Supplementary Information: Tartarus: Practical and Realistic Benchmarks for Inverse Molecular Design

S1. INTRODUCTION

Traditionally, property-guided optimization has relied on expert intuition [1] and several rounds of trial, error, and human-inspired optimization, occasionally supported by computer simulations. Alternatively, computer-assisted approaches have employed virtual screening [2] or optimization algorithms such as genetic algorithms (GAs) [3–5]. More recently, with the surge of deep learning, deep generative models have emerged, specifically designed to operate in chemical space and tackle inverse molecular design [6–8]. This has led to the development of numerous algorithmic approaches for this purpose, with the most popular including variational autoencoders (VAEs) [9, 10], generative adversarial networks (GANs) [11, 12], and reinforcement learning (RL) [13, 14].

S2. METHODS OVERVIEW

In this section, we provide an overview of the molecular generative models employed throughout this work and summarize the associated design choices we needed to make during their implementation. The molecular design algorithms we considered are VAEs, long short-term memory hill climbing (LSTM-HC) models [15–17], REINVENT [18], JANUS [19], and a graph-based genetic algorithm (GB-GA) [20]. At the core of the majority of these approaches are molecular string representations, the most commonly used of which is the Simplified Molecular Input Line Entry System (SMILES) [21]. Accordingly, many of the algorithms tested rely on predicting subsequent characters from partial strings to propose structures. However, algorithms based on SMILES can regularly produce invalid strings that do not represent molecules, which is problematic both in terms of robustness and interpretability of the corresponding methodologies [22, 23]. Consequently, this issue was addressed systematically by introducing Self-Referencing Embedded Strings (SELFIES) [22], a molecular string representation that guarantees validity. Thus, unlike for SMILES, every arbitrary combination of SELFIES characters represents a molecule. Nevertheless, its impact on structure optimization has not yet been evaluated systematically [23]. To address this issue, we modify some of the existing generative models relying on SMILES to be also compatible with SELFIES and test their performance depending on representation, similar to how it has been done recently [24].

Among the models tested, REINVENT, the VAEs, and the LSTM-HC models use recurrent neural networks (RNNs) [25] to learn the conditional probability distributions of the characters that represent molecules. RNNs are a class of artificial neural networks (ANNs) that utilize sequential information from their previous predictions and states. They have been incorporated into molecular design algorithms with special NN node architectures such as gated recurrent units (GRUs) [26] and LSTM cells [15]. The first of these models, i.e., REINVENT [18], is an RL-based approach that relies on an LSTM-based RNN as an agent which is tasked with generating molecules with desired properties. Over time and continued training, the agent learns to propose compounds with increasingly favorable target property values.

For the VAEs, we used the implementation described by Gómez-Bombarelli et al. as a starting point [10]. Therein, SMILES are converted to one-hot encodings and, subsequently, passed through a 1-dimensional convolutional neural network (CNN) encoder that generates a continuous latent space. A separate RNN decoder with GRU nodes regenerates the SMILES strings from the latent space. Molecular optimization is performed on the continuous representation of the latent space in a stochastic and iterative manner. Specifically, the best available molecule for a given task is encoded into the latent space of a trained VAE. Subsequently, random Gaussian noise is added to the obtained latent vector to produce a population of latent vectors that can be decoded to a population of new candidate structures. These structures are evaluated and the best compound obtained is used as a seed in the subsequent iteration. In our experience, this strategy leads to more stable optimization compared to direct gradient-based optimization in the latent space as described in the literature [10]. Importantly, we modified the original implementation to be compatible with both SMILES and SELFIES (cf. Computational Details in the Supporting Information). These variations will be referred to as SMILES-VAE and SELFIES-VAE, respectively.

As their name implies, the LSTM-HC models [15–17, 27] rely on LSTM cells in the NN architecture to model the character sequence probability distributions. After initial training, the resulting model is sampled using randomly truncated strings of the best currently available molecules for a given task. These truncated strings are used as seeds and completed stochastically by sampling the learned conditional probability distributions to produce a population

of candidate structures. The best new compounds obtained after property evaluation are used to retrain the model and initiate a subsequent sampling cycle. Thus, iterative sampling and retraining gradually improves the designs of the LSTM-HC models. Again, we test both the original implementation of LSTM-HC that relies on SMILES [16, 17], which will be referred to as SMILES-LSTM-HC, and a modified version making use of SELFIES, referred to as SELFIES-LSTM-HC. Notably, the SMILES-LSTM-HC model was among the best performing models in the original publication of the GuacaMol benchmarks [17] (referred to as "SMILES LSTM" in that reference).

Before the increasing adoption of ML approaches for molecular design, GAs [5, 28–30] were among the most popular computer-based methods for that purpose. Inspired by natural selection as defined in Darwinian evolution, GAs are heuristic population-based optimization algorithms. They rely on repeated stochastic generation of offspring from an existing population of candidate solutions via the genetic operations mutation and crossover. Subsequently, selection of the candidate solutions with the highest fitness determines the population to be propagated to the next generation, starting another iteration. In this work, we consider two specific implementations for inverse molecular design, GB-GA [20], and JANUS [19]. The former leverages structural information from a dataset of reference molecules by mimicking the corresponding distribution of atoms and bonds when performing genetic operations. Notably, in the original publication of the GuacaMol benchmarks [17], GB-GA achieved the best overall performance. In contrast, JANUS is a GA augmented by an NN classifier [19] (referred to as "JANUS+C" in the original publication) that actively judges the quality of a molecule before performing a full fitness evaluation. Molecules classified as likely possessing high fitness are then propagated to the next generation for subsequent evaluation. For mutation and crossover, JANUS relies on STONED-SELFIES [31], a set of algorithms based on the guaranteed validity of SELFIES to perform structural modifications. Additionally, unlike GB-GA, JANUS maintains two distinct molecule populations, one explorative and one exploitative with respect to conducted structural changes [19]. These populations exchange some members at every iteration but otherwise explore the chemical space independently [19].

When using Tartarus, the following procedures should be adopted to obtain benchmark results that are consistent with the ones provided herein. The first step for running one of the benchmarks, if necessary, is to train the generative model on the provided dataset. For all the ML models, we used the first 80% of the reference molecules for training and the remaining 20% for hyperparameter optimization. Then, the (trained) model is tasked with proposing structures to be evaluated by the objective function of the corresponding benchmark task. Notably, structure optimization was always initiated using the best reference molecule from the corresponding dataset. For the benchmarks concerned with designing photovoltaics, organic emitters, and protein ligands, structure optimization was carried out with a population size of 500 and a limit of 10 iterations, leading to a maximum number of 5,000 proposed compounds overall. For the design of chemical reaction substrates, we used the same maximum number of proposed compounds but used a population size of 100 and limited the number of iterations to 50 instead. Additionally, the associated run time was limited to 24 hours, which resulted in termination for several molecular design runs before reaching 5,000 molecule evaluations. Furthermore, to increase robustness and reproducibility of our results, we repeated each optimization run five times, allowing us to report the corresponding outcomes with both an average and a standard deviation. We believe that this resource-constrained comparison approach is necessary for fairly comparing methods and should be used as a standard by the community. A detailed account of the parameters and settings used for running each of the models is provided in the Computational Details section of the Supporting Information.

S3. ADDITIONAL RESULTS

A. Design of Organic Photovoltaics

OPVs offer simplified, cost-effective production [32, 33] and enhanced mechanical properties, particularly regarding specific weight and flexibility [34, 35]. Despite significant progress, OPVs still have lower power conversion efficiencies (PCEs) and shorter device lifetimes [36], prompting ongoing research efforts in molecular design for OSCs [35–37].

The simplest OPV cells comprise two electrodes and a photoactive layer for photoconversion, typically containing two distinct materials: donor and acceptor [33]. In bulk heterojunction cells [38], donors and acceptors are mixed, allowing nanoscale phase separation to maximize interfacial area and minimize distance within the phases [39]. Upon light exposure, excitons—neutral quasi-particles consisting of bound electron-hole pairs [33, 39, 40]—form in the photoactive layer with limited lifetimes and diffusion lengths [39]. The bulk heterojunction architecture enables excitons to reach the interface, dissociating into a free electron and hole that become charge carriers across the phases

[33, 39, 40]. These charge carriers travel through the photoactive layer to the electrodes, generating a photocurrent [33]. For exciton charge separation at the donor-acceptor interface to occur, the process must be energetically neutral or favorable [33, 40]. Therefore, the exciton energy, estimated by the HOMO-LUMO gap of the light-absorbing material, must be greater than or equal to the effective heterojunction bandgap [33, 40], approximated by the energy difference between the donor's highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital of the acceptor (LUMO) [41]. Notably, PCE is the percentage of power from the incident solar irradiation that is converted into electricity by the OPV device. We believe that the reduced reference dataset presents a more realistic scenario for new molecular design projects and is better suited for benchmarking generative models, especially when property simulations are time-consuming. We propose the following two molecular design benchmark objectives:

1. Maximize the following function:
   $PCE_{PCBM} - SAscore$.
2. Maximize the following function:
   $PCE_{PCDTBT} - SAscore$.

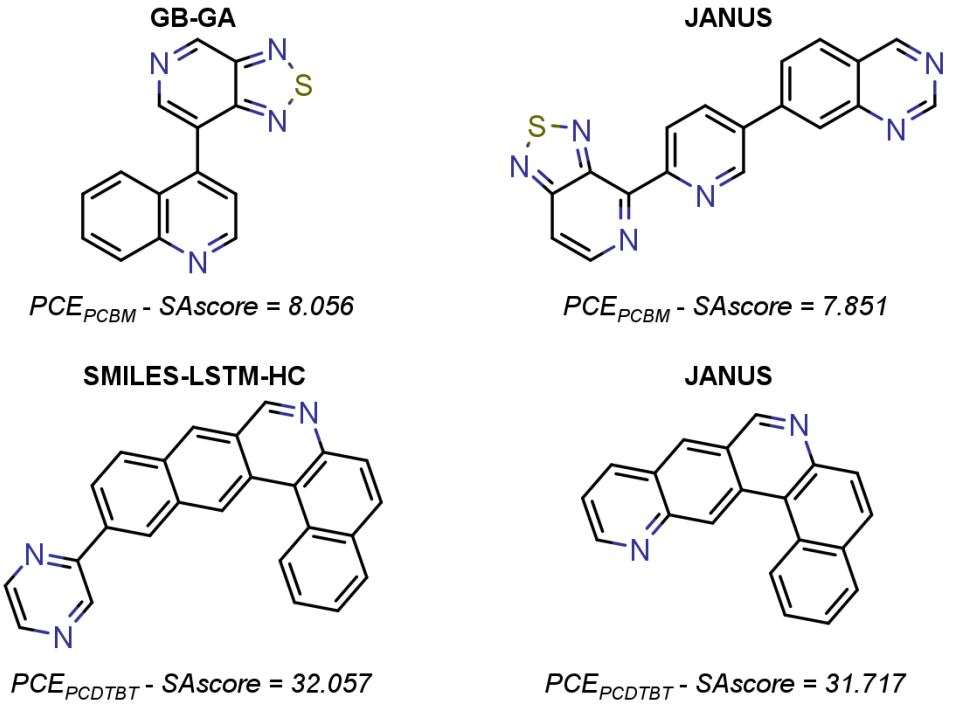

FIG. S1. Best molecules found in each of the benchmark tasks inspired by the design of organic photovoltaics. Additionally, the corresponding objective values and the molecular design models that proposed the structures are indicated.

The best molecules found in each of the design of organic photovoltaics benchmark tasks together with the corresponding objective values and the model that proposed them are depicted in Figure S1.

B. Design of Organic Emitters

The next set of benchmarks is inspired by the design of purely organic emissive materials for organic light-emitting diodes (OLEDs), which received significant attention in recent years [42–44] after the discovery of thermally activated delayed fluorescence (TADF) in the field [45]. Their main applications are digital screens and lighting devices [42]. For the former application, compared to alternative technologies, OLEDs offer improved image quality and enable both lighter and thinner devices [42]. For the latter, OLEDs are potentially more energy-efficient [42]. In OLED devices relying on TADF, after electric excitation, light generation takes place from the first excited singlet state via

fluorescence [45]. However, only 25 % of the initial excitons are excited via an electric current into singlet states and contribute directly to light emission via excited state thermal relaxation and subsequent prompt fluorescence [42–45]. The 75 % of the initial excitons that are excited into triplet states relax thermally to the corresponding first excited triplet states [42–45]. However, in ordinary organic molecules, light emission from their first excited triplet states via phosphorescence is slow, giving rise to various radiationless decay and decomposition pathways [42–44]. Consequently, both device efficiency and device lifetime are reduced considerably unless these triplet excitons can still be utilized for light production [42–44]. To achieve that, TADF emitters rely on minimizing the energy difference between the first excited singlet and triplet states, i.e., the singlet-triplet gap, allowing thermal upconversion of the triplet excitons to the first excited singlet state giving rise to delayed fluorescence [42–44]. Importantly, under the assumption of fast both forward and reverse intersystem crossing (ISC), the steady-state triplet population is governed by the singlet-triplet gap and its reduction leads to reduced triplet population and acceleration of delayed fluorescence. This increases internal quantum efficiency up to a maximum of 100 % and reduces decomposition of the emissive material [42–44]. Designing efficient emissive organic materials for blue OLEDs is of particular interest. This can be achieved by targeting organic molecules with excitation energies between their ground state and their first excited singlet states that correspond to the energy of blue light. The fitness functions of these three tasks are summarized as follows:

- Minimize the singlet-triplet gap, $\Delta E(S_1\text{-}T_1)$.
- Maximize the oscillator strength for the transition between S_1 and S_0, f_{12}.
- Maximize the following function:
  $+f_{12} - \Delta E(S_1\text{-}T_1) - |\Delta E(S_0\text{-}S_1) - 3.2\ eV|$.

The best molecules found in each of the design of organic emitters benchmark tasks together with the corresponding objective values and the model that proposed them are depicted in Figure S2.

C. Design of Protein Ligands

Notably, while docking simulations are a standard tool in virtual screening pipelines of drug discovery campaigns, their accuracy compared to experimental binding affinities is at best modest [46–49]. Thus, typical workflows use them merely for a preselection which is narrowed down further with subsequent free energy simulations [50]. Nevertheless, for molecular design benchmarks, docking still provides the best trade-off between computational efficiency and relevance to real-world molecular design. Additionally, it is important to realize that drug design requires the consideration of many other molecular design aspects that make or break a hit compound, e.g., toxicity, solubility, stability and many more. However, as these properties are significantly harder to model computationally, we decided to disregard them for our set of benchmarks. Notably, finding small molecule ligands for the selected proteins marks the first step towards the development of treatments for various important conditions.

Importantly, most likely due to its maturity, complexity, and demand for resources, drug design was probably the first chemical problem where molecular design algorithms were tested comprehensively. The use of computer algorithms has a long-standing history in medicinal chemistry and GA-based molecular design algorithms making use of full atomic representations have already been used as early as 1993 [51]. Initial toy tasks for testing these algorithms included rediscovery of known drugs via a structural similarity metric, highly simplified molecular docking of ligands to rigid protein binding sites minimizing the interaction scores, and the optimization of estimated molecular properties like the decadic logarithm of the n-octanol-water distribution coefficient (log P) [51]. Interestingly, some of the still most widely used benchmarks for generative models rely on essentially the same types of metrics as they are simple to implement and compute [17, 52–55]. More recently, the quantitative estimate of drug-likeness (QED) was introduced, which is a desirability function that uses the position of a small set of common and simple molecular descriptors relative to the corresponding distributions in a dataset of approved drugs to estimate structural resemblance to therapeutics [56]. Due to its simplicity, it found immediate application in several molecular design benchmarks [12, 57, 58]. However, using QED alone in generative models is not meaningful for finding drug candidates as it only accounts for the general structural requirements of drug-like molecules but disregards intended modes of action with respect to specific targets entirely. The specific benchmark objectives we implemented are summarized below.

- Minimize the docking score to 1SYH, ΔE_{1SYH}.

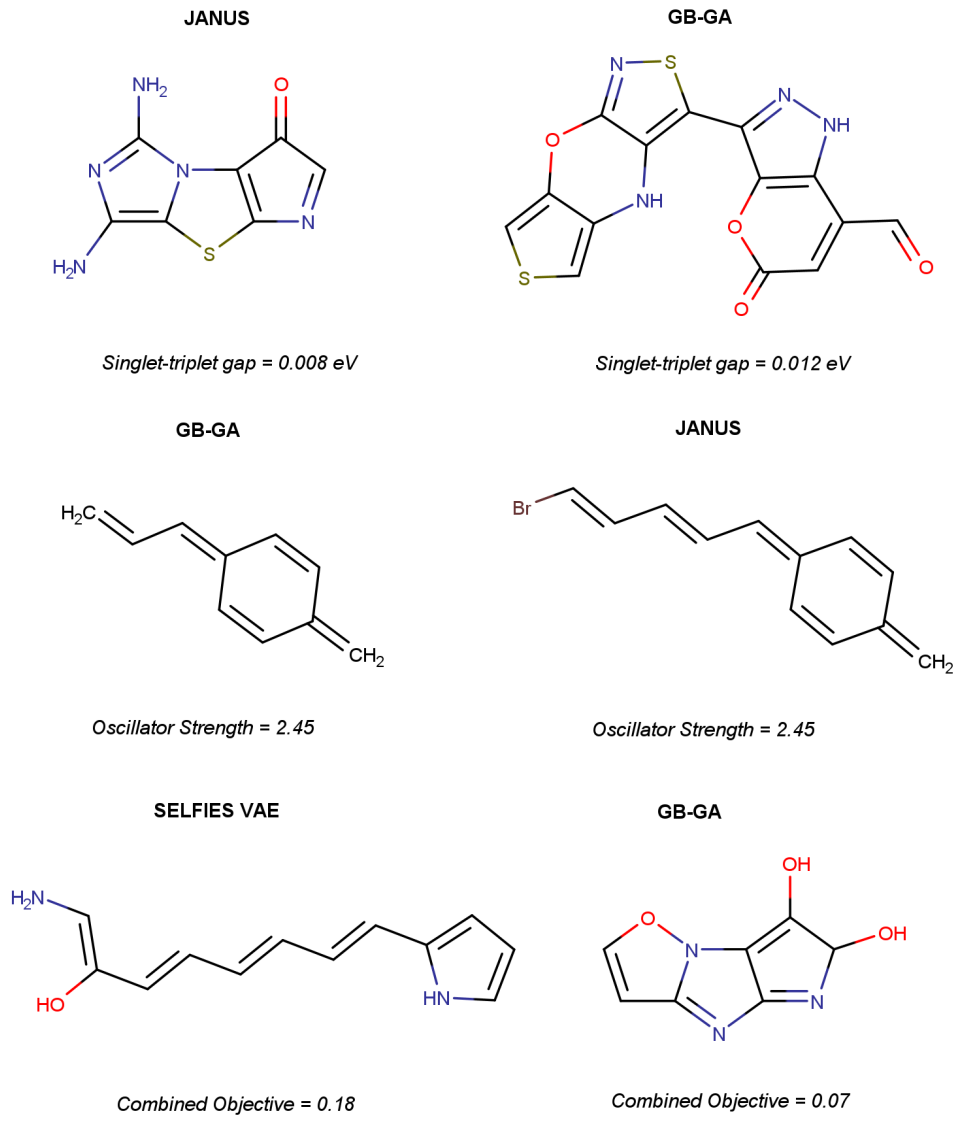

FIG. S2. Best molecules found in each of the benchmark tasks inspired by the design of organic emitters. Additionally, the corresponding objective values and the molecular design models that proposed the structures are indicated.

- Minimize the docking score to 6Y2F, ΔE_{6Y2F}.
- Minimize the docking score to 4LDE, ΔE_{4LDE}.

Notably, the corresponding objective functions do not solely consist of the docking scores but also have hard structural constraints directly incorporated. When these constraints are not fulfilled, an extremely unfavorable score of 10,000 is returned instead of the actual docking score. They consist of a set of filters that checks for the presence of unstable or reactive structural moieties and determines whether the compound in question fulfils Lipinski's Rule of Five [59]. Notably, most of these filters were developed based on our previous experience using molecular design algorithms to minimize docking scores and are tailored to avoid extremely unstable and reactive molecules that seem to be strongly favored by molecular docking simulations [19]. Additionally, the constraints avoid rings with more than 8 members as the docking approach implemented is unable to sample the corresponding conformations in a proper manner [60]. To fulfill them, the proposed structure needs to have an SAscore smaller than 4.5, which is the revised optimal threshold for that metric proposed in the literature [61], and a QED value larger than 0.3, which corresponds to the first quartile of the distribution of QED values for compounds in the CHEMBL database [56]. A list of these metrics is provided here:

- Absence of reactive groups.

- Absence of formal charges.
- Absence of radicals.
- At most 2 bridgehead atoms.
- No rings larger than 8-membered.
- Fulfils Lipinski's Rule of Five.
- $SAscore < 4.5$.
- $QED > 0.3$.
- $TPSA > 140$.
- Molecule passes the PAINS and WEHI and MCF filters.
- Molecule does not contain Si and Sn atoms.

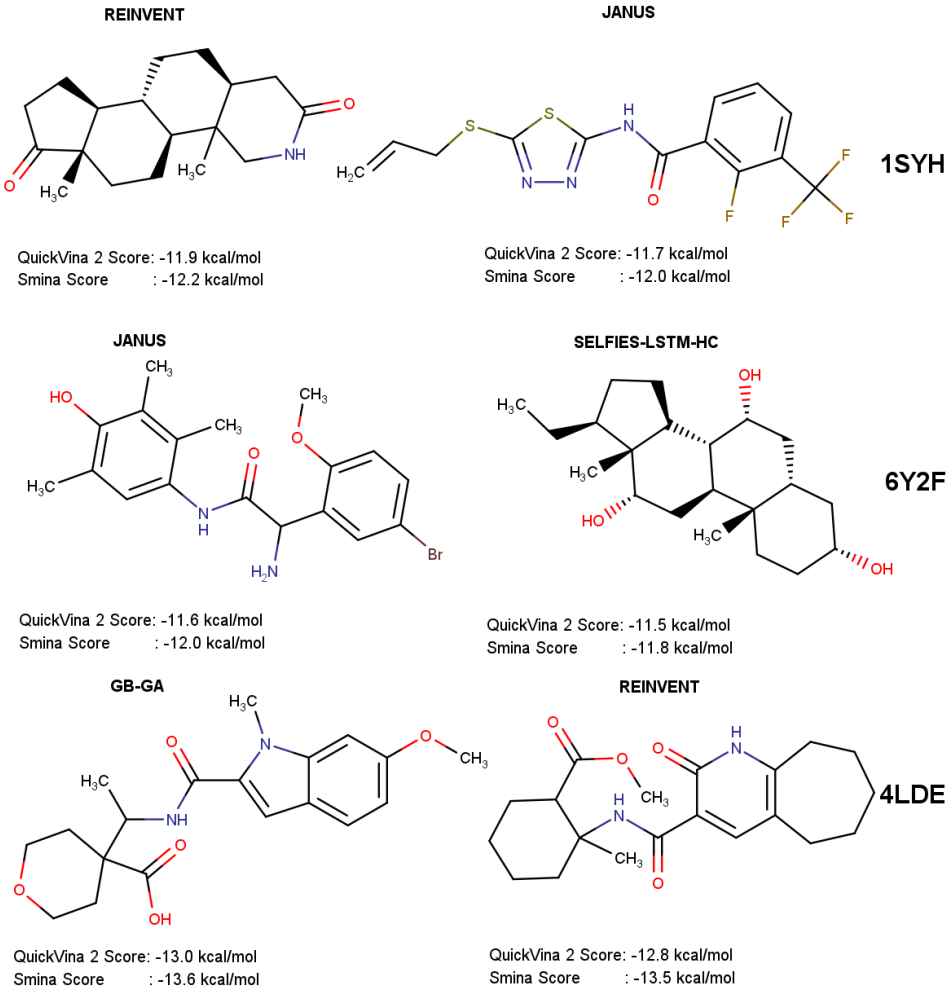

FIG. S3. Best molecules found in each of the benchmark tasks inspired by the design of protein ligands. Additionally, the corresponding objective values and the molecular design models that proposed the structures are indicated.

The best molecules found in each of the design of protein ligands benchmark tasks together with the corresponding objective values and the model that proposed them are depicted in Figure S3.

D. Design of Chemical Reaction Substrates

Whereas, classically, the optimization of reaction parameters was largely dominated by experimental work, in recent years, the significant increase in computing power and the continuous improvement of computer algorithms enabled molecular simulations to play an increasingly important part [62–65]. With the aid of transition state (TS) theory, fundamental reaction parameters such as thermodynamic feasibility, reaction rate and selectivity can be computed from first principles [66, 67]. This requires explicit modeling of the corresponding TS, a postulated state on the multi-dimensional potential energy surface (PES) of the process, which lies on a saddle point of order one [68]. Due to the difficulty of finding such saddle points in high-dimensional spaces and the often delicate electronic structure associated with the corresponding structures, in practice, automated TS optimizations often suffer from high failure rates in the range of 10–50% [69–71], and even well-behaved case studies combined with robust methodologies still have some room for improvement in that respect[72]. Additionally, they are typically carried out with relatively resource-intensive density functional approximation (DFA) calculations taking on the order of hours or even days to complete [69, 71]. Overall, these issues make them ill-suited for benchmarking molecular design algorithms or for routine application combined with generative models in computer-guided inverse molecular design campaigns. Nevertheless, GAs have been employed for computational catalyst design, particularly via the use of regression models based electronic structure descriptors derived from SQC simulations [73] and based on both structural and electronic structure descriptors obtained from DFA computations [74]. Most notably, very recently, GAs have been employed to optimize an organocatalyst for the Morita-Baylis-Hilman reaction [75].

The two force fields cross at approximately the TS geometry, and we introduce a coupling term that turns this into an avoided crossing, ensuring smoothness of the PES [76]. The resulting PES has a local maximum on the ground state surface (the TS) and a corresponding local minimum on the excited state surface, at approximately the same geometry. This allows optimizing TS geometries with robust gradient-based minimization algorithms. Thus, the SEAM force field method delivers activation energy estimates for reasonably sized molecules within minutes, and, in our hands, reaches very high success rates above 99.9% on a set of test reactions. Notably, we implemented this method in our Python package `polanyi`, which will be described in more detail in a separate publication. To generate the reference dataset for this set of benchmarks, starting from the unsubstituted reactive core structure, we performed repetitive cycles of STONED-SELFIES mutations [31] followed by removing all proposed compounds that violated the core and functional group constraints (Details in the Supporting Information). Thus, after several cycles, we obtained approximately 60,000 molecules defining the reference structures, which we refer to as SNB-60K dataset. Notably, in the selected reaction, there is only one TS connecting reactants and products in the selected reaction. Additionally, the fourth task aims to break the Bell–Evans–Polanyi principle [77–80], a linear free energy relationship that holds empirically for a large number of reactions. The following list summarizes the four benchmark tasks for chemical reactivity.

- Minimize the activation energy, $\Delta E^{\ddagger}$, and maintain the core and functional group constraints.
- Minimize the reaction energy, ΔE_r, and maintain the core and functional group constraints.
- Minimize the following function:
  $+\Delta E^{\ddagger} + \Delta E_r$, and maintain the core, functional group and SAscore constraints.
- Minimize the following function:
  $-\Delta E^{\ddagger} + \Delta E_r$, and maintain the core, functional group and SAscore constraints.

On top of these primary objectives, we added several hard constraints that the target molecules need to fulfill in order to reward generative models that propose realistic and feasible molecules. In particular, all substrates need to retain the *syn*-sesquinorbornene motif (referred to as "core constraint") which is required for the reaction to take place. Additionally, we selected a set of unstable and reactive substructures that need to be avoided (referred to as "functional group constraint", details in section S5 B 4 of the Supporting Information). Furthermore, for the two benchmark tasks with objective functions combining two target properties, we also required all proposed structures to possess an SAscore [81] of 6.0 or lower (referred to as "SAscore constraint"). The following list summarizes the four benchmark tasks for chemical reactivity.

The best molecules found in each of the design of chemical reaction substrates benchmark tasks together with the corresponding objective values and the model that proposed them are depicted in Figure S3.

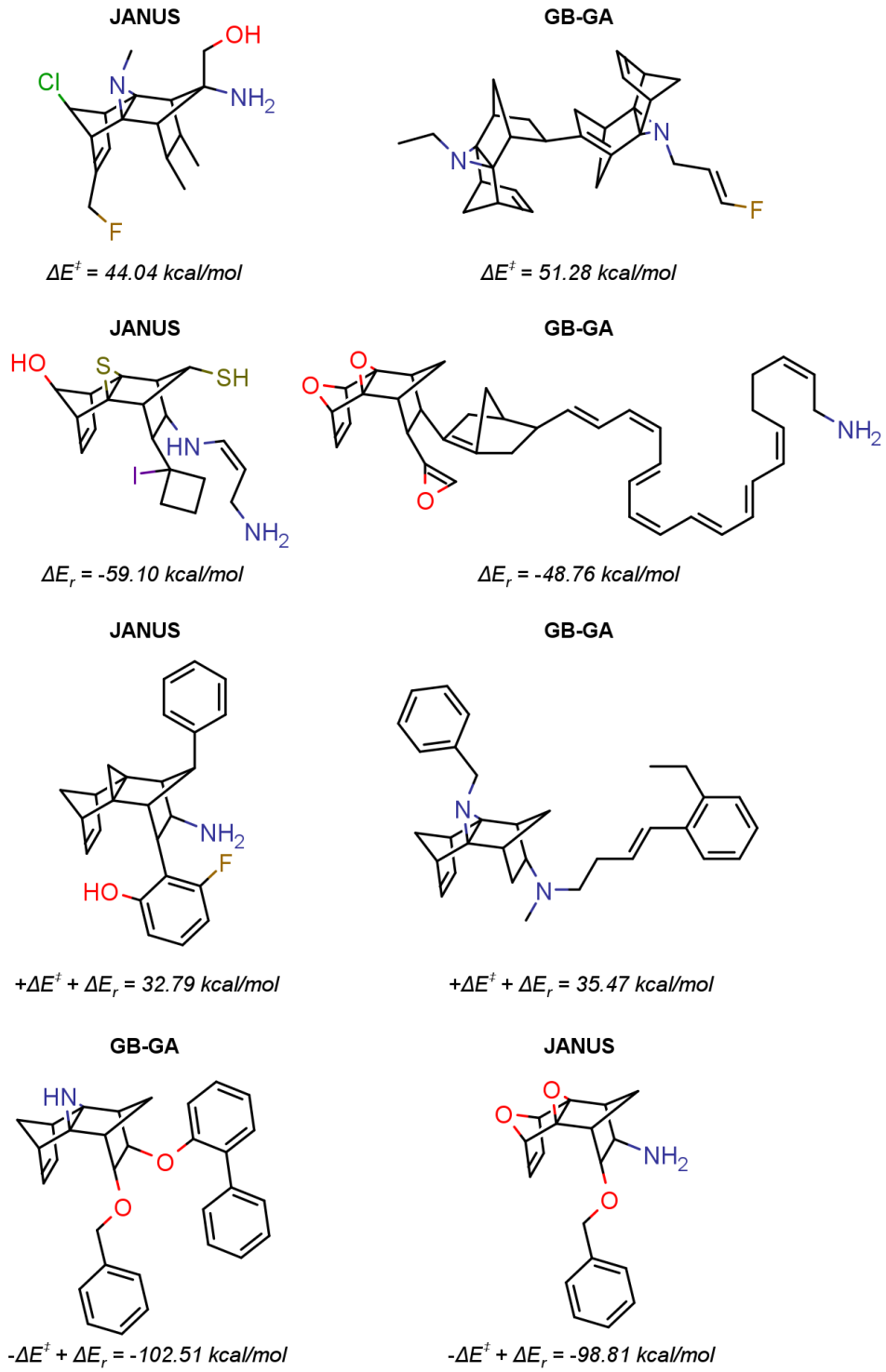

FIG. S4. Best molecules found in each of the benchmark tasks inspired by the design of chemical reaction substrates. Additionally, the corresponding objective values and the molecular design models that proposed the structures are indicated.

E. Model Timing Comparison

For ML-based molecular design algorithms, the pre-conditioning corresponds to training time and sampling time, respectively. For the GA-based algorithms considered, they translate to the duration of the derivation of genetic

operators from a reference dataset, and of applying the genetic operators to propose new candidate solutions, respectively. When the number of property evaluations is kept constant, assuming that the molecular size distribution between the generative models does not differ significantly, these steps are the major origin of timing differences between the algorithms. Notably, as was done for generating the molecular design results, we derived these timing metrics from five independent measurements and provide the results as averages with standard deviations.

Comparison of the single epoch times with and without a GPU allows estimating which models profit most from GPU usage. We find that the longer the single epoch time, the more the model profits from using a GPU which is consistent with expectations. Finally, looking at the sampling times, we find that REINVENT significantly outperforms all other methods considered needing less than 1 minute. Most of the other molecular design algorithms need between 2 and 3 minutes with GB-GA having the slowest sampling time of 6 minutes. Nevertheless, the differences between the methods are less pronounced here.

TABLE S1. Raw values for the timing benchmarks. Mean and standard deviation (mean ± s.d) timings for different models are provided based on five independent runs. N.A. means not applicable.

	Training Time [s]	**Epochs**	**Sample Time [s]**	**CPU Epoch Time [s]**	**GPU Epoch Time [s]**
SELFIES-VAE	36910 ± 912	75.6 ± 5.9	201 ± 53	29810 ± 949	535 ± 38
SMILES-VAE	34868 ± 667	74.1 ± 6.2	154 ± 79	28476 ± 724	515 ± 39
MoFlow	2804 ± 216	70.1 ± 5.4	62.41 ± 5	2131 ± 63	45 ± 8
GB-GA	N.A.	N.A.	314 ± 54	3.653 ± 0.007	N.A.
JANUS	N.A.	N.A.	147 ± 4	10.3 ± 2.4	N.A.
REINVENT	2844 ± 310	33.8 ± 1.2	33 ± 18	397 ± 14	81.7 ± 9.2
SMILES-LSTM-HC	7208 ± 1605	45.8 ± 10.5	128.0 ± 1.7	1870 ± 139	157.4 ± 5.6
SELFIES-LSTM-HC	7321 ± 1039	45.4 ± 6.3	119.3 ± 0.5	1661 ± 112	161 ± 27

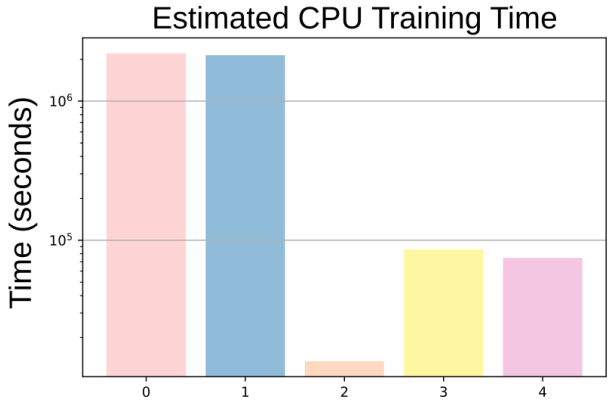

FIG. S5. Estimated CPU training time for the deep generative models based on the CPU single epoch time and the number of epochs during training with GPUs. Models are trained on a subset of the DTP Open Compound Collection. Results are provided as mean (main bar) from five independent runs. The numbers on the abscissa each refer to the following molecular design algorithms. Model 0: SELFIES-VAE; Model 1: SMILES-VAE; Model 2: REINVENT; Model 3: SMILES-LSTM-HC, Model 4: SELFIES-LSTM-HC.

The numerical results of all the timing benchmarks are provided in Table S1. The results of the estimation of total CPU training times are illustrated in Figure S5.

F. Diversity Calculation

Following the definition of diversity in the literature[82], for each task, we calculate diversity of the proposed moleucles using the following equation:

$$\text{Diversity} = 1 - \frac{2}{n(n-1)} \sum_{X,Y} \text{sim}(X, Y) \tag{1}$$

The expression $\text{sim}(X, Y)$ computes the pairwise molecular similarity for all n structures calculated as the Tanimoto distance of the Morgan fingerprints, which were obtained with a radius of size 3 and a 2048 bit size [83]. The individual results for the diversity evaluations of all the benchmark tasks are provided in Tables S2-S5. Overall, we observe mostly relatively small differences in the diversity of the proposed molecules between the different models, except for the chemical reaction substrate design benchmarks. We hypothesize that the strict structural requirement of the corresponding tasks are responsible for this observation as only molecules with the correct reactive core structure are subjected to the property simulation workflow. Additionally, we observe that all generative models except for SMILES-LSTM-HC show consistently a relatively high diversity across all four groups of benchmarks. The molecules proposed by SMILES-LSTM-HC for the organic photovoltaics and for the chemical reaction substrate benchmarks have quite a low diversity compared to the other generative models. Nevertheless, these results suggest that the diversity metric can be insightful in cases of especially low values, but otherwise does not allow to distinguish between the performance of the generative models in a reliable way.

TABLE S2. Diversity results for the organic photovoltaics benchmarks. Models are trained on a subset of the Harvard Clean Energy Project Database. Results are provided as mean and standard deviation of the best target objective values that are obtained in five independent runs in the form mean ± standard deviation. Metrics: PCE_{PCBM}: PCBM power conversion efficiency; PCE_{PCDTBT}: PCDTBT power conversion efficiency; SAscore: synthetic accessibility score.

	$PCE_{PCBM} - SAscore$	$PCE_{PCDTBT} - SAscore$
SMILES-VAE	72.4 ± 2.6%	81.6 ± 3.0%
SELFIES-VAE	**91.1 ± 0.3%**	**91.6 ± 0.3%**
MoFlow	88.9 ± 3.2%	88.0 ± 2.2%
SMILES-LSTM-HC	33.6 ± 7.2%	25.8 ± 13.5%
SELFIES-LSTM-HC	90.6 ± 0.6%	90.5 ± 1.6%
REINVENT	88.4 ± 0.1%	88.3 ± 0.1%
GB-GA	86.2 ± 0.4%	89.2 ± 0.3%
JANUS	88.2 ± 0.6%	88.9 ± 0.2%

TABLE S3. Diversity results for the organic emitters design benchmark objectives. Models are trained on a subset of the GDB-13 dataset. Results are provided as mean and standard deviation of the best objective values from five independent runs in the form mean ± standard deviation. Metrics: $\Delta E(S_1 - T_1)$: singlet-triplet gap; f_{12}: S_1 and S_0 transition oscillator strength; $\Delta E(S_0\text{-}S_1)$: S_0 and S_1 vertical excitation energy.

	$\Delta E(S_1 - T_1)$	f_{12}	$+f_{12} - \Delta E(S_1\text{-}T_1) - \lvert\Delta E(S_0\text{-}S_1) - 3.2; eV\rvert$
SMILES-VAE	90.1 ± 1.1%	78.5 ± 2.7%	80.4 ± 5.3%
SELFIES-VAE	**93.0 ± 0.2%**	89.5 ± 0.6%	92.1 ± 0.6%
MoFlow	92.2 ± 1.4%	89.1 ± 0.2%	90.3 ± 0.6%
SMILES-LSTM-HC	89.90 ± 0.01%	89.91 ± 0.01%	89.91 ± 0.01%
SELFIES-LSTM-HC	89.91 ± 0.01%	89.91 ± 0.01%	89.91 ± 0.01%
REINVENT	90.8 ± 0.2%	90.8 ± 0.1%	90.9 ± 0.1%
GB-GA	92.0 ± 0.2%	91.1 ± 0.6%	92.1 ± 0.4%
JANUS	91.7 ± 0.1%	**92.6 ± 0.1%**	**92.3 ± 0.2%**

TABLE S4. Diversity metrics for protein-ligand design benchmarks, based on models trained on a subset of the DTP Open Compound Collection. Metrics show mean and standard deviation of optimal target objective values over five independent runs (mean ± standard deviation). Benchmark metrics: ΔE_X denotes docking score for protein target X.

	ΔE_{1SYH}	ΔE_{6Y2F}	ΔE_{4LDE}
SMILES-VAE	80.2 ± 0.2%	91.0 ± 1.0%	89.9 ± 1.4%
SELFIES-VAE	**92.8 ± 0.4%**	91.9 ± 0.7%	90.6 ± 0.8%
MoFlow	92.5 ± 1.6%	**92.5 ± 1.5%**	**92.6 ± 1.6%**
SMILES-LSTM-HC	91.19 ± 0.01%	91.14 ± 0.04%	91.15 ± 0.02%
SELFIES-LSTM-HC	91.223 ± 0.001%	91.11 ± 0.01%	91.11 ± 0.01%
REINVENT	92.4 ± 0.2%	92.4 ± 0.3%	92.4 ± 0.1%
GB-GA	92.1 ± 0.3%	92.2 ± 0.2%	92.2 ± 0.1%
JANUS	91.5 ± 0.3%	91.6 ± 0.7%	90.9 ± 1.0%

TABLE S5. Diversity results for chemical reaction substrate design benchmarks. Models, trained on a benchmark dataset generated with STONED-SELFIES cycles, yield mean and standard deviation of optimal objective values over five runs (mean ± standard deviation). Benchmark metrics: $\Delta E^{\ddagger}$: Activation energy of the reaction; ΔE_r: Reaction energy.

	$\Delta E^{\ddagger}$	ΔE_r	$\Delta E^{\ddagger} + \Delta E_r$	$-\Delta E^{\ddagger} + \Delta E_r$
SMILES-VAE	70.4 ± 10.0%	81.7 ± 1.6%	66.5 ± 7.3%	64.1 ± 5.1%
SELFIES-VAE	69.2 ± 4.4%	66.9 ± 5.9%	64.3 ± 4.3%	68.4 ± 12.6%
MoFlow	**90.6 ± 1.9%**	87.1 ± 3.2%	**84.9 ± 4.4%**	**85.7 ± 3.7%**
SMILES-LSTM-HC	34.2 ± 13.2%	37.6 ± 11.2%	3.8 ± 1.5%	18.4 ± 15.4%
SELFIES-LSTM-HC	59.3 ± 11.3%	75.6 ± 5.4%	55.6 ± 19.8%	72.5 ± 5.8%
REINVENT	88.0 ± 0.4%	**88.3 ± 0.3%**	83.2 ± 0.8%	83.2 ± 0.8%
GB-GA	81.6 ± 1.1%	83.3 ± 2.2%	76.3 ± 1.6%	80.8 ± 0.6%
JANUS	77.3 ± 2.6%	81.2 ± 1.4%	68.9 ± 4.0%	71.4 ± 2.0%

S4. ADDITIONAL DISCUSSION

In this section, we will put our findings into perspective and compare them to the pre-existing knowledge in the field. Specifically, we will address two major talking points. First, we will compare the molecular design benchmarks developed as part of Tartarus to existing benchmarks that are currently being used. Second, we will investigate the results obtained in the various molecular design objectives in detail and discuss their potential implications for the current status of artificial molecular design and future directions for the field.

A. Benchmarking with TARTARUS

First, we would like to emphasize that the benchmark results provided as part of this work largely serve as a demonstration of how the objectives can be used to compare the performance of different algorithmic approaches. They should not be viewed as final performance judgement for any of the methods used as model hyperparameters were not optimized comprehensively. Additionally, changing the constraints with respect to computer time and property evaluations will likely lead to different results calling for further efforts towards better benchmark standardization and the introduction of various benchmark scenarios. We believe that explicitly considering the number of evaluations and computation times are still important to make the use of molecular design algorithms more accessible and relevant to the scientific community as long run times are detrimental to widespread application. Nevertheless, we think that useful insights are still to be garnered from our preliminary benchmark results. When looking at all the outcomes combined, one curious observation is that, unlike many previous benchmarking efforts, we do not find one algorithm to perform the best, or at least close to the best, across all the tasks considered. This suggests that the newly introduced design objectives differ significantly in their structure requirements and algorithmic demands. Thus, we believe that different domains of chemistry and material sciences have different structure and design requirements that could potentially motivate the development of algorithms tailored to specific problem domains. At the very least, it suggests that significant further developments are necessary in the field in order to create a true champion algorithm for the design tasks considered.

Next, we will investigate the results of each of the molecular design sets in detail. The benchmark tasks inspired by the design of OPVs prove extremely difficult as most models fail to improve upon the reference structures and, if they do, the improvements are very small. This suggests that the provided dataset already contains almost optimal solutions with respect to the fitness functions chosen. Nevertheless, the models should at least be able to reproduce the dataset results. We suspect that these difficulties of some of the models is likely related to uncertainties in the learned structural space [84]. Notably, the property distributions depicted in Figure S8 show that the multiobjective function is mainly dominated by the PCE values. We believe that increasing the relative importance of the SAscore in the objective function could improve this benchmark task even further. Nevertheless, our results suggest that the design of OPVs based on the Scharber model and an explicit account of synthetic accessibility is delicate and proves challenging for many common molecular design algorithms. Based on the judgement of an expert chemist, the best structures proposed by the molecular design algorithms are likely all stable and synthesizable showing the importance of accounting for that explicitly in molecular design benchmarks. It is interesting to see that the two best structures found in each of the objectives are quite similar which is likely because they do not stray far from the

best compounds already contained in the reference dataset. The best structures for the donor design task seem to prefer having a 2,1,3-thiadiazole moiety while the acceptor design task prefers extended annulated π-systems.

Looking at the objectives based on the design of OLEDs, they are all practically relevant as they are directly based on the design of TADF emitters. The molecular design algorithms showing most consistent performance across the three tasks are JANUS and GB-GA. Interestingly, SELFIES-VAE also shows good performance over all three tasks and it outperforms SMILES-VAE very significantly which could suggest difficulties of the SMILES-VAE to learn the underlying property distribution of the training dataset. Notably, none of the molecular design algorithms used outperform the best molecule in the reference data for the oscillator strength benchmark. Additionally, only three of the models, namely SELFIES-VAE, JANUS, and GB-GA, deliver better properties for the combined objective than the reference structures possess. This demonstrates the difficulty of the corresponding benchmark tasks. Looking at the best performing molecules proposed by the models, the best structures are largely stable and synthesizable. However, there are a few features that are not ideal. In particular, the two molecules with the highest oscillator strength have an oxidized benzene ring that is likely reactive. Additionally, the molecule with the highest combined objective is present in its enol tautomer but it is likely that the corresponding ketone is preferred. This suggests that incorporating the SAscore as a constraint helps increase the stability and synthesizability of the generated molecules but also suggests that additional filters are necessary for higher robustness in that regard. Moreover, some structural features required for favorable properties can be derived. In particular, for high oscillator strength values, extended π-systems are beneficial which is well-known in the field of organic electronic materials [85, 86].

Next, we need to inspect the outcomes of the protein ligand design benchmarks, which also have the lowest number of specific objectives, and is currently the only benchmark set without a multiobjective molecular design task. Most interestingly, while the general setup of all the three objectives is identical, we find the performance of the molecular design algorithms to be varying significantly depending on the specific target. This suggests that designing ligands based on molecular docking for the three proteins chosen differs significantly in terms of structural requirements. Most models succeed at proposing ligands with better docking scores than the best reference compounds for all three target proteins. In terms of the constraints, surprisingly, the VAEs show the poorest performance which hints towards difficulties during training. Otherwise, JANUS also generally shows somewhat lower success rates which is likely caused by its limited features to bias structure generation by reference compounds and its tendency to explore more diverse structural regimes to potentially find new candidate solutions. This is relevant as the success rates are not only measured on the best-performing molecules but over all the structures proposed. In contrast, the LSTM-HC models, REINVENT and GB-GA consistently reach higher success rates demonstrating that they succeed at learning to propose desirable structures based on the reference dataset. Additionally, it likely also implies that exploration far outside of the distribution of the reference dataset is avoided. Thus, overall, REINVENT and GB-GA provide the most promising results in these protein ligand design benchmarks. When looking at the best-performing structures, it is notable that molecules resembling steroids seem to be good binders for at least two of the protein targets selected. Notably, there is one important aspect to be discussed that relates specifically to the dataset of reference compounds and the simulation workflow. While the DTP Open Compound Collection is particularly attractive to be used as training dataset because all these structures have been synthesized and tested experimentally for potential use as drug candidates, it suffers a major drawback in terms of data quality. Unfortunately, the majority of structures lack proper definition of stereochemistry despite often having multiple stereogenic centers. While our computational workflow is deterministic with respect to the particular stereoisomer produced when provided with SMILES without defined stereochemistry, it is still very likely that stereoisomers were produced and simulated that do not correspond to the ones of the actual molecules in the compound collection. Thus, when SMILES without well-defined stereochemistry are submitted to the property prediction workflow, the 3-dimensional structure generated in the process needs to be inspected to derive the stereoisomer computed and guide subsequent experimental validation. For even more realistic drug design benchmarks tasks in the future, a reference dataset of structures that have both been experimentally synthesized before and have well-defined stereochemistry that reflects the experimental samples needs to be used.

When investigating the results of the chemical reaction substrate design benchmark tasks, we find that only the two GAs employed consistently outperform the best reference structures. This is particularly notable as both JANUS and GB-GA achieve large improvements over the best properties in the training set for all but the fourth benchmark objective showing that there is, in principle, significant room for further improvement. Additionally, in many cases, the proposed molecules perform significantly worse than the best-performing reference compounds suggesting that many of the algorithms used seem to have difficulties to learn from the provided dataset. This could potentially point towards peculiarities of the provided training set as it was the only one that was created in a random way from the necessary core structure. It might be too small in size paired with a high diversity of structural moieties making it challenging to extract structure-property relationships. Notably, as expected because it goes against

the Bell–Evans–Polanyi principle, the fourth benchmark objective proves to be extremely difficult with the best results providing only comparably small improvements over the reference set. The structures of the best proposed compounds for each of the design objectives provide some preliminary insights into the structural features effecting favorable properties (cf. Figure S4). The sites donating hydrogen atoms seem to prefer being more substituted for both lowering the barriers and lowering the reaction energies. This effect can be understood by a reduction of steric repulsion between these substituents and the bridge substituents when going from tetrahedral to trigonal geometry around the respective carbon atom. Particularly effective for lowering reaction energies seem moieties suitable for conjugation with the newly generated double bond in the product, owing to the stabilizing effect of extended conjugation. To break the correlation between reaction energies and activation energies, the introduction of two mesomerically electron-donating groups appear most suitable, potentially due to the opposing factors of sterics, lowering the reaction energy as outlined above, and electronics. Notably, deciphering substituent effects for these type of dyotropic reactions is known to be difficult [87], and would require more extensive and systematic analysis and corroboration which is outside the scope of this work.

The final benchmark of Tartarus to be discussed is the model timing comparison. This is important as it can provide information about whether improved optimization performances originate from the utilization of more computational resources and longer training times. It should be noted that the absolute computation times reported are not useful as they strongly depend on the computer hardware available. Hence, we recommend them to be used as a relative rather than an absolute measure. Thus, when new models are being tested for timing, some of the molecular design algorithms provided here as a reference should be run again with the same hardware as internal standards permitting a relative timing comparison to all the other algorithms. Looking at the timing results, apart from the VAEs, we consider training to be still affordable for the deep generative models tested, even without the availability of GPUs. However, without GPUs, training the VAEs is computationally too expensive and is detrimental to its widespread application, and even for the LSTM-HC models it is comparably cumbersome. Additionally, while molecule sampling time, taking roughly on the order of a few minutes in most cases, is generally reasonable, we believe that there is considerable room for further improvement. This is particularly evident from the extremely good timing performance of REINVENT taking less than one minute for that step. Nevertheless, as it is more important to propose several meaningful structures than to generate a large number of structures fast, we believe that the sampling times of the all the algorithms tested are sufficient for practical purposes.

S5. COMPUTATIONAL DETAILS

A. Molecular Design Algorithms

1. VAE

We implemented a variational autoencoder (VAE) architecture from the literature [10]. In particular, the encoder used three 1D convolutional layers with filter sizes of 9, 9, and 10 convolution kernels, followed by one fully connected layer that encodes a latent space of size 292. The decoder fed into three layers of gated recurrent unit (GRU) networks [26] with a hidden dimension of 501 each. Based on the respective training dataset to be used and the encoding used for representing the molecules (i.e., SMILES or SELFIES), we modified the alphabet size for converting the list of molecules into one-hot encodings. Additionally, the string length of the largest molecule from the dataset is used to pad sequences to the same length. Based on the dataset, these two parameters lead to minor modifications in the sizes of the first and last layers of the encoder and decoder, respectively. For each dataset, the models are trained using 24 CPUs and a single GPU, with training times ranging between 2 and 10 hours. Generally, 80% of the reference molecules are used for training, while the remaining ones are used for testing. We manually optimize the learning rate, learning rate decay rate, number of epochs for training, and the batch size based on performance on the test set. After obtaining a fully trained model, we perform structure optimization. The best known molecule for a given task, i.e., the one with the most favorable score, is converted into a one-hot encoding and passed through the encoder to the latent space. Subsequently, Gaussian noise is added to the latent space vector to produce a population of vectors. These vectors are passed through the decoder to produce molecules that are evaluated via the respective objective function. The best molecule resulting from this population is then used to repeat this procedure.

2. JANUS

Throughout the benchmarks, we used the default implementation of JANUS (`https://github.com/aspuru-guzik-group/JANUS`). We used version 1.0.3 of SELFIES, and the default valence constraints implemented in that version of SELFIES were employed. Additionally, no structure filters were employed in the genetic operators. For initiation, the file `sample_start_smiles.txt` was populated with the 1,000 best molecules from the corresponding benchmark reference dataset. The `generation_size` parameters and the other parameters that influence the generations are modified based on the respective benchmark task. An artificial neural network (ANN) classifier is used for additional selection pressure for all the tasks (cf. JANUS+C in the original publication [19]).

3. GB-GA

We employed of the implementation provided by Jensen [20] in the following repository: `https://github.com/jensengroup/GB_GA/`. For initiation, the file `ZINC_first_1000.smi` was populated with the 1,000 best molecules from the corresponding benchmark reference dataset for pre-conditioning to derive both mutation and crossover probabilities. The file `scoring_functions.py` was modified each time to incorporate the corresponding fitness functions from different benchmark tasks.

4. LSTM-HC

The long short-term memory hill climbing (LSTM-HC) approach uses a pre-trained long short-term memory (LSTM) [15] recurrent neural network (RNN) to generate sequences of molecular strings [16], and can be implemented with either SELFIES or SMILES. For each molecule, the string representation characters are transformed into one-hot encoded vectors, which are padded up to the length of the longest string in the respective reference dataset for each benchmark task. Similar to how it was implemented in the literature [16], the model consisted of 3 stacked LSTMs, each with 1024 hidden dimensions and a dropout ratio of 0.2, followed by a linear layer that outputs the logits of the next predicted character in the string sequence. That way, the output returns the probability for observing a certain subsequent character, given the preceding characters in the sequence. The model was pre-trained with a batch size of 128, using the ADAM optimizer [88] with a learning rate of 0.001. For optimization in each of the benchmark tasks, the top 2 known molecules with the best fitness values are used to create seeds for the LSTM structure generation. For SMILES, the strings are truncated randomly at 25% to 75% of the size of the original sequence. For SELFIES, the strings are first randomized 5 times using reordered SMILES that represent the same molecule before they are being truncated. Subsequently, the truncated strings are passed into the model and the next characters are generated by sampling from the probability distribution that is generated by the LSTM output. Each iteration, the model generates 500 new molecules for evaluation, and the top 2 molecules are selected for the subsequent iteration. Additionally, the model is retrained after each iteration on the set of new molecules for 10 epochs at a lower learning rate of 0.0001.

5. MoFlow

We employed the MoFlow framework developed by Zang et al. [89] for our molecular structure generation tasks. The implementation we used is publicly accessible at `https://github.com/calvin-zcx/moflow`. We trained the model on 80% of the respective dataset, using a batch size of 256 for a maximum of 200 epochs. All model parameters were selected based on the example provided in the GitHub repository. Specifically, the B-Block of the model utilized 10 flows and had 512 hidden channels in both layers. In the A-Block, the model was configured with 38 flows and had 256 hidden graph neural network channels, along with hidden linear layers of sizes 512 and 64. Masking parameters were set with both a row size list and a row stride list of 1. A noise scale of 0.6 was employed to introduce a stochastic element into the behavior of the model. To optimize properties, we implemented a hill-climbing-based algorithm. In this approach, the model was randomly sampled, and the top 250 molecules were selected from a total of 500 oracle calls for subsequent training. This cycle was repeated for up to 10 iterations, depending on the benchmarking task.

6. REINVENT

In REINVENT [18], similar to LSTM-HC, an LSTM that is pretrained on a reference dataset is used as a prior for generating new SMILES sequences. Reinforcement learning (RL) is used to optimize for a particular fitness value. The LSTM model acts as an agent, where the addition of a character is an action, and the output probability distribution is the action probability, or the policy. The action probability is augmented by a scoring function $S(m) \in [-1, 1]$, with larger scores corresponding to more desirable fitness values. SMILES that do not correspond to valid molecules are assigned a score of 0. The algorithm updates the policy to increase the expected fitness while still keeping the learned conditional sequence probabilities to produce valid SMILES that resemble the strucutural distribution of the reference dataset. The scoring function is a modified sigmoid function parameterized by the known fitness values from previous iterations. Let the set of known fitness values be F. The scoring function for a molecule m is defined as

$$S(m) = \frac{2}{1 + e^{-b(f-a)}} - 1, \tag{2}$$

where a is the average of F, and b is the slope of the sigmoid. This sigmoid is defined as

$$b = -\frac{1}{max(F) - a} \ln\left(\frac{2}{c+1} - 1\right), \tag{3}$$

where c is a threshold value set to 0.8. This means for the current fitness values in F that the best known fitness will map to 0.8 in the scoring function. The smaller the threshold c, the stronger the reward for fitness values larger than the best one contained in F. Further details about the approach and model parameters are provided in the original paper [18].

B. Benchmarks

1. Design of Organic Photovoltaics

a. Workflow To set up the property prediction workflow, we first implemented the Scharber model [41]. To do that, we constructed a simplified model to predict the short-circuit current density (J_{SC}). The canonical way to do that is via definite integration of the spectrum of the light source over the energy range of the absorbed light of the simulated device [90]. To bypass having to perform the integration in the property simulation workflow and provide the Reference Air Mass 1.5 Spectra (AM1.5G) as part of TARTARUS, we created a regression model for the short-circuit current density as a function of the band gap energy E_G. We used the AM1.5G spectrum that accounts for both direct and circumsolar irradiation and performed the integration via the trapezoidal rule as implemented in the `numpy` package. We found that a simple Gaussian function of the form

$$J_{SC} = A \cdot e^{-\frac{{E_G}^2}{B}}, \tag{4}$$

where A and B are fitting parameters, provides an excellent fit to the actual integration as demonstrated in the comparison between the original relationship derived from integration of the AM1.5G reference data (denoted "Original") and the fitted function (denoted "Fit") in Figure S6. Importantly, the regression model assumes an external quantum efficiency of 0.65, which is typically used in the Scharber model [41]. Typically, the band gap energy is approximated by the gap between the HOMO of the donor and the LUMO of the acceptor, which then, in conjunction with the regression model just described, allows to estimate the short-circuit current density that is needed to compute the power conversion efficiency (PCE) under the assumptions of the Scharber model [41].

Besides the computation of the short-circuit current density, the procedure described in the original paper was followed [41]. The open circuit voltage (V_{OC}) is determined as energy difference between HOMO of the donor and the LUMO of the acceptor minus 0.3 eV of overpotential that is typically assumed to be required for any charge separation to take place [90]. Should the open circuit voltage be formally estimated to be negative then it is assumed to be 0. In case the LUMO of the donor is less than 0.3 eV higher in energy than the LUMO of the acceptor then the open circuit voltage is also assumed to be 0. The power conversion efficiency is estimated based on the equation

$$PCE = 100\% \cdot V_{OC} \cdot FF \cdot \frac{J_{SC}}{P_{in}}, \tag{5}$$

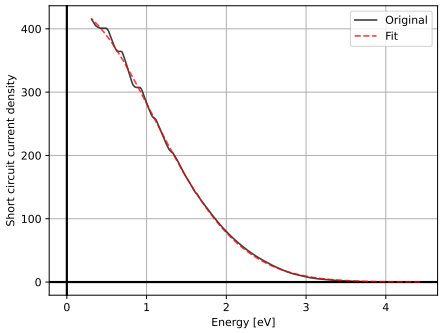

FIG. S6. Short circuit current density as a function of the band gap energy of the light-absorbing material in an organic solar cell. The original dependency results from the integration of the Reference Air Mass 1.5 Spectra (AM1.5G). The fit depicts the result of regression with a Gaussian function with two fitting parameters.

where V_{OC} is the open circuit voltage , FF is the fill factor, J_{SC} is the short-circuit current density, and P_{in} is the incident light intensity obtained by integration over the entire AM1.5G reference spectrum and therefore is constant. In the Scharber model, the fill factor is typically assumed to be optimizable to reach a value of 0.65 [91], which is thus inserted into this equation. For the first task of the organic photovoltaic design benchmarks, the goal is to design a small organic donor molecule used with [6,6]-phenyl-C61-butyric acid methyl ester (PCBM) as acceptor. The LUMO energy level of PCBM is -4.3 eV and all light is assumed to be absorbed by the donor molecule. Notably, this is a common design objective that has been pursued in the OPV literature [41, 90] and was also implemented in the Harvard Clean Energy Project (CEP) [92, 93]. For the second task, a small organic acceptor molecule is to be designed that is used with poly[N-90-heptadecanyl-2,7-carbazole-alt-5,5-(40,70-di-2-thienyl-20,10,30-benzothiadiazole)] (PCDTBT) as a donor. The HOMO energy level of PCDTBT is -5.5 eV and all light is assumed to be absorbed by the acceptor molecule. This benchmark task is based on a high-throughput virtual screening that was conducted in the literature to find new non-fullerene acceptors [94].

The computational property prediction workflow accepts a SMILES string as input and uses `Open Babel` [95] to generate initial guess Cartesian coordinates. Next, `crest` [96] is used for conformer search of the molecule via the iMTD-GC workflow at the GFN-FF level of theory [97] using the additional keywords `-mquick`, and `-noreftopo`. Subsequently, the lowest energy conformer resulting from this search is used for a geometry optimization at the GFN2-xTB level of theory [98]. The properties of interest, i.e. HOMO energy, LUMO energy, HOMO-LUMO gap and molecular dipole moment, are obtained from the minimized structure at the same level of theory.

Importantly, to obtain better PCE estimates, we calibrated the HOMO and LUMO energies obtained from GFN2-xTB simulations against the corresponding density functional approximation (DFA) results taken from the Harvard Clean Energy Project Database (CEPDB) for the reference molecules of the CEPDB subset. In particular, the median HOMO and LUMO energies of all the DFA results contained in the CEPDB were used. Both the HOMO and the LUMO energies were calibrated against the respective medians of the DFA HOMO and LUMO energies using a Theil-Sen estimator [99, 100], as implemented in the python package `scikit-learn` [101], due to its robustness with respect to outliers. The maximum number of subpopulations set to 250,000. The linear regression functions derived are illustrated in Figure S7 and are provided in the following equations:

$$E_{HOMO,calibrated} = E_{HOMO,GFN2-xTB} \cdot 0.8051 + 2.5377\ eV \tag{6}$$

$$E_{LUMO,calibrated} = E_{LUMO,GFN2-xTB} \cdot 0.8788 + 3.7913\ eV \tag{7}$$

As can be seen in Figure S7, the assumption of a linear relationship between the energies at the GFN2-xTB level of theory and at the DFA level of theory is reasonable.

b. Dataset The CEPDB subset used for training all the generative models was obtained from the following repository: `http://github.com/HIPS/neural-fingerprint`. For all molecules in this dataset, we simulated the corresponding properties of interest with the developed property simulation workflow (cf. Figure S8).

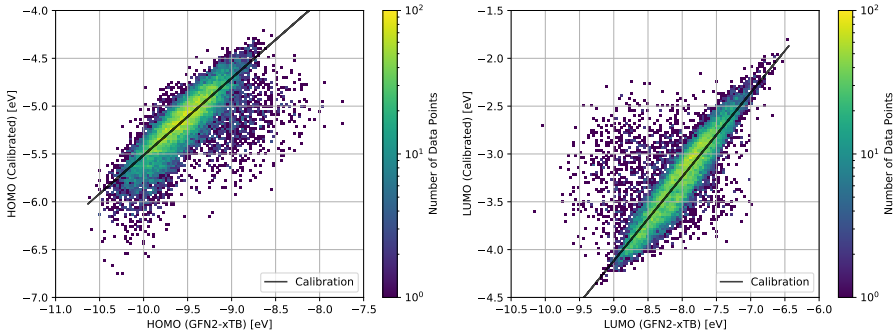

FIG. S7. Two-dimensional histograms depicting the comparison of HOMO (left) and LUMO (right) energies obtained using DFA computations taken from the CEPDB against the corresponding energies at the GFN2-xTB level of theory obtained using the property simulation workflow developed for the design of organic photovoltaics benchmarks. The black lines were obtained via Theil-Sen estimators for linear regression.

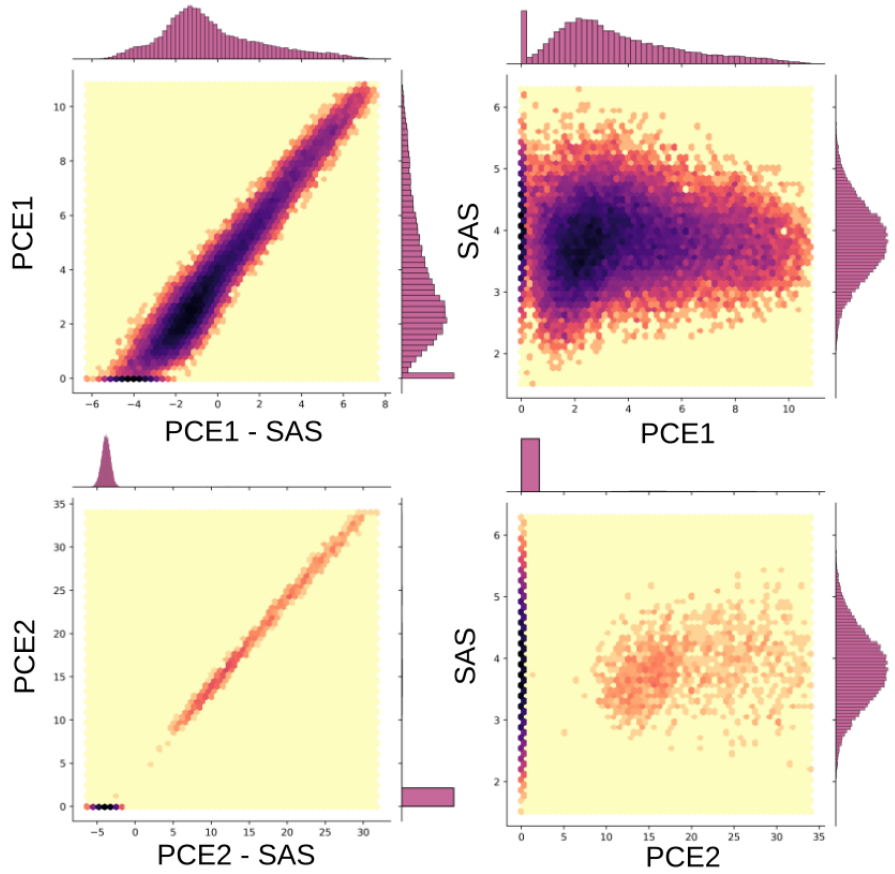

FIG. S8. Two-dimensional histograms depicting the property distributions of the provided reference dataset as obtained from the developed simulation workflow for the design of organic photovoltaics benchmark set. Top left: multiobjective target function $PCE_{PCBM} - SAscore$ against power conversion efficiency with PCBM as acceptor PCE_{PCBM}; top right: synthetic accessibility metric $SAscore$ against power conversion efficiency with PCBM as acceptor PCE_{PCBM}; bottom left: multiobjective target function $PCE_{PCDTBT} - SAscore$ against power conversion efficiency with PCDTBT as donor PCE_{PCDTBT}; bottom right: synthetic accessibility metric $SAscore$ against power conversion efficiency with PCDTBT as donor PCE_{PCDTBT}.

2. *Design of Organic Emitters*

a. Workflow The computational property prediction workflow accepts a SMILES string as input and uses both `Open Babel` [95] and `RDKit` [102] to generate initial guess Cartesian coordinates. Whichever guess coordinates have a

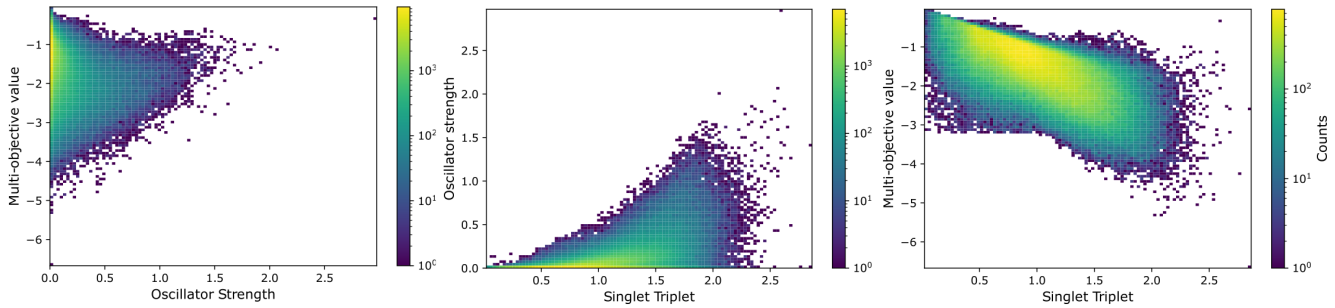

FIG. S9. Two-dimensional histograms depicting the property distributions of the provided reference dataset as obtained from the developed simulation workflow for the design of organic emitters benchmark set. Left: multiobjective target function $+f_{12} - \Delta E(S_1\text{-}T_1) - |\Delta E(S_0\text{-}S_1) - 3.2\ eV|$ against oscillator strength for the transition between S_1 and S_0, f_{12}; middle: multiobjective target function $+f_{12} - \Delta E(S_1\text{-}T_1) - |\Delta E(S_0\text{-}S_1) - 3.2\ eV|$ against singlet-triplet gap, $\Delta E(S_1 - T_1)$; right: oscillator strength for the transition between S_1 and S_0, f_{12} against singlet-triplet gap, $\Delta E(S_1 - T_1)$.

lower in energy at the GFN-FF level of theory are used as starting point for the subsequent conformer search of the molecule via the iMTD-GC workflow at the GFN-FF level of theory [97] using the additional keywords `-mquick`, and `-noreftopo`. Subsequently, the lowest energy conformer resulting from this search is used for a geometry optimization at the GFN0-xTB level of theory [103–105]. Afterwards, the optimized geometry is used to perform a TD-DFT single point calculation at the B3LYP/6-31G* level of theory [106–111] using the `PySCF` python package (version 1.7.6) [112–114]. Density fitting is utilized via the `def2-universal-jkfit` auxiliary basis set. For the excited state simulation, two excited states are generated for both the singlet and the triplet manifold. Notably, in case the molecule contains iodine, the LANL2DZ basis set [115] is used for that atom. From the corresponding results, the energies of the lowest excited singlet and triplet states, and the oscillator strength for the transition between the ground state and the lowest excited singlet state are obtained. The singlet-triplet gap is derived as energy difference between the first excited singlet and the first excited triplet state. Notably, only molecules with an SAscore below or equal to 4.5 are subjected to the property simulation workflow. All molecules above are assigned a very unfavorable fitness.

b. Dataset We developed a comprehensive set of filters to define a subset of GDB-13 [116], originally comprising more than 970 million organic molecules, with over 400,000 structures possessing cycles and a high degree of conjugation, that we herein refer to as GDB-13_{SUB}. These filters ensure that the molecules encompass cyclic π-systems to increase representation of the relevant structural space for potential organic emitters. They were implemented via RDKit [102] and are summarized in Table S6. For the forbidden substructures, we utilized the following SMARTS patterns:

```
[Cl,Br,I], *=*=*, *#*, [O,o,S,s]~[O,o,S,s], [N,n,O,o,S,s]~[N,n,O,o,S,s]~[N,n,O,o,S,s], [C,c]~N=,
:[O,o,S,s;!R], [N,n,O,o,S,s]~[N,n,O,o,S,s]~[C,c]=,:[O,o,S,s,N,n;!R], *=[NH], *=N-[*;!R],
*~[N,n,O,o,S,s]-[N,n,O,o,S,s;!R], *-[CH1]-*, *-[CH2]-*, *-[CH3]
```

TABLE S6. List of filters employed to create the π-systems subset of GDB-13, GDB-13_{SUB}. In each line, x denotes the value of the corresponding feature.

Number	Feature	Definition	Value
1	Charge	Charge of the molecule.	$x = 0$
2	Radicals	Number of radical electrons.	$x = 0$
3	Bridgehead Atoms	Number of bridgehead atoms.	$x = 0$
4	Spiro Atoms	Number of spiro atoms.	$x = 0$
5	Aromaticity Degree	Percentage of aromatic non-hydrogen atoms.	$x \geq 0.5$
6	Conjugation Degree	Percentage of conjugated bonds between non-hydrogen atoms.	$x \geq 0.7$
7	Maximum Ring Size	Size of the largest ring.	$4 \leq x \leq 8$
8	Minimum Ring Size	Size of the smallest ring.	$4 \leq x \leq 8$
9	Substructures	Presence of any forbidden substructures (see text).	FALSE

After filtering the GDB-13 dataset with these filters, and removing all molecules with an SAscore larger than 4.5, we subjected the resulting molecules to the property simulation workflow described above to obtain the reference dataset for the design of organic emitters benchmarks. The corresponding property distributions are visualized in Figure S9.

3. *Design of Protein Ligands*

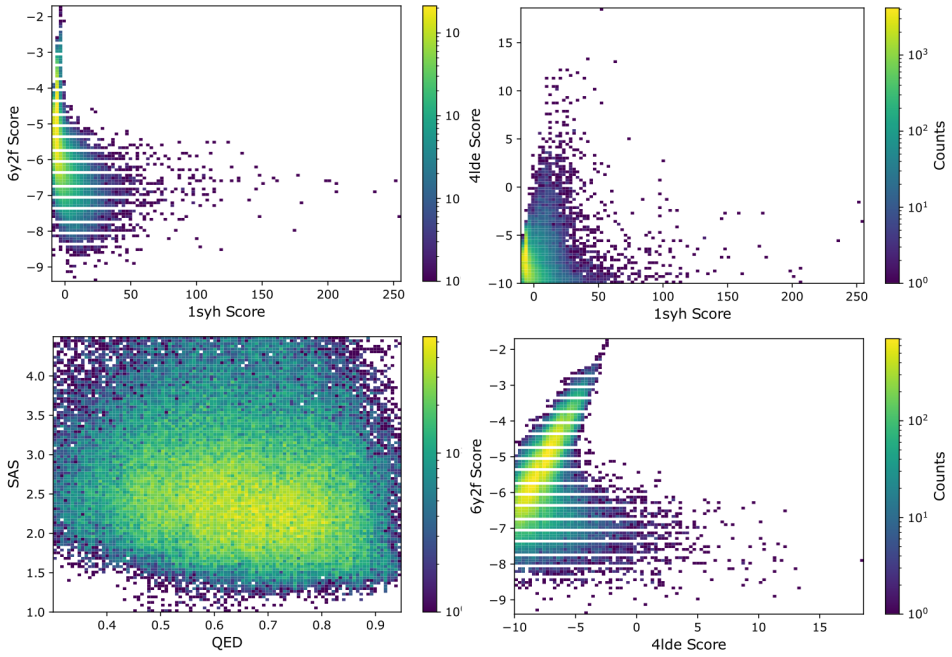

FIG. S10. Two-dimensional histograms depicting the property distributions of the provided reference dataset as obtained from the developed simulation workflow for the design of protein ligands benchmark set. Top left: docking score to 4LDE ΔE_{4LDE} against docking score to 1SYH ΔE_{1SYH}; top right: docking score to 6Y2F ΔE_{6Y2F} against docking score to 1SYH ΔE_{1SYH}; bottom left: docking score to 6Y2F ΔE_{6Y2F} against docking score to 4LDE ΔE_{4LDE}; bottom right: synthetic accessibility metric $SAscore$ against quantitative estimate of drug-likeness QED.

a. Workflow To set up the computational property prediction workflow with docking calculations, we downloaded crystal structures for the three proteins of interest from the Protein Data Bank (PDB). The corresponding PDB codes are 1SYH, 4LDE, and 6Y2F. The corresponding structures were all co-crystallized with a ligand bound to the protein. All these structures were processed using the protein preparation wizard [117] as implemented in the computer program Schrödinger (version 12.9). We created a box around the bound ligand for each of the protein structures based on re-docking accuracy. Subsequently, the native ligand was removed from the structure providing the starting setup for new ligands to be bound in the property simulation workflow. This workflow accepts a SMILES string as input, converts it into an initial 3D guess geometry using `Open Babel` [118] and exports it as a `PDBQT` file. Subsequently, the quality of the molecules is checked by running `obenergy`. During reading of the SMILES, the corresponding structure is inspected for the presence of the following SMARTS patterns:

```
*1=**=*1, *1*=*=*1, *1~*=*1, [F,Cl,Br]C=[O,S,N], [Br]-C-C=[O,S,N], [N,n,S,s,O,o]C[F,Cl,Br],
[I], [S&X3], [S&X5], [S&X6], [B,N,n,O,S]~[F,Cl,Br,I], *=*=*=*, *=[NH], [P,p]~[F,Cl,Br], SS,
C#C, C=C=C, *~[P,p](=O)~*, C=C=N, NNN, "[*;R1]1~[*]~[*]~[*]1", OOO,
[#8]1-[#6]2[#8][#6][#8][#6]12, N=C=O, C1CN1, [#6](=[#8])[F,Cl,Br,I],
[#6](=[#8])=[#6](-[#8])-[#6](=[#8])~[#8], N(-[#6])=[#7]-[#8].
```

These SMARTS patterns were developed from our experience with the use of docking in conjunction with inverse molecular design algorithms. They correspond to reactive structural moieties that can lead to very favorable docking scores and should therefore be explicitly avoided [19]. Additionally, it is also tested whether the structure is charged, possesses radical electrons, contains more than two bridgehead atoms, has at least one ring with more than 8 members

and violates Lipinski's Rule of Five [59]. If at least one of these initial checks is true, the property simulation workflow is aborted and a very bad fitness of 10,000 is returned. If all of these checks are false, the workflow is continued. Subsequently, `smina` is used to add the ligand to the prepared protein structure of interest and perform a docking simulation with an exhaustiveness setting of 10. The docked ligand structure with the lowest docking score is selected and the corresponding docking score returned by the simulation workflow. Following recent literature,[119, 120] the quality of the generated docked structure was checked using `obenergy` and only values below 10,000 were accepted.

b. Dataset We obtained a list of approximately 250,000 canonical SMILES from PubChem [121] that are part of the DTP Open Compound Collection [122, 123]. Next, all the structures that do not pass any of the structural filters described in the previous section are removed. Thus, we obtained a set of 152,296 molecules that are provided as reference dataset for the protein ligand design benchmark set and used it for training the generative models. For all molecules in this dataset, we simulated the corresponding properties of interest with the developed property simulation workflow (cf. Figure S10).

4. Design of Chemical Reaction Substrates

a. Workflow The double hydrogen transfer in *syn*-sesquinorbornenes (cf. Figure S11A) was chosen as a suitable benchmark reaction due to its robust behavior in simulations. Additionaly, it is well-described by the SEAM method as shown previously by Jensen [76]. For the benchmark, we allow substitutions at the positions indicated by **X** or **R** in Figure S11B.

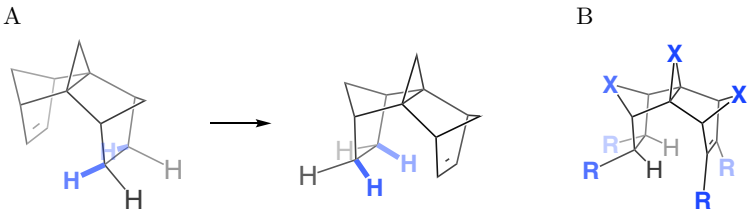

FIG. S11. Intramolecular concerted double hydrogen transfer reaction of *syn*-sesquinorbornenes (A) and positions in the *syn*-sesquinorbornene core structure allowed to be substituted (B).

The general procedure of the reactivity workflow we developed is outlined in the following steps:

1. Generation of initial conformation of the reactant.
   (a) Try to generate coordinates with `Open Babel` [95].
   (b) In case of failure, generate coordinates with `RDKit` [102].
2. MMFF [124] optimization of the reactant.
3. `crest` [96] conformer search for the reactant via the iMTD-GC workflow.
4. Constrained MMFF optimization going from the reactant to the product in conformation consistent with reactant.
5. Full MMFF optimization of the product.
6. `crest` conformer search for the product lowest conformation via the iMTD-GC workflow.
7. Crude reaction path interpolation with the geodesic interpolation method [125].
8. SEAM [76] optimization of the TS.
9. Constrained `crest` conformer search of the TS via the iMTD-GC workflow.
10. SEAM re-optimization of the lowest energy TS conformer.

11. Calculation of reaction and activation energy from the SEAM model.

Prior to performing any property simulations, the molecule is checked for the presence of the *syn*-sesquinorbornene core motif which is strictly required for the reaction of interest to occur:

```
[H][C@@]1(*)[C@;R2](*)2[C@@]34[C@;R2]5(*)[C;R1](*)=[C;R1](*)[C@;R2](*)([*;R2]5)[C@@]3([*;R1]4)
[C@](*)([*;R2]2)[C@@;R1]1([*])[H]
```

Additionally, the input structure is inspected to ensure the absence of several reactive structural moieties as described by the following SMARTS patterns:

```
[C-], [S-], [O-], [N-], ‘[*+], [*-] [PH], [pH], [N&X5], *=[S,s;!R], [S&X3], [S&X4],
[S&X5], [S&X6], [P,p], [B,b,N,n,O,o,S,s]~[F,Cl,Br,I], *=*=*, *#*, [O,o,S,s]~[O,o,S,s],
[N,n,O,o,S,s]~[N,n,O,o,S,s]~[N,n,O,o,S,s], [N,n,O,o,S,s]~[N,n,O,o,S,s]~[C,c]=,
:[O,o,S,s,N,n;!R], *=N-[*;!R], *~[N,n,O,o,S,s]-[N,n,O,o,S,s;!R]
```

A detailed description of some of the steps in the property simulation workflow for the design of chemical reaction substrates is provided in the following bullet points:

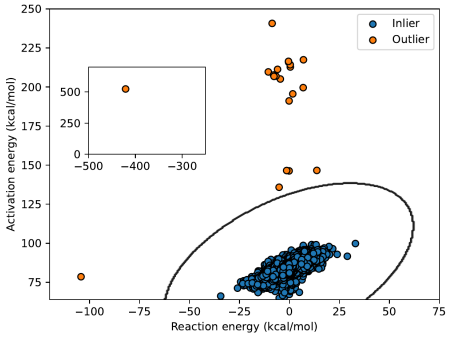

FIG. S12. Illustration of the outlier detection cutoff for the elliptic envelope used for the property simulation workflow of the design of chemical reaction substrates benchmark set. Data points appearing outside the black line are regarded as outliers and assigned a very low fitness.

1. SEAM optimization of the TS:
   The SEAM model optimizes the transition state as the crossing point between reactant and product force fields. In this work, we used the GFN-FF force field [97], and shifted the product force field to reproduce the energy difference between the reactants and products at the GFN2-xTB level of theory [98]. For the product force field shift, we used the lowest energy conformers of the reactant and product, respectively. The optimization was carried out on the upper SEAM surface by introducing a small coupling of 0.001 eV between the states. The guess structure for the TS was taken from a geodesic interpolation between reactant and product geometries [125]. The TS itself was optimized with the geomeTRIC package [126] via an interface to `pyscf` [112–114]. The subsequent `crest` transition state conformer search used constrained C—H—C bond lengths with a force constant of 1.0 a.u. in strength. All calculations were carried out using the POLANYI package, which will be reported in detail elsewhere. The activation energy was taken as the difference of the GFN-FF energy of the optimized TS structure and the energy of the lowest energy conformer of the reactant. The reaction energy was taken as the energy difference between the lowest energy conformers of the product and reactant, including the energy shift of the product force field. These energies are electronic energies from the force field. While free energies would be preferable, it would increase the complexity of the workflow, and significant error cancellation is expected for this intramolecular reaction.

2. `crest` conformer searches:
   We opted to perform the `crest` conformer searches with the keywords `-mquick`, `-gfn2//gfnff`, and `-noreftopo`. Although it would be more consistent to use `-gfnff`, we found that it produced a larger number of outliers in preliminary test calculations.

3. Outlier detection:
   Although the workflow is very robust, it does produce some outliers due to failures in the conformational search

or the TS optimization. In the worst cases, these outliers could display extreme properties such as artificially low or high reaction or activation energies, and, hence, sometimes emerge as the best performering structures. One solution is increasing the exhaustiveness of the conformational sampling, but this also leads to significantly longer computational run-times. Instead, we opted for an outlier detection model based on the computed reaction and activation energies. According to the Bell-Evans-Polanyi relationship [80], the reaction energy is expected to correlate with the activation energy. Therefore, an outlier model based on the 2D scatter plot of these two properties can be used to identify potential outliers. We used the `EllipticEnvelope` from scikit-learn [127]. The `contamination` hyperparameter was set to 0.00035 by visual inspection to allow for a considerable deviation from the correlation observed in the reference dataset, while simultaneously capturing clear outliers. This outlier detection scheme is illustrated in Figure S12. Accordingly, we incorporated this outlier detection into the property simulation workflow and assigned an extremely low fitness of -10^4 to outlier molecules.

b. Dataset We constructed a dataset of 60,850 molecules by performing multiple mutations of various SELFIES representations of the parent *syn*-sesquinorbornene. Specifically, the starting structure was used to produce 20 randomly reordered SMILES strings. The resulting SMILES were converted to SELFIES and 20 random mutations were conducted via STONED-SELFIES [31]. All mutants were subsequently converted to SMILES and checked for the presence of the core motif. Mutations that lead to structures without the core motif were discarded. Additionally, molecules that did not pass the filters were also removed. This procedure was repeated on all unique and valid mutated structures until the final number of molecules was reached. Notably, we used the default mutation settings of STONED-SELFIES [31]. When running the dataset through the property simulation workflow, only 35 molecules either did not run through the property simulation workflow properly or were outliers, corresponding to a success rate of 99.94%. The results of running the reference structures through the property simulation workflow with respect to both property distributions and computational run time are illustrated in Figure S13.

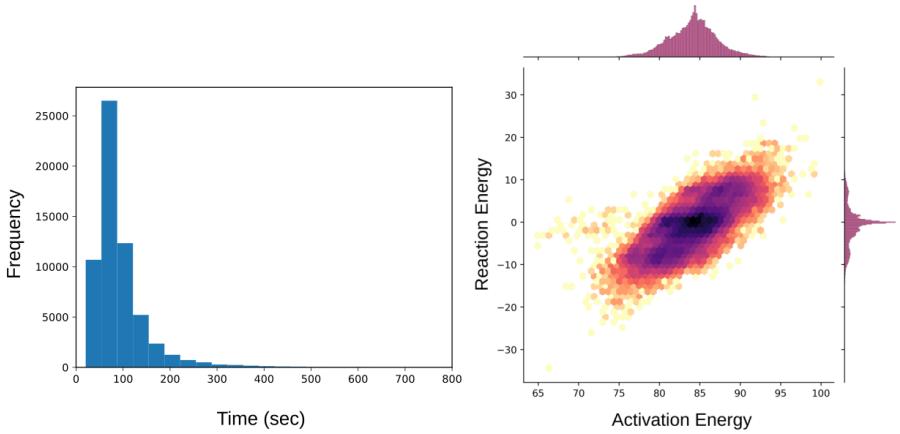

FIG. S13. Simulation results for the reference structures in the developed simulation workflow for the design of chemical reaction substrates benchmark set. Left: Histogram of the simulation workflow run times; right: two-dimensional histograms depicting the property distributions of the provided reference dataset: reaction energy ΔE_r against activation energy $\Delta E^{\ddagger}$.

S6. MODEL TIMING COMPARISON

All deep generative models (SMILES-VAE, SELFIES-VAE, MoFlow, REINVENT, SMILES-LSTM-HC and SELFIES-LSTM-HC) were trained using the first 80% of the molecules from the protein ligand design tasks. The remaining 20% of the dataset was used to assess convergence and perform manual hyperparameter optimization. The corresponding figure in the main text provides four timing benchmark metrics: (1) Training Time with GPU, (2) Sample time for 10,000 molecules, (3) Epoch time without GPU and (4) Epoch time with GPU. For each metric, 5 independent experiments were conducted to obtain both average and standard deviations of the respective values. For the first metric, all the generative models were trained using 24 CPUs (AMD Rome 7532 2.40 GHz 256M cache L3) and a single GPU (Tesla A100). For REINVENT, MoFlow, and the LSTM-HC models, a batch size of 512 was used for training. In contrast, the VAE models performed better in terms of validity and reconstruction with a smaller batch size of 8. For sampling times of 10,000 molecules, we used fully trained deep learning models as obtained after the training time measurements and determined the time required for generating 10,000 molecules in the presence

of 24 CPUs (AMD Rome 7532 @ 2.40 GHz 256M cache L3) and a single GPU (Tesla A100). The batch size was 512 during sampling for all deep generative models. For GAs, the molecule sampling time was measured only in the presence of 24 CPUs as both JANUS and GB-GA lack GPU support. Notably, in the models that use SELFIES as a molecular representation, the time needed for translating SELFIES to SMILES was included in the final reported time. For JANUS and GB-GA, this metric was obtained by increasing the population size to 10,000 and by using single generation runs and a dummy fitness function that returns a value of 1 for any molecule. For the single epoch times, we report the time required for training models in the presence of either 24 CPUs (AMD Rome 7532 @ 2.40 GHz 256M cache L3) or 24 CPUs (AMD Rome 7532 @ 2.40 GHz 256M cache L3) with a GPU (Tesla A100). For the deep generative models, a batch size of 512 was used. Notably, for both GB-GA and JANUS, the CPU epoch time corresponds to the total time the models need to precondition the genetic operators based on a dataset of reference structures. Additionally, for both these models, preconditioning was performed by providing 1,000 random SMILES from the reference dataset of the protein ligand design task to the molecular design algorithms. Finally, using the CPU epoch times and the number of epochs needed for training when also using a GPU, we estimated the total CPU training times for all the deep generative models by multiplying these two numbers. Notably, we decided not to estimate the corresponding estimated errors via error propagation as these values are merely rough estimates.

[1] Vikas Nanda and Ronald L Koder. Designing artificial enzymes by intuition and computation. *Nature chemistry*, 2(1):15–24, 2010.

[2] 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(1):195–216, 2015.

[3] Michel Gendreau and Potvin Jean-Yves. *Handbook of metaheuristics*. Springer Vol. 2, 2010.

[4] Christian Blum and Andrea Roli. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Comput. Surv.*, 35(3):268–308, 2003.

[5] Gisbert Schneider and Uli Fechner. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 4(8):649–663, August 2005.

[6] Robert Pollice, Gabriel dos Passos Gomes, Matteo Aldeghi, Riley J Hickman, Mario Krenn, Cyrille Lavigne, Michael Lindner-D'Addario, AkshatKumar Nigam, Cher Tian Ser, Zhenpeng Yao, et al. Data-driven strategies for accelerated materials design. *Accounts of Chemical Research*, 54(4):849–860, 2021.

[7] Daniel Schwalbe-Koda and Rafael Gómez-Bombarelli. Generative models for automatic chemical design. In *Machine Learning Meets Quantum Physics*, pages 445–467. Springer, 2020.

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

[9] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[10] 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, Jan 2018.

[11] 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.

[12] Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.

[13] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[14] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):1–10, 2019.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] 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.

[17] 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.

[18] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):1–14, 2017.

[19] AkshatKumar Nigam, Robert Pollice, and Alán Aspuru-Guzik. Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. *Digital Discovery*, page Advance Article, 2022.

[20] 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.

[21] David Weininger, Arthur Weininger, and Joseph L. Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, 1989.

[22] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[23] Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, et al. Selfies and the future of molecular string representations. *arXiv preprint arXiv:2204.00056*, 2022.

[24] Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor W Coley. Sample efficiency matters: A benchmark for practical molecular optimization. *arXiv preprint arXiv:2206.12411*, 2022.

[25] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[26] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[27] Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. Language models can learn complex molecular distributions. *Nature Communications*, 13(1):1–10, 2022.

[28] John H Holland. Genetic algorithms. *Scientific american*, 267(1):66–73, 1992.

[29] Markus Hartenfeller and Gisbert Schneider. Enabling future drug discovery by de novo design. *WIREs Computational Molecular Science*, 1(5):742–759, 2011.

[30] Seyedali Mirjalili. Genetic algorithm. In *Evolutionary algorithms and neural networks*, pages 43–55. Springer, 2019.

[31] AkshatKumar Nigam, Robert Pollice, Mario Krenn, Gabriel dos Passos Gomes, and Alan Aspuru-Guzik. Beyond generative models: Superfast traversal, optimization, novelty, exploration and discovery (stoned) algorithm for molecules using selfies. *Chemical Science*, 2021.

[32] Christoph J Brabec, Jens A Hauch, Pavel Schilinsky, and Christoph Waldauf. Production aspects of organic photovoltaics and their impact on the commercialization of devices. *MRS bulletin*, 30(1):50–52, 2005.

[33] Alexander W. Hains, Ziqi Liang, Michael A. Woodhouse, and Brian A. Gregg. Molecular semiconductors in organic photovoltaic cells. *Chemical Reviews*, 110(11):6689–6735, 2010. PMID: 20184362.

[34] Hongseok Youn, Hui Joon Park, and L. Jay Guo. Organic photovoltaic cells: From performance improvement to manufacturing processes. *Small*, 11(19):2228–2246, 2015.

[35] Moritz Riede, Donato Spoltore, and Karl Leo. Organic solar cells—the path to commercial success. *Advanced Energy Materials*, 11(1):2002653, 2021.

[36] Lin X. Chen. Organic solar cells: Recent progress and challenges. *ACS Energy Letters*, 4(10):2537–2539, 2019.

[37] Rongming Xue, Jingwen Zhang, Yaowen Li, and Yongfang Li. Organic solar cell materials toward commercialization. *Small*, 14(41):1801793, 2018.

[38] Carsten Deibel, Vladimir Dyakonov, and Christoph J. Brabec. Organic bulk-heterojunction solar cells. *IEEE Journal of Selected Topics in Quantum Electronics*, 16(6):1517–1527, 2010.

[39] Rene AJ Janssen, Jan C Hummelen, and N Serdar Sariciftci. Polymer–fullerene bulk heterojunction solar cells. *MRS bulletin*, 30(1):33–36, 2005.

[40] Brian A Gregg. The photoconversion mechanism of excitonic solar cells. *MRS bulletin*, 30(1):20–22, 2005.

[41] M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger, and C. J. Brabec. Design rules for donors in bulk-heterojunction solar cells—towards 10 % energy-conversion efficiency. *Advanced Materials*, 18(6):789–794, 2006.

[42] Michael Y. Wong and Eli Zysman-Colman. Purely organic thermally activated delayed fluorescence materials for organic light-emitting diodes. *Advanced Materials*, 29(22):1605444, 2017.

[43] Zhiyong Yang, Zhu Mao, Zongliang Xie, Yi Zhang, Siwei Liu, Juan Zhao, Jiarui Xu, Zhenguo Chi, and Matthew P. Aldred. Recent advances in organic thermally activated delayed fluorescence materials. *Chem. Soc. Rev.*, 46:915–1016, 2017.

[44] Yuchao Liu, Chensen Li, Zhongjie Ren, Shouke Yan, and Martin R Bryce. All-organic thermally activated delayed fluorescence materials for organic light-emitting diodes. *Nature Reviews Materials*, 3(4):1–20, 2018.

[45] Hiroki Uoyama, Kenichi Goushi, Katsuyuki Shizu, Hiroko Nomura, and Chihaya Adachi. Highly efficient organic light-emitting diodes from delayed fluorescence. *Nature*, 492(7428):234–238, 2012.

[46] Kelly L. Damm-Ganamet, Richard D. Smith, James B. Dunbar, Jeanne A. Stuckey, and Heather A. Carlson. Csar benchmark exercise 2011–2012: Evaluation of results from docking and relative ranking of blinded congeneric series. *Journal of Chemical Information and Modeling*, 53(8):1853–1870, 2013.

[47] Yan Li, Zhihai Liu, Jie Li, Li Han, Jie Liu, Zhixiong Zhao, and Renxiao Wang. Comparative assessment of scoring functions on an updated benchmark: 1. compilation of the test set. *Journal of Chemical Information and Modeling*, 54(6):1700–1716, 2014.

[48] Yan Li, Li Han, Zhihai Liu, and Renxiao Wang. Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *Journal of Chemical Information and Modeling*, 54(6):1717–1736, 2014.

[49] Zhe Wang, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.*, 18:12964–12975, 2016.

[50] Zoe Cournia, Bryce K. Allen, Thijs Beuming, David A. Pearlman, Brian K. Radak, and Woody Sherman. Rigorous free energy simulations in virtual screening. *Journal of Chemical Information and Modeling*, 60(9):4153–4169, 2020.

[51] Tony Slater and Dave Timms. Meeting on binding sites: Characterizing and satisfying steric and chemical restraints. University of York, 28–30 March 1993. *Journal of Molecular Graphics*, 11(4):248–251, December 1993.

[52] 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. *arXiv*, page 1610.02415v1, 2016.

[53] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2323–2332. PMLR, 10–15 Jul 2018.

[54] 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.

[55] Miguel García-Ortegón, Gregor N. C. Simm, Austin J. Tripp, José Miguel Hernández-Lobato, Andreas Bender, and Sergio Bacallado. Dockstring: Easy molecular docking yields better benchmarks for ligand design. *Journal of Chemical Information and Modeling*, 0(0):null, 2022.

[56] 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.

[57] 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. *arXiv*, page 1610.02415v2, 2017.

[58] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In *International Conference on Machine Learning*, pages 4849–4859. PMLR, 2020.

[59] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

[60] Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021.

[61] Milan Voršilák, Michal Kolář, Ivan Čmelo, and Daniel Svozil. Syba: Bayesian estimation of synthetic accessibility of organic compounds. *Journal of cheminformatics*, 12(1):1–13, 2020.

[62] K. N. Houk and Fang Liu. Holy grails for computational organic chemistry and biochemistry. *Accounts of Chemical Research*, 50(3):539–543, 2017.

[63] Carl Poree and Franziska Schoenebeck. A holy grail in chemistry: Computational catalyst design: Feasible or fiction? *Accounts of Chemical Research*, 50(3):605–608, 2017.

[64] Sharon Hammes-Schiffer. Catalysts by design: The power of theory. *Accounts of Chemical Research*, 50(3):561–566, 2017.

[65] Todd J. Martínez. Ab initio reactive computer aided molecular design. *Accounts of Chemical Research*, 50(3):652–656, 2017.

[66] Keith J. Laidler and M. Christine King. Development of transition-state theory. *The Journal of Physical Chemistry*, 87(15):2657–2664, 1983.

[67] Donald G. Truhlar, Bruce C. Garrett, and Stephen J. Klippenstein. Current status of transition-state theory. *The Journal of Physical Chemistry*, 100(31):12771–12800, 1996.

[68] A. D. McNaught, A. Wilkinson, and S. J. Chalk. *IUPAC. Compendium of Chemical Terminology, (the "Gold Book"), Online Version.* Blackwell Scientific Publications, Oxford, second edition, 2019.

[69] Pascal Friederich, Gabriel dos Passos Gomes, Riccardo De Bin, Alán Aspuru-Guzik, and David Balcells. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chemical Science*, 11(18):4584–4601, 2020.

[70] Colin A. Grambow, Adeel Jamal, Yi-Pei Li, William H. Green, Judit Zádor, and Yury V. Suleimanov. Unimolecular reaction pathways of a γ-ketohydroperoxide from combined application of automated reaction discovery methods. *Journal of the American Chemical Society*, 140(3):1035–1048, 2018.

[71] Lagnajit Pattanaik, John B. Ingraham, Colin A. Grambow, and William H. Green. Generating transition states of isomerization reactions with deep learning. *Physical Chemistry Chemical Physics*, 22(41):23618–23626, 2020.

[72] Kjell Jorner, Tore Brinck, Per-Ola Norrby, and David Buttar. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.*, 12:1163–1175, 2021.

[73] Yunhan Chu, Wouter Heyndrickx, Giovanni Occhipinti, Vidar R. Jensen, and Bjørn K. Alsberg. An evolutionary algorithm for de novo optimization of functional transition metal compounds. *Journal of the American Chemical Society*, 134(21):8885–8895, May 2012.

[74] Ruben Laplaza, Simone Gallarati, and Clemence Corminboeuf. Genetic optimization of homogeneous catalysts. *Chemistry–Methods*, 2(6), June 2022.

[75] Julius Seumer and Jan H Jensen. Computational evolution of new catalysts for the morita–baylis–hillman reaction. *ChemRxiv*, 2022.

[76] Frank Jensen. Locating minima on seams of intersecting potential energy surfaces. An application to transition structure modeling. *Journal of the American Chemical Society*, 114(5):1596–1603, February 1992.

[77] J. N. Brønsted and K. J. Pedersen. Die katalytische zersetzung des nitramids und ihre physikalisch-chemische bedeutung. *Zeitschrift für Phys. Chemie, Stöchiometrie und Verwandtschaftslehre*, 108:185–235, 1924.

[78] Ronald Percy Bell. The theory of reactions involving proton transfers. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 154(882):414–429, 1936.

[79] M. G. Evans and M. Polanyi. Further considerations on the thermodynamics of chemical equilibria and reaction rates. *Trans. Faraday Soc.*, 32:1333–1360, 1936.

[80] Joseph R. Murdoch. A simple relationship between empirical theories for predicting barrier heights of electron-, proton-, atom-, and group-transfer reactions. *Journal of the American Chemical Society*, 105(8):2159–2164, April 1983.

[81] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1(1):8, 2009.

[82] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *International conference on machine learning*, pages 4839–4848. PMLR, 2020.

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

[84] AkshatKumar Nigam, Robert Pollice, Matthew F. D. Hurley, Riley J. Hickman, Matteo Aldeghi, Naruki Yoshikawa, Seyone Chithrananda, Vincent A. Voelz, and Alán Aspuru-Guzik. Assigning confidence to molecular property prediction. *Expert Opinion on Drug Discovery*, 16(9):1009–1023, 2021.

[85] Thomas J. Penfold. On predicting the excited-state properties of thermally activated delayed fluorescence emitters. *The Journal of Physical Chemistry C*, 119(24):13535–13544, 2015.

[86] Youichi Tsuchiya, Keita Tsuji, Ko Inada, Fatima Bencheikh, Yan Geng, H. Shaun Kwak, Thomas J. L. Mustard, Mathew D. Halls, Hajime Nakanotani, and Chihaya Adachi. Molecular design based on donor-weak donor scaffold for blue thermally-activated delayed fluorescence designed by combinatorial dft calculations. *Frontiers in Chemistry*, 8, 2020.

[87] Israel Fernández, Fernando P. Cossío, and Miguel A. Sierra. Dyotropic reactions: Mechanisms and synthetic applications. *Chemical Reviews*, 109(12):6687–6711, 2009. PMID: 19746971.

[88] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[89] Chengxi Zang and Fei Wang. Moflow: an invertible flow model for generating molecular graphs. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 617–626, 2020.

[90] Tayebeh Ameri, Gilles Dennler, Christoph Lungenschmied, and Christoph J. Brabec. Organic tandem solar cells: A review. *Energy Environ. Sci.*, 2:347–363, 2009.

[91] Markus C Scharber, David Mühlbacher, Markus Koppe, Patrick Denk, Christoph Waldauf, Alan J Heeger, and Christoph J Brabec. Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency. *Advanced materials*, 18(6):789–794, 2006.

[92] 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.

[93] Johannes Hachmann, Roberto Olivares-Amaya, Adrian Jinich, Anthony L. Appleton, Martin A. Blood-Forsythe, László R. Seress, Carolina Román-Salgado, Kai Trepte, Sule Atahan-Evrenk, Süleyman Er, Supriya Shrestha, Rajib Mondal, Anatoliy Sokolov, Zhenan Bao, and Alán Aspuru-Guzik. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the harvard clean energy project. *Energy Environ. Sci.*, 7:698–704, 2014.

[94] Steven A. Lopez, Benjamin Sanchez-Lengeling, Julio de Goes Soares, and Alán Aspuru-Guzik. Design principles and top non-fullerene acceptor candidates for organic photovoltaics. *Joule*, 1(4):857–870, 2017.

[95] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, December 2011.

[96] Philipp Pracht, Fabian Bohle, and Stefan Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics*, 22(14):7169–7192, 2020.

[97] Sebastian Spicher and Stefan Grimme. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angewandte Chemie International Edition*, 59(36):15665–15673, September 2020.

[98] 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, March 2019.

[99] Henri Theil. A rank-invariant method of linear and polynomial regression analysis. *Indagationes mathematicae*, 12(85):173, 1950.

[100] Pranab Kumar Sen. Estimates of the regression coefficient based on kendall's tau. *Journal of the American Statistical Association*, 63(324):1379–1389, 1968.

[101] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

[103] Stefan Grimme, Christoph Bannwarth, and Philip Shushkov. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (z = 1–86). *Journal of Chemical Theory and Computation*, 13(5):1989–2009, 2017. PMID: 28418654.

[104] Philipp Pracht, Eike Caldeweyher, Sebastian Ehlert, and Stefan Grimme. A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for large Molecules. *ChemRxiv*, June 2019.

[105] Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. Extended tight-binding quantum chemistry methods. *WIREs Computational Molecular Science*, 11(2):e1493, 2021.

[106] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098–3100, Sep 1988.

[107] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789, Jan 1988.

[108] Axel D. Becke. Density-functional thermochemistry. iii. the role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993.

[109] R. Ditchfield, W. J. Hehre, and J. A. Pople. Self-consistent molecular-orbital methods. ix. an extended gaussian-type basis for molecular-orbital studies of organic molecules. *The Journal of Chemical Physics*, 54(2):724–728, 1971.

[110] W. J. Hehre, R. Ditchfield, and J. A. Pople. Self-consistent molecular orbital methods. xii. further extensions of gaussian-type basis sets for use in molecular orbital studies of organic molecules. *The Journal of Chemical Physics*, 56(5):2257–2261, 1972.

[111] Praveen C Hariharan and John A Pople. The influence of polarization functions on molecular orbital hydrogenation energies. *Theoretica chimica acta*, 28(3):213–222, 1973.

[112] Qiming Sun. Libcint: An efficient general integral library for gaussian basis functions. *Journal of Computational Chemistry*, 36(22):1664–1671, 2015.

[113] Qiming Sun, Timothy C. Berkelbach, Nick S. Blunt, George H. Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D. McClain, Elvira R. Sayfutyarova, Sandeep Sharma, Sebastian Wouters, and Garnet Kin-Lic Chan. Pyscf: the python-based simulations of chemistry framework. *WIREs Computational Molecular Science*, 8(1):e1340, 2018.

[114] Qiming Sun, Xing Zhang, Samragni Banerjee, Peng Bao, Marc Barbry, Nick S. Blunt, Nikolay A. Bogdanov, George H. Booth, Jia Chen, Zhi-Hao Cui, Janus J. Eriksen, Yang Gao, Sheng Guo, Jan Hermann, Matthew R. Hermes, Kevin Koh, Peter Koval, Susi Lehtola, Zhendong Li, Junzi Liu, Narbe Mardirossian, James D. McClain, Mario Motta, Bastien Mussard, Hung Q. Pham, Artem Pulkin, Wirawan Purwanto, Paul J. Robinson, Enrico Ronca, Elvira R. Sayfutyarova, Maximilian Scheurer, Henry F. Schurkus, James E. T. Smith, Chong Sun, Shi-Ning Sun, Shiv Upadhyay, Lucas K. Wagner, Xiao Wang, Alec White, James Daniel Whitfield, Mark J. Williamson, Sebastian Wouters, Jun Yang, Jason M. Yu, Tianyu Zhu, Timothy C. Berkelbach, Sandeep Sharma, Alexander Yu. Sokolov, and Garnet Kin-Lic Chan. Recent developments in the pyscf program package. *The Journal of Chemical Physics*, 153(2):024109, 2020.

[115] P. Jeffrey Hay and Willard R. Wadt. Ab initio effective core potentials for molecular calculations. potentials for k to au including the outermost core orbitals. *The Journal of Chemical Physics*, 82(1):299–310, 1985.

[116] Lorenz C Blum and Jean-Louis Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *Journal of the American Chemical Society*, 131(25):8732–8733, 2009.

[117] G Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of computer-aided molecular design*, 27(3):221–234, 2013.

[118] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):1–14, 2011.

[119] Christoph Gorgulla, Andras Boeszoermenyi, Zi-Fu Wang, Patrick D Fischer, Paul W Coote, Krishna M Padmanabha Das, Yehor S Malets, Dmytro S Radchenko, Yurii S Moroz, David A Scott, et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, 580(7805):663–668, 2020.

[120] Christoph Gorgulla, AkshatKumar Nigam, Matt Koop, Suleyman S Cinaroglu, Christopher Secker, Mohammad Haddadnia, Abhishek Kumar, Yehor Malets, Alexander Hasson, Roni Levin-Konigsberg, et al. Virtualflow 2.0-the next generation drug discovery platform enabling adaptive screens of 69 billion molecules. *bioRxiv*, pages 2023–04, 2023.

[121] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

[122] Johannes H. Voigt, Bruno Bienfait, Shaomeng Wang, and Marc C. Nicklaus. Comparison of the nci open database with seven large chemical structural databases. *Journal of Chemical Information and Computer Sciences*, 41(3):702–712, 2001.

[123] Wolf-Dietrich Ihlenfeldt, Johannes H. Voigt, Bruno Bienfait, Frank Oellien, and Marc C. Nicklaus. Enhanced cactvs browser of the open nci database. *Journal of Chemical Information and Computer Sciences*, 42(1):46–57, 2002.

[124] Thomas A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519, April 1996.

[125] Xiaolei Zhu, Keiran C. Thompson, and Todd J. Martínez. Geodesic interpolation for reaction pathways. *The Journal of Chemical Physics*, 150(16):164103, April 2019.

[126] Lee-Ping Wang and Chenchen Song. Geometry optimization made simple with translation and rotation coordinates. *The Journal of Chemical Physics*, 144(21):214108, June 2016.

[127] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.