3834-6985

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c01086

## Notes

The authors declare no competing financial interest.

Data and Software Availability: Developed code is available at the GitHub storage of the Laboratory of Chemoinformatics: https://github.com/Laboratoire-de-Chemoinformatique/ACoVAE. The data used for the model training and validation are available at https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/ILWSLF.

## ■ ACKNOWLEDGMENTS

The High Performance Computing (HPC) Center of the Strasbourg University is acknowledged for technical support.

## ■ ABBREVIATIONS

ACoVAE, attention-based conditional variational autoencoder; DNN, deep neural network; GA, genetic algorithm; GTM, generative topographic map; QSA/PR, quantitative structure−activity/property relationships; SVR, support vector regression; VS, virtual screening

## ■ REFERENCES

(1) Dudek, A.; Arodz, T.; Galvez, J. Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. *Comb. Chem. High Throughput Screening* **2006**, *9*, 213−228.

(2) Hansch, C. H.; Leo, A. J. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society, 1995; Vol. *1*.

(3) Korotcov, A.; Tkachenko, V.; Russo, D.; Ekins, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharm.* **2017**, *14*, 4462−4475.

(4) Varnek, A.; Baskin, I. Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis? *J. Chem. Inf. Model.* **2012**, *52*, 1413−1437.

(5) Jin, Y.; Wang, H.; Sun, C. *Introduction to Machine Learning*; Springer International Publishing, 2021; Vol. *975*.

(6) Smola, A.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199−222.

(7) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(8) Skvortsova, M.; Fedyaev, K.; Palyulin, V.; Zefirov, N. Inverse Structure−Property Relationship Problem for the Case of a Correlation Equation Containing the Hosoya Index. *Dokl. Chem.* **2001**, *379*, 191−195.

(9) Skvortsova, M.; Stankevich, I.; Zefirov, N. Generation of molecular structures of polycondensed benzenoid hydrocarbons using the randic index. *J. Struct. Chem.* **1992**, *33*, 416−422.

(10) Skvortsova, M.; Baskin, I.; Slovokhotova, O.; Palyulin, V.; Zefirov, N. Inverse Problem in QSAR/QSPR Studies for the Case of Topological Indices Characterizing Molecular Shape (Kier Indices). *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 630−634.

(11) Lin, A.; Horvath, D.; Marcou, G.; Beck, B.; Varnek, A. Multi-task generative topographic mapping in virtual screening. *J. Comput.-Aided Mol. Des.* **2019**, *33*, 331−343.

(12) Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today* **2011**, *16*, 372−376.

(13) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: Reaction-Driven de novo Design of Bioactive Compounds. *PLoS Comput. Biol.* **2012**, *8*, No. e1002380.

(14) Mauser, H.; Guba, W. Recent developments in de novo design and scaffold hopping. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 365−374.

(15) Hartenfeller, M.; Proschak, E.; Schüller, A.; Schneider, G. Concept of Combinatorial De Novo Design of Drug-like Molecules by Particle Swarm Optimization. *Chem. Biol. Drug Des.* **2008**, *72*, 16−26.

(16) Sattarov, B.; Baskin, I.; Horvath, D.; Marcou, G.; Bjerrum, E.; Varnek, A. De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *J. Chem. Inf. Model.* **2019**, *59*, 1182−1196.

(17) Gómez-Bombarelli, R.; Wei, J.; Duvenaud, D.; Hernández-Lobato, J.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T.; Adams, R.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(18) Zhou, Z.; Kearnes, S.; Li, L.; Zare, R.; Riley, P. Optimization of Molecules via Deep Reinforcement Learning. *Sci. Rep.* **2019**, *9*, 10752.

(19) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.

(20) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De Novo Design through Deep Reinforcement Learning. *J. Cheminf.* **2017**, *9*, 48.

(21) Prykhodko, O.; Johansson, S.; Kotsias, P.; Arús-Pous, J.; Bjerrum, E.; Engkvist, O.; Hongming, C. A de novo molecular generation method using latent vector based generative adversarial network. *J. Cheminf.* **2019**, *11*, 74.

(22) Arús-Pous, J.; Patronov, A.; Bjerrum, E.; Tyrchan, C.; Reymond, J.-L.; Hongming, C.; Engkvist, O. SMILES-based deep generative scaffold decorator for de-novo drug design. *J. Cheminf.* **2020**, *12*, 38.

(23) Cova, T.; Pais, A. Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns. *Front. Chem.* **2019**, *7*, 809.

(24) Gupta, M. K.; Gupta, S.; Rawal, R. Impact of Artificial Neural Networks in QSAR and Computational Modeling. In *Artificial Neural Network for Drug Design, Delivery and Disposition*; Academic Press, 2016; pp 153−179.

(25) Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 468−481.

(26) Fjodorova, N.; Vračko, M.; Jezierska, A.; Novič, M. Counter propagation artificial neural network categorical models for prediction of carcinogenicity for non-congeneric chemicals. *SAR QSAR Environ. Res.* **2010**, *21*, 57−75.

(27) Ajmani, S.; Viswanadhan, V. N. A Neural Network-Based QSAR Approach for Exploration of Diverse Multi-Tyrosine Kinase Inhibitors and its Comparison with a Fragment- Based Approach. *Curr. Comput.-Aided Drug Des.* **2013**, *9*, 482−490.

(28) Myint, K.-Z.; Wang, L.; Tong, Q.; Xie, X. Molecular Fingerprint-Based Artificial Neural Networks QSAR for Ligand Biological Activity Predictions. *Mol. Pharmaceutics* **2012**, *9*, 2912−2923.

(29) Rana, A.; Rawat, A.; Bijalwan, A.; Bahuguna, H. Application of Multi Layer (Perceptron) Artificial Neural Network in the Diagnosis System: A Systematic Review. In *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*; IEEE, 2018; pp 1−6.

(30) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Neural Networks in Building QSAR Models. In *Artificial Neural Networks: Methods and*

*Applications*; Livingstone, D. J., Ed.; Humana Press, 2009; pp 133−154.

(31) Sabando, M. V.; Ponzoni, I.; Milios, E.; Soto, A. Using molecular embeddings in QSAR modeling: Does it make a difference? *Briefings Bioinf.* **2022**, *23*, bbab365.

(32) Muratov, E. N.; Bajorath, J.; Sheridan, R.; Tetko, I.; Filimonov, D.; Poroikov, V.; Oprea, T.; Baskin, I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525−3564.

(33) Sousa, T.; Correia, J.; Pereira, V.; Rocha, M. Generative Deep Learning for Targeted Compound Design. *J. Chem. Inf. Model.* **2021**, *61*, 5343−5361.

(34) Weininger, D.; Weininger, A.; Weininger, J. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(35) Kotsias, P.-C.; Arús-Pous, J.; Chen, C.; Engkvist, O.; Tyrchan, C.; Bjerrum, E. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* **2020**, *2*, 254−265.

(36) Ucak, U. V.; Ashyrmamatov, I.; Lee, J. Reconstruction of lossless molecular representations, SMILES and SELFIES, from fingerprints. *ChemRxiv* **2022**, DOI: 10.26434/chemrxiv-2022-tqv76-v2.

(37) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I.; Marcou, G. ISIDA—Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191−198.

(38) Varnek, A.; Fourches, D.; Solov'ev, V.; Klimchuk, O.; Ouadi, A.; Billard, I. Successful "In Silico" Design of New Efficient Uranyl Binders. *Solvent Extr. Ion Exch.* **2007**, *25*, 433−462.

(39) Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D. Mappability of drug-like space: Towards a polypharmacologically competent map of drug-relevant compounds. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 1087−1108.

(40) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-labelled fragment descriptors. *Mol. Inf.* **2010**, *29*, 855−868.

(41) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−169.

(42) *LeadIT*; BioSolveIT GmbH: Sankt Augustin, Germany, 2022. www.biosolveit.de.

(43) Hoffer, L.; Chira, C.; Marcou, G.; Varnek, A.; Horvath, D. S4MPLE—Sampler for Multiple Protein-Ligand Entities: Methodology and Rigid-Site Docking Benchmarking. *Molecules* **2015**, *20*, 8997−9028.

(44) Mehta, S.; Ghazvininejad, M.; Iyer, S.; Zettlemoyer, L.; Hajishirzi, H. DeLighT: Deep and Light-weight Transformer. **2020**, arxiv:2008.00623. arXiv preprint.

(45) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *31st Conference on Neural Information Processing Systems 2017*, 2017; pp 5999−6009.

(46) Lin, Z.; Winata, G. I.; Xu, P.; Liu, Z.; Fung, P. Variational Transformers for Diverse Response Generation. **2020**, arxiv:2003.12738. arXiv preprint.

(47) Davidson, T. R.; Falorsi, L.; De Cao, N.; Kipf, T.; Tomczak, J. M. Hyperspherical variational auto-encoders. *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2018; Vol. *2*, pp 856−865.

(48) De Cao, N.; Aziz, W. The Power Spherical Distribution. **2020**, arxiv:2006.04437. arXiv preprint.

(49) Mehta, S.; Ghazvininejad, M.; Iyer, S.; Zettlemoyer, L.; Hajishirzi, H. DeLighT: Deep and Light-weight Transformer. **2021**, arXiv:2008.00623. arXiv preprint.

(50) Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). **2016**, arxiv:1606.08415. arXiv preprint.

(51) Chieng, H. H.; Wahid, N.; Pauline, O.; Perla, S. R. K. Flatten-t swish: A thresholded relu-swish-like activation function for deep learning. *Int. J. Adv. Intell. Inform.* **2018**, *4*, 76−86.

(52) Horvath, D.; Brown, J.; Marcou, G.; Varnek, A. An Evolutionary Optimizer of libsvm Models. *Challenges* **2014**, *5*, 450−472.

(53) Bishop, C. M.; Svensen, M.; Williams, C. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10*, 215−234.

(54) Dragos, H.; Gilles, M.; Alexandre, V. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J. Chem. Inf. Model.* **2009**, *49*, 1762−1776.

(55) Horvath, D.; Jeandenans, C. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces—A novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680−690.

(56) Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A. Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *J. Chem. Inf. Model.* **2019**, *59*, 564−572.

(57) Nugmanov, R. I.; Mukhametgaleev, R.; Akhmetshin, T.; Gimadiev, T.; Afonina, V.; Madzhidov, T.; Varnek, A. CGRtools: Python Library for Molecule, Reaction, and Condensed Graph of Reaction Processing. *J. Chem. Inf. Model.* **2019**, *59*, 2516−2521.

(58) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.

(59) Kosugi, T.; Ohue, M. Quantitative Estimate Index for Early-Stage Screening of Compounds Targeting Protein-Protein Interactions. *Int. J. Mol. Sci.* **2021**, *22*, 10925.

(60) *ChemAxon Standardizer*, Version 5.12; ChemAxon, Ltd.: Budapest, Hungary, 2012.

(61) O'Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. *ChemRxiv* **2018**, DOI: 10.26434/chemrxiv.7097960.v1.

(62) Berenger, F.; Tsuda, K. Molecular generation by Fast Assembly of (Deep)SMILES fragments. *J. Cheminf.* **2021**, *13*, 88.

(63) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.

(64) Ahmed, S.; Prabahar, A. E.; Saxena, A. K. Molecular docking-based interactions in QSAR studies on Mycobacterium tuberculosis ATP synthase inhibitors. *SAR QSAR Environ. Res.* **2022**, *33*, 289−305.