Guided Diffusion for Inverse Molecular Design

Tomer Weiss,[†] Luca Cosmo,[‡] Eduardo Mayo Yanes,[¶] Sabyasachi Chakraborty,[¶] Alex M. Bronstein,[*,†] and Renana Gershoni-Poranne[*,¶]

[†]*Department of Computer Science, Technion - Israel Institute of Technology, Haifa 32000, Israel*

[‡]*University Ca' Foscari of Venice, Venice, Italy*

[¶]*Schulich Faculty of Chemistry, Technion - Israel Institute of Technology, Haifa 32000, Israel*

E-mail: bron@cs.technion.ac.il; rporanne@technion.ac.il

Abstract

The holy grail of materials science is *de novo* molecular design – i.e., the ability to engineer molecules with desired characteristics. Recently, this goal has become increasingly achievable thanks to developments such as equivariant graph neural networks that can better predict molecular properties, and to the improved performance of generation tasks, in particular of conditional generation, in text-to-image generators and large language models. Herein, we introduce GaUDI, a guided diffusion model for inverse molecular design, which combines these advances and can generate novel molecules with desired properties. GaUDI decouples the generator and the property-predicting models and can be guided using both point-wise targets and open-ended targets (e.g., minimum/maximum). We demonstrate GaUDI's effectiveness using single- and multiple-objective tasks applied to newly-generated data sets of polycyclic aromatic systems, achieving nearly 100% validity of generated molecules. Further, for some tasks, GaUDI discovers better molecules than those present in our data set of 475k molecules.

Introduction

The development of new technologies often hinges on the ability to source new functional molecules. Yet, molecular discovery remains an open challenge for chemists and materials scientists, due to the difficulty in (accurately) modeling molecular and material properties. This problem is exacerbated by the fact that such molecules must often fulfill multiple requirements, which can sometimes be contradictory or even mutually exclusive, e.g., the need for a catalyst to be both stable and active.[1] The key, therefore, is to find the optimal trade-off between multiple molecular properties, such that a given molecule may provide the desired function(s).

Finding this "sweet-spot" first requires identifying the relationships between the structure of the molecule and its various properties. To do so, traditional approaches to molecular design rely on manually-constructed heuristics and chemical intuition. In addition to being slow and arduous, this approach is usually limited to relatively simple structure-property relationships that are relevant within a small chemical space. In recent years, generative models[2–4] – which formulate this chemical challenge as an inverse design problem – have been introduced as an alternative approach and have become increasingly powerful tools for identifying new candidate structures for various applications, ranging from drug design[5,6] to fluorescent molecules[7] to peptides.[8,9]

Within the realm of generative approaches, diffusion models have recently become the lead-

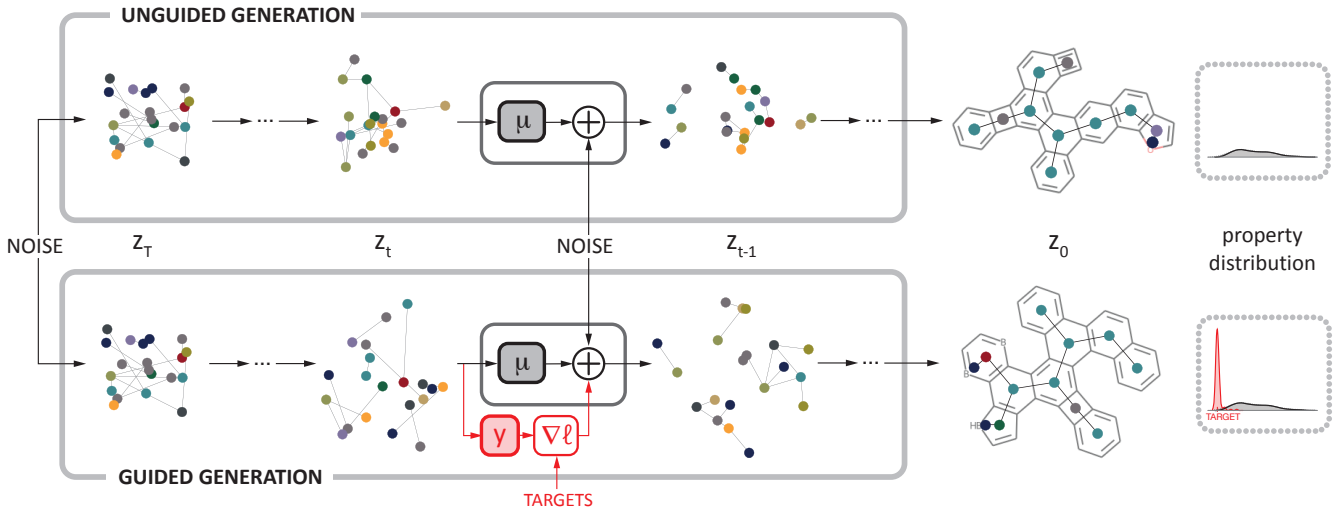

Figure 1: Generation workflow. Top: standard diffusion generation process, the noise is iteratively denoised using a neural model $\boldsymbol{\mu}$ until a clean sample is generated. Bottom: the guiding mechanism, at each iteration the prediction model $\mathbf{y}$ estimates the molecular properties, which are then used to calculate the target function ℓ. The gradients of the target function are combined with the output of the denoiser for guidance. The graph-of-rings representation of the polyaromatic chemical space is shown.

ing method for many generation tasks, such as image,[10] video,[11] and text[12] generation. In the context of chemistry, they have also shown great promise: Hoogeboom *et al* implemented an equivariant diffusion model for molecule generation that outperformed all previously reported methods.[13] Another interesting advance in this area is the ability to guide a diffusion model to sample from a conditional distribution.[14–16] In this work, we combine these developments and describe the design and implementation of a novel guided diffusion model for the generative design of molecules with targeted properties. The name GaUDI combines the acronym of *GU*ided *DI*ffusion with a nod to the famous Catalan architect and designer (of buildings, rather than molecules), Antoni Gaudí.

To demonstrate the performance and efficiency of GaUDI, we apply it to the use-case of polycyclic aromatic systems (PASs) – molecules constructed from multiple aromatic rings of varying sizes and atomic compositions. PASs are highly prevalent in nature and in the man-made world, comprising two-thirds of known molecules,[17] and have a broad impact on fields as diverse as the environment[18] and astrochemistry.[19] Most notably, PASs are the cornerstone of organic electronics, as they form the vast majority of organic semiconductors that are the molecular backbones of these devices.[20–22] New functional PASs are crucial for technologies such as organic light-emitting diodes and organic photovoltaics, as well as many others.[23,24]

To test the capabilities of GaUDI, we focus on two types of data sets and perform both single- and multiple-objective generation tasks. These examples showcase the advantages of GaUDI, including exceptionally good conditional sampling performance and versatility of the target function. Not only can GaUDI be directed toward specific numerical target values, but it can also be tasked with open-ended targets, e.g., finding a minimum/maximum value of the target property even when such a value is not known *a priori*. Indeed, any differentiable target function of single or multiple properties can be used to condition the generation process. Moreover, when used in combination with the graph-of-rings representation (GOR,[25] *vide infra*), almost 100% of the molecules generated by GaUDI are valid, novel, and unique.

Results

Workflow

Our method uses two pre-trained models to design molecules: the first is a generative diffusion model trained to generate unconditional samples from a given data distribution, and the second is a prediction model trained to predict molecular properties.

The diffusion model samples from some trackable source of noise and then iteratively denoises the signal, as in standard diffusion sampling. It is important to emphasize that the generative model should use a representation suitable for the target chemical space. We demonstrate the importance of this aspect by sampling molecules in the GOR molecular representation, which guarantees that the generated structures remain in the PAS chemical space.

In addition, the intermediate outputs of the generative model are fed to the prediction model, which predicts a predefined set of properties; the gradients of a target function of those properties are used to "guide" the sampling process by adding a correction term in each iteration (Figure 1). In this way, the diffusion generation is biased towards molecules with low target function values, a process that is equivalent to sampling from a conditional distribution with almost arbitrarily complex conditioning (*vide infra*).

Unguided molecular generation

We start by demonstrating the ability of the diffusion model and our GOR molecular representation to capture the existing data distribution and generate new structures within the chosen chemical space. We trained two Euclidean-equivariant diffusion models (EDM)[13] on two data sets, respectively: COMPAS-1x containing *cata*-condensed polybenzenoid hydrocarbons (cc-PBHs), and a PAS data set comprising a diverse set of heterocycle-containing PASs, and generated 1000 molecules from each model. The success of the generation was evaluated in three aspects: a) ***validity*** – the percentage of valid molecules as measured by RDKit;[26] b) ***novelty*** – the percentage of valid molecules not found in the training set; c) ***uniqueness*** – the percentage of unique molecules among the valid molecules.

As shown in Table 1, both of our trained models captured the data distribution well. Furthermore, nearly 100% of the generated molecules were valid, which is higher than reported for the original implementation.[13] The difference most likely stems from our use of the GOR as the chemical representation, which simplifies the learning (further details in the Supporting Information). It is unsurprising that the novelty of generated cc-PBHs is low, as the size of this chemical space is smaller and 80% of the molecules in this class already appear in the training set. In contrast, both the novelty and uniqueness of the generated PASs are 100%, which is again unsurprising, considering the vastness of this chemical space.

Table 1: **Performance of unguided generation for batches of 1000 molecules generated for each of the data sets.**

Dataset	Valid	Novel	Unique
cc-PBH	99.21%	23.75%	93.41%
PAS	99.71%	100.00%	100.00%

Guided design of cc-PBHs

Single-objective target

Having demonstrated that the combination of the EDM and the GOR is capable of generating valid and unique molecules in the chemical space of polycyclic molecules, we proceeded to the next goal of molecular generation: design of molecules with desired properties, i.e., guided generation with GaUDI. As an initial proof-of-concept, we focused on the simpler class of molecules, cc-PBHs, for which the COMPAS-1x data set contains a variety of molecular properties, including highest occupied molecular orbital (HOMO) energy, lowest unoccupied molecular orbital (LUMO) energy, HOMO-LUMO gap (HLG), relative energy (Erel), adiabatic ionization potential (IP),

Table 2: **Guided generation performance of the different models and of GaUDI with various gradient scaling values.**

	Joint distribution		Marginal distribution		Test case	
	Valid	MAE	Valid	MAE	Valid	MAE
pcEDM	59.3%	0.090	12.9%	0.133	0%	-
EEGSDE[27]	84%	0.158	78%	0.149	90.1%	0.301
GaUDI (s=30)	96.8%	0.109	84.3%	0.165	91.1%	0.256
GaUDI (s=100)	77.3%	0.074	80.4%	0.131	87.3%	0.241
GaUDI (s=300)	75%	0.056	65.2%	0.119	82.2%	0.211
GaUDI (s=1000)	36.7%	0.039	42.1%	0.107	40.6%	0.183

and adiabatic electron affinity (EA).

We compared the performance of GaUDI to two other methods for conditioning diffusion models: a) **Pointwise conditional EDM (pcEDM)** – a straightforward approach for conditioning a diffusion model, which conditions the denoiser with the ground-truth properties of the molecules at training, with the desired target properties during the generation; b) **EEGSDE**[27] – an approach for conditioning the diffusion process using score-based generative modeling through stochastic differential equations.[16] Additionally, for GaUDI, we evaluated the effect of the gradient scaling s, which allows us to tune the strength of the guidance.

We conditioned all three models on LUMO, HLG, Erel, IP, and EA, and tasked the models with generating 10-ring cc-PBHs with various combinations of target values for these properties, at varying levels of difficulty:

1. ***Joint distribution*** (easy) – a set of properties sampled from molecules in the test set.

2. ***Marginal distribution*** (harder) – a set of desired properties sampled from the product of marginal distributions of each property as estimated on the training set. This is a harder task, as the combination of the marginal property values might be infeasible.

3. ***Real test case*** (hard) – the properties of pentacene (detailed in Figure 2A). This is a difficult task because the likelihood of locating a 10-ring system with similar properties is small, as some of the properties are size-dependent.

Table 2 details the evaluation using two metrics: the validity of the generated molecules and the mean absolute error (MAE) relative to the respective desired properties (similarly to Hoogeboom et al., we calculated properties using a property-prediction network, described in the Methods section). The results show that the standard conditional method produced a relatively low percentage of valid molecules and failed completely when conditioned on harder targets, whereas both EEGSDE and GaUDI succeeded in generating molecules even when provided with difficult targets. Additionally, GaUDI significantly outperformed the two other methods in terms of the MAE and successfully found molecules with the closest properties to the desired ones in all cases. Table 2 also clearly depicts the trade-off of the gradient scaling s: increasing the scaling reduces the number of valid molecules but decreases their MAE. Our experience showed that using high values of s and sampling multiple molecules helps to find the best molecules.

Global minimum target

One of the main advantages of GaUDI is its unique ability to be guided not only toward a specific value (point-wise conditioning), but also toward any differentiable function of one or more properties or their combination, e.g., minimum/maximum. The COMPAS-1x data set includes all of the cc-PBH molecules containing up to 10 rings, which allowed us to de-

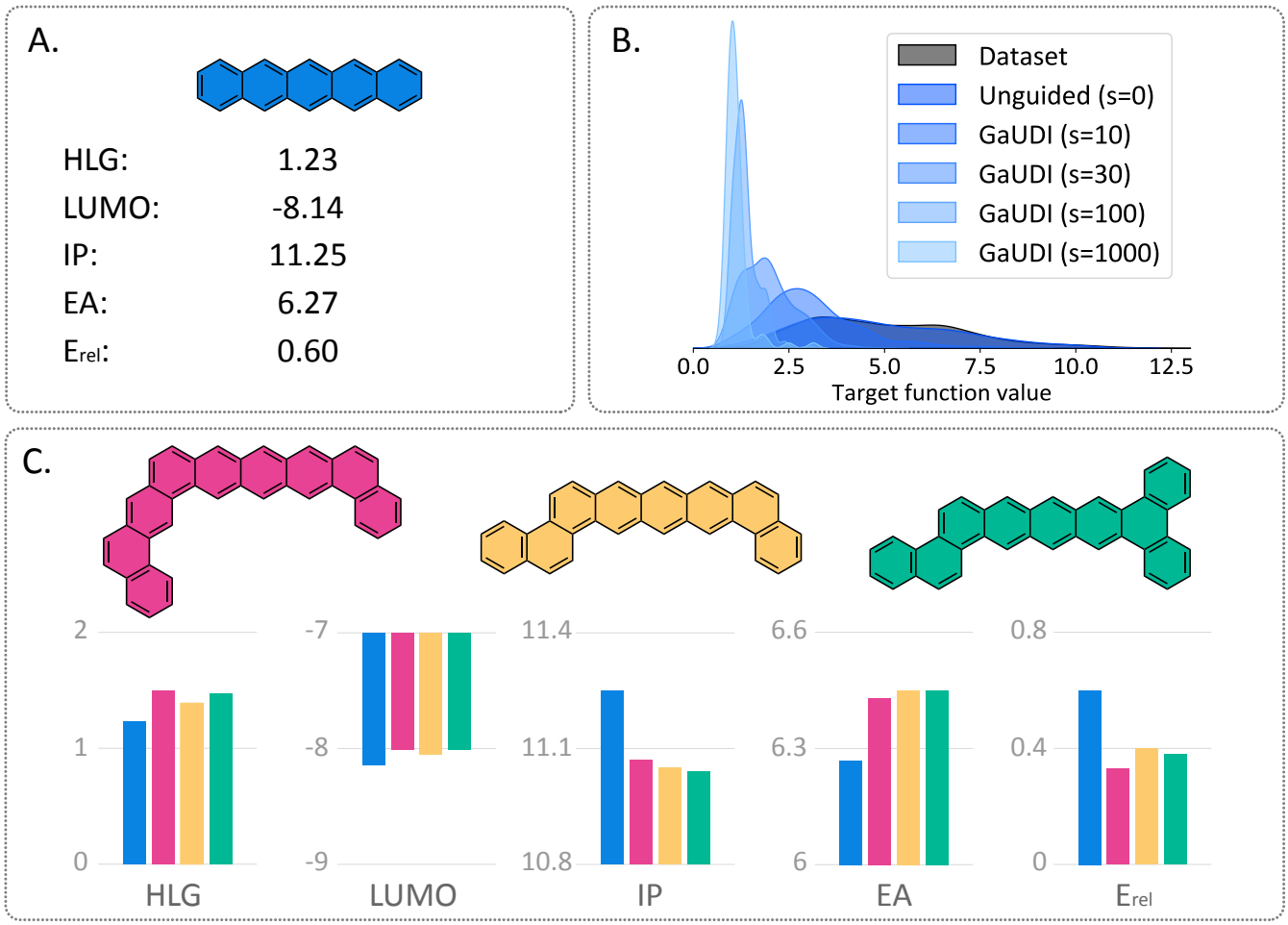

Figure 2: Guided generation of cc-PBHs molecules to global minimum. (A) Pentacene and its molecular properties, calculated with GFN2-xTB. (B) Distributions of the target-function values for the data set and for samples generated by GaUDI with different gradient scalar values. (C) Selected examples of GaUDI-generated cc-PBHs at the global minimum of the target function, which aims for properties similar to pentacene and minimal Erel (using $s = 100$). The individual properties of each molecule are denoted on the bar plots, color coded in the same colors as the molecules; the left-most bar (blue) is the value of pentacene.

sign a control experiment to evaluate the performance of GaUDI in finding molecules at the global minimum of a defined target function. To provide a relevant example, we chose pentacene (Figure 2A), one of the most commonly used cc-PBHs in organic electronics, as our target. We tasked GaUDI with discovering a cc-PBH molecule containing six or more rings with the electronic properties of pentacene but with increased stability (i.e., lower Erel). The target function for this purpose was defined as the Mean Square Error (MSE) of the properties LUMO, HLG, IP, and EA between the generated molecule and pentacene plus Erel.

Prior to generation, we identified the ten molecules in the entire data set with the lowest target-function values and removed them from the training sets of the diffusion model and of the prediction model. We then had GaUDI generate a sample of 512 cc-PBHs using the described target function and a gradient scaling of $s = 100$. Gratifyingly, all 10 molecules with the lowest target-function values were present in this sample, indicating that GaUDI did indeed reach the global minimum of the declared target. In addition, we generated a series of samples using different values of gradient scaling s (Figure 2B). As seen in the distribution plots, setting s to zero (i.e., unguided generation/no conditioning) affords a distribution that is al-

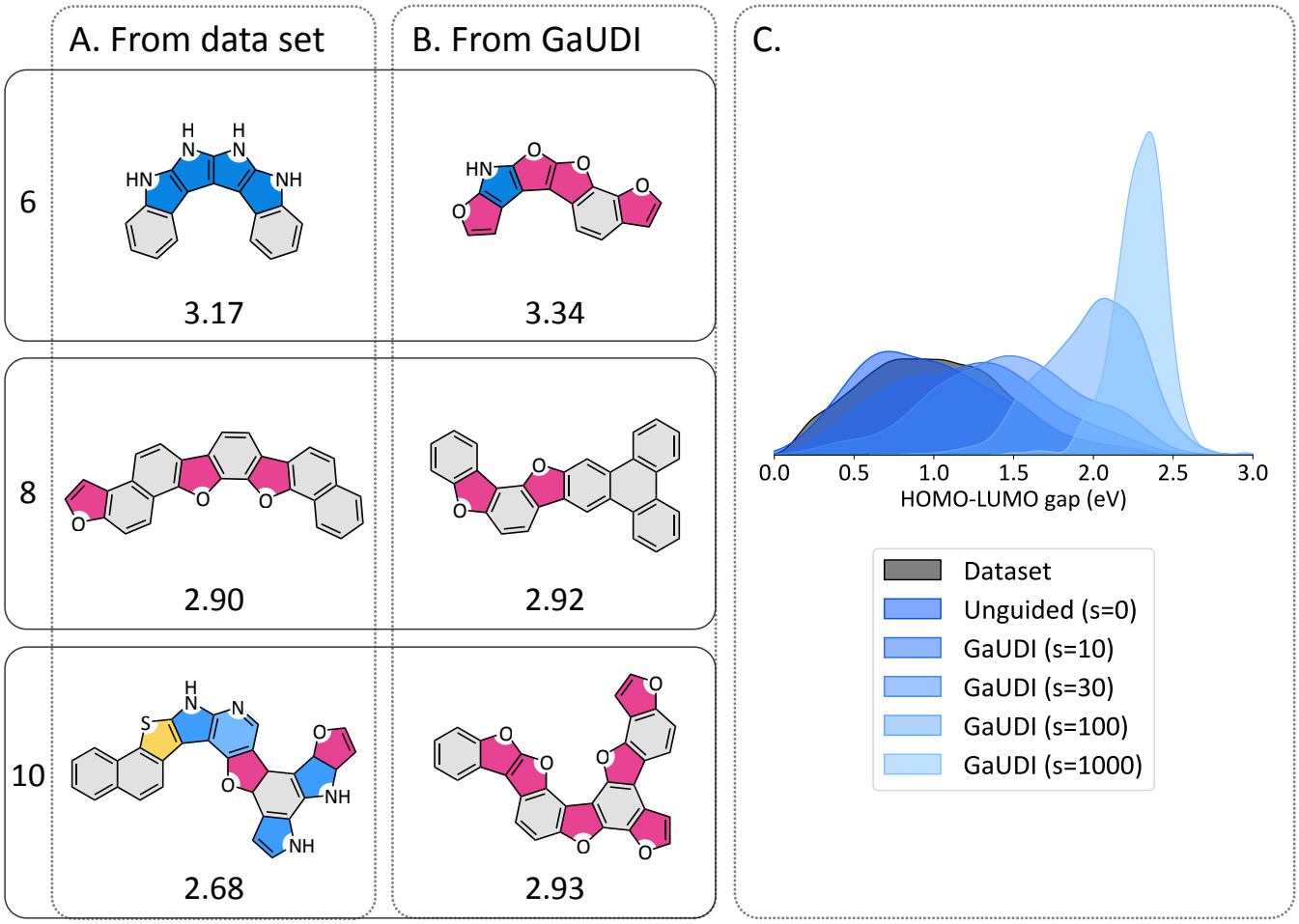

Figure 3: Guided design of PASs with high HLG values. (A) The 6-, 8-, and 10-ring PASs with the highest HLG values in the data set. (B) Selected examples of 6-, 8- & 10-ring PASs with high HLG designed by GaUDI (C) Distributions of HLG values for the data set and for samples generated by GaUDI with different gradient scalar values.

most equal to the data set distribution, and as the value of s is increased, the distributions shift toward increasingly lower target-function values. These results demonstrate both that GaUDI successfully captures the true data distribution and that the gradient scaling s can be used to guide the generation to molecules with properties closer to the desired values. In Figure 2C we present a few selected examples of the molecules designed by GaUDI with the described target function and $s = 100$. It is worth noting that all generated molecules contain a pentacene substructure. This is not surprising, as we have previously shown that the majority of electronic properties of cc-PBHs are determined by the longest linear motif.[25,28,29]

Guided design of PASs

Out-of-distribution generation

Whereas cc-PBHs contain only one type of aromatic ring (benzene) and all isomers can be easily enumerated, heterocycle-containing PASs are a vastly larger group, which is infeasible to enumerate exhaustively. The PAS data set we generated, which contains approximately 475K molecules, covers only a tiny fraction of this chemical space. Thus, PASs present a much greater challenge for both the learning and the generation processes, but also provide much more potential for the design of interesting and functional molecules.

The true test for GaUDI is whether it can design better molecules than the ones found by combining high-throughput calculation and

screening. In other words, can it generate molecules that have properties outside the distribution of the original data set? To investigate this, we first focused on a single property, the HLG, and tasked GaUDI with generating molecules with high HLG values. To avoid bias in terms of molecular size, we constrained the generation to structures with different numbers of rings. In Figure 3A and B, we show three pairs of PASs of equal size (6-, 8-, and 10-ring systems, respectively); for each pair, the molecule on the left is the PAS with the highest HLG located in the pre-generated data set, and the molecule on the right is the PAS with the highest HLG designed by GaUDI (the values presented were obtained with GFN2-xTB, the same level as the original data set). It can be observed that, for each molecular size specified, GaUDI consistently returned novel structures with higher HLGs than found in the data set. The generality of these results can be seen in Figure 3C, where we show the effect of the scaling factor s on a series of generation batches (for uniformity of the comparison, we focused on 10-ring systems). As seen in the distribution plots of the various batches, increasing s afforded molecules with increasingly higher HLG values (i.e., lower target-function values). Thus, GaUDI is capable of designing valid novel molecules beyond the boundaries of the property distribution. Interestingly, it appears that the presence of five-membered heterocycles pushes the HLG up. In particular, multiple furan moieties show up as recurring motifs in the high-HLG structures. Oligofurans molecules have been recognized as promising compounds for organic electronics.[30,31]

Multi-property target

An even more challenging task is to optimize several properties at once. To test GaUDI, we tasked it with generating molecules with a small HLG, low IP, and high EA. This combination of properties is relevant for narrow band-gap molecules potentially suitable for use in photonics.[32] Therefore, we defined the target function for this purpose as $\ell(\mathrm{HLG}, \mathrm{IP}, \mathrm{EA}) = 3 \cdot \mathrm{HLG} + \mathrm{IP} - \mathrm{EA}$, using a factor of 3 for the HLG property in order to better balance the properties, which have different value ranges. In Figure 4C we present selected examples of GaUDI-designed PASs and their target-function values. In contrast to the previous experiment, GaUDI was not able to generate out-of-distribution molecules. However, it was able to generate vast numbers of molecules with low target-function values. For example, out of all the 10-ring PASs in our data set (70k molecules), only 25 have target-function values below 3 (0.036%). In a single generation batch of 512 molecules, GaUDI generated 159 new molecules with similar target-function values (31%). Thus, GaUDI produces a $\times 861$ enrichment, substantially increasing the likelihood of identifying promising new narrow-band gap candidate molecules that may be functional in optoelectronic applications. It is interesting to note the increased prevalence of boron atoms in the generated structures. Boron substitution has been recognized as a LUMO-lowering mechanism and, in recent years, boron-doped PASs have been incorporated in numerous organic-electronic applications.[33–36]

Discussion

Despite being a relatively recent development, diffusion models have already shown promise for generative molecular design.[13,37] The inherent characteristics of chemical structures – discrete chemical space, bonding rules, etc. – make this an exceedingly challenging inverse design problem. Indeed, previous reports have shown that, even when the generative model succeeds in biasing the generation, the majority of generated molecules are invalid. Within this context, our current work provides several advantages.

Firstly, and most importantly, we achieved guided molecular generation toward desired properties in a truly scalable and flexible way. Standard approaches to conditional generation are typically limited to point-wise conditioning (i.e., generate a structure with a property equal to some target value) and require the *(structure, property)* pairs during training. Consequently, the required training set size grows exponen-

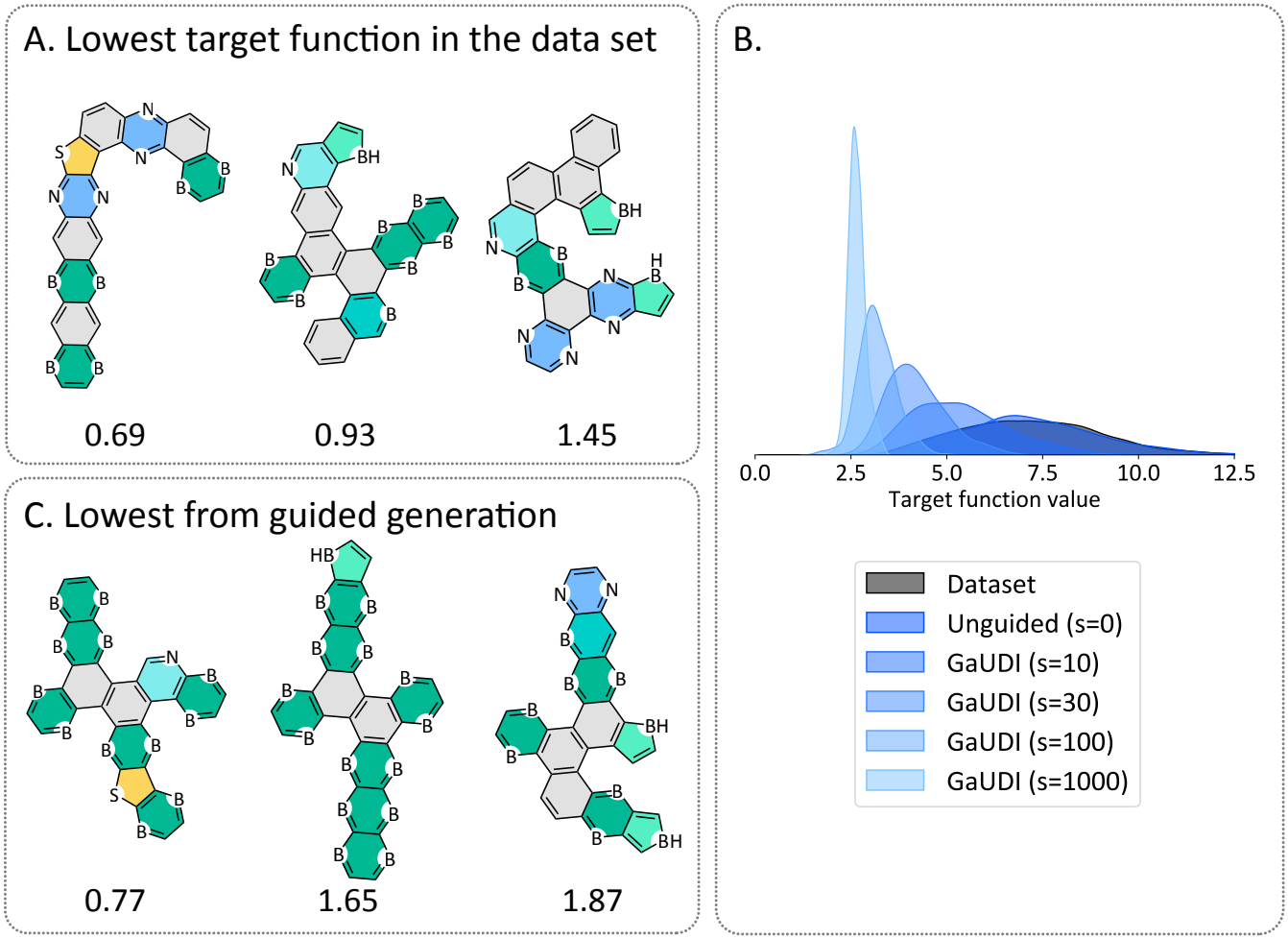

Figure 4: Guided design of narrow band-gap molecules. (A) The molecules with the lowest target-function values in the PAS data set (B) Distributions of the target function for the data set and for GaUDI-generated batches with different gradient scalar values. (C) The molecules with the lowest target-function values designed by GaUDI.

tially with the dimension of the property vector. Furthermore, point-wise conditioning does not allow, e.g., finding molecules with the maximum HLG, unless the maximum value is known in advance. In contrast, GaUDI guides the conditional generation through an arbitrary differentiable scalar loss function that can combine multiple properties and can include non-point-wise operations, such as minimum or maximum. In addition to the clear advantage this presents in terms of defining targets for molecular generation, we emphasize that this feature is extremely important from another aspect: it allows the use of inexpensively generated data sets, in which the trends are correct but the numerical values are not. For example, in the current work, we used GFN2-xTB, a quick and inexpensive method, to construct our large PAS data set. Our previous work has shown that xTB reveals the same structure-property trends as DFT, but the property values themselves are in significantly different value ranges.[28] By allowing minimum/maximum targets, rather than numerical target values, GaUDI can leverage this wealth of data, which might otherwise be meaningless or require expensive correction schemes.

Another novelty of our approach is that it fully decouples the training of an unconditional generator and a property predictor, which can then use different training sets and incorporate distinct inductive assumptions (e.g., some properties may be inferred from local substructures and thus their predictor may be trained on a set of small molecular structures).

Finally, to the best of our knowledge, we

demonstrated the first generative design of PAS molecules, which was enabled by our COMPAS data generation pipeline.[28] We note that, despite beginning from a complicated distribution, the generation achieved a remarkably high percentage of valid molecules (nearly 100%). This is thanks to the GOR representation, which makes the rules the model needs to learn much simpler. In contrast, using a graph of atoms representation resulted in much lower validity scores, and many of the (formally) valid molecules obtained were not, in fact, within the PAS chemical space (see Supporting Information for further details). As proposed by Westermayr et al,[38] our method may also be further developed by implementing an iterative process, whereby the properties and structures of generated molecules are calculated on-the-fly by an inexpensive method and added to the training set, or by incorporating an evaluator that scores the generated structures on their feasibility for synthesis. In this regard, we note that synthesizability scores have not yet been developed for PASs. Nevertheless, many of the molecules proposed by GaUDI are relatively simple and appear to be reasonably feasible to synthesize (e.g., Figure 3B).

Importantly, both the model and the representation we described in this work can be generalized for other tasks. The conditioning method we introduce to guide the molecular design can be used to turn any unconditional diffusion model into a controllable conditional generative model and can be useful in many tasks in computer vision, natural language processing, etc. The GOR representation can be easily adapted to other molecular families by defining different building blocks as the graph node features, which can enable molecular design in other chemical spaces. Thus, GaUDI contributes to acceleration of molecular design and discovery in numerous areas of interest, including but not limited to organic electronics and optoelectronics.

Methods

Data

Two data sets were used: the COMPAS-1x data set[28] from the COMPAS project and a new PAS data set we prepared for this work. COMPAS-1x contains the GFN2-xTB-calculated structures and properties of ∼34K *cata*-condensed PBHs (cc-PBH) comprising 1–11 rings. The new PASs data set contains the GFN2-xTB[39,40]-calculated structures and properties of ∼475K polycyclic aromatic systems (PASs) comprising 1–10 rings. The PASs in this data set are built from 11 types of aromatic rings, including heterocyclic components. For further details on the PASs data set, we refer the reader to the Supplementary Materials.

Molecular Representation

In the field of chemistry, the majority of approaches applying graph neural networks use a molecular graph as the molecular input representation. In such graphs, the atoms are the nodes, and the bonds are the edges (i.e., graph of atoms, or GOA). In our previous work,,[25] we introduced the graph of rings (GOR) representation for PBHs. In the GOR (which can be seen in Figure 1 and in the Supporting Information), each node represents a ring (the coordinates of the node are the centroid of the rings). In the current work, we extended the GOR representation to heterocyclic-containing systems by setting the ring type as a node feature. In addition, we introduced an additional node for each ring, situated at the location of the heteroatom, to note the orientation of each ring within the PAS. In the case of two heteroatoms in a single ring, e.g., pyrazine, only one of the heteroatoms is indicated. This is sufficient because our data only contains rings in which the two heteroatoms are at *para* position to one another. In contrast to our previous work, the current representation does not include any information on the connectivity of rings. This modification is crucial to allow the inverse design to learn any connectivity between the rings.

Using the GOR, rather than the GOA, allows the diffusion model to learn a much simpler distribution, because the rules the model needs to learn are much simpler than the collection of bonding rules required to construct a graph of atoms. This leads to a substantial improvement in performance and a reduction in the required computational resources. It also leads to a much higher percentage of valid molecules generated by the model. Importantly, while the GOR representation reduces the complexity of the graph, it retains important chemical information and provides an adequate representation of the molecule, as demonstrated in this work and in our previous report.[25]

We will denote by the matrix $\mathbf{X} \in \mathbb{R}^{n\times 3}$ the coordinates of the n nodes, and by $\mathbf{H} \in \mathbb{R}^{n\times c}$ the corresponding node attributes encoded as one-hot vectors, with c being the number of classes.

Equivariant diffusion model

Diffusion models[10] are a class of powerful likelihood-based generative models that have recently been shown to outperform generative adversarial networks (GANs)[41] in image generation tasks.[14] Diffusion models generate samples by gradually removing noise from a signal and their training objective can be expressed as a reweighed variational lower bound.[10]

During sample generation (after the model is trained), we start from sampling from $q(\mathbf{z}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ with $\mathbf{z} = (\mathbf{X}, \mathbf{H})$ collectively denoting both the coordinates and the attributes of a molecule representation with a fixed number of nodes. A sequence of $\mathbf{z}_t$'s is then sampled backwards in time from a Markov process described by the transition probability density $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$, until reaching $\mathbf{z}_0 \sim q_0$. The transition probability $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$ is approximated using a neural network of the form

$$p_{\boldsymbol{\theta}}(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t, t), \boldsymbol{\Sigma}_t), \quad (1)$$

where the vector $\boldsymbol{\theta}$ denotes the learnable parameters of the neural network $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, denotes the Gaussian density with location $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ evaluated at point $\mathbf{z}$. An isotropic sequence of covariances, $\boldsymbol{\Sigma}_t = \beta_t \mathbf{I}$, is typically asserted. A detailed derivation of the training and generation algorithm is available in the Supporting Information.

The probability distribution q_0 embodied by the diffusion model from which the node coordinates $\mathbf{X}$ and attributes $\mathbf{H}$ are sampled must satisfy two fundamental properties: (1) *permutation invariance*, implying that any permutation of the columns of $\mathbf{X}$ and $\mathbf{H}$ is equiprobable; and (2) $E(3)$ *invariance* implying that any Euclidean transformation (translation and rotation) of $\mathbf{X}$ is equiprobable.

We chose to use the $E(3)$ Equivariant Diffusion Model (EDM)[13] employing the $E(n)$-Equivariant Graph Neural Network (EGNN)[42] to satisfy the desired properties of $p_{\boldsymbol{\theta}}(\mathbf{z}_{t-1}|\mathbf{z}_t)$ and, consequently, of q_0.

Conditional generation

In order to bias (guide) the generation process toward desired molecular properties $\mathbf{y}$, one can attempt sampling from a conditional distribution $q_0(\mathbf{z}|\mathbf{y})$. This can be achieved by providing the values of $\mathbf{y}$ for every training sample during training. Hoogeboom et al. showed that, in practice, such an approach has ample space for improvement.[13] One of the reasons for its lack of success is the fact that conditional distributions are much harder to model. Another major shortcoming of the method is that the type of conditioning needs to be known at training. Here, we focused on an approach for conditioning the sampling process on any target function of $\mathbf{y}$ post-training.

In developing our method, we were inspired by the *classifier-guidance* proposed by Dhariwal and Nichol and adopted it due to its simplicity.[14] Nevertheless, it is important to note that Song et al. developed a similar approach from a very different perspective.[16] In *classifier-guidance*, in order to sample from the conditional distribution $p(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{y})$, one can use the Bayes rule to show that

$$p(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{y}) \propto p(\mathbf{z}_{t-1}|\mathbf{z}_t) p(\mathbf{y}|\mathbf{z}_{t-1}). \quad (2)$$

It is typically intractable to sample from this

Algorithm 1 Guided diffusion sampling, given a diffusion model $(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t), \boldsymbol{\Sigma}_t)$, $f(\mathbf{z}, t)$, and gradient scale s.

$\mathbf{z}_T \leftarrow$ sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, T-1, \ldots, 1$ **do**
  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_t)$
  $\mathbf{g} \leftarrow -\nabla_{\mathbf{z}_{t-1}} f(\mathbf{z} = \boldsymbol{\mu}, t)$
  $\mathbf{z}_{t-1} \leftarrow$ sample from $\mathcal{N}(\boldsymbol{\mu} + s\boldsymbol{\Sigma}_t\mathbf{g}, \boldsymbol{\Sigma}_t)$
end for
return x_0

distribution exactly, but it has been shown that it can be approximated as a perturbed Gaussian distribution:[43] instead of predicting the previous timestep $\mathbf{z}_{t-1}$ from timestep $\mathbf{z}_t$ using a Gaussian distribution

$$p_{\boldsymbol{\theta}}(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

one can transform it using (2) into

$$p_{\boldsymbol{\theta}}(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{g}, \boldsymbol{\Sigma}), \tag{4}$$

where $\mathbf{g} = \nabla_{\mathbf{z}_{t-1}} \log p(\mathbf{y}|\mathbf{z} = \boldsymbol{\mu})$. For a full derivation, refer to Section 4 in Dhariwal and Nichol.[14]

Dhariwal and Nichol only considered the case of guiding the generation toward a desired class and, therefore, use the logits of a classifier network as $\log p(\mathbf{y}|\mathbf{x}_t)$. We extend this formulation to any differentiable target function $f(\mathbf{z}, t)$ we want to minimize by defining $\log p(\mathbf{y}|\mathbf{z}) = -f(\mathbf{z}, t) + \text{const}$, where the constant is due to the density normalization factor and can be ignored when considering the gradient $\mathbf{g} = -\nabla_{\mathbf{z}} f(\mathbf{z}, t)$ evaluated at $\mathbf{z} = \boldsymbol{\mu}$. The entire conditional sampling process using our guidance method is summarized in Algorithm 1. Note that we include an optional scaling factor s for the gradients. Observe that $s\nabla_{\mathbf{z}} \log p(\mathbf{y}|\mathbf{z}) = \nabla_{\mathbf{z}} \log p(\mathbf{y}|\mathbf{z})^s + \text{const}$. When $s > 1$, this distribution becomes sharper than the original $p(\mathbf{y}|\mathbf{z})$.

Target function

To guide the molecular generation towards desired properties, we use a target function of the form $f(\mathbf{z}_t, t) = \ell(\hat{\mathbf{y}}(\mathbf{z}_t, t))$, where $\hat{\mathbf{y}}$ is a (forward) model that receives the molecular representation and predicts its property $\mathbf{y}$, and ℓ is a loss function that assigns lower values to molecules satisfying the desired properties. Note that the target function is conditioned on the time and, thus, needs to be able to assign scores to noisy inputs at any timestamp during the denoising process. Therefore, we train a time-conditioned structure-property prediction model $\hat{\mathbf{y}}(\mathbf{z}_t, t)$ on noisy samples using the same noise scheduler of the diffusion model.

In all our experiments, we implemented the time-conditioned prediction model using the same EGNN[42] architecture as the network used to approximate the diffusion dynamics, and trained it by minimizing

$$\mathbb{E}_{t\sim\mathcal{U}[0,T],(\mathbf{z}_0,\mathbf{y})\sim q_0(\mathbf{z},\mathbf{y}),\mathbf{z}_t\sim q_t(\mathbf{z}_t|\mathbf{z}_0)}\, \ell(\hat{\mathbf{y}}_{\boldsymbol{\phi}}(\mathbf{z}_t, t)) \tag{5}$$

over a set of parameters $\boldsymbol{\phi}$. Note that the unconditional generator is pre-trained and the predictor is trained once to predict a set of desired properties. Then, any combination of target properties can be used to guide conditional sampling as long as the conditioning can be expressed through a loss function ℓ.

Data Availability

All data for cc-PBHs used in this project was obtained from the COMPAS Project,[28] a freely available data repository at https://gitlab.com/porannegroup/compas. All PAS data is available free of charge at https://gitlab.com/porannegroup/gaudi.

Code Availability

All code used to train the models and generate molecules is provided free of charge at https://gitlab.com/porannegroup/gaudi.

Supporting Information Available

The following files are available free of charge.

- Supporting Information PDF file: all additional figures mentioned in this text, including textual description and relevant discussion.

- GitLab repository.

Acknowledgement E.M.Y., S.C., and R.G.P. are grateful for the financial support of the Branco Weiss Fellowship. R.G.P. is a Branco Weiss Fellow and a Horev Fellow. A.M.B. and T.W. are supported by the ERC StG EARS. L.C. is supported by the IRIDE grant from DAIS, Ca' Foscari University of Venice.

References

(1) Hwang, J.; Rao, R. R.; Giordano, L.; Katayama, Y.; Yu, Y.; Shao-Horn, Y. Perovskites in catalysis and electrocatalysis. *Science* **2017**, *358*, 751–756.

(2) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022**, *12*, e1608.

(3) Fuhr, A. S.; Sumpter, B. G. Deep generative models for materials discovery and machine learning-accelerated innovation. *Frontiers in Materials* **2022**, 182.

(4) Walters, W. P.; Barzilay, R. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research* **2020**, *54*, 263–270.

(5) Zeng, X.; Wang, F.; Luo, Y.; Kang, S.-g.; Tang, J.; Lightstone, F. C.; Fang, E. F.; Cornell, W.; Nussinov, R.; Cheng, F. Deep generative molecular design reshapes drug discovery. *Cell Reports Medicine* **2022**, 100794.

(6) Martinelli, D. Generative machine learning for de novo drug discovery: A systematic review. *Computers in Biology and Medicine* **2022**, 105403.

(7) Tan, Z.; Li, Y.; Wu, X.; Zhang, Z.; Shi, W.; Yang, S.; Zhang, W. De novo creation of fluorescent molecules via adversarial generative modeling. *RSC Advances* **2023**, *13*, 1031–1040.

(8) Wan, F.; Kontogiorgos-Heintz, D.; de la Fuente-Nunez, C. Deep generative models for peptide design. *Digital Discovery* **2022**, *1*, 195–208.

(9) Lin, E.; Lin, C.-H.; Lane, H.-Y. De novo peptide and protein design using generative adversarial networks: an update. *Journal of Chemical Information and Modeling* **2022**, *62*, 761–774.

(10) Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **2020**, *33*, 6840–6851.

(11) Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; Fleet, D. J. Video diffusion models. *arXiv preprint arXiv:2204.03458* **2022**,

(12) Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; van den Berg, R. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems* **2021**, *34*, 17981–17993.

(13) Hoogeboom, E.; Satorras, V. G.; Vignac, C.; Welling, M. Equivariant diffusion for molecule generation in 3d. International Conference on Machine Learning. 2022; pp 8867–8887.

(14) Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* **2021**, *34*, 8780–8794.

(15) Ho, J.; Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* **2022**,

(16) Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* **2020**,

(17) Balaban, A. T.; Oniciu, D. C.; Katritzky, A. R. Aromaticity as a Cornerstone of Heterocyclic Chemistry. *Chemical Reviews* **2004**, *104*, 2777–2812, DOI: 10.1021/cr0306790.

(18) Premnath, N.; Mohanrasu, K.; Rao, R. G. R.; Dinesh, G.; Prakash, G. S.; Ananthi, V.; Ponnuchamy, K.; Muthusamy, G.; Arun, A. A crucial review on polycyclic aromatic Hydrocarbons-Environmental occurrence and strategies for microbial degradation. *Chemosphere* **2021**, *280*, 130608.

(19) McGuire, B. A.; Loomis, R. A.; Burkhardt, A. M.; Lee, K. L. K.; Shingledecker, C. N.; Charnley, S. B.; Cooke, I. R.; Cordiner, M. A.; Herbst, E.; Kalenskii, S., et al. Detection of two interstellar polycyclic aromatic hydrocarbons via spectral matched filtering. *Science* **2021**, *371*, 1265–1269.

(20) Li, Q.; Zhang, Y.; Xie, Z.; Zhen, Y.; Hu, W.; Dong, H. Polycyclic aromatic hydrocarbon-based organic semiconductors: ring-closing synthesis and optoelectronic properties. *Journal of Materials Chemistry C* **2022**, *10*, 2411–2430.

(21) Aumaitre, C.; Morin, J.-F. Polycyclic aromatic hydrocarbons as potential building blocks for organic solar cells. *The Chemical Record* **2019**, *19*, 1142–1154.

(22) Kilaru, S.; Gade, R.; Tripathi, A.; Chetti, P.; Pola, S., et al. Organic materials based on hetero polycyclic aromatic hydrocarbons for organic thin-film transistor applications. *Materials Science in Semiconductor Processing* **2022**, *147*, 106730.

(23) Omar, Ö. H.; Del Cueto, M.; Nematiaram, T.; Troisi, A. High-throughput virtual screening for organic electronics: a comparative study of alternative strategies. *Journal of Materials Chemistry C* **2021**, *9*, 13557–13583.

(24) Das, S.; Bhauriyal, P.; Pathak, B. Polycyclic aromatic hydrocarbons as prospective cathodes for aluminum organic batteries. *The Journal of Physical Chemistry C* **2020**, *125*, 49–57.

(25) Weiss, T.; Wahab, A.; Bronstein, A. M.; Gershoni-Poranne, R. Interpretable Deep-Learning Unveils Structure–Property Relationships in Polybenzenoid Hydrocarbons. *The Journal of Organic Chemistry* **2023**, DOI: 10.1021/acs.joc.2c02381.

(26) Landrum, G., et al. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**, *8*.

(27) Bao, F.; Zhao, M.; Hao, Z.; Li, P.; Li, C.; Zhu, J. Equivariant Energy-Guided SDE for Inverse Molecular Design. *arXiv preprint arXiv:2209.15408* **2022**,

(28) Wahab, A.; Pfuderer, L.; Paenurk, E.; Gershoni-Poranne, R. The COMPAS Project: A Computational Database of Polycyclic Aromatic Systems. Phase 1: cata-Condensed Polybenzenoid Hydrocarbons. *J. Chem. Inf. Model.* **2022**, *62*, 3704–3713.

(29) Fite, S.; Wahab, A.; Paenurk, E.; Gross, Z.; Gershoni-Poranne, R. Text-based representations with interpretable machine learning reveal structure–property relationships of polybenzenoid hydrocarbons. *J. Phys. Org. Chem.* **2022**, e4458.

(30) Gidron, O.; Dadvand, A.; Sheynin, Y.; Bendikov, M.; Perepichka, D. F. Towards "green" electronic materials. α-Oligofurans as semiconductors. *Chemical communications* **2011**, *47*, 1976–1978.

(31) Gidron, O.; Bendikov, M. α-oligofurans: an emerging class of conjugated oligomers

for organic electronics. *Angewandte Chemie International Edition* **2014**, *53*, 2546–2555.

(32) Li, X.-H.; Guo, Y.-X.; Ren, Y.; Peng, J.-J.; Liu, J.-S.; Wang, C.; Zhang, H. Narrow-bandgap materials for optoelectronics applications. *Frontiers of Physics* **2022**, *17*, 1–33.

(33) Agnoli, S.; Favaro, M. Doping graphene with boron: a review of synthesis methods, physicochemical characterization, and emerging applications. *Journal of Materials Chemistry A* **2016**, *4*, 5002–5025.

(34) Kahan, R. J.; Hirunpinyopas, W.; Cid, J.; Ingleson, M. J.; Dryfe, R. A. Well-defined boron/nitrogen-doped polycyclic aromatic hydrocarbons are active electrocatalysts for the oxygen reduction reaction. *Chemistry of Materials* **2019**, *31*, 1891–1898.

(35) Stoycheva, J.; Tadjer, A.; Garavelli, M.; Spassova, M.; Nenov, A.; Romanova, J. Boron-doped polycyclic aromatic hydrocarbons: A molecular set revealing the interplay between topology and singlet fission propensity. *The Journal of Physical Chemistry Letters* **2020**, *11*, 1390–1396.

(36) Kothavale, S. S.; Lee, J. Y. Three- and four-coordinate, boron-based, thermally activated delayed fluorescent emitters. *Advanced Optical Materials* **2020**, *8*, 2000922.

(37) Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; Tang, J. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923* **2022**,

(38) Westermayr, J.; Gilkes, J.; Barrett, R.; Maurer, R. J. High-throughput property-driven generative design of functional organic molecules. *Nature Computational Science* **2023**, 1–10.

(39) Grimme, S.; Bannwarth, C.; Shushkov, P. 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* **2017**, *13*, 1989–2009.

(40) Bannwarth, C.; Ehlert, S.; Grimme, S. 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* **2019**, *15*, 1652–1671.

(41) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Communications of the ACM* **2020**, *63*, 139–144.

(42) Satorras, V. G.; Hoogeboom, E.; Welling, M. E (n) equivariant graph neural networks. International conference on machine learning. 2021; pp 9323–9332.

(43) Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. International Conference on Machine Learning. 2015; pp 2256–2265.