

Impact of Applicability Domains to Generative Artificial Intelligence

Maxime Langevin, Christoph Grebner, Stefan Güssregen, Susanne Sauer, Yi Li, Hans Matter, and Marc Bianciotto*

Cite This: *ACS Omega* 2023, 8, 23148–23167

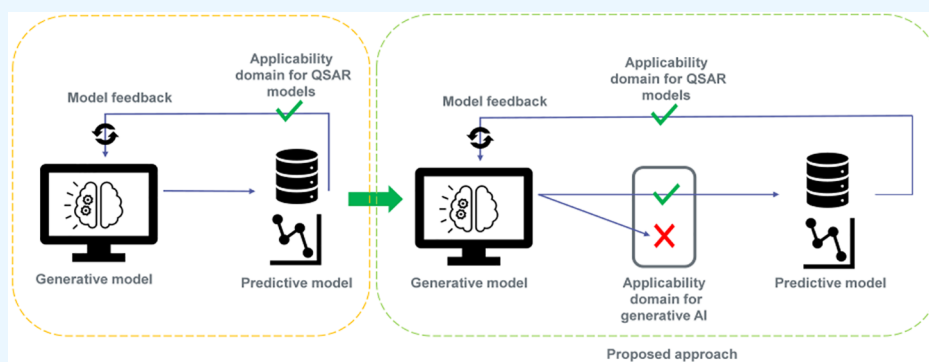
Read Online

ACCESS |

Metrics & More

Article Recommendations

Supporting Information



ABSTRACT: Molecular generative artificial intelligence is drawing significant attention in the drug design community, with several experimentally validated proof of concepts already published. Nevertheless, generative models are known for sometimes generating unrealistic, unstable, unsynthesizable, or uninteresting structures. This calls for methods to constrain those algorithms to generate structures in drug-like portions of the chemical space. While the concept of applicability domains for predictive models is well studied, its counterpart for generative models is not yet well-defined. In this work, we empirically examine various possibilities and propose applicability domains suited for generative models. Using both public and internal data sets, we use generative methods to generate novel structures that are predicted to be actives by a corresponding quantitative structure–activity relationships model while constraining the generative model to stay within a given applicability domain. Our work looks at several applicability domain definitions, combining various criteria, such as structural similarity to the training set, similarity of physicochemical properties, unwanted substructures, and quantitative estimate of drug-likeness. We assess the structures generated from both qualitative and quantitative points of view and find that the applicability domain definitions have a strong influence on the drug-likeness of generated molecules. An extensive analysis of our results allows us to identify applicability domain definitions that are best suited for generating drug-like molecules with generative models. We anticipate that this work will help foster the adoption of generative models in an industrial context.

INTRODUCTION

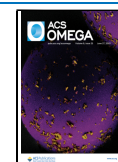
Recent years have seen growing interest in generative models for drug design.^{1–3} The first application of generative models is the design of libraries of compounds with desired physicochemical properties.³ This approach, which is referred to as distribution learning,⁴ can be seen as an alternative to virtual screening. Generative models have also been used for designing molecules with a desired profile according to predictive models, e.g., molecules with high predicted activity on a therapeutic target.² This approach is referred to as goal-directed generation. Goal-directed generation searches for compounds that maximize a user-defined scoring function, which reflects the desirability of a compound in a drug discovery project. The scoring function is usually a combination of predicted biological and physicochemical properties, determined using both machine learning models and computed properties (e.g., clogP^5). Nonetheless, this focus on generating high-scoring molecules has sometimes been made

at the detriment of generating molecules that would be considered for synthesis in a drug design project (for reasons relating directly to synthesizability or apparently unstable, toxic, or reactive moieties). Low drug-likeness of generated compounds has been reported for SMILES-based methods and genetic algorithms,⁶ as well as for graph-based methods (e.g., in Mercado et al.,⁷ Figure I2, where molecules with conjugated non aromatic rings are shown). Indeed, generated molecules can contain reactive fragments, long heteroatom chains, or macrocycles.^{6,8} While some works^{9,10} have included metrics such as

Received: April 17, 2023

Accepted: May 26, 2023

Published: June 12, 2023



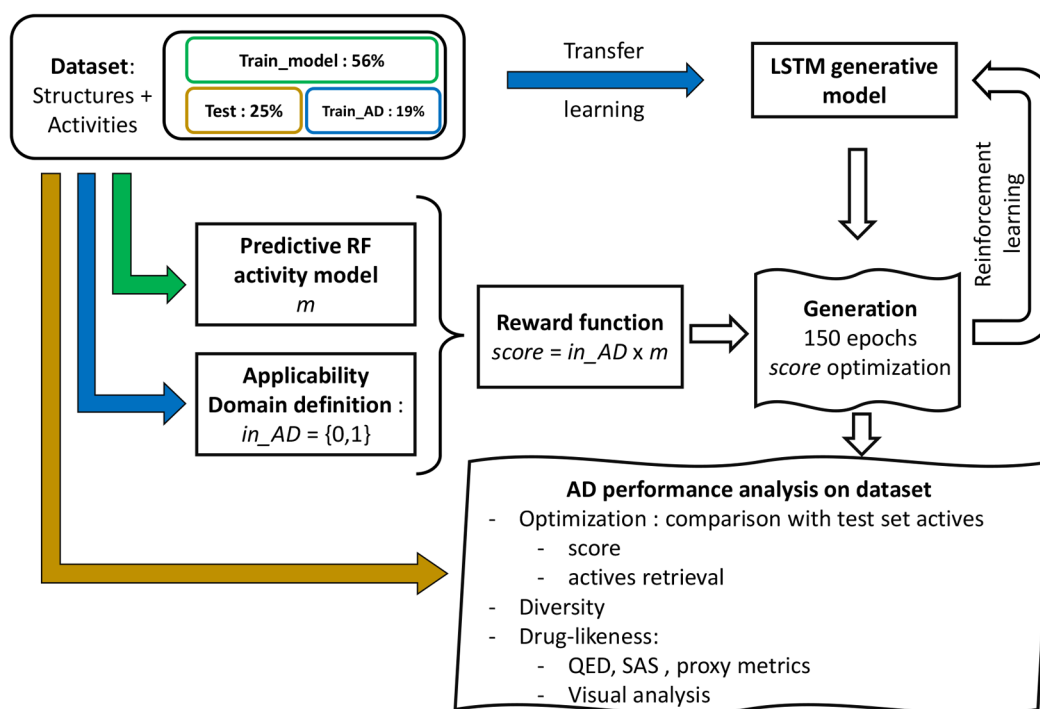


Figure 1. Overview of the workflow used for evaluating an AD. During generation, the AD is taken into account in the reward function as a multiplicative term that yields 0 if the molecule generated is out of the AD and 1 otherwise. The colors used for the arrows are related to the different subsets of the data set that are used at different stages of the process: green for the activity model training set, blue for set used for AD definition and generative model pretraining, and orange for the test set.

the quantitative estimate of drug-likeness¹¹ (QED) or the percentage of de novo fingerprints bits¹² to ensure the generation of drug-like molecules while generating molecules with good docking scores, concerns remain regarding the drug-likeness of molecules generated using artificial intelligence algorithms. As we will discuss below, drug-likeness is difficult to define in absolute terms.¹³ Nevertheless, preventing the generation of the most blatant non-drug-like molecules would be welcome and would foster the adoption of generative models in drug discovery. This leads us to search for a way to enforce the generation of drug-like molecules by translating the applicability domain (AD) concept, which is generally associated with quantitative structure–activity relationship (QSAR) models, to generative models. Applicability domains for QSAR and ML predictive models are well-studied.^{14–17} In the context of QSAR modeling, it is defined as follows: “The applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability”.¹⁸ The need for QSAR models to be associated with an applicability domain is the third principle of the five OECD Principles for (Q)SAR validation¹⁸ and was formalized in the REACH initiative for toxicology assessment of new chemical entities.¹⁹

In this work, we evaluate AD definitions in the context of goal-directed generation, based not their ability to define whether a reliable prediction can be made but instead on their ability to discriminate against non-drug-like molecules in the context of generative modeling. By analogy with the definition of AD for QSAR models, we investigate the relevance of a generative applicability domain defined as the chemical structure space in which the generative model makes structures with a given reliability in terms of drug-likeness. As a start, it is appealing to use the ADs of the QSAR models that are used for orienting the

generation and to apply them directly to generative algorithms. This can even happen silently if the QSAR models used to bias the generation give a null prediction to generated molecules that fall beyond their AD, leading the generated structure to be discarded. Nonetheless, this method might not be sufficient to guarantee that the generated structures are drug-like: when QSAR models are used in a lead optimization context, they are applied on molecules that are generally proposed or selected by experienced researchers and notions of reactivity, synthetic access, and drug-likeness were already taken into account when designing the molecules, while it is not the case when molecules come from generative models. On the other hand, as some non-drug-like features of generated molecules can be related to a specific part of its structure, if the AD is based on similarity metrics, their similarity to the training set molecules can be high enough for them to end up in the AD of the QSAR model.

Focusing on goal-directed generation algorithms in the context of lead optimization, we thus investigate in this work the validity of the concept of applicability domains for generative models. First, we highlight several difficulties associated with the drug-likeness concept and how the drug-likeness of generated molecules can be evaluated. Then, by combining different criteria used for QSAR AD model definitions, we derive several generative AD definitions. Studying data sets that span diverse lead-optimization scenarios, we use these definitions to constrain the output of generative algorithms. We then analyze the influence of these definitions on the molecules sets that are generated, especially in order to identify whether those definitions are suited for generating drug-like molecules in the context of lead optimization. An overview of the workflow used for this evaluation is provided Figure 1. Finally, we validate our analysis on the difference in drug-likeness between different AD by performing a molecular Turing test on one of our data sets.

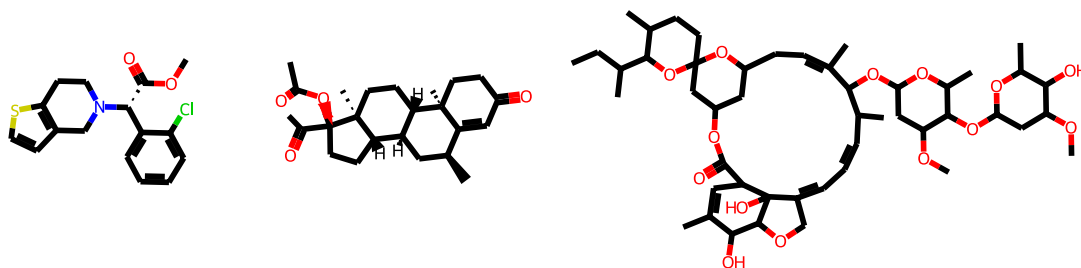


Figure 2. From left to right: chemical structures of clodigogrel, medroxyprogesterone acetate (a steroid), and ivermectin (a drug derived from a natural product). The diversity in chemical features of these three drugs (e.g., presence of macrocycles, number of cycles, and number of chiral centers) shows how much drug-likeness is context-dependent.

Drug-Likeness. The understanding of drug likeness²⁰ and its prediction²¹ is a long-standing research topic in drug design. However, as stated by Shultz, “The term ‘drug-like’ is universally known, utilized by many and precisely defined by no one. There are several inherent problems with the term ‘drug-like’, including the definition of ‘drug’ and what level of similarity is implied by ‘like’.”²² Approaches for drug-likeness prediction include the comparison of drug databases to selections of chemicals that are supposed not to be drug-like (typically random subsets of the ACD²¹) and the identification of a drug-like property space as in the case of the “rule of 5”²³ (keeping in mind that this rule reflects statistical trends on the permeation or absorption of oral drugs at the end of the twentieth century rather than an absolute rule,^{22,24} especially about drug-likeness in general), which can be combined with a desirability function for building the QED metric,¹¹ with functional group filters as in REOS,²⁵ or using a fragment based approach.²⁰ Another approach has been to use the feedback from several chemists about the drug-likeness of a large set of molecules for building classification models.²⁶ In the specific context of de novo design, the evaluation of drug-likeness is often associated with computing the quantitative estimate of drug-likeness (QED),²⁷ the Synthetic Accessibility Score (SAS),²⁸ or the synthetic complexity.²⁹

Drug-likeness is also a context-dependent concept. In Figure 2, the rightmost molecule (ivermectin³⁰) is a marketed drug derived from a natural product. Despite being a marketed drug, its macrocycle would disqualify it for many contemporary small-molecule drug discovery programs. The molecule at the center (medroxyprogesterone acetate, a steroid) is also a marketed drug; however, the steroid scaffold bears a risk of biological unspecificity, which could disqualify this structure as a starting point for current medicinal chemistry programs. Those examples show that whether a molecule is drug-like depends on the context of the drug discovery program.

Finally, drug-likeness is also dependent on the personal experience of researchers who assess the molecule and is prone to subjective bias;^{26,31,32} to the point that, according to Bickerton et al., “As beauty is in the eye of the beholder, so chemical attractiveness is in the eye of the chemist”.¹¹ In this Article, we will use “drug-likeness” in the following sense: drug-like molecules are molecules likely to be considered for synthesis by a drug discovery team, which includes both assessment of synthesizability and medicinal chemistry heuristics. It is context-dependent (depending on the training set) and prone to subjectivity (assessment of drug-likeness can vary from one therapeutic area and one researcher to another). We can nonetheless measure it using a molecular Turing test. The molecular Turing test has already been described by researchers from GSK.³³ In this procedure, generated molecules are shown

to researchers along the training set, and they are asked whether they would consider them for synthesis or not. It is then possible to compare the acceptance rate of molecules generated compared to the acceptance rate of test set compounds, which are designed by humans. Drug-like compounds, in the context of this work, refer to molecules whose acceptance rates in this molecular Turing test are of the same magnitude as those of human-generated compounds.

METHODS

An empirical approach to identifying AD definitions for generative models requires making several modeling choices. In this section, we explain the rationale behind the choices we make regarding AD definitions studied, data sets, generative models, and evaluation of the results.

Existing Applicability Domain Definitions. The main methods for defining a predictive QSAR AD can be classified as follows: chemical-physical, structural, fragment-based, and those based on the response domain.¹⁷ Most methods based on chemical-physical and structural description rely in fine on some form of similarity metric with respect to the model’s training set. Appreciation of the structural similarity to the training set is dependent both on the molecular descriptors and on the measure of similarity used. To restrict our search space of an AD definition for generative models, we make several choices to focus on the descriptors and AD definitions that we estimate to be the more relevant in practice. For example, some classical approaches we have left apart for AD definition, such as using structural fragment-based approaches or k-nearest neighbor similarity,¹⁷ could also be of interest for improving the drug-likeness during generation.

Descriptors. We explore three main families of molecular descriptors that relate to different kinds of molecular descriptions. The first ones are extended-connectivity fingerprints (ECFP),³⁴ implemented within the RDKit³⁵ using the Morgan algorithm.³⁶ In those fingerprints, features represent the presence or absence of given substructures in the molecule. ECFP descriptors can be either count-based or binary. When count-based, a feature takes the value of the number of times the substructure is found in the molecule. When binary, the features can be set to either 0 or 1, denoting only the presence or absence of the substructure. As shown in Figure 3, count-based fingerprints can discriminate between molecular structures that binary fingerprints cannot discriminate. We explore both count-based and binary fingerprints, in addition to different radii for the Morgan algorithm (2 and 3, which correspond to ECFP4 and ECFP6 descriptors, respectively) for the fingerprints. The second kind of molecular descriptors we consider are atom-pair (AP) fingerprints³⁷ as implemented in the RDKit. In AP

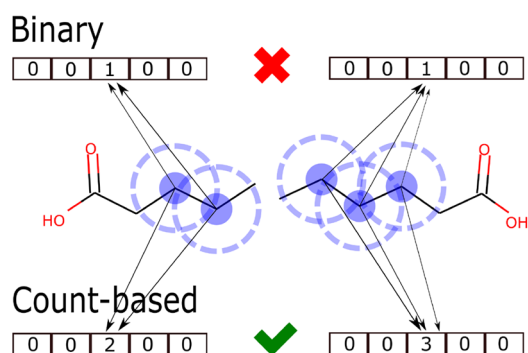


Figure 3. Limitations of binary fingerprints to discriminate unusual chemical moieties. (Top) Binary. (Bottom) Count.

fingerprints, features denote the presence or absence of a given pair of atoms. We also assess the quantitative estimate of drug-likeness score¹¹ as a potential descriptor. The QED is a scale for gauging the drug-likeness of a molecular structure that is widely used,⁹ even if it is known that lower QEDs in drugs do not necessarily translate into worse PK parameters.³⁸ Finally, we also explore molecular descriptors based on a combination of physicochemical descriptors: the number of hydrogen bonds donors and acceptors, the number of rings, the number of rotatable bonds, the total polar surface area, Crippen descriptors⁵ (clogP and molar refractivity), the molecular weight, the fraction of sp³ carbons, the ratio of atoms in the Murcko scaffold on the total number of heavy atoms, the number of heavy atoms, the maximum and minimum ring size, the minimal, maximal and total charge, and the number of chiral centers.

Similarity Measures. Several similarity measures have been explored for defining applicability domains based on chemical-physical and structural description in the context of QSAR modeling.^{16,17} Here, we explore both range-based and distance-based methods. As an exhaustive exploration of all similarity measures is not feasible, we restrict ourselves to two similarity measures. The principal range-based AD definitions are the bounding box approach and the convex hull approach.¹⁶ The bounding-box approach checks whether the individual values of each feature lie within the range observed for this given feature in the training set, while the convex hull method defines the interpolation space as the smallest convex area that contains the descriptors from the training set. In practice, computing the convex-hull has an algorithmic complexity in $\Omega(n^d)$, where n is the number of molecules and d the number of descriptors.³⁹ This makes it quickly impractical as soon as the number of descriptors used grows. Thus, the bounding-box approach is often used to approximate the convex hull.⁴⁰ This led us to select the bounding-box approach as a representative range-based AD. By default, we defined the bounding box using the minimum and maximum value of the descriptors, but it is also possible to use measures more robust to outliers such as percentiles (e.g., first through 99th percentiles). For distance-based approaches, we use an AD definition based on the Tanimoto similarity, which is widely used in cheminformatics.⁴¹ We determine whether a molecule is within the AD by assessing whether the maximum similarity value to molecules of the training set is above a prespecified threshold.

Applicability Domains. Using these descriptors and similarity measures, we define several applicability domains using either a similarity measure and a set of descriptors (e.g.,

range of ECFP4) or a combination of a pair of descriptors and similarity measures (e.g., range of ECFP4 and range of physicochemical descriptors). We also investigate SMILES validity as a baseline applicability domain. Indeed, SMILES-based generative models can produce invalid SMILES chains, which do not correspond to a valid molecular structure. The simplest applicability domain for generative models is therefore an applicability domain that only checks whether the SMILES chain corresponds to a valid structure. As a second baseline, we also include for JAK2 an AD that checks the presence of undesirable fragments published with the ChEMBL database⁴² and curated by P. Walters.⁴³ The full list of applicability domain definitions and the names by which we refer to them is given in Table 1.

Table 1. Applicability Domain Definitions Used Throughout This Work

name	definition
SMILES validity	proposed SMILES string corresponds to a valid molecular structure, as computed by the RDKit ⁴³
filters validity	structure contains none of the substructures flagged in ChEMBL ⁴²
maxsim ECFP4	maximum Tanimoto similarity (Morgan fingerprints, radius 2) between a structure and the training set is >0.5
maxsim ECFP6	maximum Tanimoto similarity (Morgan fingerprints, radius 3) between a structure and the training set is >0.5
maxsim AP	maximum Tanimoto similarity (atom-pair fingerprints) between a structure and the training set is >0.5
range QED	QED of the structure is within the range of the QED of the training set
range ECFP4	Morgan fingerprints of radius 2 of the structure are within the range of those of the training set
range ECFP4 counts	count-based Morgan fingerprints of radius 2 of the structure are within the range of those of the training set
range physchem	physicochemical descriptors of the structure are within the range of those of the training set
range physchem + range AP	checks if the structure is in "range physchem" and atom-pair fingerprint bits of the structure are within the range of those of the training set
range physchem + range ECFP4	checks if the structure is in "range physchem" and in "range ECFP4"
range physchem + range ECFP4 counts	checks if the structure is in "range physchem" and in "range ECFP4 counts"
range physchem + range ECFP6	checks if the structure is in "range physchem" and in "range ECFP6"
range physchem + maxsim ECFP4	checks if the structure is in "range physchem" and valid according to "maxsim ECFP4"
range physchem + maxsim ECFP6	checks if the structure is in "range physchem" and valid according to "maxsim ECFP6"
range physchem + maxsim AP	checks if the structure is in "range physchem" and valid according to "maxsim AP"

Experimental Setting. The generative model we focus on is a SMILES-based long short-term memory network (LSTM).^{44,45} A SMILES-based LSTM models the conditional probability distribution over SMILES strings (conditioned on the beginning of a SMILES). It can be used to generate sequentially novel SMILES strings. The LSTM is optimized with a hill-climbing algorithm.⁴ At each time step, the LSTM generates novel structures and the hill-climbing algorithm fine-tunes the LSTM on the best molecules of the batch. Molecules that fall outside of the applicability domain have their score set at 0, which effectively discards them from being selected by the hill-climbing algorithm. This generative approach was chosen

because it is widely studied, popular in the field, and prone to generating compounds that would not be considered as drug-like.⁶ Therefore, this approach is adapted for the development of a generative applicability domain. In order to assess whether our results are transferable to other generative methods, we also ran our study using two of the generative approaches from the Guacamol baselines⁴ on the JAK2 data set.

For the predictive models, we use random forest classifiers,⁴⁶ where the predicted probability of being active is used as the reward for the reinforcement learning of the generative model, but the generation can be performed with predictive models coming from other machine learning algorithms. Random forest models have been shown to be robust to adversarial examples.⁴⁷ The descriptors used are folded ECFP fingerprints³⁴ of size 1024 and radius 2. The implementation is done with the python library scikit-learn⁴⁸ and the RDKit.³⁵ As described in Figure 1, the data sets are split at random in a 75:25 fashion to build a training set and a test set. The training set itself is split in a 75:25 fashion; the largest set is used to build the QSAR model, and the smallest is used to pretrain the generative algorithm and define the applicability domain. The test set is therefore unseen by the generative algorithm during the whole process, and the AD definition is not related to the training set of the activity model.

We study potential generative applicability domains on three lead optimization-like data sets and one hit-finding like data set, whose characteristics are listed in Table 2. The first one, the

Table 2. Datasets Used for Benchmarking Generative Applicability Domains^a

	JAK2	Renin	11 β HSD	ChEMBL11 β HSD
molecules	667	142	1409	166
actives	140	32	792	8
origin	ChEMBL	Sanofi	Sanofi	ChEMBL
internal diversity	0.47	0.54	0.79	0.82

^aInternal diversity is computed as the average of the intermolecule Tanimoto distance on Morgan fingerprints (with 1024 bits and a radius of 2) for the dataset.

JAK2 data set, is a public data set extracted from ChEMBL.⁴⁹ Molecules with a pIC₅₀ greater than 8 were labeled as actives. The second and the third ones, the Renin data set^{50–52} and the 11 β HSD data set, respectively, are extracted from internal Sanofi research programs. These three data sets are diverse in size (small for Renin, medium sized for JAK2, and large for 11 β HSD), in type of therapeutic target (JAK2 is a kinase, Renin a protease, and 11 β HSD a dehydrogenase), and in origins (coming from both public and corporate databases). Furthermore, the JAK2 and Renin data set are comprised of a single chemical series, while the 11 β HSD data set contains two distinct chemical series, ureas and oxathiazines. By studying these three data sets, we provide insight as to how the generative applicability domain performs in different lead optimization situations. As our results can be of interest for applications of generative AI to other stages of drug discovery than lead optimization, we also investigate a 11 β HSD data set extracted from ChEMBL. It features some of the molecules from the internal data set that were published but spans a more heterogeneous chemical space. Every AD definition is assessed on all data sets, except for the filters validity. Indeed, as this AD produces molecules that are not drug-like on the JAK2 data set, and as its runtime is long compared to the other AD definitions, we do not assess it on the other tasks.

Evaluating Applicability Domains. To evaluate each AD definition, we first perform 10 runs where the 128 best scoring molecules are kept for each of the 150 epochs of the generation. We then compute several metrics related to diversity, recovered actives, physicochemical descriptors, and drug-likeness metrics. It is important to note that those metrics are not necessarily objectives in themselves (otherwise we could include them directly in the scoring function) but rather serve as proxies to evaluate whether generated molecules are drug-like or not. While our final objective is to evaluate the capacity to generate molecules that would be selected for synthesis by a drug discovery project team, running a molecular Turing test to measure this outcome is time- and resource-consuming and cannot be performed for each AD definition on each data set. Thus, this set of metrics allows us to evaluate extensively all the AD on all data sets before confirming the results with a molecular Turing test. In the next section, we will describe how the evaluation of the applicability domains were performed and illustrate it with examples from the JAK2 data set.

Visualization of Generated Molecule Sets. The most basic way to evaluate the generated molecules as a set is the comparison of this set with the training data set in a common chemical space. A principal component analysis (PCA) is performed on the Morgan fingerprints of the data set. PCA is a dimensionality reduction technique that allows us to project the high-dimensional fingerprints to a space of reduced dimension for visualization. It was performed using the scikit-learn⁴⁸ implementation with default parameters and the number of components set to 2. Then, both the data set and generated molecules are projected on the first two principal components. This allows for visual comparison of the generated molecules with the original molecules, as shown in Figure 4. This visualization is nonetheless dependent on the descriptors used and might not be sufficient to assess the drug-likeness and diversity of generated molecules. Indeed, the addition of an incorrect small substituent to a correct structure will likely not be captured in this visualization. Nevertheless, the three ADs represented in the top row show a different distribution than those in the middle and bottom rows; this indicates that more diverse molecules are generated by the former compared to the latter, and that some of the generated molecules come close to, or fall into, three clusters where known actives are located.

Diversity. An important aspect of evaluating the output of generative models is the diversity of generated compounds.⁵³ In order to quantify the diversity of the output when using each applicability domain, we first measure its internal diversity based on the Tanimoto similarity coefficient (computed for ECFP fingerprints of radius 2 and with 1024 bits). Internal diversity of a set of molecules is defined as 1 minus the average Tanimoto similarity between two molecules of the set. While this measure indicates the chemical diversity of the set of molecules, it also bears inherent limitations, as it summarizes the distribution of the intermolecule Tanimoto similarity to its mean. It is also dependent on the similarity measure used. We also define a second diversity metric described in Figure 5 to evaluate the generated sets. The training set is clustered with the k-means algorithm.⁵⁴ After the parameter search on the three data sets, the number of clusters was set to five, as this hyperparameter yielded a qualitatively reasonable clustering on the three data sets. Then, each generated molecule is assigned to one of the clusters obtained. The entropy of the generated molecules' repartition within the clusters is used as another measure of diversity. As shown in Figure 5, the less even the distribution of

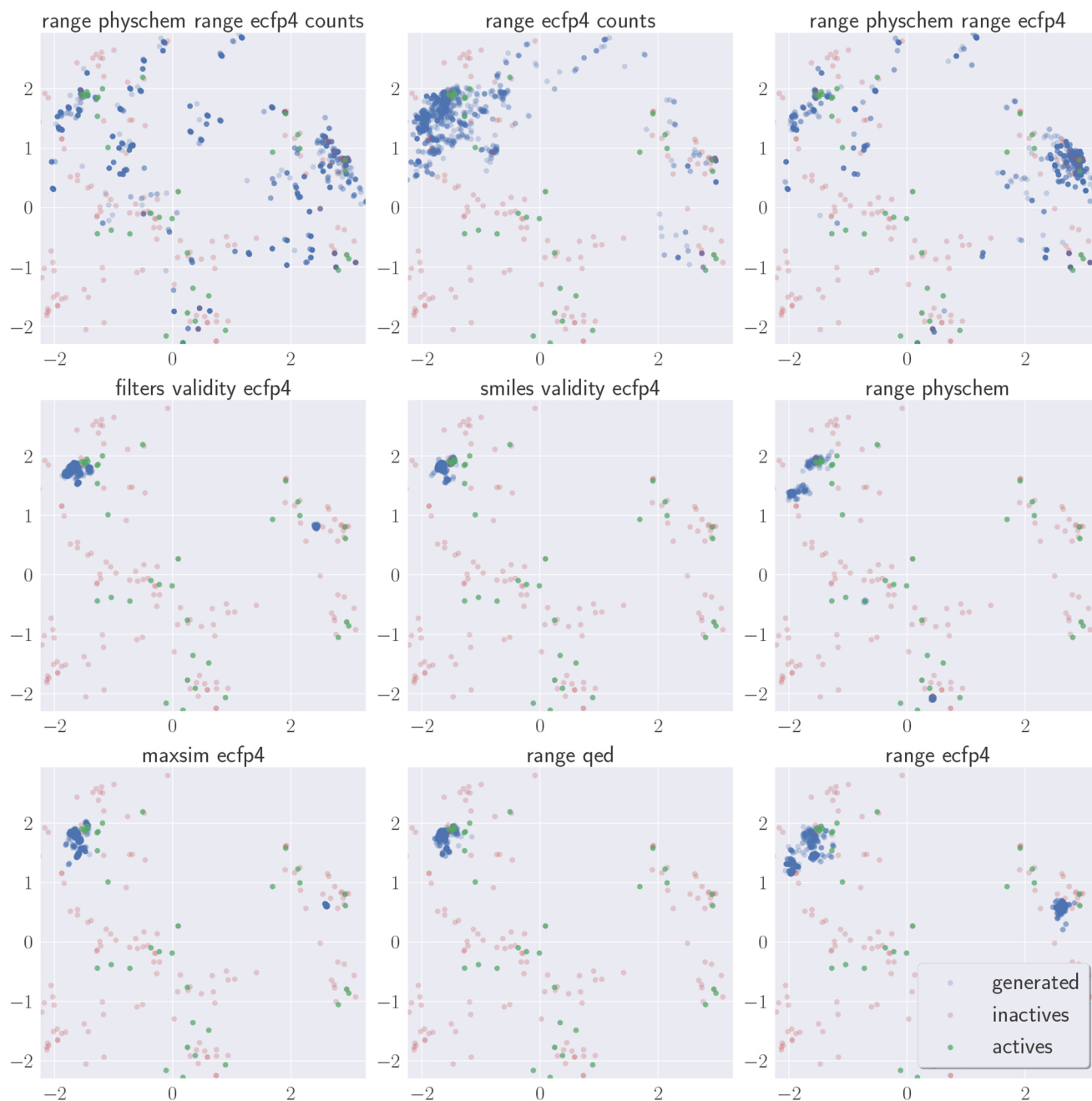


Figure 4. JAK2: Projection of molecules generated with the LSTM-HC model on the original data set using the first two dimensions of the PCA of their Morgan fingerprints. Blue dots represent generated molecules, green dots represent actives from the test set, and red dots represent inactives. The three AD metrics in the top row lead to more diverse molecules than the ones in the middle and bottom rows.

molecules within the clusters, the lower the entropy. Maximal entropy is reached when generated molecules are distributed evenly across the clusters and therefore span roughly the same chemical space as the training set.

Finally, we also compute the number of clusters found on the generated set using the Butina algorithm⁵⁵ with a 0.5 Tanimoto cutoff. As for internal diversity, these measures are limited both by the descriptors used to represent molecules and by the clustering algorithm used. Nonetheless, coupling the three measures yields good insight on the molecular diversity of the generated sets.

Recovered Actives. Goal-directed generation in the context of lead optimization is aimed at discovering novel bioactive

molecules. For each generative applicability domain, we assess whether the generative algorithm can retrieve actives from the test set (that were held-out during the training of the QSAR model and the pretraining and reinforcement of the generative algorithm). We report the percentage of actives for which there was a generated molecule with a Tanimoto similarity above 0.9 (i.e., the generative algorithm found a very close analog) and with a Tanimoto similarity of 1 (exact match). Assessing whether the generative algorithm can retrieve unseen known actives constitutes the best proxy that we have for its ability to generate novel actives.

Measures of Drug-Likeness. Different scores have been developed to assess the drug-likeness of a molecule in a

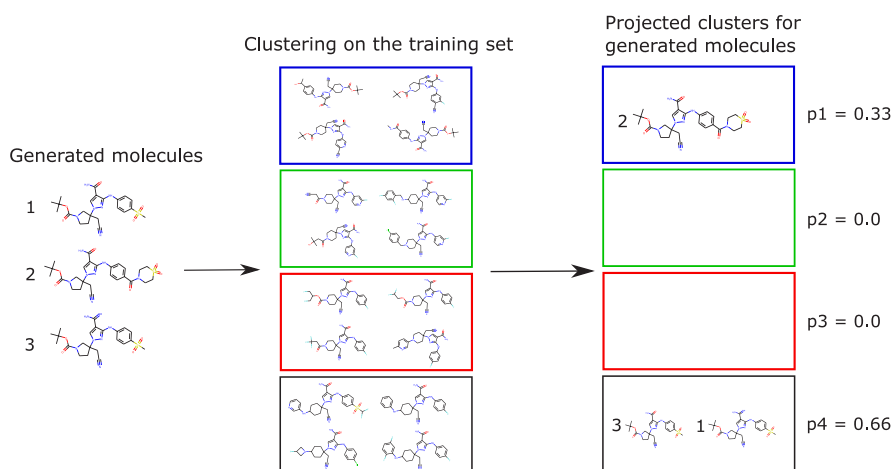


Figure 5. Entropy as a measure of the coverage of the training set diversity. The more homogeneous the repartition of the generated molecules between clusters is, the higher the entropy will be. Generated molecules too far from the training set are set apart. This example displays molecules from the JAK2 data set.

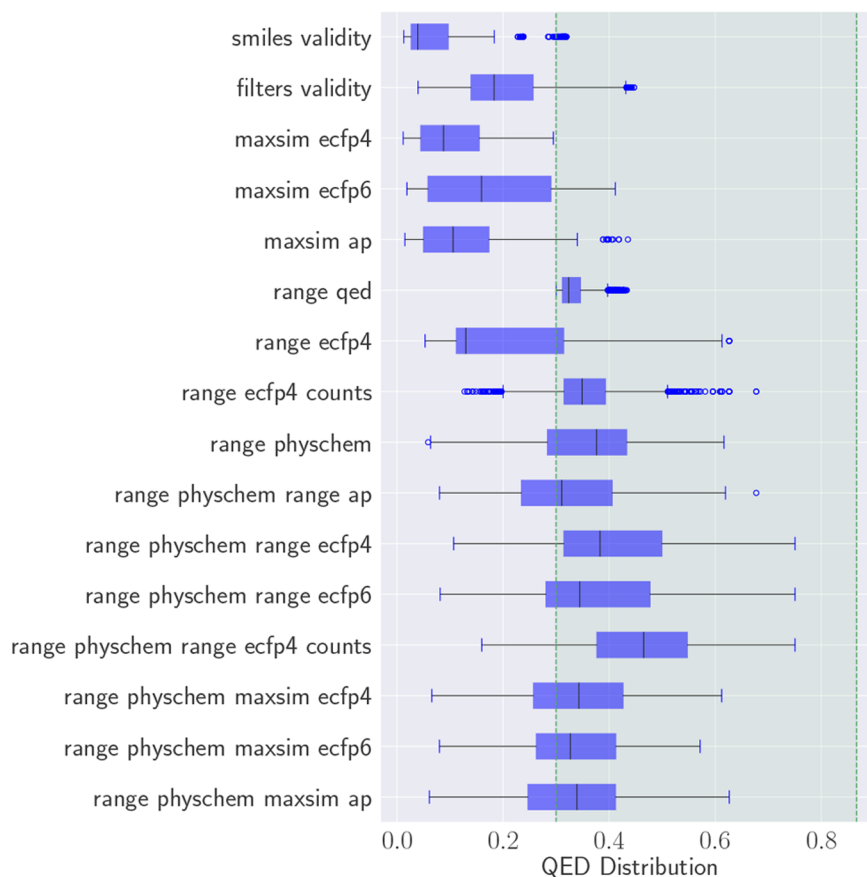


Figure 6. QED distribution of generated molecules for different applicability domain definitions on the JAK2 data set. The vertical green lines correspond to the minimum and maximum values for QED found in the training set, where higher is better.

medicinal chemistry context. We use the QED¹¹ and the SAS²⁸ to evaluate the generated compounds. As these values are dependent on the chemical series explored by the generative algorithm, we compute the most extreme values found in the data set both for SAS and QED and compare them with the distribution SAS and QED of all the molecules generated during the different runs, as described in Figure 6. We also report the mean and the standard deviation of the percentage of molecules generated at each run that fall between these extreme values.

Proxy Drug-Likeness Descriptors. We examined the 50 best scoring compounds generated by the generative process using the different ADs. Through this qualitative analysis, we identified several simple descriptors that were markedly different between the training set and the generated set when an inappropriate AD was used during generation on JAK2. Those descriptors are the following: the number of halogen atoms, the number of sulfur atoms, the number of unpaired electrons in their valence shell, the number of heteroatom–heteroatom

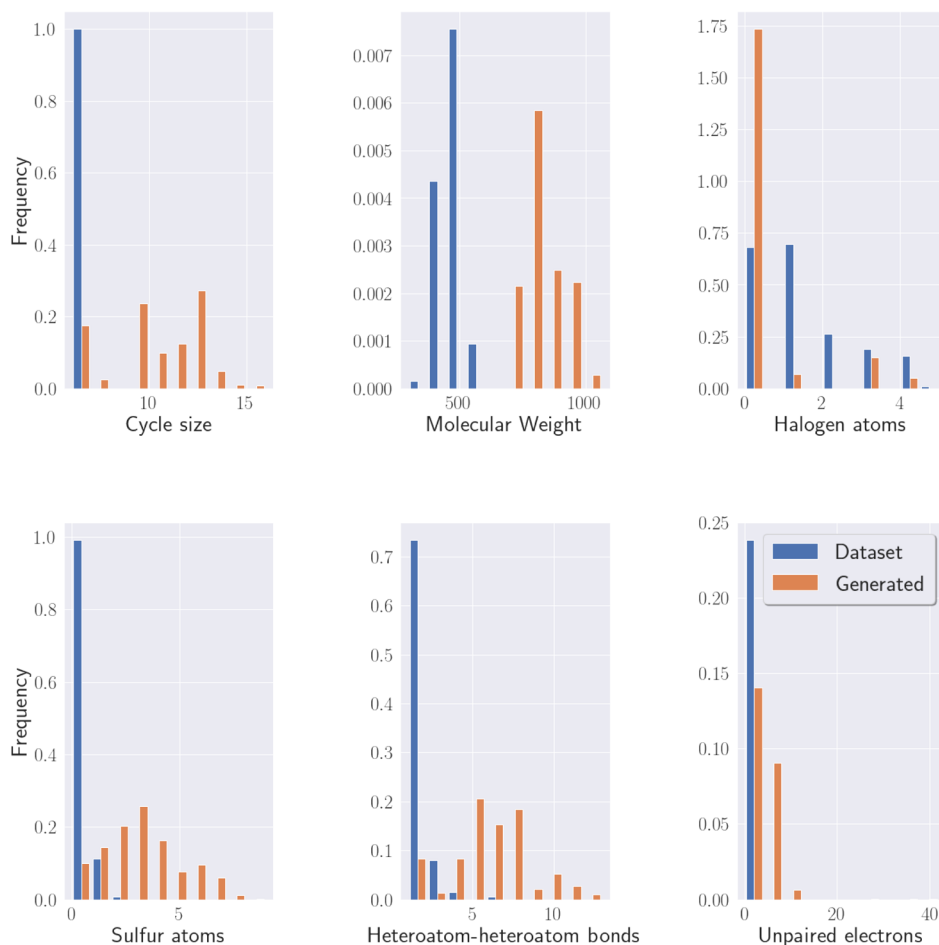


Figure 7. Comparison of data set and generated molecules on the JAK2 test case (here with the maximum similarity on atom-pair descriptors) with the distributions for properties identified as important through qualitative analysis.

bonds, the maximum ring size, and the molecular weight (see Figure 7). For each AD, we computed the distribution of those properties. In a similar fashion to SAS and QED, we also compute the most extreme values found in the data set for each of those properties and report the mean and standard deviation of the percentage of molecules generated at each run that fall between those values. It is to note that the presence of heteroatom-heteroatom bonds in a molecule is not a problem per se but that bad AD might lead to great number of them in the generated structures.

The distributions of these descriptors have been used to quickly sort out bad AD for all data sets before double-checking the relevance of the AD and of the proxy descriptors by visual inspection of the generated structures. Figure 7 shows the results produced by a bad applicability domain definition. One can see that molecules generated have on average a molecular weight close to 1000 Da, while it is 500 Da in the data set. Figure 7 also shows a much higher number of sulfur atoms and heteroatom-heteroatom bonds in generated molecules. Those results point to the fact that the generated molecules are poorly drug-like. This first analysis allows us to discard obviously bad AD definitions in order to focus on those that pass this filter. It is to note that most applicability domains, even the best ones, can sometimes generate peculiar patterns of problematic structures, and that bad ADs very often lead to the generation of specific structural patterns as illustrated in Table S1 and Table S2.

RESULTS

In this section, we report the systematic evaluation of the different applicability domains definitions on each data set. For each data set, we identify the best applicability domains according to the evaluation metrics described in the previous section.

JAK2 Data Set. On the JAK2 data set (see Table 3 for the full results), most AD definitions fail to generate molecules with molecular weights, ring sizes, number of heteroatoms, and number of radicals within the range of the training set. Three ADs stand out: “range physchem + range ECFP4”, “range physchem + range ECFP6” and “range physchem + range ECFP4 counts”. These three methods produce overall molecules in the range of the training set for the different proxy metrics and retrieve the same proportion of unseen actives. The AD definition “range physchem + range ECFP4 counts” stands out with the highest percentage of generated molecules in the same QED range as the training set and a high diversity of generated molecules.

For the JAK2 data set, we also explore the impact of AD for other goal-directed generation approaches from the set of Guacamol baselines.⁴

Graph GA. For the Graph GA algorithm (Table 4), there are more AD definitions for which the generated set scores highly on the evaluation metrics. Noteworthy, we see that the metrics identified in section Proxy Drug-Likeness Descriptors are almost

Table 3. Evaluation of the Molecule Sets Generated with the Different Applicability Domains on the JAK2 Dataset^a

	Number of clusters	Entropies	Internal diversity	% actives recovered at d < 0.1	% actives recovered	% valid SAS	% valid QED	% valid cycle sizes	% valid MW	% valid het-het bonds	% closed shell
smiles validity	1.0 (0.0)	0.0 (0.0)	0.23	0.0 (0.0)	0.0 (0.0)	0.0	0.0	0.8	0.0	3.4	49.7
			(0.04)			(0.0)	(0.0)	(1.6)	(0.0)	(10.1)	(49.7)
filters validity	1.0 (0.0)	0.0 (0.0)	0.22	0.0 (0.0)	0.0 (0.0)	2.81	18.05	18.8	0.0	100.0	55.2
			(0.03)			(0.0)	(0.0)	(16.53)	(38.46)	(35.4)	(0.0)
maxsim ecfp4	1.0 (0.0)	0.0 (0.0)	0.2	0.0 (0.0)	0.0 (0.0)	0.0	0.0	10.0	0.0	34.7	69.9
			(0.05)			(0.0)	(0.0)	(0.0)	(0.0)	(30.0)	(0.0)
maxsim ecfp6	1.0 (0.0)	0.0 (0.0)	0.21	0.0 (0.0)	0.0 (0.0)	0.0	6.33	0.0	0.0	30.5	53.0
			(0.07)			(0.0)	(0.0)	(0.0)	(24.35)	(0.0)	(0.0)
maxsim ap	1.0 (0.0)	0.0 (0.0)	0.25	0.0 (0.0)	0.0 (0.0)	0.0	0.94	17.6	0.0	64.6	22.3
			(0.03)			(0.0)	(0.0)	(0.0)	(9.64)	(35.6)	(0.0)
range qed	1.0 (0.0)	0.0 (0.0)	0.22	0.0 (0.0)	0.0 (0.0)	0.0	16.48	0.0	0.0	66.2	49.8
			(0.04)			(0.0)	(0.0)	(0.0)	(37.1)	(0.0)	(0.0)
range ecfp4	2.0 (0.45)	(0.05)	0.05	0.0 (0.0)	0.0 (0.0)	32.66	14.77	27.8	1.6	100.0	100.0
			(0.08)			(0.0)	(0.0)	(46.9)	(35.48)	(34.6)	(2.9)
range ecfp4 counts	2.7 (0.9)	(0.03)	0.39	0.0 (0.0)	0.0 (0.0)	61.72	29.45	71.7	19.1	100.0	100.0
			(0.05)			(0.0)	(0.0)	(48.61)	(45.58)	(35.5)	(11.3)
range physchem	1.3 (0.46)	0.0 (0.0)	0.29	0.0 (0.0)	0.0 (0.0)	22.97	51.02	100.0	100.0	99.8	8.6
			(0.06)			(0.0)	(0.0)	(42.06)	(49.99)	(0.0)	(0.0)
range physchem range ap	2.2 (0.4)	(0.04)	0.34	0.0 (0.0)	0.0 (0.0)	51.72	37.81	100.0	100.0	99.8	29.3
			(0.04)			(0.0)	(0.0)	(49.97)	(48.49)	(0.0)	(0.0)
range physchem range ecfp4	3.1 (0.94)	(0.01)	0.4	19.2 (3.8)	19.2 (3.8)	94.92	53.12	100.0	100.0	100.0	99.8
			(0.03)			(0.0)	(0.0)	(21.96)	(49.9)	(0.0)	(0.0)
range physchem range ecfp6	4.2 (1.25)	(0.02)	0.43	16.2	14.6 (8.7)	90.7	44.45	100.0	100.0	100.0	92.7
			(0.04)			(10.6)	(0.0)	(29.04)	(49.69)	(0.0)	(0.0)
range physchem range ecfp4 counts	6.5 (0.92)	(0.01)	0.49	30.8 (0.0)	23.1 (0.0)	85.31	76.25	100.0	100.0	100.0	99.9
			(0.01)			(0.0)	(0.0)	(35.4)	(42.56)	(0.0)	(0.0)
range physchem maxsim ecfp4	1.4 (0.8)	0.0 (0.0)	0.29	0.0 (0.0)	0.0 (0.0)	15.47	39.84	100.0	100.0	99.5	14.6
			(0.07)			(0.0)	(0.0)	(36.16)	(48.96)	(0.0)	(0.0)
range physchem maxsim ecfp6	1.1 (0.3)	0.0 (0.0)	0.28	0.0 (0.0)	0.0 (0.0)	21.25	40.0	100.0	100.0	100.0	13.7
			(0.05)			(0.0)	(0.0)	(40.91)	(48.99)	(0.0)	(0.0)
range physchem maxsim ap	1.4 (0.66)	0.0 (0.0)	0.3	0.0 (0.0)	0.0 (0.0)	16.41	36.02	100.0	100.0	99.7	16.4
			(0.03)			(0.0)	(0.0)	(37.03)	(48.0)	(0.0)	(0.0)

^aThe three first columns are measures of the diversity of generated molecules, while the following columns are measures of the quality of generated molecules. For all columns, higher is better, and therefore darker shades indicate better performances. For each column with a name that starts with “% valid”, the reported number is the percentage of molecules for which the value for the property falls in the range observed in the training set for this property. The standard deviations over ten replica are indicated between parentheses. We see that applicability domains that combine a range of physicochemical descriptors and a range of Morgan fingerprints perform best.

satisfied for 100% of molecules for every AD. Overall, in addition to the ones identified above for the LSTM-HC method, the “range ECFP4”, “range ECFP4 counts”, and “range physchem + range AP” AD definitions show good results. This suggests that the efficiency of AD in keeping the generation in a drug-like chemical space is in part dependent on the generative method.

SMILES GA. Interestingly, the results for the SMILES GA algorithm (Table 5) show the same AD definitions performing well as for the Graph GA algorithm.

Overall, results across Guacamol’s goal-directed generation algorithms suggest that the most stringent AD definitions (e.g., “range physchem + range ECFP4 counts”) seem to perform well across algorithms. They also show that the LSTM-HC approach (perhaps due to its high flexibility as it operates directly on the space of all SMILES strings) requires a more stringent AD than its graph or SMILES based genetic algorithm counterparts.

On the other hand, the evolution of scores when generating molecules under the constraint of a stringent AD (“range physchem + range ECFP4 counts”, see Figure 8) shows that LSTM-HC reaches far higher scores than the Graph GA and SMILES GA. Generative models have been suspected⁶ of overoptimizing their scoring functions, thereby producing unrealistic molecules. Generation under the constraint of a well-chosen AD (as displayed in Figure 8) prevents this behavior and shows the ability of a generative approach to generate

optimized molecules while maintaining their drug-likeness. In this respect, LSTM-HC seems to have an edge over its counterparts, reaching higher scores. On the other hand, the other goal-directed generation algorithms generally display better statistics than LSTM-HC with respect to the proportion of actives recovered and in terms of valid compounds generated.

Renin Data Set. On the Renin data set (see Table 6), the same three ADs (“range physchem + range ECFP4”, “range physchem + range ECFP6”, and “range physchem + range ECFP4 counts”) still lead to the highest values on the evaluation metrics. Nonetheless, two other AD, “range of physchem + range AP” and “range ECFP4 counts”, also lead to molecules sets that are well scored using the proxy descriptors and are largely in the same range of SAS and QED as the training set. Among those different applicability domains definitions, “range physchem + range ECFP4 counts” and “range of physchem + range AP” stand out due to the high percentage of identified actives and high diversity of generated molecules.

On the Renin data set, there are more applicability domain definitions that produce drug-like molecule sets in comparison to the JAK2 data set. This could be explained by the smaller size of the Renin data set. Indeed, with less molecules in the training set, an applicability domain definition will become more stringent. This can explain that an AD that is not sufficiently

Table 4. Evaluation of the Molecule Sets Generated with the Graph GA Algorithm and the Applicability Domains on the JAK2 Dataset^a

	Number of clusters	Entropies	Internal diversity	% actives recovered at d < 0.1	% actives recovered	% valid SAS	% valid QED	% valid cycle sizes	% valid MW	% valid het-het bonds	% closed shell
smiles validity	1.5 (0.67)	0.01	0.3	1.5 (3.1)	1.5 (3.1)	17.4	20.1	100.0	77.7	99.7	100.0
		(0.03)	(0.05)			(37.91)	(40.07)	(0.0)	(16.1)	(0.5)	(0.0)
maxsim ecfp4	1.4 (0.66)	0.02	0.34	2.3 (3.5)	2.3 (3.5)	14.4	22.6	100.0	80.6	99.2	100.0
		(0.05)	(0.08)			(35.11)	(41.82)	(0.0)	(8.1)	(2.4)	(0.0)
maxsim ecfp6	1.6 (0.66)	0.02	0.36	2.3 (3.5)	2.3 (3.5)	31.6	19.4	100.0	82.6	100.0	100.0
		(0.05)	(0.1)			(46.49)	(39.54)	(0.0)	(13.9)	(0.0)	(0.0)
maxsim ap	1.8 (0.6)	0.01	0.33	3.1 (3.8)	3.1 (3.8)	4.4	23.9	100.0	83.8	99.6	100.0
		(0.02)	(0.06)			(20.51)	(42.65)	(0.0)	(4.7)	(1.2)	(0.0)
range qed	1.4 (0.92)	0.01	0.32	3.8 (3.8)	3.8 (3.8)	11.8	33.0	100.0	82.5	99.9	100.0
		(0.02)	(0.04)			(32.26)	(47.02)	(0.0)	(11.6)	(0.3)	(0.0)
range ecfp4	3.3 (0.64)	0.22	0.47	17.7 (8.5)	17.7 (8.5)	74.8	82.9	100.0	99.6	100.0	100.0
		(0.01)	(0.01)			(43.42)	(37.65)	(0.0)	(0.5)	(0.0)	(0.0)
range ecfp4 counts	3.7 (1.0)	0.23	0.47	16.2 (5.4)	16.2 (5.4)	77.5	90.0	100.0	100.0	100.0	100.0
		(0.01)	(0.01)			(41.76)	(30.0)	(0.0)	(0.0)	(0.0)	(0.0)
range physchem	2.1 (1.14)	0.06	0.4	6.9 (2.3)	6.9 (2.3)	33.3	39.7	100.0	100.0	100.0	100.0
		(0.07)	(0.07)			(47.13)	(48.93)	(0.0)	(0.0)	(0.0)	(0.0)
range physchem range ap	4.7 (0.78)	0.21	0.45	20.0 (6.2)	20.0 (6.2)	84.1	91.4	100.0	100.0	100.0	100.0
		(0.02)	(0.01)			(36.57)	(28.04)	(0.0)	(0.0)	(0.0)	(0.0)
range physchem range ecfp4	3.7 (1.1)	0.21	0.46	14.6 (5.4)	14.6 (5.4)	80.3	89.2	100.0	100.0	100.0	100.0
		(0.01)	(0.01)			(39.77)	(31.04)	(0.0)	(0.0)	(0.0)	(0.0)
range physchem range ecfp6	3.5 (0.81)	0.22	0.45	14.6 (5.4)	14.6 (5.4)	81.4	86.8	100.0	100.0	100.0	100.0
		(0.01)	(0.01)			(38.91)	(33.85)	(0.0)	(0.0)	(0.0)	(0.0)
range physchem range ecfp4 counts	3.9 (1.04)	0.23 (0.0)	0.47	16.9 (5.8)	16.9 (5.8)	76.1	93.6	100.0	100.0	100.0	100.0
		0.03	(0.01)			(42.65)	(24.48)	(0.0)	(0.0)	(0.0)	(0.0)
range physchem maxsim ecfp4	1.8 (0.6)	0.03	0.4	7.7 (0.0)	7.7 (0.0)	17.4	48.1	100.0	100.0	99.1	100.0
		(0.05)	(0.04)			(37.91)	(49.96)	(0.0)	(0.0)	(2.7)	(0.0)
range physchem maxsim ecfp6	1.8 (0.75)	0.04	0.37	7.7 (0.0)	7.7 (0.0)	11.4	42.7	100.0	100.0	99.9	100.0
		(0.06)	(0.08)			(31.78)	(49.46)	(0.0)	(0.0)	(0.3)	(0.0)
range physchem maxsim ap	2.4 (1.2)	0.04	0.38	7.7 (0.0)	7.7 (0.0)	13.0	38.8	100.0	100.0	100.0	100.0
		(0.05)	(0.06)			(33.63)	(48.73)	(0.0)	(0.0)	(0.0)	(0.0)

^aFor the “range QED” AD definition, the score for the QED is below 100%. This is due to the fact that molecules should fall between the 5th and 95th percentiles to have a valid QED, while the AD checks if the QED is between the most extreme values. Thus, the AD definition is less stringent than the evaluation metric.

Table 5. Evaluation of the Molecule Sets Generated with the SMILES GA Algorithm and the Different Applicability Domains on the JAK2 Dataset

	Number of clusters	Entropies	Internal diversity	% actives recovered at d < 0.1	% actives recovered	% valid SAS	% valid QED	% valid cycle sizes	% valid MW	% valid het-het bonds	% closed shell
smiles validity	3.2 (0.75)	0.11	0.49	0.0 (0.0)	0.0 (0.0)	49.8	15.6	100.0	71.7	100.0	97.0
		(0.05)	(0.06)			(50.0)	(36.29)	(0.0)	(10.0)	(0.0)	(5.8)
maxsim ecfp4	2.9 (0.3)	0.11	0.51	0.0 (0.0)	0.0 (0.0)	58.7	19.3	100.0	74.7	100.0	97.9
		(0.04)	(0.03)			(49.24)	(39.47)	(0.0)	(9.8)	(0.0)	(3.8)
maxsim ecfp6	2.9 (0.7)	0.1 (0.05)	0.48	0.0 (0.0)	0.0 (0.0)	50.5	15.5	100.0	72.1	100.0	97.4
		(0.07)	(0.07)			(50.0)	(36.19)	(0.0)	(10.9)	(0.0)	(4.9)
maxsim ap	4.3 (1.0)	0.11	0.52	0.0 (0.0)	0.0 (0.0)	52.4	15.5	100.0	73.5	99.2	98.4
		(0.04)	(0.05)			(49.94)	(36.19)	(0.0)	(8.7)	(2.1)	(3.5)
range qed	4.3 (0.64)	0.13	0.55	5.4 (6.9)	5.4 (6.9)	71.1	51.1	100.0	95.7	100.0	96.1
		(0.02)	(0.02)			(45.33)	(49.99)	(0.0)	(3.1)	(0.0)	(5.5)
range ecfp4	5.5 (0.92)	0.21	0.47	100.0	100.0	85.8	93.4	100.0	98.9	100.0	99.6
		(0.01)	(0.01)	(0.0)	(0.0)	(34.91)	(24.83)	(0.0)	(1.4)	(0.0)	(0.5)
range ecfp4 counts	5.1 (0.7)	0.21	0.47	100.0	100.0	85.9	92.6	100.0	98.9	100.0	100.0
		(0.01)	(0.01)	(0.0)	(0.0)	(34.8)	(26.18)	(0.0)	(1.5)	(0.0)	(0.0)
range physchem	5.5 (0.92)	0.14	0.57	13.8 (7.5)	13.8 (7.5)	68.0	45.8	100.0	99.6	100.0	92.8
		(0.01)	(0.02)			(46.65)	(49.82)	(0.0)	(0.7)	(0.0)	(9.5)
range physchem range ap	4.7 (1.1)	0.2 (0.02)	0.48	94.6	94.6	73.4	90.3	100.0	99.7	100.0	98.0
		(0.01)	(0.0)	(11.9)	(11.9)	(44.19)	(29.6)	(0.0)	(0.5)	(0.0)	(2.0)
range physchem range ecfp4	4.9 (1.04)	0.21	0.46	100.0	100.0	88.5	94.8	100.0	99.9	100.0	99.9
		(0.01)	(0.0)	(0.0)	(0.0)	(31.9)	(22.2)	(0.0)	(0.3)	(0.0)	(0.3)
range physchem range ecfp6	4.3 (0.46)	0.22 (0.0)	0.46	100.0	100.0	87.9	95.1	100.0	100.0	100.0	99.8
		(0.01)	(0.0)	(0.0)	(0.0)	(32.61)	(21.59)	(0.0)	(0.0)	(0.0)	(0.4)
range physchem range ecfp4 counts	4.8 (1.08)	0.21	0.46	100.0	100.0	87.9	94.4	100.0	100.0	100.0	100.0
		(0.01)	(0.01)	(0.0)	(0.0)	(32.61)	(22.99)	(0.0)	(0.0)	(0.0)	(0.0)
range physchem maxsim ecfp4	5.2 (1.47)	0.13	0.56	13.1 (9.1)	13.1 (9.1)	63.9	41.1	100.0	99.8	99.7	87.9
		(0.03)	(0.03)			(48.03)	(49.2)	(0.0)	(0.6)	(0.9)	(11.6)
range physchem maxsim ecfp6	5.2 (1.47)	0.12	0.56	13.1 (7.7)	13.1 (7.7)	61.8	40.9	100.0	99.8	99.9	90.6
		(0.02)	(0.05)			(48.59)	(49.16)	(0.0)	(0.6)	(0.3)	(13.2)
range physchem maxsim ap	5.9 (2.17)	0.13	0.58	13.8 (4.6)	13.8 (4.6)	61.4	38.2	100.0	99.4	100.0	95.5
		(0.04)	(0.03)			(48.68)	(48.59)	(0.0)	(0.5)	(0.0)	(5.6)

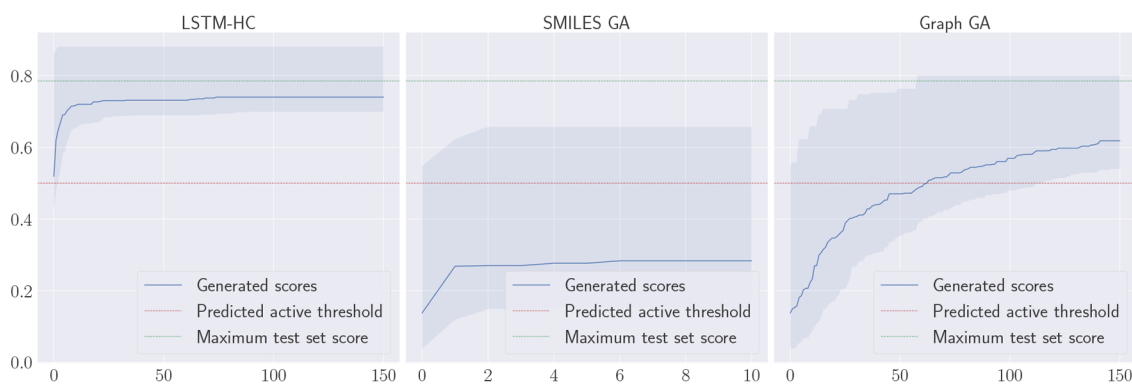


Figure 8. Comparison of scores reached by different algorithms when optimizing JAK2 predicted activity while staying in the “range physchem + range ECFP4 counts” AD. LSTM-HC reaches higher optimization scores and recovers slightly more active compounds than Graph GA and SMILES GA.

Table 6. Evaluation of the Molecule Sets Generated with the Different Applicability Domains on the Renin Dataset

	Number of clusters	Entropies	Internal diversity	% actives recovered at d < 0.1	% actives recovered	% valid SAS	% valid QED	% valid cycle sizes	% valid MW	% valid het-het bonds	% valid closed shell
smiles validity	1.0 (0.0)	0.02 (0.04)	0.21 (0.03)	0.0 (0.0)	0.0 (0.0)	0.08 (2.79)	0.0 (0.0)	40.1 (48.9)	0.0 (0.0)	9.5 (18.6)	89.7 (24.7)
maxsim ecfp4	1.0 (0.0)	0.02 (0.04)	0.18 (0.08)	0.0 (0.0)	0.0 (0.0)	12.58 (33.16)	0.0 (0.0)	39.6 (47.9)	0.0 (0.0)	42.0 (44.2)	70.4 (45.2)
maxsim ecfp6	1.0 (0.0)	0.03 (0.04)	0.16 (0.08)	0.0 (0.0)	0.0 (0.0)	14.45 (35.16)	0.08 (2.79)	70.0 (45.8)	0.1 (0.2)	51.7 (45.8)	41.6 (46.9)
maxsim ap	1.0 (0.0)	0.01 (0.01)	0.25 (0.06)	0.0 (0.0)	0.0 (0.0)	0.23 (4.84)	0.0 (0.0)	20.0 (40.0)	0.0 (0.0)	11.2 (29.7)	38.6 (47.3)
range qed	1.0 (0.0)	0.01 (0.01)	0.25 (0.03)	0.0 (0.0)	0.0 (0.0)	1.72 (13.0)	6.17 (24.06)	0.3 (0.7)	2.6 (2.6)	25.9 (37.9)	1.6 (2.3)
range ecfp4	1.0 (0.0)	0.0 (0.0)	0.01 (0.0)	0.0 (0.0)	0.0 (0.0)	2.34 (15.13)	1.56 (12.4)	0.8 (0.0)	1.6 (0.0)	100.0 (0.0)	100.0 (0.0)
range ecfp4 counts	6.0 (0.0)	0.25 (0.0)	0.51 (0.0)	100.0 (0.0)	100.0 (0.0)	95.31 (21.14)	64.92 (47.72)	65.5 (0.5)	97.2 (0.6)	100.0 (0.0)	100.0 (0.0)
range physchem	1.0 (0.0)	0.01 (0.01)	0.27 (0.01)	0.0 (0.0)	0.0 (0.0)	93.44 (24.76)	15.62 (36.31)	100.0 (0.0)	100.0 (0.0)	47.0 (4.2)	39.8 (40.7)
range physchem range ap	7.7 (1.35)	0.28 (0.01)	0.53 (0.01)	100.0 (0.0)	100.0 (0.0)	92.89 (25.7)	95.0 (21.79)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
range physchem range ecfp4	7.0 (0.45)	0.24 (0.0)	0.52 (0.0)	50.0 (0.0)	50.0 (0.0)	90.94 (28.71)	83.2 (37.38)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
range physchem range ecfp6	6.0 (0.0)	0.22 (0.01)	0.51 (0.0)	100.0 (0.0)	100.0 (0.0)	92.19 (26.84)	87.11 (33.51)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
range physchem range ecfp4 counts	6.9 (0.94)	0.27 (0.01)	0.54 (0.0)	100.0 (0.0)	100.0 (0.0)	90.08 (29.9)	88.2 (32.26)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)
range physchem maxsim ecfp4	1.0 (0.0)	0.01 (0.01)	0.28 (0.02)	0.0 (0.0)	0.0 (0.0)	97.11 (16.75)	17.5 (38.0)	100.0 (0.0)	100.0 (0.0)	44.4 (9.6)	40.1 (39.8)
range physchem maxsim ecfp6	1.0 (0.0)	0.0 (0.0)	0.24 (0.0)	0.0 (0.0)	0.0 (0.0)	98.91 (10.4)	6.95 (25.44)	100.0 (0.0)	100.0 (0.0)	45.9 (5.4)	68.9 (40.9)
range physchem maxsim ap	1.0 (0.0)	0.01 (0.01)	0.26 (0.01)	0.0 (0.0)	0.0 (0.0)	95.94 (19.74)	17.73 (38.2)	100.0 (0.0)	100.0 (0.0)	45.6 (6.8)	34.1 (31.0)

strict on the JAK2 data set can still lead to the generation of drug-like molecules on the Renin data set.

11 β HSD Data Set. As the 11 β HSD data set is made of two chemical series, three series of results are presented, two featuring an AD defined using representatives of one of the two series and one where the training set of the AD contains representatives of both series. On the urea chemical series (see Table 7 for full results), only “range physchem + range ECFP4” and “range physchem + range ECFP4 counts” ADs yield the highest evaluation metrics, with the latter giving better results on the SAS and QED metrics.

On the oxathiazine series (see Table 8), the two same ADs produce high evaluation metrics, as well as the “range ECFP4 counts” AD.

Surprisingly, on the full data set (see Table 9 for full results), even the best applicability domain (“range physchem + range

ECFP4 counts”) yields molecules with low SAS and QED metrics and low diversity. This could be explained by the fact that an AD defined with two distinct chemical series leads to a broader chemical space, imposing less constraints on the generative model and allowing the generation of non-drug-like structures.

ChEMBL 11 β HSD Data Set. The ChEMBL 11 β HSD data set allows us to evaluate AD definitions in a different context than lead-optimization. It could for instance mimic a hit-finding scenario, starting from a diverse data set to design novel actives. A first analysis of the results from Table 11 yields surprising results: the drug-likeness of generated molecules using the best AD definitions, according to the various proxies, is better than that for the 11 β HSD internal data set. The fact that the ChEMBL data set is more diverse, with various chemotypes represented, would suggest the opposite.

Table 7. Evaluation of the Molecule Sets Generated with the Different Applicability Domains on the 11 β HSD Urea Series Dataset^a

	Number of clusters	Entropies	Internal diversity	% actives recovered at d < 0.1	% actives recovered	% valid SAS	% valid QED	% valid cycle sizes	% valid MW	% valid het-het bonds	% closed shell
smiles validity	1.0 (0.0)	0.0 (0.0)	0.24	0.0 (0.0)	0.0 (0.0)	5.31	1.02	25.9	8.6	36.6	0.0
			(0.04)			(22.43)	(10.03)	(38.8)	(22.3)	(43.3)	(0.0)
maxsim ecfp4	1.0 (0.0)	0.0 (0.0)	0.19	0.0 (0.0)	0.0 (0.0)	3.44	0.0	20.0	6.7	11.9	20.0
			(0.04)			(18.22)	(0.0)	(40.0)	(20.2)	(23.3)	(40.0)
maxsim ecfp6	1.0 (0.0)	0.0 (0.0)	0.22	0.0 (0.0)	0.0 (0.0)	1.3	0.0	20.3	6.3	25.3	0.0
			(0.07)			(11.34)	(0.0)	(35.3)	(10.4)	(39.5)	(0.0)
maxsim ap	1.0 (0.0)	0.0 (0.0)	0.22	0.0 (0.0)	0.0 (0.0)	0.31	0.0	69.5	0.0	12.9	0.0
			(0.03)			(5.58)	(0.0)	(45.5)	(0.0)	(30.3)	(0.0)
range qed	1.0 (0.0)	0.0 (0.0)	0.18	0.0 (0.0)	0.0 (0.0)	1.56	0.0	0.7	5.7	65.7	0.0
			(0.06)			(12.4)	(0.0)	(1.7)	(15.8)	(44.3)	(0.0)
range ecfp4	1.0 (0.0)	0.0 (0.0)	0.13	0.4 (0.5)	0.4 (0.5)	0.47	13.36	1.1	12.7	99.0	80.0
			(0.05)			(6.83)	(34.02)	(2.3)	(19.4)	(3.0)	(40.0)
range ecfp4 counts	1.0 (0.0)	0.0 (0.0)	0.32	2.0 (0.8)	1.1 (0.4)	22.58	55.55	36.0	99.0	100.0	85.4
			(0.03)			(41.81)	(49.69)	(26.2)	(3.0)	(0.0)	(30.4)
range physchem	1.0 (0.0)	0.0 (0.0)	0.28	0.0 (0.0)	0.0 (0.0)	0.0	0.16	100.0	100.0	88.6	0.0
			(0.05)			(0.0)	(3.95)	(0.0)	(0.0)	(29.6)	(0.0)
range physchem range ap	1.0 (0.0)	0.0 (0.0)	0.33	0.0 (0.0)	0.0 (0.0)	0.08	26.95	100.0	100.0	100.0	0.1
			(0.02)			(2.79)	(44.37)	(0.0)	(0.0)	(0.0)	(0.2)
range physchem range ecfp4	1.0 (0.0)	0.0 (0.0)	0.29	1.1 (0.6)	0.9 (0.5)	29.22	46.95	100.0	100.0	100.0	77.2
			(0.04)			(45.48)	(49.91)	(0.0)	(0.0)	(0.0)	(39.4)
range physchem range ecfp6	1.0 (0.0)	0.0 (0.0)	0.3	0.0 (0.0)	0.0 (0.0)	4.84	18.83	100.0	100.0	98.2	21.6
			(0.02)			(21.47)	(39.09)	(0.0)	(0.0)	(3.2)	(37.3)
range physchem range ecfp4 counts	1.0 (0.0)	0.0 (0.0)	0.31	3.9 (3.5)	3.1 (3.4)	79.45	91.09	100.0	100.0	100.0	79.2
			(0.02)			(40.4)	(28.48)	(0.0)	(0.0)	(0.0)	(39.6)
range physchem maxsim ecfp4	1.0 (0.0)	0.0 (0.0)	0.28	0.0 (0.0)	0.0 (0.0)	0.08	4.61	100.0	100.0	73.0	0.0
			(0.02)			(2.79)	(20.97)	(0.0)	(0.0)	(33.2)	(0.0)
range physchem maxsim ecfp6	1.0 (0.0)	0.0 (0.0)	0.29	0.0 (0.0)	0.0 (0.0)	0.0	0.0	100.0	100.0	98.0	0.0
			(0.02)			(0.0)	(0.0)	(0.0)	(0.0)	(5.1)	(0.0)
range physchem maxsim ap	1.0 (0.0)	0.0 (0.0)	0.28	0.0 (0.0)	0.0 (0.0)	0.0	5.62	100.0	100.0	82.9	0.0
			(0.05)			(0.0)	(23.04)	(0.0)	(0.0)	(33.6)	(0.0)

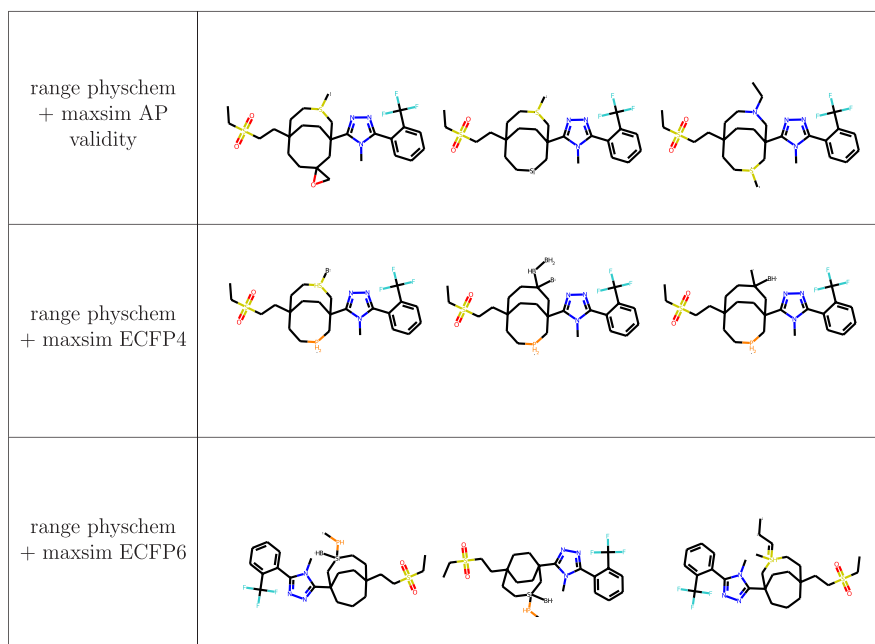
^aAll generated molecules belong to the same cluster, so entropies are 0.

Table 8. Evaluation of the Molecule Sets Generated with the Different Applicability Domains on the 11 β HSD Oxathiazine Series Dataset

	Number of clusters	Entropies	Internal diversity	% actives recovered at d < 0.1	% actives recovered	% valid SAS	% valid QED	% valid cycle sizes	% valid MW	% valid het-het bonds	% closed shell
smiles validity	1.0 (0.0)	0.01	0.28	0.0 (0.0)	0.0 (0.0)	0.09	0.0	77.8	0.0	0.0 (0.0)	0.0
			(0.02)			(0.04)	(2.94)	(0.0)	(41.6)	(0.0)	(0.0)
maxsim ecfp4	1.0 (0.0)	0.04	0.25	0.0 (0.0)	0.0 (0.0)	0.0	0.0	90.0	0.0	0.0 (0.0)	0.0
			(0.04)			(0.05)	(0.0)	(0.0)	(30.0)	(0.0)	(0.0)
maxsim ecfp6	1.0 (0.0)	0.05	0.26	0.0 (0.0)	0.0 (0.0)	0.08	0.0	100.0	0.0	0.0 (0.0)	0.0
			(0.05)			(0.04)	(2.79)	(0.0)	(0.0)	(0.0)	(0.0)
maxsim ap	1.0 (0.0)	0.01	0.24	0.0 (0.0)	0.0 (0.0)	0.0	0.0	90.0	0.0	0.0 (0.0)	0.0
			(0.04)			(0.05)	(0.0)	(0.0)	(30.0)	(0.0)	(0.0)
range qed	1.0 (0.0)	0.03	0.28	0.0 (0.0)	0.0 (0.0)	6.56	0.23	34.6	18.3	24.3	0.0
			(0.04)			(0.04)	(24.76)	(4.84)	(40.3)	(29.7)	(37.8)
range ecfp4	1.0 (0.0)	0.11	0.26	0.0 (0.0)	0.0 (0.0)	99.92	0.16	100.0	0.2	100.0	100.0
			(0.04)			(0.03)	(2.79)	(3.95)	(0.0)	(0.3)	(0.0)
range ecfp4 counts	1.9 (0.7)	0.01	0.43	0.2 (0.4)	0.0 (0.0)	99.61	26.56	100.0	82.0	100.0	98.5
			(0.01)			(0.02)	(6.24)	(44.17)	(0.0)	(9.2)	(0.0)
range physchem	1.0 (0.0)	0.02	0.26	0.0 (0.0)	0.0 (0.0)	28.2	0.39	100.0	100.0	76.1	0.0
			(0.06)			(0.05)	(45.0)	(6.24)	(0.0)	(0.0)	(36.1)
range physchem range ap	1.6 (0.8)	0.09	0.4	0.0 (0.0)	0.0 (0.0)	81.33	42.19	100.0	100.0	99.8	30.0
			(0.05)			(0.05)	(38.97)	(49.39)	(0.0)	(0.0)	(0.5)
range physchem range ecfp4	1.6 (0.49)	0.14	0.39	0.0 (0.0)	0.0 (0.0)	99.06	33.28	100.0	100.0	100.0	99.1
			(0.01)			(0.02)	(9.64)	(47.12)	(0.0)	(0.0)	(0.0)
range physchem range ecfp6	3.0 (1.0)	0.15	0.44	0.0 (0.0)	0.0 (0.0)	95.55	45.08	100.0	100.0	100.0	78.5
			(0.02)			(0.04)	(20.63)	(49.76)	(0.0)	(0.0)	(0.0)
range physchem range ecfp4 counts	3.7 (1.49)	0.16	0.46	4.7 (4.3)	4.0 (4.0)	98.75	55.7	100.0	100.0	100.0	100.0
			(0.02)			(0.02)	(11.11)	(49.67)	(0.0)	(0.0)	(0.0)
range physchem maxsim ecfp4	1.0 (0.0)	0.03	0.26	0.0 (0.0)	0.0 (0.0)	65.0	2.97	100.0	100.0	95.5	0.0
			(0.03)			(0.03)	(47.7)	(16.97)	(0.0)	(0.0)	(7.2)
range physchem maxsim ecfp6	1.0 (0.0)	0.08	0.28	0.0 (0.0)	0.0 (0.0)	77.03	22.27	100.0	100.0	81.8	20.0
			(0.06)			(0.03)	(42.06)	(41.6)	(0.0)	(0.0)	(29.1)
range physchem maxsim ap	1.0 (0.0)	0.07	0.3	0.0 (0.0)	0.0 (0.0)	53.2	0.23	100.0	100.0	96.6	0.0
			(0.09)			(0.05)	(49.9)	(4.84)	(0.0)	(0.0)	(8.5)

Table 9. Evaluation of the Molecule Sets Generated with the Different Applicability Domains on the Full 11 β hsd Dataset

	Number of clusters	Entropies	Internal diversity	% actives recovered at d < 0.1	% actives recovered	% valid SAS	% valid QED	% valid cycle sizes	% valid MW	% valid het-het bonds	% valid closed shell
smiles validity	1.0 (0.0)	0.0 (0.0)	0.19	0.0 (0.0)	0.0 (0.0)	(16.97)	(2.79)	10.6	1.3	13.6	30.0
			(0.04)					(29.8)			
maxsim ecfp4	1.0 (0.0)	0.0 (0.0)	0.25	0.0 (0.0)	0.0 (0.0)	(2.79)	(0.0)	40.0	2.1	19.8	10.0
			(0.05)					(49.0)			
maxsim ecfp6	1.0 (0.0)	0.0 (0.0)	0.22	0.0 (0.0)	0.0 (0.0)	(15.85)	(0.0)	32.0	8.6	45.3	19.1
			(0.06)					(33.7)			
maxsim ap	1.1 (0.3)	0.0 (0.0)	0.25	0.0 (0.0)	0.0 (0.0)	2.5	0.86	30.2	5.4	29.5	17.2
			(0.04)					(42.8)			
range qed	1.0 (0.0)	0.0 (0.0)	0.26	0.0 (0.0)	0.0 (0.0)	(7.88)	(6.24)	24.5	41.7	52.1	0.0
			(0.04)					(38.5)			
range ecfp4	1.2 (0.4)	0.0 (0.0)	0.21	0.0 (0.0)	0.0 (0.0)	(4.84)	(2.79)	22.8	14.1	72.0	59.8
			(0.04)					(39.0)			
range ecfp4 counts	1.1 (0.3)	0.0 (0.0)	0.28	1.6 (1.3)	1.6 (1.3)	(18.61)	(19.74)	40.1	46.7	100.0	68.6
			(0.03)					(40.9)			
range physchem	1.0 (0.0)	0.0 (0.0)	0.29	0.0 (0.0)	0.0 (0.0)	0.08	1.64	100.0	100.0	93.5	27.8
			(0.04)					(0.0)			
range physchem range ap	1.0 (0.0)	0.0 (0.0)	0.28	0.0 (0.0)	0.0 (0.0)	(10.4)	(28.48)	100.0	100.0	(0.2)	(28.8)
			(0.03)					(0.0)			
range physchem range ecfp4	1.1 (0.3)	0.0 (0.01)	0.27	0.0 (0.0)	0.0 (0.0)	(34.11)	(22.11)	100.0	100.0	100.0	61.4
			(0.03)					(0.0)			
range physchem range ecfp6	1.0 (0.0)	0.0 (0.0)	0.32	0.0 (0.0)	0.0 (0.0)	(29.47)	(5.58)	100.0	100.0	100.0	40.0
			(0.04)					(0.0)			
range physchem range ecfp4 counts	1.2 (0.4)	0.0 (0.0)	0.27	1.7 (1.3)	1.6 (1.3)	(43.96)	(46.51)	100.0	100.0	100.0	70.9
			(0.05)					(0.0)			
range physchem maxsim ecfp4	1.1 (0.3)	0.0 (0.0)	0.08	0.0 (0.0)	0.0 (0.0)	0.31	0.31	89.8	100.0	89.8	0.1
			(0.04)					(0.0)			
range physchem maxsim ecfp6	1.0 (0.0)	0.0 (0.0)	0.26	0.0 (0.0)	0.0 (0.0)	(2.79)	(5.58)	100.0	100.0	(18.7)	(0.2)
			(0.04)					(0.0)			
range physchem maxsim ap	1.0 (0.0)	0.0 (0.0)	0.3	0.0 (0.0)	0.0 (0.0)	(19.0)	(16.97)	100.0	100.0	100.0	9.9
			(0.04)					(0.0)			
range physchem maxsim ap	1.1 (0.3)	0.0 (0.0)	0.3	0.0 (0.0)	0.0 (0.0)	0.39	1.17	100.0	100.0	86.5	20.0
			(0.03)					(0.0)			
range physchem maxsim ap	1.1 (0.3)	0.0 (0.0)	0.03	0.0 (0.0)	0.0 (0.0)	(6.24)	(10.76)	100.0	100.0	(25.4)	(40.0)
			(0.03)					(0.0)			

Table 10. ChEMBL-Extracted 11 β HSD Dataset: Typical Problematic Structures Generated with the Three ADs That Are Able to Optimize the Scoring Function while Producing High Scores on Evaluation Metrics^a

^aAs reported in Table 11.

Nonetheless, if we look at the optimization scores displayed in Figure 9, we see that scores do not even cross the threshold for being predicted active for all AD definitions that feature the range of physicochemical descriptors and the range of fingerprints. Additionally, while AD definitions combining range of physicochemical descriptors with similarity using

fingerprints produce good results according to Table 11, visual inspection shows molecules that are not drug-like (Table 10). AD definitions are therefore either too stringent (leading to a failure of optimization) or not stringent enough, leading to molecules that are obviously non-drug-like, such as the ones displayed in Table 10.

Table 11. Evaluation of the Molecule Sets Generated with the Different Applicability Domains on the ChEMBL 11/ β HSD Dataset

	Number of clusters	Entropies	Internal diversity	% actives recovered at d < 0.1	% actives recovered	% valid SAS	% valid QED	% valid cycle sizes	% valid MW	% valid het-het bonds	% closed shell
smiles validity	1.4 (0.49)	0.0 (0.0)	0.24	40.0	40.0	0.78	0.31	18.6	0.3	60.3	59.1
			(0.05)	(49.0)	(49.0)	(8.8)	(5.58)	(36.8)	(0.4)	(48.6)	(41.3)
maxsim ecfp4	1.8 (0.4)	0.01 (0.0)	0.2	80.0	80.0	1.88	0.62	13.6	0.6	20.8	65.9
			(0.04)	(40.0)	(40.0)	(13.56)	(7.88)	(26.0)	(0.3)	(39.6)	(42.7)
maxsim ecfp6	1.8 (0.4)	0.01 (0.0)	0.19	80.0	80.0	0.62	0.62	30.2	0.6	20.6	45.9
			(0.05)	(40.0)	(40.0)	(7.88)	(7.88)	(37.6)	(0.3)	(39.7)	(44.7)
maxsim ap	1.6 (0.49)	0.01 (0.0)	0.22	60.0	60.0	3.12	0.47	40.2	0.5	44.8	45.0
			(0.03)	(49.0)	(49.0)	(17.4)	(6.83)	(48.9)	(0.4)	(45.7)	(44.6)
range qed	1.8 (0.4)	0.01 (0.01)	0.26	80.0	80.0	0.78	0.78	20.0	0.8		80.5
			(0.04)	(40.0)	(40.0)	(8.8)	(8.8)	(38.4)	(0.5)	0.8 (0.5)	(39.1)
range ecfp4	2.2 (0.4)	0.02 (0.0)	0.24	100.0	100.0	86.72	4.38	85.2	5.3	83.8	100.0
			(0.04)	(0.0)	(0.0)	(33.94)	(20.45)	(26.7)	(0.6)	(21.4)	(0.0)
range ecfp4 counts	10.0 (1.26)	0.19 (0.03)	0.72	100.0	100.0	100.0	82.19	91.4	92.8	100.0	100.0
			(0.02)	(0.0)	(0.0)	(0.0)	(38.26)	(2.6)	(2.4)	(0.0)	(0.0)
range physchem	2.0 (0.63)	0.01 (0.0)	0.33	100.0	100.0	33.12	25.62	100.0	31.6	92.5	19.5
			(0.03)	(0.0)	(0.0)	(47.07)	(43.66)	(0.0)	(4.9)	(10.2)	(26.4)
range physchem range ap	2.2 (1.17)	0.06 (0.07)	0.42	100.0	100.0	58.59	70.94	100.0	96.7	100.0	54.4
			(0.03)	(0.0)	(0.0)	(49.26)	(45.41)	(0.0)	(5.5)	(0.0)	(37.9)
range physchem range ecfp4	3.8 (1.6)	0.08 (0.04)	0.54	100.0	100.0	100.0	67.03	100.0	94.8	100.0	100.0
			(0.08)	(0.0)	(0.0)	(0.0)	(47.01)	(0.0)	(4.1)	(0.0)	(0.0)
range physchem range ecfp6	8.2 (2.14)	0.08 (0.01)	0.63	100.0	100.0	80.31	73.75	81.2	78.9	81.2	87.3
			(0.08)	(0.0)	(0.0)	(39.76)	(44.0)	(23.7)	(23.3)	(23.7)	(15.9)
range physchem range ecfp4 counts	10.8 (0.75)	0.17 (0.02)	0.71	100.0	100.0	99.69	90.31	100.0	96.4	100.0	100.0
			(0.02)	(0.0)	(0.0)	(5.58)	(29.58)	(0.0)	(1.6)	(0.0)	(0.0)
range physchem maxsim ecfp4	1.6 (0.49)	0.01 (0.0)	0.3	100.0	100.0	36.25	8.75	100.0	22.8	85.6	23.1
			(0.02)	(0.0)	(0.0)	(48.07)	(28.26)	(0.0)	(8.9)	(12.2)	(38.5)
range physchem maxsim ecfp6	1.8 (0.4)	0.01 (0.0)	0.28	100.0	100.0	53.44	43.12	100.0	47.5	96.4	7.2
			(0.04)	(0.0)	(0.0)	(49.88)	(49.53)	(0.0)	(12.9)	(3.7)	(4.3)
range physchem maxsim ap	1.6 (0.49)	0.01 (0.0)	0.33	100.0	100.0	37.5	25.94	100.0	29.4	90.6	6.6
			(0.02)	(0.0)	(0.0)	(48.41)	(43.83)	(0.0)	(12.0)	(12.8)	(2.3)

Molecular Turing Test. We validate our results by running a molecular Turing test. We restricted ourselves to the JAK2 data set, and to four different AD definitions: “range QED”, “maxsim ECFP4”, “range physchem + maxsim ECFP4”, and “range physchem + range ECFP4 counts”. We then asked Sanofi researchers working in early drug discovery (e.g., drug discovery project team leaders, medicinal chemists, and computational chemists) to discriminate, among molecules generated with those four AD as well as for molecules from the test set, those that they would not consider for synthesis in a drug discovery program. Twenty molecules were selected at random in each set. The 100 structures were shuffled and shown to the anonymous participants. The rejection rates are shown as a boxplot for each set of molecules in Figure 10. The molecular Turing test confirms that ADs traditionally used in QSAR modeling (“maxsim ECFP4”) or based on existing drug-likeness scores (“range QED”) fail at generating drug-like molecules. Indeed, their average rejection rate is 100%. Results for the “range physchem + maxsim ECFP4” AD are also poor, with a rejection rate on average of 90%, showing that ADs with intermediate performance according to our proxy metrics (Table 3) also fail to generate drug-like molecules. Finally, the AD definition that we identified as the best (“range physchem + range ECFP4 counts”) shows rejection rates similar to those of molecules from the data set. Interestingly, rejection rate of molecules from the test set (that were thus selected, synthesized and tested in a drug discovery project) is around 15%, confirming that a large part of subjectivity remains when choosing which molecules to select for synthesis.

DISCUSSION

Our results highlight the influence of the applicability domain definition on the drug-likeness of the molecules generated de novo. Indeed, some of the classic applicability domain definitions used in QSAR modeling, such as ensuring similarity to the training set (corresponding to the applicability domain definitions “maxsim ECFP4”, “maxsim ECFP6”, and “maxsim ap”), fail to generate an high enough proportion of drug-like molecules. Other approaches such as filtering unwanted substructures or using drug-likeness scores (“filters validity” and “range qed” ADs) are not very efficient either. More stringent definitions based on bounding-box approaches (“range ECFP4”, “range ECFP4 counts”, “range physchem”) can sometimes lead to drug-like molecule sets, but they do not perform well across all tasks. The best results are obtained through a combination of physicochemical descriptors and fingerprints (“range physchem + range ECFP4”, “range physchem + range ECFP6”, “range physchem + range ECFP4 counts”, “range physchem + maxsim ECFP4”, “range physchem + maxsim ECFP6”, and “range physchem + maxsim ap”), while the only applicability domain that performs well throughout all tasks is the “range physchem + range ECFP4 counts” definition.

Figure 11 represents a tree map⁵⁶ of the results obtained on the Renin data set with a good AD definition (i.e., “range physchem + range ECFP4 counts”) and a bad AD definition (i.e., “Maxsim ECFP4”) together with the original Renin data set. In terms of chemical diversity, good applicability domain definitions tend to generate sets of molecules that are qualitatively similar to the original data set used, while applicability domains that are not stringent enough generate sets of molecules that overfit to specific regions of chemical

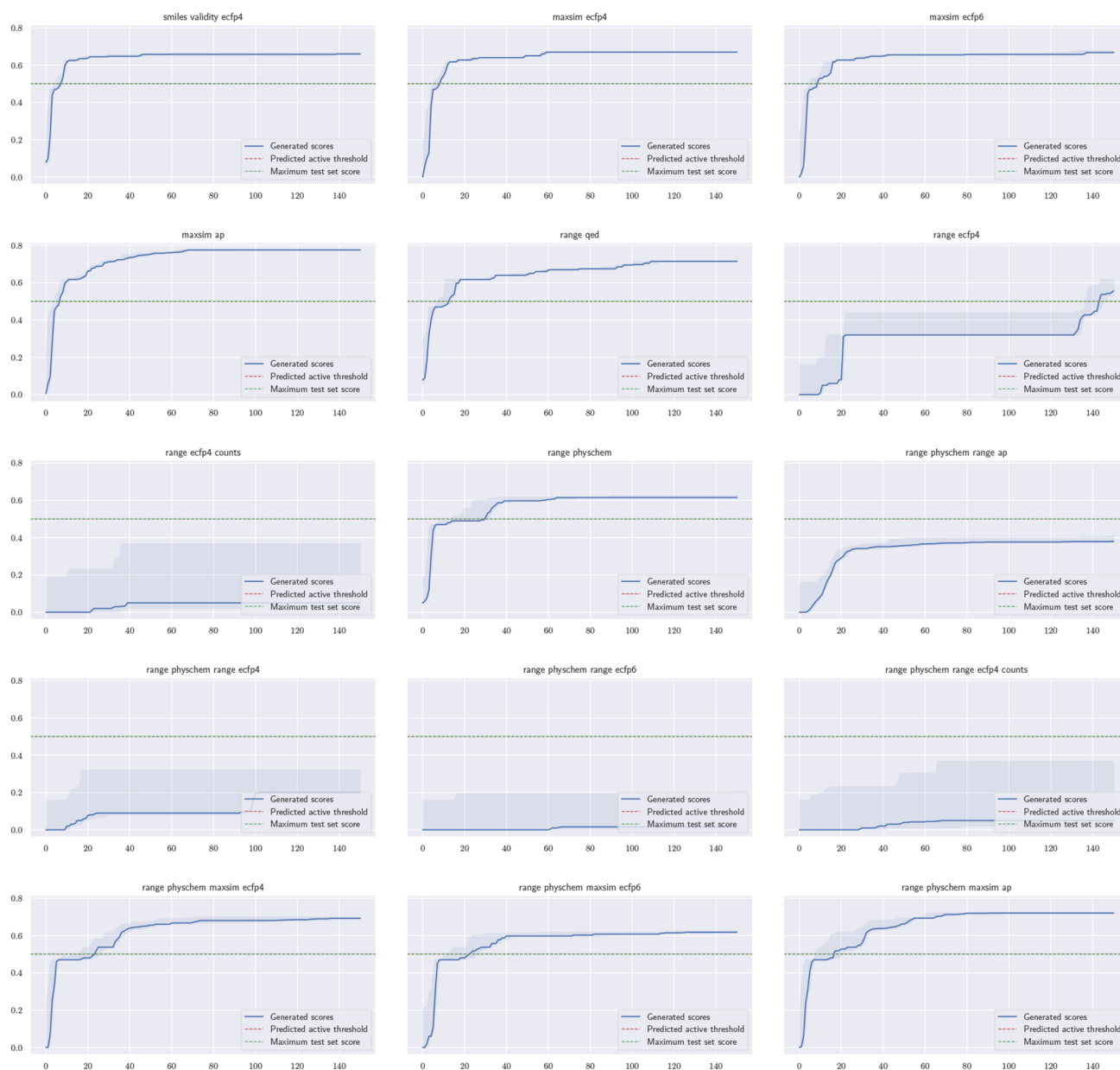


Figure 9. Comparison of scores reached by the LSTM-HC under the constraint of different AD definitions on the ChEMBL 11 β HSD data set.

space. Interestingly, good applicability domain definitions lead to higher diversity in addition to better drug-likeness for generated molecules.

Figure 12, which shows the fraction of actives and inactives close to generated compounds for different similarity thresholds, also highlights the better coverage of the training set's chemical space by good AD definitions. Good applicability domains explore the vicinity of the original training set, with close neighbors of generated molecules showing an enrichment toward the actives of the data set.

Counterintuitively, good applicability domains are associated with lower scores, as shown in Figure 13. In this figure, the evolution of scores throughout training of the generative model is shown for both a good AD definition and a bad AD definition. These results suggest that an applicability domain definition that is not sufficient to constrain generative algorithms leads to the overexploitation of narrow regions of the chemical space, often

with the addition of non-drug-like patterns to a high-scoring molecular structure. This reward hacking behavior⁵⁷ is prevented by good AD definitions: the constraints associated with good applicability domains prevent generative algorithms for overexploiting a specific region of chemical space, leading to a wider distribution of scores.

Another interesting takeaway from our results is that different well performing applicability domain definitions can lead to the exploration of different regions of the chemical space. Figure 14 compares the average Tanimoto similarity between sets generated with good AD definitions and shows that different well performing applicability domains lead to the exploration of different regions of chemical space.

Table 12 displays two patented molecules on the oxathiazine series of the 11 β HSD data set that were not seen during model building or generation, as well as highly similar generated molecules. This showcases the fact that good applicability

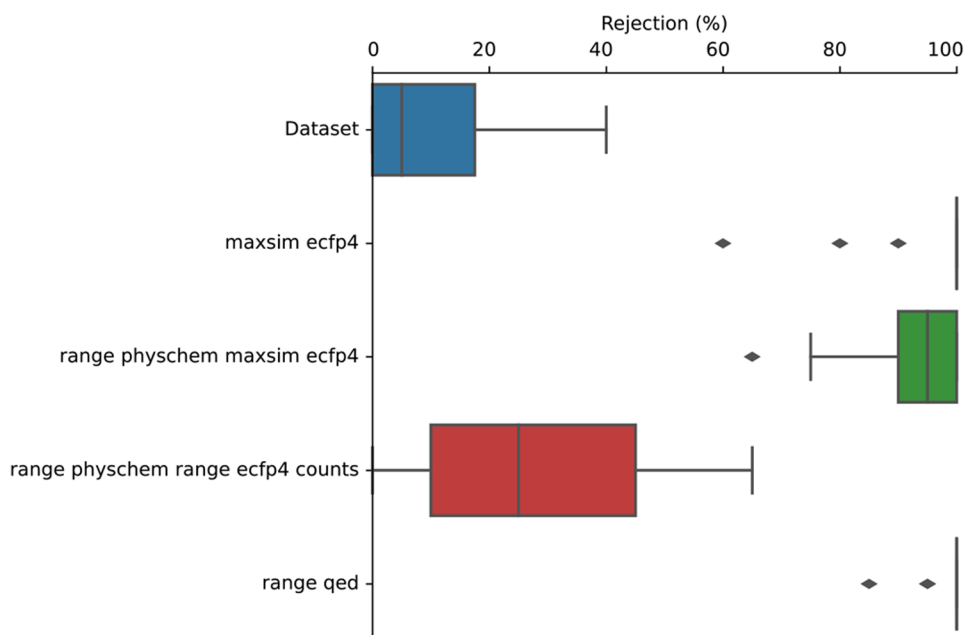


Figure 10. Results of the molecular Turing test for each of four different AD definitions (“range QED”, “maxsim ECFP4”, “range physchem + maxim ECFP4”, and “range physchem + range ECFP4 counts”) and for the JAK2 training set. The black bar denotes the mean, and the box denotes an interval with 90% of the values. Results were obtained with 15 different participants.

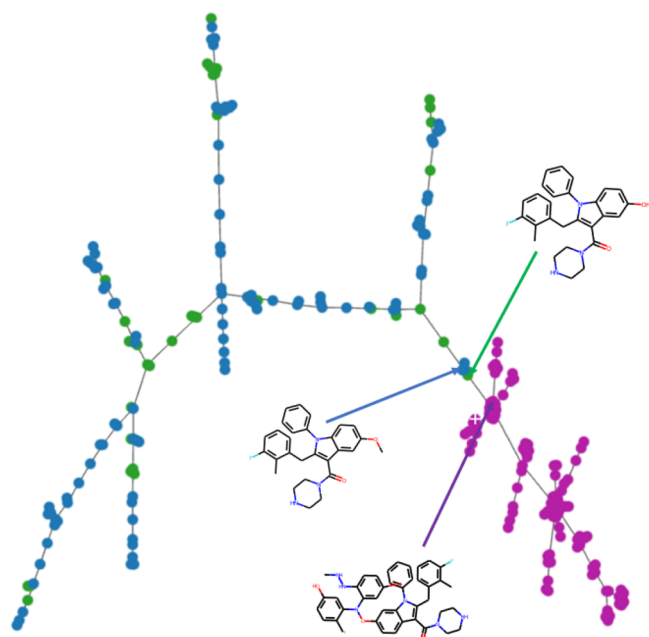


Figure 11. Tree map plot of the Renin test set (in green), molecules generated with a good applicability domain (“range physchem + range ECFP4 counts”, in blue), and molecules generated by an applicability domain showing poor results (“maxsim ECFP4”, in purple). The molecules from the good AD and the test data set (blue and green, respectively) mainly fall on the same trees and share connections, suggesting that the two sets could be similar and the generated molecules relevant. In contrast, the molecules from the bad AD are all located on a separate tree, suggesting that they are dissimilar from the test set and probably irrelevant. The molecule generated with the bad AD that is closest to the test data set (purple arrow) is still dissimilar to the closest molecules from the test data set (green arrow) or from those generated with a good AD (blue arrow). The tree map is generated using the TMAP library⁵⁶ with default settings.

domain definitions are also capable of retrieving unseen actives. This highlights the ability of generative algorithms to produce valuable molecules in an industrial drug discovery context.

Finally, our results also highlight the importance of choosing a carefully curated training set. Indeed, the set that we used for training the generator and the QSAR models was also used to define the applicability domain for generation. This implies that the training set needs to have only molecules that a practitioner considers as drug-like. If this is not the case (for instance, if the training set includes results from a high-throughput screening campaign, all of which would not be deemed acceptable starting points for a lead-optimization program), the applicability domains defined with respect to the training set might be insufficient to generate only drug-like molecules. Such an effect can be seen in the full 11 β HSD data set (see Table 7): as the training set is composed of two distinct chemical series, the applicability domains defined are less restrictive and lead to an overall lower drug-likeness of generated compounds. Those findings underline the importance of curating an adequate training set for generative models.

Finally, another approach for generating drug-like compounds would be to include the output of more sophisticated retrosynthesis tools^{59,60} within our scoring function. A limitation of this approach is the computing time associated with the evaluation of a single molecule, although this could be addressed through enhanced sample efficiency.⁶¹ Retrosynthesizability could also be achieved using generative models of forward synthesis pathways.^{62,63} While the integration of synthesizability is a very promising avenue of research,⁶² it is beyond the scope of this work. We acknowledge that several choices we made in this work are ad hoc and would deserve to be explored further, such as the other AD definitions, the choice of other metrics or similarity thresholds. Nevertheless, when we have employed the best-performing AD in the four data sets and in lead optimization projects, the qualitative improvement in the generated structures was such that it discouraged investigating systematically other possible options.



Figure 12. Fraction of actives and inactives among the Renin training set molecules close to the generated molecules at different similarity thresholds. Results are shown for the “range physchem + range ECFP4 counts” and “maxsim ECFP4” applicability domains. They show that good applicability domains generate molecules closer to the actual training set, with a clear enrichment toward active molecules.

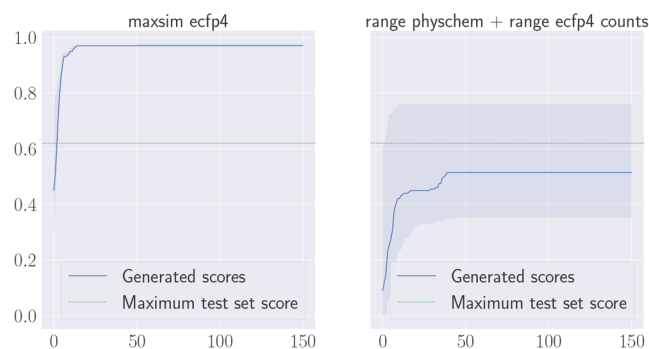


Figure 13. Renin data set: evolution of generated molecules' scores throughout the optimization epochs for a good applicability domain (“range physchem + range ECFP4 counts”) and for an applicability domain showing poor results (“maxsim ECFP4”). The scores of the molecules generated with the good applicability domain are more spread out and lower than those generated with the other applicability domain (while most are still in the correct range, between the predicted active threshold and the maximum score among the test set molecules). This illustrates that poorly performing ADs leave room for reward hacking by the generator.⁵⁷

CONCLUSION

In this work, we highlight methods that improve the drug-likeness of molecules designed by generative models. For this, we defined several different applicability domains and evaluated them on different data sets. An analysis allows us to identify valuable AD definitions for molecular generative algorithms. Our analysis shows that classic applicability domains metrics used in QSAR modeling (e.g., Tanimoto similarity on ECFP4 fingerprints) are not sufficient to discriminate molecules generated using generative artificial intelligence algorithms. Furthermore, measures of drug-likeness commonly used, such as QED, can be optimized in unintended ways and do not constitute a valid applicability domain for generative models either. Even applicability domains based on the combination of physicochemical descriptors, which intuitively could be most adapted to distinguish between drug-like and non-drug-like molecules, fail at this task. This highlights the need to distinguish

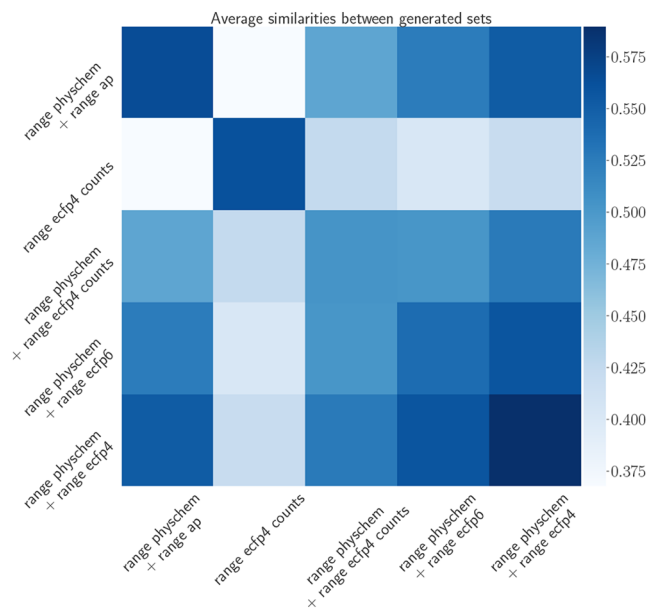
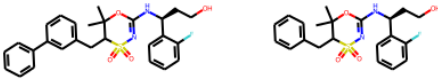
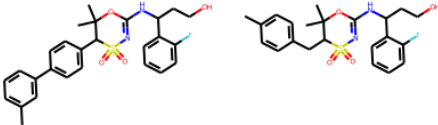


Figure 14. Average Tanimoto similarities (computed on ECFP4 fingerprints) for generated sets of molecules using a good applicability domain definition on the JAK2 data set. Different applicability domains can lead to the exploration of different portions of chemical space.

between applicability domains used in QSAR modeling and applicability domains used for the generative algorithm, as applicability domains performing well in the case of QSAR modeling do not necessarily perform well for generative models. Nonetheless, by combining physicochemical descriptors and fingerprint-based similarity in the “range physchem + range ECFP4 counts” AD, it is possible to obtain a definition that performs well for the different data sets we have considered. The molecules generated using this applicability domain definition are generally drug-like, and moreover show a higher diversity and coverage of chemical space than using other applicability domain definitions. It is an important feature as novelty is often a challenge in lead-optimization. We leave for future work the exploration of how other parameters of the AD (such as using

Table 12. Patented 11 β HSD Molecules⁵⁸ (Top) for Which We Found Generated Molecules (Bottom) With a Similarity Higher than 0.8^a

Patented molecules	
Generated molecules	

^aThe applicability domain that generated the molecules is the “range physchem + range of ECFP4 counts”. This showcases the ability of generative algorithms to propose structures of interest for a drug discovery project.

percentiles instead of span to define the AD) might impact the generation results. As our results were obtained in a lead-optimization context, where the chemical space explored is focused on a narrow set of chemical series, the applicability domains that we explore and identify in this work as yielding the largest proportion of drug-like molecules could behave differently in a different context. For instance, when using generative algorithms in distribution learning tasks,⁴ where the goal is to generate libraries of compounds for downstream task, other applicability domains definitions could be more suitable. For example, while novelty (in the sense of generating molecules with a similarity to the training set below a given threshold) is not a critical feature in the context of lead optimization, it could be desirable in distribution learning tasks. Indeed, lead optimization searches for novel molecules within a very limited chemical space, while distribution learning aims at exploring a diverse chemical space. Nevertheless, our results show that using an adequate applicability domain definition for generative models can greatly improve the drug-likeness of the structures generated. As this is a key aspect for the adoption of generative models in a drug discovery setting, we hope that our results will benefit practical applications of generative models for drug design.

■ ASSOCIATED CONTENT

Data Availability Statement

The full open source code needed to reproduce the results from the manuscript and Supporting Information is available at: <https://github.com/Sanofi-Public/IDD-papers-generative-applicability-domains>. The data sets (.csv format) and results of experiments are also available except for the internal 11 β HSD data sets, which are proprietary and not published yet. The code to reproduce plots and results regarding these data sets is nonetheless made available in the github repository.

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsomega.3c00883>.

PCA projection of generated molecules, QED and SAS distribution of generated molecules, enrichment in actives/inactives close to generated molecules, score distributions, average Tanimoto similarities between generated sets, and examples of typical problematic structures in generated molecules for Renin and JAX2 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Marc Bianciotto – Molecular Design Sciences–Integrated Drug Discovery, R&D, Sanofi, 94400 Vitry-sur-Seine, France; orcid.org/0000-0002-4345-995X; Email: marc.bianciotto@sanofi.com

Authors

Maxime Langevin – PASTEUR, Département de Chimie, École Normale Supérieure, PSL University, Sorbonne Université, CNRS, 75005 Paris, France; Molecular Design Sciences–Integrated Drug Discovery, R&D, Sanofi, 94400 Vitry-sur-Seine, France; Present Address: Owkin, 14-16 Bd Poissonnière, 75009 Paris, France; orcid.org/0000-0002-5498-4661

Christoph Grebner – Molecular Design Sciences–Integrated Drug Discovery, R&D, Sanofi, 65929 Frankfurt-am-Main, Germany

Stefan Güssregen – Molecular Design Sciences–Integrated Drug Discovery, R&D, Sanofi, 65929 Frankfurt-am-Main, Germany; orcid.org/0000-0002-1868-9614

Susanne Sauer – Molecular Design Sciences–Integrated Drug Discovery, R&D, Sanofi, 65929 Frankfurt-am-Main, Germany

Yi Li – Molecular Design Sciences–Integrated Drug Discovery, R&D, Sanofi, Waltham, Massachusetts 02451, United States
Hans Matter – Molecular Design Sciences–Integrated Drug Discovery, R&D, Sanofi, 65929 Frankfurt-am-Main, Germany; orcid.org/0000-0002-0249-6025

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acsomega.3c00883>

Notes

The authors declare the following competing financial interest(s): All authors are or have been employed by Sanofi and may hold shares and/or stock options in the company.

ACKNOWLEDGMENTS

The French National Association of Research and Technology (ANRT) is gratefully acknowledged for supporting M.L. (contract 2019/0821).

REFERENCES

- (1) Schneider, P.; et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discovery* **2020**, *19*, 353–364.
- (2) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, *9*, 48.
- (3) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Central Science* **2018**, *4*, 120–131.
- (4) Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096–1108.
- (5) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.
- (6) Renz, P.; Van Rompaey, D.; Wegner, J. K.; Hochreiter, S.; Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discov. Today: Technol.* **2019**, *32–33*, 55–63.
- (7) Mercado, R.; Rastemo, T.; Lindelöf, E.; Klambauer, G.; Engkvist, O.; Chen, H.; Bjerrum, E. J. Graph networks for molecular design. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 025023.
- (8) Gao, W.; Coley, C. W. *Synthesizability of Molecules Proposed by Generative Models* **2020**, *60*, 5714–5723.
- (9) Polykovskiy, D.; et al. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology* **2020**, *11*, 565644.
- (10) Steinmann, C.; Jensen, J. H. Using a genetic algorithm to find molecules with good docking scores. *PeerJ. Physical Chemistry* **2021**, *3*, e18.
- (11) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90–98.
- (12) Walters, P. *silly_walks*. https://github.com/PatWalters/silly_walks/, 2021.
- (13) Ajay; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between “drug-like” and “nondrug-like” molecules? *J. Med. Chem.* **1998**, *41*, 3314–3324.
- (14) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361–1375.
- (15) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504.
- (16) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.
- (17) Gadaleta, D.; Mangiatordi, G. F.; Catto, M.; Carotti, A.; Nicolotti, O. Applicability Domain for QSAR Models: Where Theory Meets Reality. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)* **2016**, *1*, 45–63.
- (18) Netzeva, T. I.; et al. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. *Alternatives to Laboratory Animals* **2005**, *33*, 155–173.
- (19) Schoeters, G. The Reach Perspective: Toward a New Concept of Toxicity Testing. *Journal of Toxicology and Environmental Health, Part B* **2010**, *13*, 232–241.
- (20) Ursu, O.; Rayan, A.; Goldblum, A.; Oprea, T. I. Understanding drug-likeness. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 760–781.
- (21) Walters, W.; Murcko, M. A. Prediction of ‘drug-likeness’. *Adv. Drug Delivery Rev.* **2002**, *54*, 255–271.
- (22) Shultz, M. D. Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs. *J. Med. Chem.* **2019**, *62*, 1701–1714.
- (23) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
- (24) Hartung, I. V.; Huck, B. R.; Crespo, A. Rules were made to be broken. *Nature Reviews Chemistry* **2023**, *7*, 3–4.
- (25) Walters, W. P.; Murcko, M. A.; Murcko, M. A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **1999**, *3*, 384–387.
- (26) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists’ intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269–1275.
- (27) Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; Klambauer, G. Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J. Chem. Inf. Model.* **2018**, *58*, 1736–1741.
- (28) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8.
- (29) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58*, 252–261.
- (30) Campbell, W. C. History of avermectin and ivermectin, with notes on the history of other macrocyclic lactone antiparasitic agents. *Curr. Pharm. Biotechnol.* **2012**, *13*, 853–865.
- (31) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* **2004**, *47*, 4891–4896.
- (32) Kutchukian, P. S.; Vasilyeva, N. Y.; Xu, J.; Lindvall, M. K.; Dillon, M. P.; Glick, M.; Coley, J. D.; Brooijmans, N. Inside the mind of a medicinal chemist: the role of human bias in compound prioritization during drug discovery. *PLoS one* **2012**, *7*, e48476.
- (33) Bush, J. T.; et al. A Turing Test for Molecular Generators. *J. Med. Chem.* **2020**, *63*, 11964–11971.
- (34) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (35) Landrum, G. *RDKit: Open-Source Cheminformatics*. <http://www.rdkit.org>.
- (36) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, *5*, 107–113.
- (37) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
- (38) Ritchie, T. J.; Macdonald, S. J. How drug-like are ‘ugly’ drugs: do drug-likeness metrics predict ADME behaviour in humans? *Drug Discovery Today* **2014**, *19*, 489–495.

- (39) Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.* **1996**, *22*, 469–483.
- (40) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set in descriptor space: a review. *Altern. Lab. Anim.* **2005**, *33*, 445–459.
- (41) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **2015**, *7*, 20.
- (42) 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 Res.* **2012**, *40*, D1100–D1107.
- (43) Walters, P. *rd_filters*. https://github.com/PatWalters/rd_filters.
- (44) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Central Science* **2018**, *4*, 120–131.
- (45) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **2017**, *9*, 48.
- (46) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.
- (47) Ding, Y.; Wang, L.; Zhang, H.; Yi, J.; Fan, D.; Gong, B. Defending Against Adversarial Attacks Using Random Forest. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach CA, June 16–20, 2019; IEEE, 2019; pp 105–114.
- (48) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (49) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.
- (50) Scheiper, B.; Matter, H.; Steinhagen, H.; Stilz, U.; Böcskei, Z.; Fleury, V.; McCort, G. Discovery and optimization of a new class of potent and non-chiral indole-3-carboxamide-based renin inhibitors. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 6268–6272.
- (51) Matter, H.; Scheiper, B.; Steinhagen, H.; Böcskei, Z.; Fleury, V.; McCort, G. Structure-based design and optimization of potent renin inhibitors on 5- or 7-azaindole-scaffolds. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 5487–5492.
- (52) Scheiper, B.; Matter, H.; Steinhagen, H.; Böcskei, Z.; Fleury, V.; McCort, G. Structure-based optimization of potent 4- and 6-azaindole-3-carboxamides as renin inhibitors. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 5480–5486.
- (53) Blaschke, T.; Engkvist, O.; Bajorath, J.; Chen, H. Memory-assisted reinforcement learning for diverse molecular de novo design. *J. Cheminform.* **2020**, *12*, 68.
- (54) Duda, R. O.; Hart, P. E.; Stork, D. G.; Duda, C. R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification, 2nd Ed*; Wiley, 2001.
- (55) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.
- (56) Probst, D.; Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **2020**, *12*, 12.
- (57) Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete Problems in AI Safety. *arXiv (Computer Science.Artificial Intelligence)*, June 21, 2016, 1606.06565. DOI: 10.48550/arXiv.1606.06565v2.
- (58) Boehme, T.; Engel, C.; Guessregen, S.; Haack, T.; Ritter, K.; Tschank, G. Di and tri- substituted oxathiazine derivatives, method for the production, method for the production thereof, use thereof as medicine and drug containing said derivatives and use thereof. US 8710050 B2.
- (59) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminform.* **2020**, *12*, 70.
- (60) Fortunato, M. E.; Coley, C. W.; Barnes, B. C.; Jensen, K. F. Data Augmentation and Pretraining for Template-Based Retrosynthetic Prediction in Computer-Aided Synthesis Planning. *J. Chem. Inf. Model.* **2020**, *60*, 3398–3407. PMID: 32568548.
- (61) Gao, W.; Fu, T.; Sun, J.; Coley, C. W. Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization. *arXiv (Computer Science.Computational Engineering, Finance, and Science)*, June 22, 2022, 2206.12411. DOI: 10.48550/arXiv.2206.12411v1.
- (62) Gao, W.; Mercado, R.; Coley, C. W. Amortized tree generation for bottom-up synthesis planning and synthesizable molecular design. *arXiv (Computer Science.Machine Learning)*, October 12, 2021, 2110.06389. DOI: 10.48550/arXiv.2110.06389v2.
- (63) Bradshaw, J.; Paige, B.; Kusner, M. J.; Segler, M. H. S.; Hernández-Lobato, J. M. Barking up the right tree: an approach to search over molecule synthesis DAGs. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, December 6–12, 2020; Curran Associates, Inc.: Red Hook, NY, 2020; Vol. 33, pp 6852–6866.