# Beyond Generative Models: Superfast Traversal, Optimization, Novelty, Exploration and Discovery (STONED) Algorithm for Molecules using SELFIES

AkshatKumar Nigam,[1,2] Robert Pollice,[1,2] Mario Krenn,[1,2,3]
Gabriel dos Passos Gomes,[1,2] and Alán Aspuru-Guzik[1,2,3,4,*]

[1] *Department of Computer Science, University of Toronto, Canada.*
[2] *Department of Chemistry, University of Toronto, Canada.*
[3] *Vector Institute for Artificial Intelligence, Toronto, Canada.*
[4] *Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), 661 University Ave, Toronto, Ontario M5G, Canada.*

**Abstract** Inverse design allows the design of molecules with desirable properties using property optimization. Deep generative models have recently been applied to tackle inverse design, as they possess the ability to optimize molecular properties directly through structure modification using gradients. While the ability to carry out direct property optimizations is promising, the use of generative deep learning models to solve practical problems requires large amounts of data and is very time-consuming. In this work, we propose STONED – a simple and efficient algorithm to perform interpolation and exploration in the chemical space, comparable to deep generative models. STONED bypasses the need for large amounts of data and training times by using string modifications in the SELFIES molecular representation. We achieve comparable performance on typical benchmarks without any training. We demonstrate applications in high-throughput virtual screening for the design of drugs, photovoltaics, and the construction of chemical paths, allowing for both property and structure-based interpolation in the chemical space. We anticipate our results to be a stepping stone for developing more sophisticated inverse design models and benchmarking tools, ultimately helping generative models achieve wide adoption.

## I. INTRODUCTION

Generative models are a class of techniques which have emerged with applications in inverse molecular design [1]. Among them, variational autoencoders (VAEs) [2, 3], generative adversarial networks (GANs) [4, 5], recurrent neural networks (RNNs) [6, 7], deep reinforcement learning (DRL) [8, 9] and genetic algorithms (GAs) [10–12] have been applied to the design of molecules. Importantly, the choice of molecular representation employed in these approaches impacts performance dramatically. Deep generative models trained on molecular representations form low dimensional latent spaces enabling the sampling of unseen molecules. This allows for exploration in the chemical space and interpolation by chemical path formation [3]. In contrast to genetic algorithms with the SMILES string representation[13, 14], a unique aspect of these deep learning techniques is that the generation of new molecules does not require the design of hand-crafted rules. However, they can require access to large datasets and expensive computational resources to offset large training times. Furthermore, with fragile representations such as SMILES, large areas of a latent space can correspond to invalid molecules [3]. Alternatively, deep generative models using molecular graphs as adjacency matrices have also been demonstrated with applications in drug design [15, 16]. Recently, the development and application of a 100% valid strings representation – SELFIES [17] has been demonstrated for inverse design [18]. Compared to SMILES and adjacency matrices, the use of SELFIES in generative models overcomes the problem of generating invalid molecules.

In this work, using SELFIES as a robust molecular representation, we propose an efficient algorithm (STONED) to perform exploration and interpolation in the chemical space (Section II A). These tasks are commonly addressable by expensive deep generative models. Our algorithm avoids the need for extensive training times, large datasets, and hand-crafted rules for obtaining novel molecules. We achieve this using string manipulations of SELFIES and demonstrate the ability to form local chemical subspaces (Section II B), allowing for local optimization, and obtain chemical paths (Sections II C, II D), enabling interpolation between structures. Additionally, we demonstrate applications in designing molecules for material science (Section II E) and drug development (Section II D 2). On established benchmarks, our algorithm achieves results comparable to the state of the art in generative modeling. The ease of obtaining molecules for local optimization and interpolation via chemical paths allows for our methods to be used in high-throughput virtual screening in materials science [19], catalysis [20], and drug design [21]. Moreover, the simplicity of our technique highlights deficiencies in current molecular design benchmarks for deep generative models. We anticipate that our results will stimulate more powerful models, more meaningful benchmarks, and more widespread use of generative models in general.

* Correspondence to: alan@aspuru.com
Full code is available at:
`https://github.com/aspuru-guzik-group/stoned-selfies`

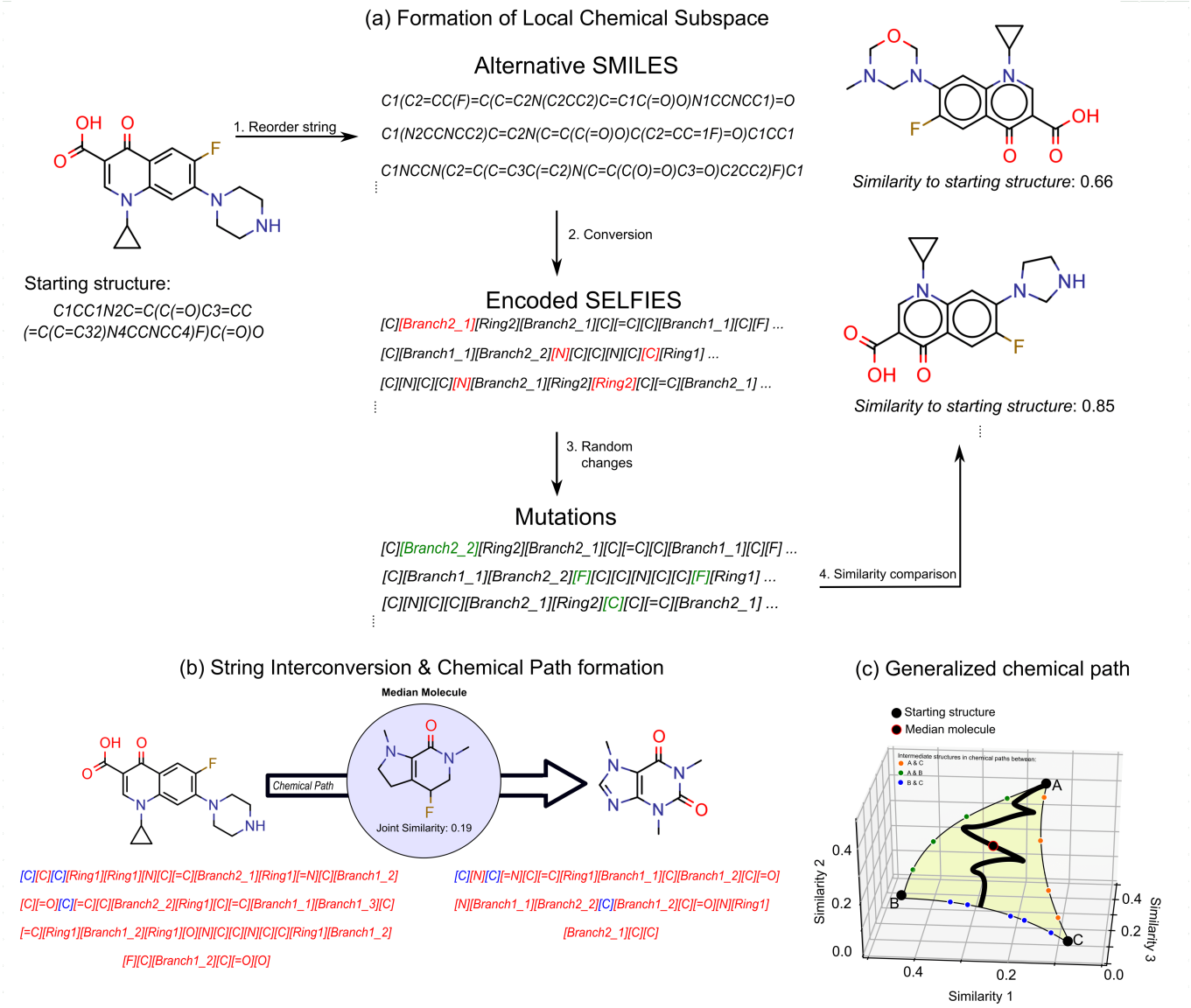


FIG. 1. Illustration of string manipulations within STONED to form local chemical subspaces (a - Section II B) for a given structure, discovering median molecules on the chemical path between two structures (b - Section II D) and formation of generalized chemical paths between more than two molecules (c - Section II E)

## II. RESULTS AND DISCUSSION

### A. Algorithmic Overview

In this work, we show that random changes within the SELFIES molecular representation are a powerful tool for performing structural and property-based changes to molecules. Akin to deep generative models, these changes can be utilized for forming local chemical subspaces of molecules (Figure 1(a)), forming chemical paths between known molecules (Figure 1(b-c)) and obtaining a molecule representative of multiple structures (median molecules – Figure 1(b)). We make use of three important techniques within STONED. Firstly, within SELFIES, random character changes always correspond to valid molecules. Unlike other molecular representations, this allows us to perform random changes to molecules without taking validity into account. Secondly, every molecule can be represented with multiple SMILES strings, and multiple corresponding SELFIES. Since a single SELFIES has a limited number of possible character changes, we enhance diversity of generated structures within STONED by considering multiple representations for the same molecule. Lastly, we use the efficiency of fingerprint comparisons as a tool to enforce structural similarity because edit distances within SELFIES do not reflect it. With these techniques, we can form local chemical subspaces, discover median molecules and form chem-

TABLE I. Number and percentage of unique molecules obtained within different fingerprint-based similarity thresholds ($\delta$) of the starting structures. The molecules were generated using random SELFIES mutations of starting structures. The technique trivially achieves perfect GuacaMol benchmark scores.

| Starting Structure | Fingerprint Type | $\delta > 0.75$ | $\delta > 0.60$ | $\delta > 0.40$ | GuacaMol Score |
|---|---|---|---|---|---|
| Aripirazole | ECFP4 | 513 (0.25%) | 4,206 (2.15%) | 34,416 (17.66%) | 1.000 |
| Albuterol | FCFP4 | 587 (0.32%) | 4,156 (2.33%) | 16,977 (9.35%) | 1.000 |
| Mestranol | AP | 478 (0.22%) | 4,079 (1.90%) | 45,594 (21.66%) | 1.000 |
| | | Number of molecules (and percentage) | | | |

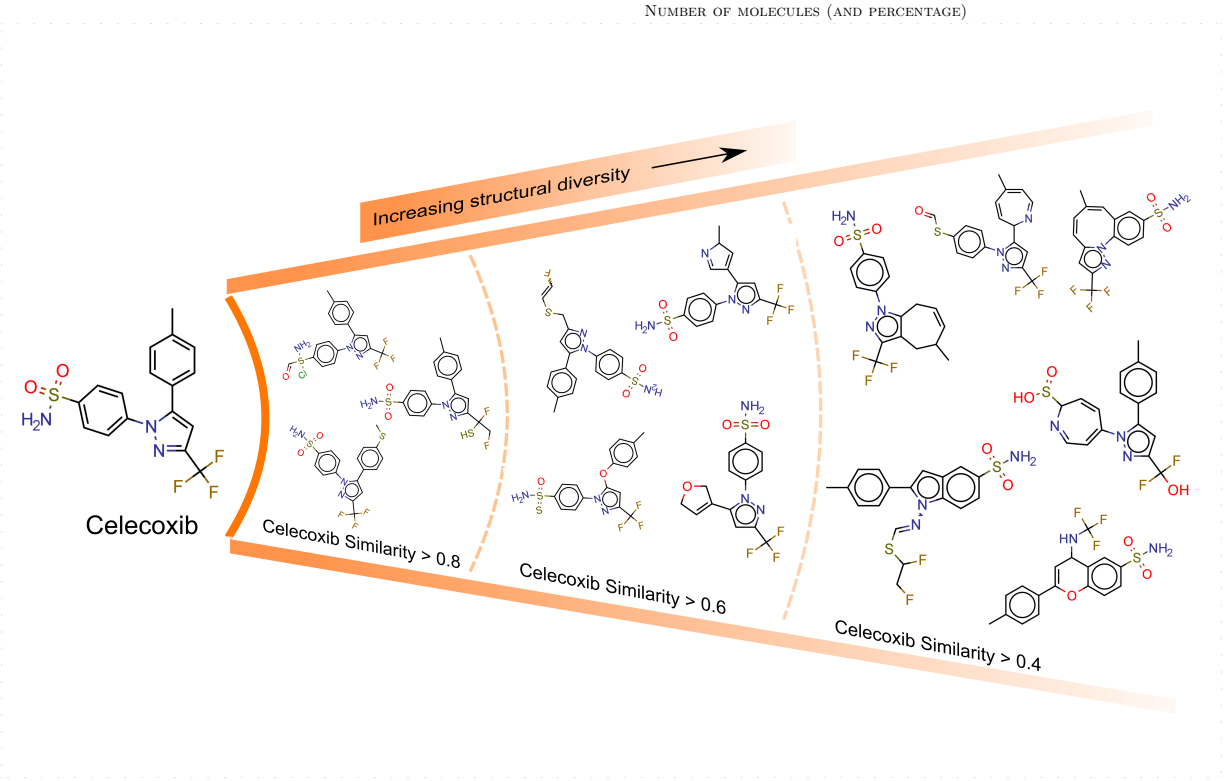


FIG. 2. Systematic local chemical space exploration of Celecoxib using mutations of different SELFIES representation. The similarity is calculated using the Tanimoto distance of the ECFP4 fingerprint between Celecoxib and the generated structures.

ical paths for interpolation.

### B. Formation of Local Chemical Spaces

The ability to generate the structural neighborhood of known molecules allows for local optimization. In drug discovery, molecular libraries are typically designed based on similarity to known active compounds to improve properties [22, 23]. The formation of these local chemical subspaces is usually achieved with predefined rules [24, 25]. However, the design of domain-specific rules for structure modification is time-consuming, non-trivial, and application-dependent. Hence, systematic methods for forming local chemical subspaces with minimal bias that can be used for any class of molecules are important. Additionally, on-the-fly structure generation has recently been considered as a benchmark to evaluate generative molecular design models in GuacaMol [26] and MOSES [27]. In these benchmarks, model quality is determined by the number of unique molecules generated within predefined fingerprint similarity thresholds. Notably, for deep generative models, the generation of unique molecules close to a target is biased by the resemblance between molecules of an independent training dataset and the target structure.

We started this work by performing point mutations of the molecules aripiprazole, albuterol, and mestranol [26] in the SELFIES representation to generate local chemical subspaces. A point mutation in the SELFIES representation corresponds to a single character addition, deletion or replacement. The corresponding results are provided in Table I. We achieve perfect results in the GuacaMol benchmarking objective, indicating that our algorithm is comparable to deep generative models for forming local chemical subspaces. In comparison to the state of the art in deep generative modeling for molecular design, our algorithm requires access to only one datapoint and is an order of magnitude faster.

Additionally, Figure 2 illustrates the algorithm's ability

to generate diverse structures in the neighborhood of the known drug Celecoxib [28]. As expected, we observe that the fraction of unique molecules obtained decreases with more stringent structure-based fingerprint similarity requirements. While the success rate of mutations leading to structurally similar molecules is low (Table I), our approach is extremely efficient, with the entire experiment running in just a few minutes on an ordinary laptop at the time of writing (Intel i7-8750H CPU, 2.20GHz). In particular, the most time-consuming experiment in Table I was the formation of Aripiprazole's subspace, completing in 500 seconds. The most expensive step in this experiment involved performing multiple SELFIES mutations and subsequently converting all mutated strings into SMILES, taking 400 seconds. Importantly, this step can be made more efficient by conducting mutations on different strings using parallel workers. Hence, this algorithm also possesses extensive parallelizability. The speed and scalability of our method suggest that it can be readily applied to extend datasets used in machine learning for creating more robust generalizable models.
Notably, in this experiment, we performed mutations solely on the starting structure. A natural extension is to repeat the procedure on all distinct neighbors to extend the subspace search significantly. Furthermore, we hypothesize that the 2D structure-based fingerprints can be replaced with efficient property-based molecular descriptors [29–31] or 3D fingerprints to form property-based or geometry-based chemical subspaces, respectively, for systematic chemical space exploration.

### C. Chemical Paths and Rediscovery

Another benchmark considered for generative modeling in GuacaMol [26] is rediscovery. The goal is to generate a predefined structure using the extended connectivity fingerprint (ECFPs) [32] similarity as a guide. Again, the performance of models biased by data, such as deep generative models, depends on the similarity between the training data and the target molecule. Usually, rediscovery is initiated with a given structure, which is iteratively modified to increase similarity with the target. This leads to the formation of chemical paths [33], a series of molecules, where each successive member is increasingly similar to the target.
Within the SELFIES universe, i.e. the set of all strings composed of SELFIES characters, the notion of path formation has a unique formulation. Using character replacements, deletions, and additions as possible mutations, for any given pair of SELFIES representing two distinct molecules, a finite number of modifications exist that interconvert them. We define every successive molecule encountered in this interconversion as within a path. Every one of these mutated SELFIES corresponds to a valid molecule. While this interconversion can in principle be achieved with any string-based molecular representation like SMILES or DeepSMILES [34], random modifications will very likely lead to the formation of syntactically or semantically invalid molecules [17]. Hence, there can be specific islands of valid molecules embedded within a sea of invalid strings. For example, between the SMILES strings *CCC1CCC1CCC* and *CCCCCCCCC*, no single mutations that correspond to an increase in Levenstein similarity form valid molecules, leading to a string without a valid chemical structure in the corresponding path.

Accordingly, in the next experiment, we optimize molecules in the SELFIES representation to rediscover celecoxib, troglitazone, and tiotixene [26]. However, instead of maximizing the fingerprint similarity between the initial structures and the targets, we task a genetic algorithm (GA) [35, 36] with maximizing the corresponding Levenshtein similarity [37]. Given two SELFIES, we define Levenshtein similarity as the relative number of matching SELFIES characters at corresponding indices normalized by the larger string length. While the use of fingerprint similarity to guide optimization has previously been established, their use can have varied success rates, depending on the choice of molecular representation and the target molecule to be discovered [33].

The use of Levenshtein similarity, in contrast, has a 100% success rate. In both Levenshtein and fingerprint similarity, one needs explicit knowledge of the target structure to perform comparisons and guide optimization – making the approaches essentially identical. However, contrary to Levenshtein similarity, an increase in fingerprint similarity always corresponds to structural changes making the initial molecule more similar in structure to the target molecule. As such, for every molecule on the path, we show the ECFP4 similarity to the target structure (Figure 3). Most importantly, all trajectories lead to perfect rediscovery. When guided by Levenshtein similarity, which only increases when equivalent characters are at equivalent positions, any mutation leading to an increase in similarity corresponds to placing the correct character at the correct position. This needs to be repeated until all the SELFIES characters have been changed to the target string. Thus, rediscovery guided by an objective that contains the full solution (i.e. the entire molecular graph, within representation such as SMILES, SELFIES, or adjacency matrices), as proposed by benchmarks in the literature, is trivial.

To summarize, we have shown that rediscovery and forming paths between two structures in the SELFIES universe is simple, independent of any datasets, and a trivial task rendering it inappropriate as a benchmark. In the subsequent sections, we analyze the properties of molecules encountered along chemical paths and show their application for efficient interpolation of both structure and property.

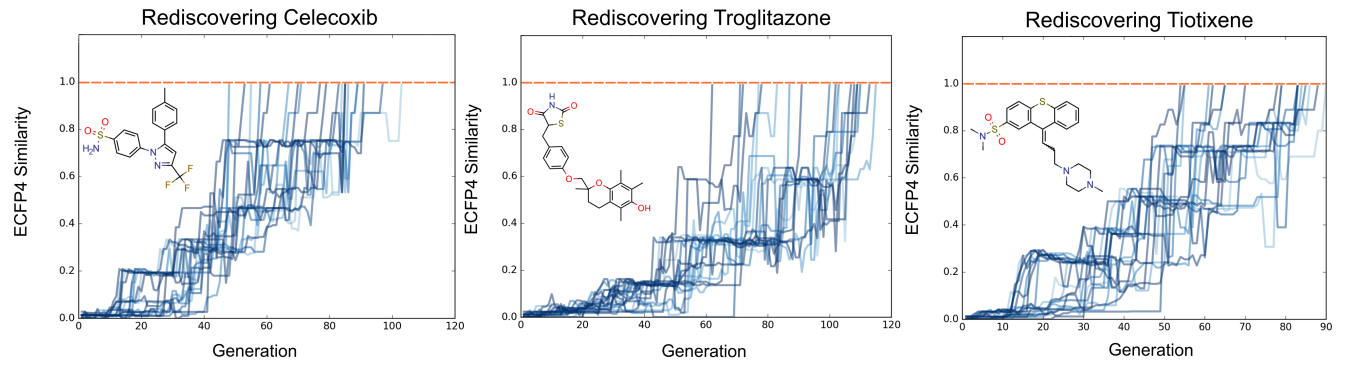


FIG. 3. ECFP4 similarity score of the best molecule found at the end of each iteration (generation). A perfect similarity is achieved for all 20 runs. Genetic algorithm optimization is performed using the Levenshtein similarity between the SELFIES strings. A value of 1.0 indicates perfect molecular rediscovery. Each run is seeded by a random SELFIES.

### D. Properties of Structures Along Chemical Paths

#### 1. *Measuring joint molecular similarity*

A median molecule of a given set of reference molecules is a molecule that resembles all the reference molecules simultaneously based on a selected similarity metric [38]. Recently, generation of median molecules has been proposed as a benchmarking objective within GuacaMol [26, 39]. In this benchmark, termed the median molecule discovery objective, the goal is to maximize the similarity to a predefined set of structures simultaneously, i.e. the joint molecular similarity. The problem can be viewed as identifying the largest fragments that are identical in a set of molecules. Notably, when the mutual similarity between the reference structures is small, the generation of median molecules can be challenging leading to low joint similarity metrics.

The similarity of proposed median structures to the references can be gauged via structure-based fingerprint similarity measures. In GuacaMol, a median molecule (i.e. $m$) of two known structures (i.e. $m'$, $m''$) is assessed based on the geometric mean of the respective fingerprint similarities to the two reference structures. The higher the geometric mean, the better the median molecule. However, we observe that maximizing the geometric mean of fingerprint similarities does not capture joint molecular similarity satisfactorily. The metric can return large values despite possessing high similarity only to one structure (see Section S1).

Therefore, we propose to redefine joint similarity for an arbitrary number of reference molecules $M = \{m_1, m_2, ...\}; n = |M|$, which is discussed in detail in the supplementary information (Sec. S1) :

$$F(m) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{sim}(m_i, m) - [\max_i(\mathrm{sim}(m_i, m)) - \min_i(\mathrm{sim}(m_i, m))] \quad (1)$$

In the subsequent sections, we investigate the behaviour of this joint molecular similarity along a chemical path between molecules which inadvertently leads to the generation of median molecules.

#### 2. *Interpolation via Chemical Path formation*

Previously, we analyzed fingerprint similarity to the target structure along the path between two molecules, and the molecules along the paths were generated via random SELFIES character mutations within a GA, replicating the benchmarking setup of GuacaMol. While a monotonically increasing fingerprint similarity score is not observed, one can extract chemical paths by requiring fingerprint similarities to increase. For a faster generation of chemical paths, we disregard randomness in mutations of SELFIES characters (see Section S2) leading to a speedup of more than one order of magnitude.

Because of the speed and parallelizability of chemical path generation, motivated by the idea that similarity in structure can correspond to similarity in properties, we looked into properties of molecules along a chemical path. As an initial test, we considered the water-octanol partition coefficient (logP) [40] and the quantitative estimate of drug-likeness (QED) [41] in paths between the known drugs Tadalafil and Sildenafil (Figure 4(a)) estimated using RDKit [42].

Moreover, we analyzed the binding affinity estimated via docking [43] in chemical paths between Dihydroergotamine and Prinomastat as a more challenging property to optimize (Figure 4(b)). Dihydroergotamine and Prinomastat have been discussed in the literature as potential inhibitors for the protein structures of Serotonin (5-HT1B) [44] and P450 2D6 (CYP2D6) [45]. The 5-HT1B receptor is a target for antimigraine drugs such as ergotamine and dihydroergotamine [44]. P450 2D6, on the other hand, contributes to the metabolism and elimination of more than 15% of drugs used in clinical practice. Among individuals, considerable variations exist in the efficacy and amount of CYP2D6 enzyme production. As a result, a clinical drug dose may need to be altered

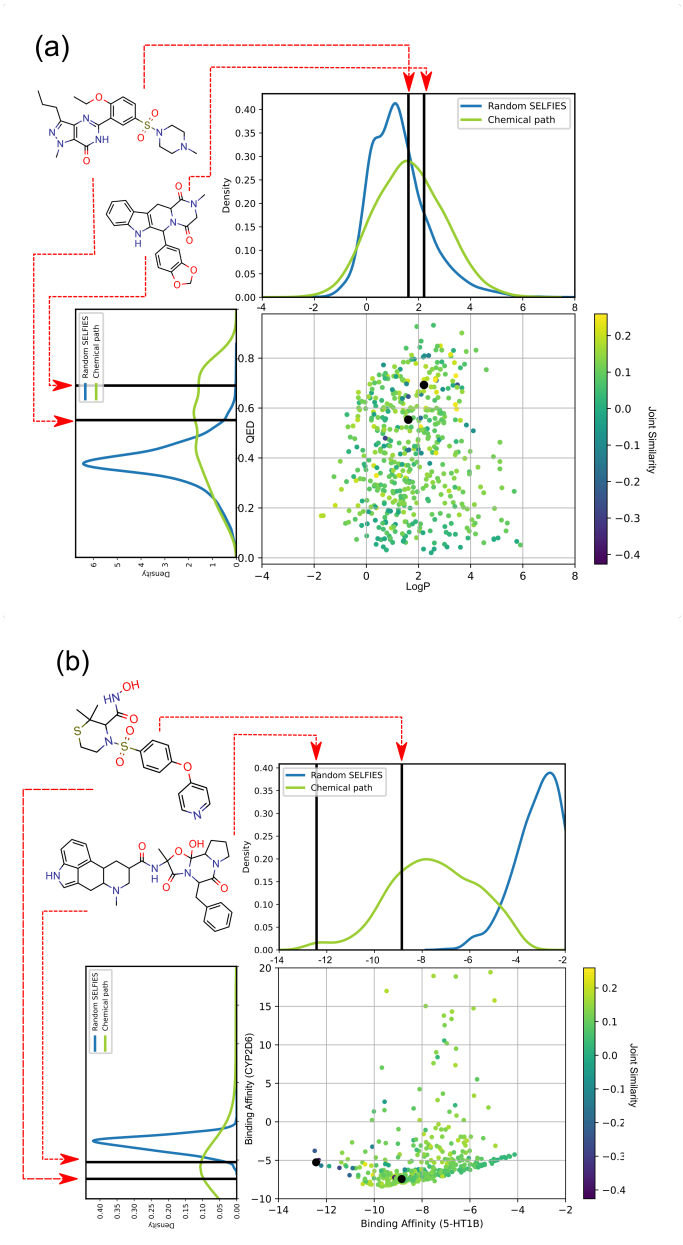


FIG. 4. (a) LogP and QED values of molecules encountered along chemical paths between Tadalafil and Sildenafil. (b) Ligand binding affinities of molecules encountered along chemical paths between Dihydroergotamine and Prinomastat. For both subfigures, the corresponding reference properties are indicated by black lines.

to account for the metabolization speed of CYP2D6 [46]. Prinomastat, as an inhibitor, decreases enzyme production, thereby allowing increased efficacy of certain drugs. Our goal in this experiment is to find molecules encountered during the paths that can simultaneously bind (i.e. possess large negative binding affinity values) to both proteins (see Figure 4(b)). One selected chemical path is depicted in Figure 5. Notably, some of the molecules obtained have unstable functional groups or would tautomerize in solution to a different structure. To improve both their stability and synthetic feasibility, rules based on domain knowledge can be implemented to modify the structures as little as possible.

Importantly, this experiment demonstrates the ability to achieve efficient structural interpolation between molecules without the need for forming continuous representations within deep generative models. Our simple algorithm for obtaining chemical paths possesses considerable potential for parallelization and does not need a large number of data points as input. Particularly, Cieplinski et al. [47] noted that with realistic training set sizes (i.e., consisting of a few thousand points), deep generative models have difficulty optimizing docking scores. In contrast, our approach to forming chemical paths between two known ligands yields several structures with non-trivial binding affinities to both proteins.

### E. Median Molecules for Photovoltaics

As pointed out previously, forming chemical paths between two molecules inadvertently leads to the generation of median molecules. Next, we generalized the concept of a chemical path to potentially having more than two reference molecules (see Section S2). As an application example, we considered the organic photovoltaic dataset form the Harvard Clean Energy (HCE) project [25], and identified 100 sets of three molecules (triplets) such that the first has a high lowest unoccupied molecular orbital (LUMO) energy, the second a high dipole moment, and the third a high energy difference between the highest occupied molecular orbital (HOMO) and LUMO energies (HOMO-LUMO gap), while having low values for the respective other two properties. This choice of properties reflects potential design objectives for organic photovoltaics [48]. HOMO-LUMO gap and LUMO energies determine the energy of light absorption and acceptor ability, respectively, while dipole moment can be considered a crude proxy for intermolecular interaction strength. We simulated these properties using the semiempirical GFN2-xTB quantum chemistry method [49] (details in the Methods Section).

We compared the ability of the obtained median molecules to resemble the triplet in structure (Figure 6(left)) and property (Figure 6(right)) of the references. For each triplet identified from the HCE database, we used the 100 median molecules with the highest joint similarities to the reference structures from chemical paths between three reference structures (*Unfiltered Medians*). We observed that many of these median molecules possessed bridgehead atoms, an undesirable structural feature for organic photovoltaics [50]. To remedy this problem, we added a simple filter discarding these molecules (*Filtered Medians*). In Figure 6(left), higher joint similarities indicate that the median molecules resemble the triplets more closely in structure. However, in Figure 6(right), low values of the normalized property distance

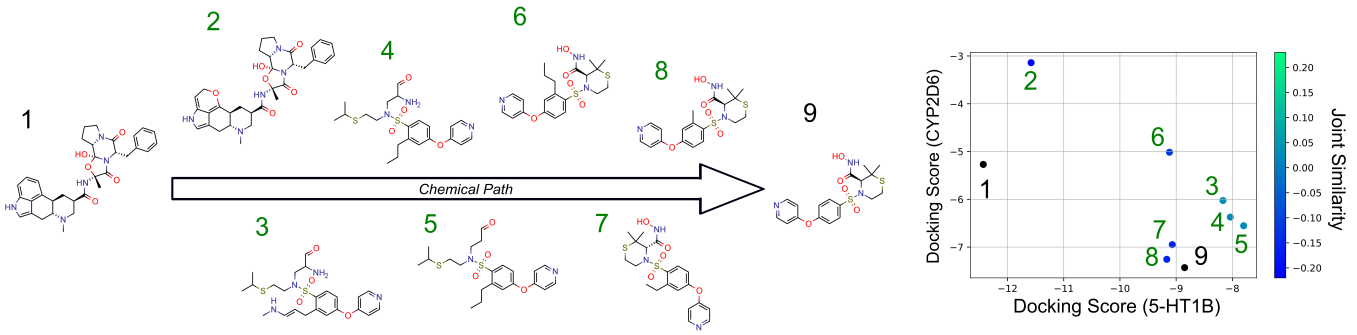


FIG. 5. Example of a chemical path between Dihydroergotamine (binder for 5-HT1B) and Prinomastat (binder for CYP2D6). Docking scores for the intermediate structures on both proteins and their joint similarity to the starting and target structures are provided in the diagram to the right.

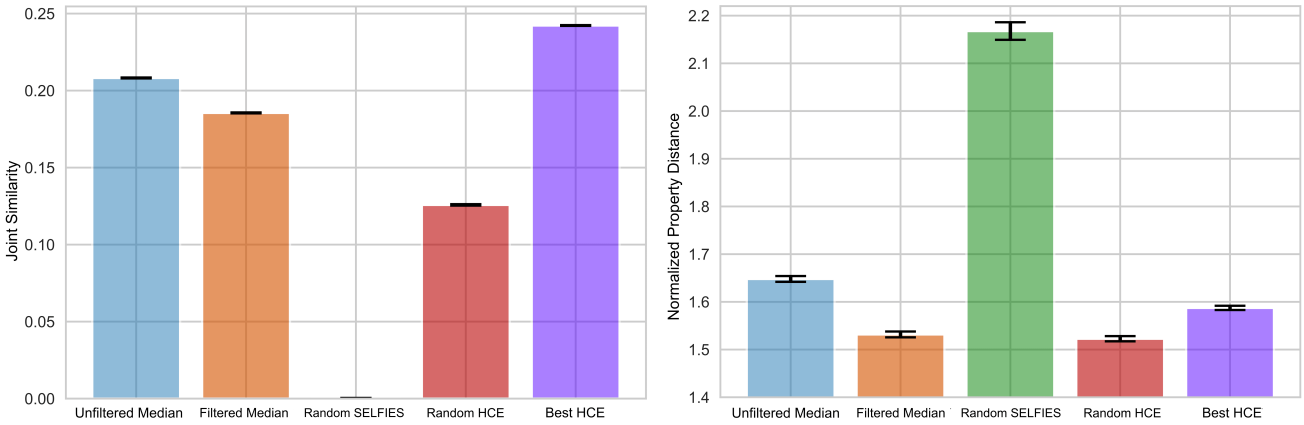


FIG. 6. Multi-objective property optimization of potential molecules of interest for photovoltaics. Structural (left) and property similarity (right) of generated median molecules compared to specific sets of three molecules taken from the Harvard Clean Energy (HCE) database. Bar plots for the mean, and error bars for the standard deviation of the mean (2 standard deviations) are shown for the joint similarity and the normalized property distance of the 100 median structures with highest joint similarities to the references, with (*Filtered Median*) and without (*Unfiltered Median*) a bridgehead atom filter. They are compared to *Random SELFIES* and to molecules from the HCE database (*Random HCE* and *Best HCE*). The obtained median molecules are very close to *Best HCE* in joint similarity and slightly better in the properties.

indicate that the median molecules have properties closer to the respective reference structures.

In Figure 6, *Random HCE* refers to sampling 100 random structures from the HCE database for each triplet, while *Best HCE* refers to the 100 molecules with the highest joint similarities to the reference structures available within the database. Importantly, we found that the median molecules are significantly closer in both structure and target properties to the respective triplets compared to *Random HCE*. In addition, they are also closer to the respective triplets in the investigated properties compared to *Best HCE* showing that generating median molecules can be an effective strategy for performing multi-objective property optimization (See Figure S3 & Table S2 for detailed statistics). Importantly, this task is a complicated multi-objective optimization in a chemical subspace tailored for a very specific application. Our method is able to produce molecules that are similar in structure to three molecules simultaneously. In that regard, our method produces structures similar in both structural similarity and property compared to a database of molecules obtained using a building block approach based on expert knowledge. Hence, our results are very promising for fully automated exploration of chemical subspaces based on a few reference structures without defining building blocks and rules to construct molecules.

Expert rules-based systems can yield median molecules [38, 51, 52], but their use can be application-dependent. For example, a potential algorithm could disassemble the reference structures into fragments by breaking rotatable bonds and then recombine the fragments in a building block approach. However, this technique would not be generalizable to molecules without rotatable bonds, such as fused polyaromatics, and more sophisticated algorithms would be required. Our method differs in that it requires no expert knowledge and relies solely on the graph representation of molecules within SELFIES and necessarily leads to a median molecule. Deep generative

models can be used to avoid such problems, with expert knowledge being derived solely from a known dataset. However, they require many training examples. Our approach is both rules-free and training-free.

## III. CONCLUSION AND OUTLOOK

In this work, we have introduced the STONED algorithm to perform simple, efficient exploration and interpolation in the chemical space. We demonstrate the simplicity of forming local chemical subspaces and obtaining chemical paths using SELFIES as molecular representation, readily achieving perfect performance metrics in the corresponding benchmarks. Furthermore, by redefining joint molecular similarity, we show that chemical path formation can be used as an efficient heuristic algorithm to find median molecules. Additionally, we showcase applications of STONED for molecular design in both drug discovery and material science.

The speed, parallelizability, and performance of STONED suggests that it can be used for practical tasks such as high throughput virtual screening [53]. In optimization algorithms such as genetic algorithms, we believe that median molecule generation through our approach can be used as a general crossover rule. The current evaluation standard for deep generative modeling includes producing valid molecules that resemble specific datasets [26, 27]. With the guarantee of molecular validity in SELFIES by design, perfect results in the validity benchmark can be trivially achieved. Furthermore, we demonstrate the simplicity of generating multiple structures that resemble a known set of molecules. Among other benchmarks, properties such as penalized logP and QED do not represent the complexity of molecular design, making them an insignificant benchmarking objective. Accordingly, we also demonstrated application to more complicated multi-objective property optimizations including protein docking, LUMO energies and HOMO-LUMO gaps as target properties. By introducing STONED, a fast class of algorithms that can compete with deep generative models on recently introduced benchmarks, we believe that we provide a stepping stone to improve generative modeling for molecular design and its benchmarking.

## IV. METHODS

**Formation of Local Chemical Spaces.** Starting from a molecule, we obtain 50,000 SMILES orderings representing the same structures, convert all to the SELFIES representation, and perform 1-5 mutations. A single mutation consists of a SELFIES character replacements, deletions, and additions at random positions of the string. The process is repeated to perform multiple mutations. All the mutated structures are subsequently converted back to SMILES to calculate their similarity to the original molecule based on different fingerprints. In GuacaMol, a perfect score of one is awarded to an algorithm if it can generate 100 structures possessing a fingerprint similarity greater than 0.75. Our algorithm trivially achieves a perfect score on this benchmark, generating more than 100 molecules possessing fingerprint similarity greater than 0.75.

**Chemical Paths, Rediscovery Interpolations.** In Section II C, we modify the code from Nigam et al. [18] to operate only in the
selfies universe without a neural network discriminator. We keep only the best performing molecule between iterations, i.e., the SELFIES string with the highest Levenshtein similarity, and repopulate the remaining members with single random mutations of the best. Single mutations lead to molecular strings that differ in exactly one position, forming molecules that possess high Levenshtein similarity to the original structure. The Levenshtein similarity score is then scaled between 0 and 1 by division with the length of the number of SELFIES characters in the larger string. Suppose that exactly $t$ characters differ in the corresponding indices of two SELFIES strings. Then there exist exactly $t!$ paths depending on which SELFIES characters are selected for mutation. The length of all such paths is $t$ as successive improvements are performed to the previous SELFIE string encountered in the path. Furthermore, similar to SMILES representations, a molecule can have multiple representations allowing multiple paths between any two given molecules. Considering $k$ representations of the target structure, each of which has $e_1, e_2, ..., e_k$ corresponding starting SELFIE characters different, the total number of paths becomes $\sum_{i=1}^{n} e_i$. It is worth noting that within GuacaMol, for rediscovery/path formation, Before any optimization, screening takes place, where the top-100 scoring, most similar structures from a 1.6 million subset of the ChEMBL database are provided as seeds to an algorithm to perform rediscover. Using Levenshtein similarity as a guide, we can perform guaranteed rediscovery starting from any molecule.

In Section II D 2, within a path, we randomly sample molecules that necessarily increase fingerprint similarity, depending on the previous sample, allowing for the formation of a chemical path. LogP QED values of molecules in a path are estimated using RDKit [42]. The docking scores are estimated with the SMINA open-source software [54] with the setup proposed by [47]. Namely, The crystal structures for 5-HT1B and CYP3D6 docking were obtained from the PDB database (4IAQ and 3QM4), the binding sites are selected manually, and the score of the top 5 best-scoring binding poses are averaged (for consistency of results). In both experiments, we consider different smile orderings of the starting and target molecule and, between each pair, repeat the experiment a few times, leading to different results, such

that approximately 800 unique molecules from the paths are obtained.
For path and chemical path formation, between two SELFIES, we pad the string to the same length with a dummy character. The dummy character is removed from the SELFIES when converting to SMILES.

**Median Molecules for Photovoltaics.** The molecules of the HCE database were ordered based on their ability to simultaneously maximize one property, while minimize the other. The top 100 structures from this ordered list were selected for our experiment in Section II E. In the formation of generalized paths, the starting molecule is selected randomly and 10,000 paths are obtained between randomized orderings of the SMILES string. We run calculations to obtain the dipole moment, LUMO and HOMO-LUMO for the HCE database and the top-100 unique median structures using GFN2-xTB quantum chemistry method [49].

## ACKNOWLEDGEMENTS

The authors thank Dr. Cyrille Lavigne for insightful discussions and for proof-reading the manuscript. R.P. acknowledges funding through a Postdoc.Mobility fellowship by the Swiss National Science Foundation (SNSF, Project No. 191127). M.K. acknowledges support from the Austrian Science Fund (FWF) through the Erwin Schrödinger fellowship No. J4309. G.P.G gratefully acknowledges the Natural Sciences and Engineering Research Council of Canada (NSERC) for the Banting Postdoctoral Fellowship. A.A.-G. thanks Anders G. Frøseth for his generous support. A.A.-G. acknowledges the generous support of Natural Resources Canada and the Canada 150 Research Chairs program. Computations were performed on the Béluga supercomputer situated at the École de technologie supérieure in Montreal. In addition, we acknowledge support provided by Compute Ontario and Compute Canada.

[1] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[3] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[5] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[7] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131, 2018.

[8] Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

[9] D. Neil, Marwin H. S. Segler, Laura Guasch, M. Ahmed, Dean Plumbley, Matthew Sellwood, and N. Brown. Exploring deep recurrent models with reinforcement learning for molecule design. In *ICLR*, 2018. `https://openreview.net/forum?id=BkOxiI1Dz`.

[10] R Vasundhara Devi, S Siva Sathya, and Mohane Selvaraj Coumar. Evolutionary algorithms for de novo drug design–a survey. *Applied Soft Computing*, 27:543–552, 2015.

[11] Jan H Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019.

[12] Naruki Yoshikawa, Kei Terayama, Masato Sumita, Teruki Homma, Kenta Oono, and Koji Tsuda. Population-based de novo molecule generation, using grammatical evolution. *Chemistry Letters*, 47(11):1431–1434, 2018.

[13] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[14] Yongbeom Kwon and Juyong Lee. Molfinder: An efficient global molecular property optimization and search algorithm using smiles. 2020. `https://chemrxiv.org/articles/preprint/MolFinder_An_Efficient_Global_Molecular_Property_Optimization_and_Search_Algorithm_Using_SMILES/13106891/1`.

[15] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. *arXiv preprint arXiv:1802.04364*, 2018.

[16] Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in neural information processing systems*, pages 6410–6421, 2018.

[17] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-referencing em-

bedded strings (selfies): A 100% robust molecular string representation. *arXiv preprint arXiv:1905.13741*, 2019.

[18] AkshatKumar Nigam, Pascal Friederich, Mario Krenn, and Alán Aspuru-Guzik. Augmenting genetic algorithms with deep neural networks for exploring the chemical space. *arXiv preprint arXiv:1909.11655*, 2019.

[19] Radislav Potyrailo, Krishna Rajan, Klaus Stoewe, Ichiro Takeuchi, Bret Chisholm, and Hubert Lam. Combinatorial and high-throughput screening of materials libraries: review of state of the art. *ACS combinatorial science*, 13(6):579–633, 2011.

[20] Gabriel dos Passos Gomes, Robert Pollice, and Alan Aspuru-Guzik. Navigating through the maze of homogeneous catalyst design with machine learning. 2020. `https://chemrxiv.org/articles/preprint/Navigating_through_the_Maze_of_Homogeneous_Catalyst_Design_with_Machine_Learning/12786722/1`.

[21] Vincent Zoete, Aurélien Grosdidier, and Olivier Michielin. Docking, virtual high throughput screening and in silico fragment-based drug design. *Journal of cellular and molecular medicine*, 13(2):238–248, 2009.

[22] Andreas Bender and Robert C Glen. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2(22):3204–3218, 2004.

[23] Hanna Eckert and Jürgen Bajorath. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug discovery today*, 12(5-6):225–233, 2007.

[24] Eric M Gordon, Ronald W Barrett, William J Dower, Stephen PA Fodor, and Mark A Gallop. Applications of combinatorial technologies to drug discovery. 2. combinatorial organic synthesis, library screening strategies, and future directions. *Journal of medicinal chemistry*, 37(10):1385–1401, 1994.

[25] Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S Sánchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M Brockway, and Alán Aspuru-Guzik. The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *The Journal of Physical Chemistry Letters*, 2(17):2241–2251, 2011.

[26] Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108, 2019.

[27] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *arXiv preprint arXiv:1811.12823*, 2018.

[28] Delyth Clemett and Karen L Goa. Celecoxib. *Drugs*, 59(4):957–980, 2000.

[29] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors*, volume 11. John Wiley & Sons, 2008.

[30] Gabriella Graziano. Fingerprints of molecular reactivity. *Nature Reviews Chemistry*, 4(5):227–227, 2020.

[31] Gaspar Cano, Jose Garcia-Rodriguez, Alberto Garcia-Garcia, Horacio Perez-Sanchez, Jón Atli Benediktsson, Anil Thapa, and Alastair Barr. Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Systems with Applications*, 72:151–159, 2017.

[32] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[33] Emilie S. Henault, Maria H. Rasmussen, and Jan H. Jensen. Chemical space exploration: how genetic algorithms find the needle in the haystack. 2:e11. Publisher: PeerJ Inc.

[34] Noel O'Boyle and Andrew Dalke. Deepsmiles: An adaptation of smiles for use in machine-learing chemical structures. *ChemRxiv*, 2018. `https://chemrxiv.org/articles/preprint/DeepSMILES_An_Adaptation_of_SMILES_for_Use_in_Machine-Learning_of_Chemical_Structures/7097960/1`.

[35] David E Goldberg. *Genetic algorithms*. Pearson Education India, 2006.

[36] Venkat Venkatasubramanian, King Chan, and James M Caruthers. Evolutionary design of molecules with desired properties using the genetic algorithm. *Journal of Chemical Information and Computer Sciences*, 35(2):188–195, 1995.

[37] Frederic P. Miller, Agnes F. Vandome, and John McBrewster. Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau?levenshtein distance, spell checker, hamming distance. 2009. isbn: 6130216904, Alpha Press.

[38] Nathan Brown, Ben McKay, François Gilardoni, and Johann Gasteiger. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *Journal of chemical information and computer sciences*, 44(3):1079–1087, 2004.

[39] Xiaoyi Jiang, A. Munger, and H. Bunke. An median graphs: properties, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1144–1151, 2001.

[40] Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873, 1999.

[41] G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

[42] Greg Landrum et al. Rdkit: Open-source cheminformatics. 2006.

[43] Tatu Pantsar and Antti Poso. Binding affinity via docking: fact and fiction. *Molecules*, 23(8):1899, 2018.

[44] Chong Wang, Yi Jiang, Jinming Ma, Huixian Wu, Daniel Wacker, Vsevolod Katritch, Gye Won Han, Wei Liu, Xi-Ping Huang, Eyal Vardy, et al. Structural basis for molecular recognition at serotonin receptors. *Science*, 340(6132):610–614, 2013.

[45] An Wang, Uzen Savas, Mei-Hui Hsu, C David Stout, and Eric F Johnson. Crystal structure of human cytochrome p450 2d6 with prinomastat bound. *Journal of Biological Chemistry*, 287(14):10834–10843, 2012.

[46] Lay Kek Teh and Leif Bertilsson. Pharmacogenomics of cyp2d6: molecular genetics, interethnic differences and clinical importance. *Drug metabolism and pharmacokinetics*, pages 1112190300–1112190300, 2011.

[47] Tobiasz Cieplinski, Tomasz Danel, Sabina Podlewska, and Stanislaw Jastrzebski. We should at least be able to design molecules that dock well. *arXiv preprint*

*arXiv:2006.16955*, 2020.

[48] Florian Häse, Loïc M Roch, Pascal Friederich, and Alán Aspuru-Guzik. Designing and understanding light-harvesting devices with machine learning. *Nature Communications*, 11(1):1–11, 2020.

[49] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15(3):1652–1671, 2019.

[50] Tayebeh Ameri, Gilles Dennler, Christoph Lungenschmied, and Christoph J Brabec. Organic tandem solar cells: A review. *Energy & Environmental Science*, 2(4):347–363, 2009.

[51] Nathan Brown, Ben McKay, and Johann Gasteiger. The de novo design of median molecules within a property range of interest. *Journal of computer-aided molecular design*, 18(12):761–771, 2004.

[52] Jonas Verhellen and Jeriek Van den Abeele. Illuminating elite patches of chemical space. *Chemical Science*, 11(42):11485–11491, 2020.

[53] Edward O Pyzer-Knapp, Changwon Suh, Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, and Alán Aspuru-Guzik. What is high-throughput virtual screening? a perspective from organic materials discovery. *Annual Review of Materials Research*, 45:195–216, 2015.

[54] David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of chemical information and modeling*, 53(8):1893–1904, 2013.

# Supplementary Information: Efficient Interpolation and Exploration in the Chemical Space Using String Representations

## S1. ANALYSIS OF JOINT SIMILARITY FUNCTIONS

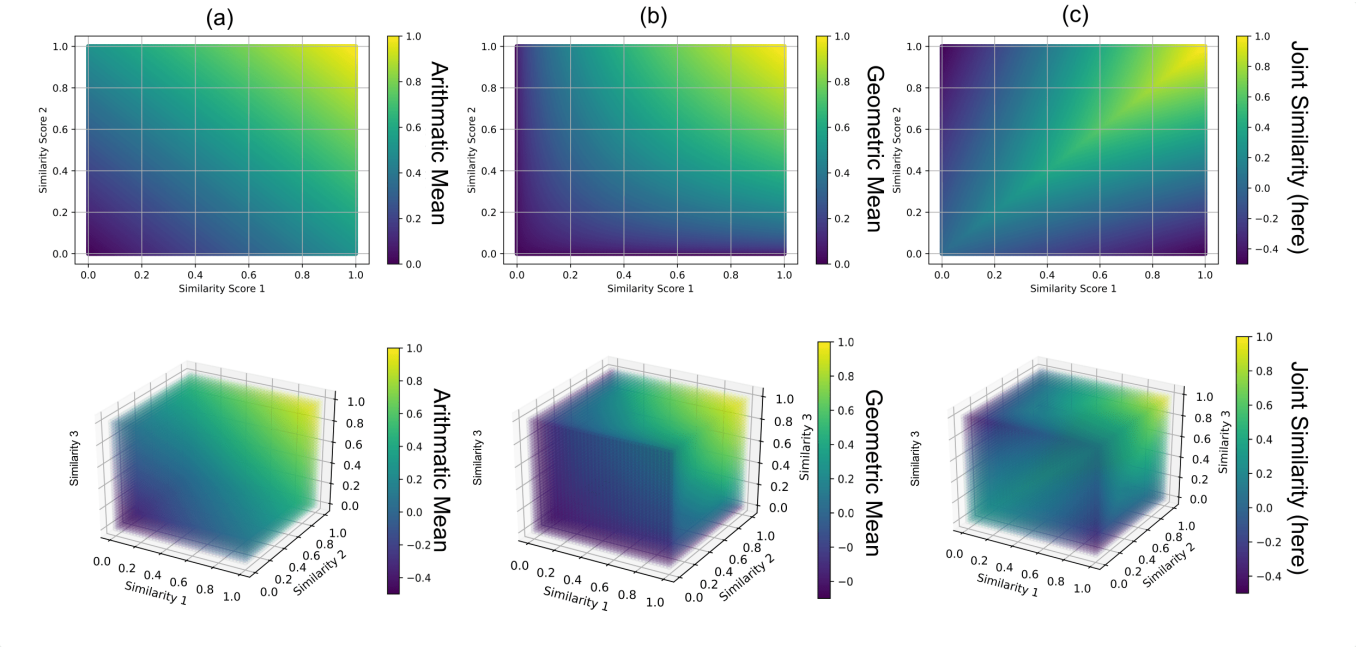


FIG. S1. Joint molecular similarity, calculated with (a) arithmetic mean, (b) geometric mean, and (c) Equations 1 for a set of two (top row) and three structures (bottom row). The axis indicates the Tanimoto similarity of the extended connectivity fingerprints, between a molecule and a reference structure, within the set.

We begin by analyzing the use of geometric mean for measuring join similarity (suggested in GuacaMol), for molecule $m$, with reference to $m'$ and $m'$. Say sim($m$, $m'$)=0.1, and sim($m$, $m''$)=0.9, then the geometric mean is 0.3. Alternatively, if sim($m$, $m'$)=0.3, and sim($m$, $m''$)=0.3, the geometric mean is 0.3 as well. Naturally, the molecule in the first example is more biased to just one structure, while in the second example, the structure is more representative of both. We plot the value of the geometric mean for the cases of two and three reference molecules in Figure S1(b). This problem becomes more prominent when the arithmetic mean (Figure S1(a)) is used instead of the geometric mean – in cases where the $m$ is the same as $m'$ or $m''$, and there is no similarity between $m'$ and $m''$, the score trivially reaches 0.5. This motivated our development of Equation 1 (Figure S1(c)).

Equation 1 shows the following boundary conditions:

1. When molecule $m$ is perfectly similar to all the molecules of the set $M = m', m'', ...$, F($m$) computes to 1.
2. When the molecule $m$ is similar to none of the structures of $M$, F($m$) computes to 0.
3. When the molecule is similar to only one structure in $M$. The minimum of the function is achieved, because all similarity scores range from 0 to 1. The value is obtained as:

$$F(m) = \frac{1}{n} - 1 = \frac{1}{n} - \frac{n}{n} = \frac{(1-n)}{n}; \text{Where } n = |m| \tag{2}$$

For intuitiveness, we fit a degree 3 polynomial through the F($m$) with the above three values (namely: 0, 1, and $\frac{(1-n)}{n}$) and assign them to 0, 1, and -1. Consequently, we observe an increasing gradual movement from:
(1) similar to only one structure in $M$ (joint similarity of -1), (2) similar to no molecule or strongly resemblance to one structure compared to the other (joint similarity close to 0, (3) and similar to all molecules (joint similarity of 1). The polynomial can be computed on the fly depending on the number of molecules n and is uniquely defined by the

three boundary conditions explained above and a local maximum at the point (1,1). Equation 1 shows the following boundary conditions:

- similar to only one structure in M (joint similarity of -1),
- similar to no molecule or strongly resemblance to one structure compared to the other (joint similarity close to 0,
- and similar to all molecules (joint similarity of 1).

## S2. REDUCED RANDOMNESS IN CHEMICAL PATH FORMATION & GENERALIZED CHEMICAL PATH

In a GA, given the top-performing molecule of a generation (i.e., the structure with the highest Levenshtein similarity), one can derive the next iterations best string by selecting a random index of the molecule's SELFIES string and mutating it to the right character of the target molecule's SELFIES. Essentially, this removes randomness with regards to which SELFIES character to use, thereby making obtaining chemical paths significantly faster.

To form a generalized path between a molecule $m$, and a set of molecules $M$, we randomly select an index (say $i$) in the SELFIES representation of $m$ and perform distinct mutations, yielding $|M|$ different SELFIES strings. The distinct strings are obtained by selecting the i-th character within the SELFIES representing the molecules of $M$, and mutating the SELFIES character at index $i$ in $m$. Among these $|M|$ distinct SELFIES, the joint similarity is calculated after conversion to SMILES, and the molecule with the largest joint similarity is identified. The process is repeated with this new string until all distinct indices are covered.

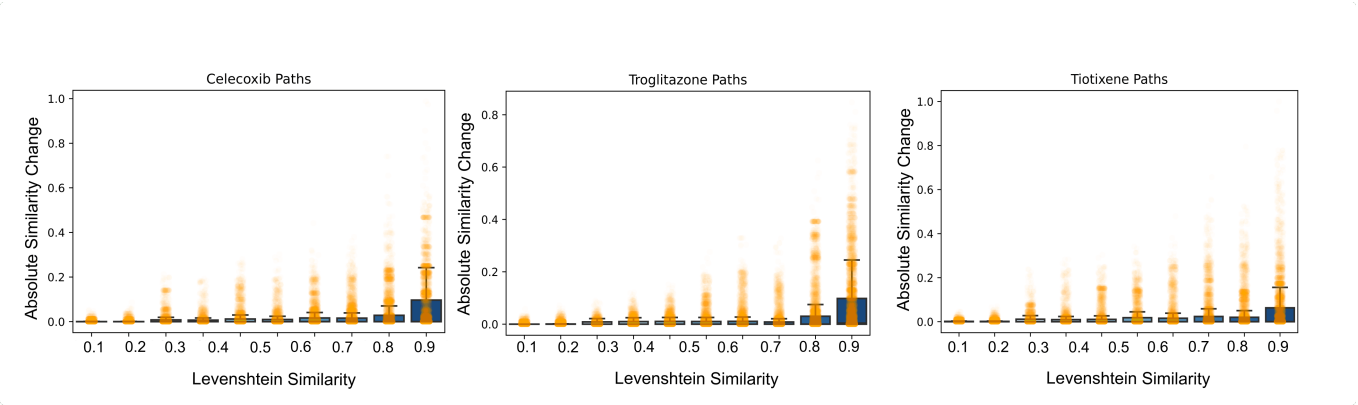


FIG. S2. A box plot of the change in Fingerprint similarity across 500 paths obtained from the setup of Figure 3 as a function of Levenshtein similarity. In the early stages of the paths, there is hardly any change in fingerprint similarity with increasing Levenshtein similarity, while towards the end, large fluctuations in fingerprint similarity are observed. This suggests that fingerprint-based rediscovery in the SELFIES universe is highly non-uniform and challenging.

TABLE S2. Joint similarity and normalized property distance of the most similar (top) median and the 100 most similar medians produced on 100 triplets of the harvard clean energy benchmark introduced in Section 2.4. Mean and standard deviation of the mean are provided.

| | **Top Median** | | **Top 100 Medians** | |
|---|---|---|---|---|
| | Joint Similarity | Normalized distance | Joint Similarity | Normalized distance |
| Unfiltered Median | 0.242±0.0035 | 0.668±0.0210 | 0.242±0.0035 | 1.648±0.0061 |
| Filtered Median | 0.226±0.0035 | 0.638±0.0221 | 0.186±0.0003 | 1.532±0.0061 |
| Random SELFIES | 0.017±0.0006 | 0.633±0.0236 | 0.000±0.0000 | 2.174±0.0198 |
| Random HCE | 0.222±0.0026 | 0.646±0.0221 | 0.126±0.0005 | 1.516±0.0056 |
| Best HCE | 0.281±0.0028 | 0.712±0.0253 | 0.242±0.0003 | 1.587±0.0045 |

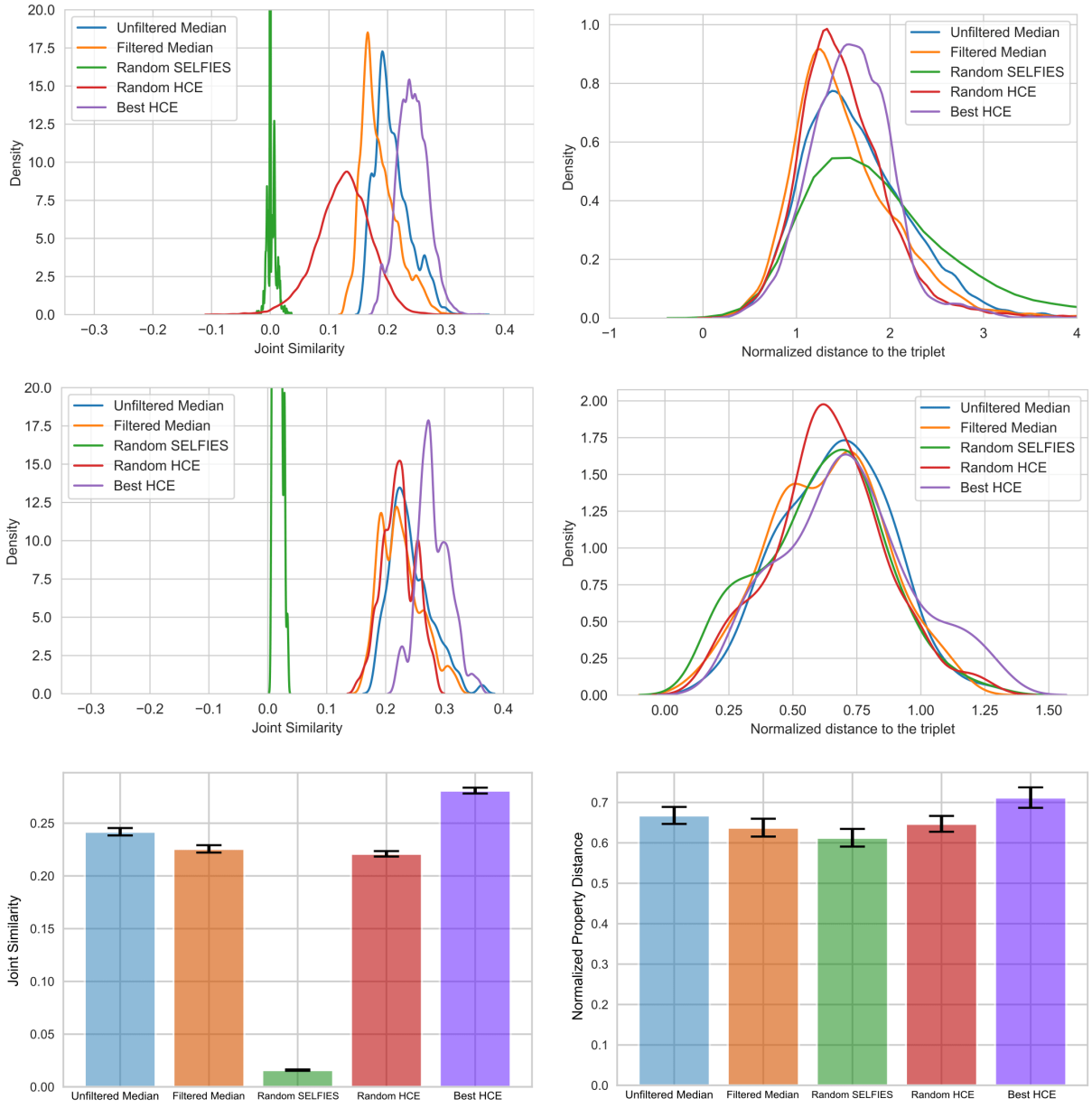


FIG. S3. Structural (left) and property similarity (right) of generated median molecules compared to specific triplets of molecules collected from the Harvard Clean Energy database. Density plots are shown for the joint similarity and the normalized property distance of the best median structures, for the best 100 medians (top row) and top median (middle row). Bar plots for the mean, and error bars for the standard deviation of the mean (2 standard deviations) and are shown for the joint similarity and the normalized property distance of the best median structure (bottom row).