nature computational science

Article

https://doi.org/10.1038/s43588-023-00532-0

Guided diffusion for inverse molecular design

Received: 4 April 2023

Accepted: 6 September 2023

Published online: 5 October 2023

Check for updates

Tomer Weiss [1], Eduardo Mayo Yanes [2], Sabyasachi Chakraborty [2], Luca Cosmo [3], Alex M. Bronstein [1] ✉ & Renana Gershoni-Poranne [2] ✉

The holy grail of materials science is de novo molecular design, meaning engineering molecules with desired characteristics. The introduction of generative deep learning has greatly advanced efforts in this direction, yet molecular discovery remains challenging and often inefficient. Herein we introduce GaUDI, a guided diffusion model for inverse molecular design that combines an equivariant graph neural net for property prediction and a generative diffusion model. We demonstrate GaUDI's effectiveness in designing molecules for organic electronic applications by using single- and multiple-objective tasks applied to a generated dataset of 475,000 polycyclic aromatic systems. GaUDI shows improved conditional design, generating molecules with optimal properties and even going beyond the original distribution to suggest better molecules than those in the dataset. In addition to point-wise targets, GaUDI can also be guided toward open-ended targets (for example, a minimum or maximum) and in all cases achieves close to 100% validity of generated molecules.

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 is often exacerbated by the requirement to fulfill multiple demands, which can sometimes be contradictory or even mutually exclusive, for example, 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 for molecular design rely on manually constructed heuristics and chemical intuition. In addition to being slow and arduous, these are 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. As summarized in a recent perspective[5], generative molecular design has been achieved for different applications using a variety of approaches, including reinforcement learning[6,7], variational autoencoders[8,9], generative adversarial networks[10,11], genetic algorithms[12–14], normalizing flows[15] and, most recently, diffusion models[16].

Diffusion models have become the leading method for many generation tasks, such as image[17], video[18] and text[19] generation. Diffusion models have already shown great promise in chemistry too. Some notable examples are Hoogeboom and colleagues' equivariant diffusion model (EDM) for molecular generation[16], Xu and co-workers' GeoDiff model for molecular conformation generation[20], Thygesen and colleagues' combined crystal diffusion and variational autoencoder model for generating 2D materials[21] and Jaakkola and co-workers' DiffDock model, which treats molecular docking as a generative problem[22]. Nevertheless, the full capabilities of diffusion models have not yet been tapped as this is still a minimally explored area[5]. Furthermore, those of the existing diffusion models that also perform conditional generation use the so-called standard approach[16], which has difficulty in learning the conditional distribution. They are also limited to only point-wise targets, must be retrained to add new properties and cannot train the

[1]Department of Computer Science, Technion—Israel Institute of Technology, Haifa, Israel. [2]Schulich Faculty of Chemistry, Technion—Israel Institute of Technology, Haifa, Israel. [3]University Ca' Foscari of Venice, Venice, Italy. ✉e-mail: bron@cs.technion.ac.il; rporanne@technion.ac.il

generator and predictor on different datasets. The ability to guide a diffusion model to sample from a conditional distribution[23–25] has not yet been fully tested in a chemical context.

In this work, we bridge this gap by designing and implementing a guided diffusion model, GaUDI, for the generative design of molecules with targeted properties. The name GaUDI combines the acronym of guided diffusion with a nod to the famous Catalan architect and designer (of buildings, rather than molecules), Antoni Gaudí.

We demonstrate the performance of GaUDI on the use-case of polycyclic aromatic systems (PASs)—molecules constructed from multiple aromatic rings of varying sizes and atomic compositions. Polycyclic aromatic systems, which comprise two-thirds of known molecules[26], are the cornerstone of organic electronics as they form the vast majority of organic semiconductors[27–29]. New PASs with specific properties are therefore crucial for advancing technologies such as organic light-emitting diodes, field effect transistors, photovoltaics and other optoelectronics[30,31].

Trained on a newly generated 475,000 PAS dataset, GaUDI outperforms other leading diffusion models at both single- and multiple-objective generation tasks, both in terms of validity and in terms of the mean average error. GaUDI affords novel molecules with optimal properties, even going beyond the distribution of the original dataset. Moreover, when used with the graph of rings (GOR) representation[32] (see below), almost 100% of the molecules generated by GaUDI are valid, novel and unique. Furthermore, as opposed to many existing methods, GaUDI offers high target function versatility and can be tasked with any differentiable target function of single or multiple properties, including open-ended targets, for example, finding a minimum/maximum value of the target property even when such a value is not known a priori. In this work, we leveraged this feature to train GaUDI on data obtained with an inexpensive computational method, which captures the same structure–property trends despite having different numerical values.

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.

As in standard diffusion sampling, the diffusion model samples from some tractable source of noise and then iteratively denoises the signal; however, in contrast to the standard unconditional models, in GaUDI 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 these properties are then used to guide the sampling process by adding a correction term in each iteration (Fig. 1). In this way, the diffusion generation is biased toward molecules with low target-function values (that is, closest to the target), a process that is equivalent to sampling from a conditional distribution with almost arbitrarily complex conditioning (see below).

Unguided molecular generation

We first demonstrate that the combination of the EDM[16] and our GOR molecular representation can capture the existing data distribution and generate new structures within the chosen chemical space. We trained two EDMs on two datasets, respectively, using the GOR as the chemical representation: the COMPAS-1x dataset containing *cata*-condensed polybenzenoid hydrocarbons (cc-PBHs)[33] and a newly generated PAS dataset comprising a diverse set of heterocycle-containing PASs (see 'PAS dataset generation' section in Methods for further details). We then generated 1,000 molecules from each model and evaluated the success of the generation by: (1) validity—the percentage of valid molecules as measured by RDKit[34] (examples of invalid molecules can be seen in Supplementary Section 8); (2) novelty—the percentage of valid molecules not found in the training set; (3) uniqueness—the percentage of unique molecules among the valid molecules (this reflects the extent of repeated molecules in the batch, 100% uniqueness indicates each molecule appears only one time). Both novelty and uniqueness were calculated with the InChI representations, obtained by converting the GOR to a full molecular graph.

As shown in Table 1, both of our trained models succeeded in generating new molecules for each chemical space. Furthermore, nearly 100% of the generated molecules were valid, which is an improvement over the original implementation[16]. The difference is probably due to the GOR representation, which simplifies the learning (see Supplementary Sections 3 and 4 for more details). 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 unsurprising, considering the vastness of this chemical space.

We selected REINVENT[35] and GraphGA[14] to compare the performance of our EDM with other generative models; these were recently determined to be the two best-performing models for generative design of molecules with pharmaceutical relevance[36] (we note that this was evaluated by the accuracy versus the number of oracle calls, a metric that is not applicable in our case). Trained on the same PBH dataset, REINVENT achieved 65% validity and GraphGA achieved 11% validity of generated molecules. To provide additional context, we detail here the reported results of other generative models trained and tested on different chemical spaces: the unconditional generator G-SchNet[37] achieved 66–77% validity; crystal diffusion and variational autoencoder model[21], which also used a diffusion model, achieved validity of 88.9% of the generated molecules, of which 56.3% were unique; Laino and colleagues' variational autoencoder for catalyst generation[38], which adds a separate predictor network, achieved 84% validity and novelty; and Westermayr and colleagues' high-throughput iterative generative method[39] generally did not exceed 50% validity per iteration step. As each of these methods employed a different model, which was trained on and optimized for a different chemical space, these values only attest to the overall performance of the respective approaches and do not reflect a direct comparison of the models (such a comparison can be found in ref. 16).

Guided design of cc-PBHs

Single-objective target. In the next step, we used GaUDI to perform guided generation, that is, generation of molecules with desired properties. As an initial proof of concept, we focused on the simpler class, cc-PBHs, for which the COMPAS-1x dataset contains a variety of molecular properties, including the highest occupied molecular orbital (HOMO) energy, lowest unoccupied molecular orbital (LUMO) energy, HOMO–LUMO gap (HLG), relative energy (E_{rel}), adiabatic ionization potential (IP) and adiabatic electron affinity (EA).

To evaluate the conditioning method itself, we compared the performance of GaUDI with two other guided diffusion methods (all three methods used the combination of the basic EDM and GOR representation, allowing us to compare only the guidance algorithm): (1) point-wise conditional EDM (that is, standard conditioning)—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, and (2) equivariant energy-guided stochastic differential equations (EEGSDE)[40]—an approach for conditioning the diffusion process using score-based generative modeling through stochastic differential equations[25]. For GaUDI, we also evaluated the effect of the gradient scaling s, which allows us to tune the strength of the guidance.

All three models were conditioned on LUMO, HLG, E_{rel}, IP and EA, and tasked with generating ten-ring cc-PBHs with various combinations of target values for these properties, at different levels of difficulty:

https://doi.org/10.1038/s43588-023-00532-0

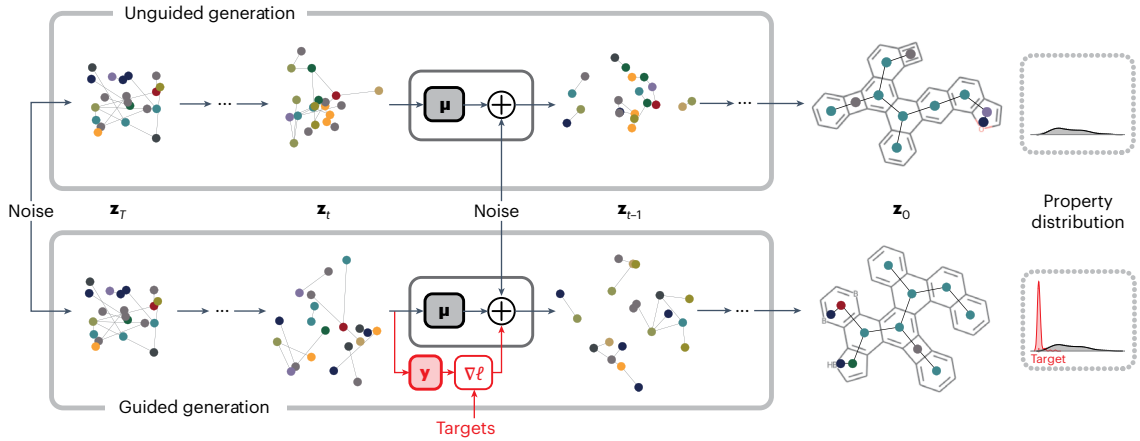

Fig. 1 | Generation workflow. Top: standard diffusion generation process, the noise is iteratively denoised using a neural model $\boldsymbol{\mu}$ until a clean sample $\mathbf{z}_0$ is generated. The iterates $\mathbf{z}_t$ collectively describe n rings and comprise coordinates determining the ring centroid and orientation (six values per ring) and hot-encoded categorical labels (visualized by different colors). 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 gradient $\nabla\ell$ of the target function is combined with the output of the denoiser for guidance. The graph of rings representation of the polycyclic aromatic chemical space is shown.

Table 1 | Performance of unguided generation

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

Reported are EDM models with GOR representation for batches of 1,000 molecules generated for each of the datasets.

1. **Joint distribution** (easy): a set of properties sampled from molecules in the test set.
2. **Marginal distribution** (moderate): a set of properties sampled from the product of marginal distributions of each property, as estimated on the training set. This is a harder task because the combination of the marginal property values might be infeasible.
3. **Real test case** (hard): the properties of pentacene (detailed in Fig. 2a). This is a difficult task because the likelihood of locating a ten-ring system with similar properties is small, as some of the properties are size-dependent.

Table 2 details the evaluation using the validity of the generated molecules and the mean absolute error (MAE) relative to the respective desired properties (calculated using a property-prediction network, described in Methods). 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. GaUDI also substantially 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, for example, the minimum or maximum. The COMPAS-1x dataset includes all of the cc-PBH molecules containing up to 11 rings, which allowed us to design 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 (Fig. 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, which would lead to a lower E_{rel} (E_{rel} is calculated for each PBH as the energy of the molecule relative to its lowest-energy isomer[33]; for PBHs of the same size, the reference molecule is the same). The target function for this purpose was defined as the mean square error of the properties LUMO, HLG, IP and EA between the generated molecule and pentacene plus E_{rel}:

$$L = (\mathrm{LUMO} - \mathrm{LUMO}_{\mathrm{target}})^2 + (\mathrm{HLG} - \mathrm{HLG}_{\mathrm{target}})^2 + (\mathrm{IP} - \mathrm{IP}_{\mathrm{target}})^2 + (\mathrm{EA} - \mathrm{EA}_{\mathrm{target}})^2 + E_{\mathrm{rel}}.$$

Our expectation, based on our experience with cc-PBHs, was that the generated molecules would contain a pentacene moiety (five linearly annulated rings) because, as we have previously shown, the majority of electronic properties of cc-PBHs are determined by the longest linear motif[32,33,41].

As shown in Fig. 2b, the target-function distribution ranges from 0 to 12.5. Prior to generation, we arbitrarily selected a cutoff value of 0.5 and identified the molecules with target-function values lower than this cutoff (a total of ten molecules, 0.03% of the dataset). We then removed these ten molecules with the lowest target-function values from the training sets of the diffusion and prediction models. Importantly, the same molecules had both the lowest calculated and predicted target-function values. We then had GaUDI generate samples of 512 cc-PBHs using the described target function and various gradient scaling values. As seen in Fig. 2b, setting s to zero (meaning, unguided generation/no conditioning) afforded a distribution almost equal to the dataset distribution, and the distribution shifted toward increasingly lower target-function values as the value of s was increased. These results demonstrate 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. As can be seen, the molecules generated by GaUDI (Fig. 2c, generated using the described target function and $s = 1$) do indeed have the expected pentacene moiety. Furthermore, all ten molecules with the

https://doi.org/10.1038/s43588-023-00532-0

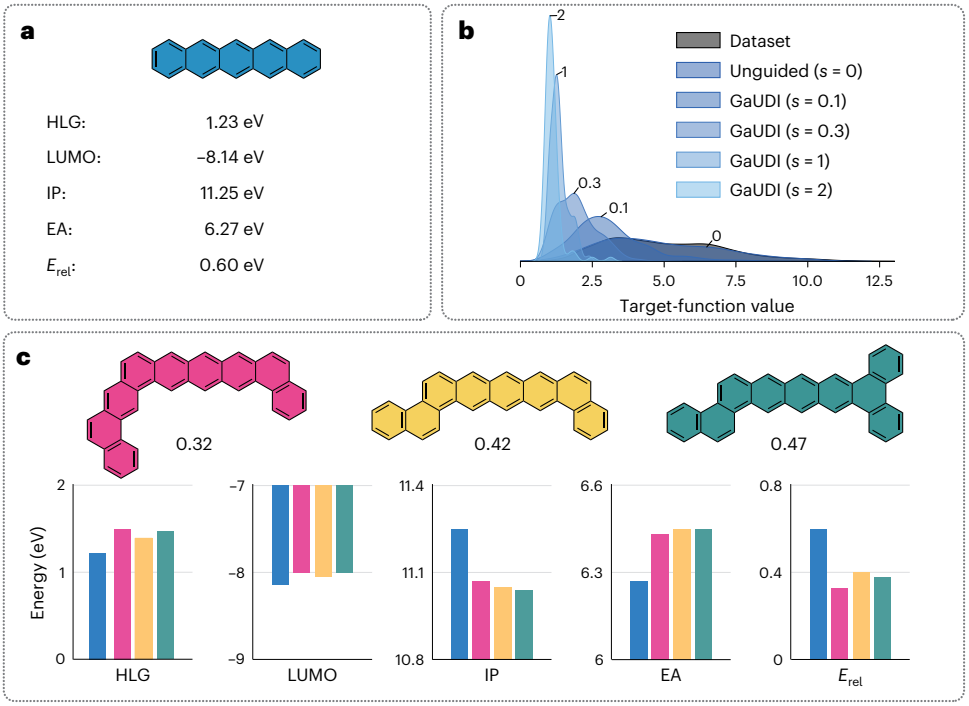

Fig. 2 | Guided generation of cc-PBH molecules to global minimum. **a**, Pentacene and its molecular properties calculated with GFN2-xTB. **b**, Distributions of the target-function values for the COMPAS-1x dataset and for cc-PBH samples generated by GaUDI with different gradient scalar values (individual distributions are normalized to have a unit sum). **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 E_{rel} (using $s = 1$). The target-function value of each molecule is displayed below the molecule. The individual properties of each molecule are denoted on the bar plots, color-coded in the same colors as the molecules; the left-most (blue) bar represents the value of pentacene.

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
EDM standard	59.3%	0.090	12.9%	0.133	0%	—
EEGSDE[40]	84%	0.158	78%	0.149	90.1%	0.301
GaUDI (s=0.1)	96.8%	0.109	84.3%	0.165	91.1%	0.256
GaUDI (s=0.3)	77.3%	0.074	80.4%	0.131	87.3%	0.241
GaUDI (s=1)	75%	0.056	65.2%	0.119	82.2%	0.211
GaUDI (s=2)	36.7%	0.039	42.1%	0.107	40.6%	0.183

MAE values are the average of the individual MAEs for the respective properties in the respective normalized spaces and are unitless.

lowest target-function values were present in the sample obtained with $s = 1$, indicating that GaUDI did in fact reach the global minimum of the declared target.

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. The PAS dataset we generated contains approximately 475,000 molecules and covers only a tiny fraction of this chemical space, which is infeasible to enumerate exhaustively; PASs therefore 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 dataset? To investigate this, we first focused on a single property, the HLG, and tasked GaUDI with generating molecules with high HLG values. We performed several runs to avoid bias in terms of molecular size, each time constraining the generation to structures with a different numbers of rings. In Fig. 3a,b, we show three pairs of PASs of equal size (six-, eight- and ten-ring systems, respectively); for each pair, the molecule on the left is the best PAS found in the dataset and the molecule on the right is the best PAS designed by GaUDI (with 'best' defined as having the highest HLG; properties were calculated with GFN2-xTB, the same level as the original dataset). GaUDI consistently returned novel structures with higher HLGs than found in the dataset. The generality of these results can be seen in Fig. 3c, where we show the effect of the scaling factor s on a series of generation batches (only ten-ring systems were generated for uniformity between the comparisons). The distribution plots show that increasing s afforded molecules with increasingly higher HLG values (meaning, lower target-function values). Notably, the distribution can indeed be pushed beyond the boundaries of the property distribution of the dataset, thus GaUDI can design better molecules than those in the original data. Interestingly, it seems that the presence of five-membered heterocycles pushes the HLG up. In particular, multiple furan moieties are recurring motifs in the high-HLG structures. Oligofuran molecules have already been recognized as promising compounds for organic electronics[42,43].

Multiproperty target. Having shown successful single-property guided design, we moved to the more challenging task of optimizing several properties at once, tasking GaUDI with generating molecules with a small HLG, low IP and high EA (a combination of properties relevant to narrow-band-gap molecules potentially suitable for photonics)[44]. We therefore defined the target function for this purpose as $\ell(\mathrm{HLG}, \mathrm{IP}, \mathrm{EA}) = 3 \times \mathrm{HLG} + \mathrm{IP} - \mathrm{EA}$, using a factor of 3 for the HLG

https://doi.org/10.1038/s43588-023-00532-0

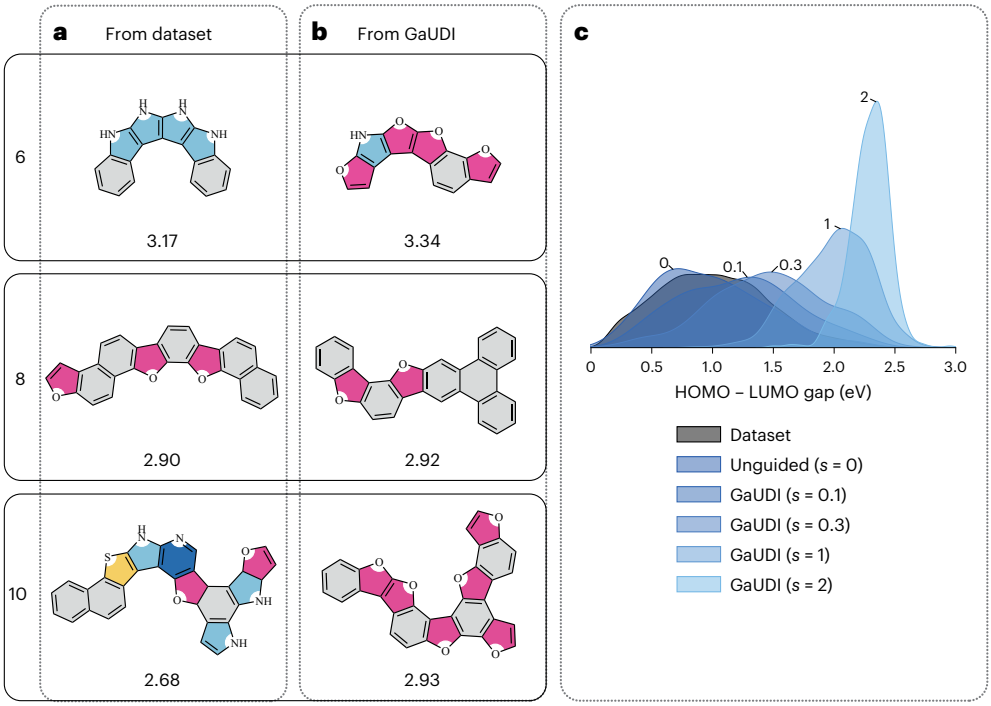

Fig. 3 | Guided design of PASs with high HLG values. a, The six-, eight- and ten-ring PASs with the highest HLG values in the dataset. **b**, Selected examples of six-, eight- and ten-ring PASs with a high HLG designed by GaUDI. Gray, benzene; blue, pyrrole; light blue, pyrazine; magenta, furan; yellow, thiophene. **c**, Distributions of HLG values for the PAS dataset and for samples generated by GaUDI with different gradient scalar values (individual distributions are normalized to have a unit sum). The distributions plotted include data only from valid molecules.

property to better balance the properties, which have different value ranges (IP = 10–12 eV; EA = 5–7 eV; HLG = 0–3 eV). There are no specific guidelines to constructing the target function. It is possible to change the factors to vary the relative importance of the different properties and thus achieve different generation outcomes. We note also that the E_{rel} property is not applicable for the PASs due to the heterogeneity in the numbers of atoms and elements. In Fig. 4 we compare the best molecules in the dataset with GaUDI-designed PASs (as determined by their target-function value). 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 ten-ring PASs in our dataset (70,000 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 ×861 enrichment, substantially increasing the likelihood of identifying promising candidate molecules for 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[45–48].

Discussion

Molecular design has many inherent challenges—such as discrete chemical space, bonding rules and so on—that make it an exceedingly difficult inverse design problem. Our presented method, GaUDI, provides an efficient tool for design of unknown molecules with desired molecular properties.

As with many generative models, one of the limitations of GaUDI is the feasibility of the designed molecules. It seems that many of the molecules proposed by GaUDI are relatively simple and reasonably feasible to synthesize (for example, Fig. 3b). Further improvements could be made by adding constraints to avoid known problematic motifs and/or by incorporating a synthesizability score in the target function (we note no such score currently exists for PASs). Furthermore, to improve the performance of the guided generation, an iterative process (as in ref. 39) could be implemented, whereby the properties and structures of generated molecules are calculated on-the-fly by an inexpensive method and added to the training set.

Many generative models have been proposed and have shown varying degrees of success. The advantages of GaUDI are the flexibility of its target function, its decoupling of the generator from the property predictor network and its remarkably high validity of generated molecules. This high validity is largely thanks to the GOR representation, which makes the rules the model needs to learn much simpler. As an added advantage, this means that fewer data are needed for training. We emphasize that the GOR is, in essence, a coarse-graining approach and, as such, is generalizable to other chemical systems for which a library of building blocks can be defined. For example, defining the rings as amino acids or carbohydrates would allow the user to generate oligopeptides or glycans, following suitable training. It is also possible to define a library of varying motifs (for example, a combination of rings, functional groups and individual atoms). As GaUDI only receives the centroids of the rings as input, it has no chemical knowledge pertaining to the types of bonds between the nodes. One must insert that information when translating the GOR back into a complete molecular structure. Further features such as orientation or functionalization/substituents can also be incorporated into the representation. The approach can therefore be generalized to other chemical spaces. Similarly, the model described in this work can also 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 and so on.

https://doi.org/10.1038/s43588-023-00532-0

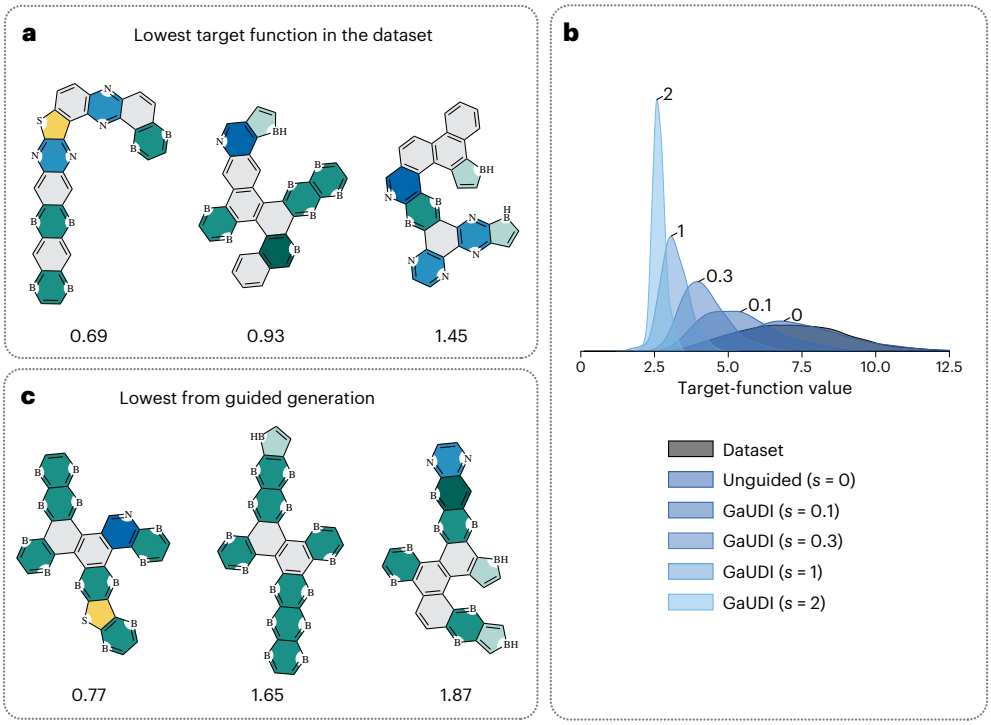

Fig. 4 | Guided design of narrow-band-gap molecules. a, The molecules with the lowest target-function values in the PAS dataset. Gray, benzene; dark blue, pyridine; light blue, pyrazine; dark green, borinine; green, 1,4-diborinine; light green, borole; yellow, thiophene. **b**, Distributions of the target function for the dataset and for GaUDI-generated batches with different gradient scalar values (individual distributions are normalized to have unit sum). The distributions plotted include data only from valid molecules. **c**, The molecules with the lowest target-function values designed by GaUDI.

GaUDI's ability to propose new molecules with desired properties, even beyond those in the initial training set, contributes to the acceleration of molecular design and discovery in numerous areas of interest, including but not limited to organic electronics and optoelectronics. Future directions include applying GaUDI to design functionalized PASs and *peri*-condensed PASs. We are also exploring an alternative approach in which GaUDI completes a given substructure to a final molecule with targeted properties.

Methods

Data

Two datasets were used: the COMPAS-1x dataset[33] from the COMPAS project and a new PAS dataset we prepared for this work. COMPAS-1x contains the GFN2-xTB-calculated structures and properties of 34,072 cc-PBHs comprising 1–11 rings. Due to the relatively small chemical space and high homogeneity of the cc-PBHs, COMPAS-1x contains all *cata*-condensed isomers that are possible for a given number of rings, up to 11 rings, using the enumeration scheme implemented by Brinkmann and colleagues in the CaGe software[49]. The data in this dataset have been described in further detail elsewhere[33]. The PAS dataset contains the GFN2-xTB-calculated structures and properties[50,51] of 474,174 PASs comprising 1–10 rings. The PASs in this dataset are built from 11 types of aromatic rings, including heterocyclic components. We refer the reader to the next section for further details on the PAS dataset.

PAS dataset generation

The PAS dataset was constructed using a library of 11 types of aromatic and heteroaromatic rings, ranging in size from four- to six-membered: cyclobutadiene, 1H-borole, pyrrole, furan, thiophene, borinine, 1,4-diborinine, 1,4-dihydro-1,4-diborinine, pyridine, pyrazine and benzene (see Supplementary Fig. 1). These building blocks were chosen due to their prevalence in organic molecules, in particular those displaying favorable optoelectronic activity.

To limit the data generation to a tractable region of chemical space, we focused only on molecules containing up to ten rings and only on *cata*-condensed systems (that is, each carbon can belong to at most two rings). We also enabled the placement of heteroatoms only on non-fused bonds (this limits the generated chemical spaces and simplifies the subsequent analysis by keeping a consistent formal charge for all systems). To generate the PAS molecules, a SMARTS-based[52] (SMILES[53] arbitrary target specification) enumeration was devised to incorporate these design principles and construct these molecules in a memory-efficient manner. To ensure diversity and remove bias, we randomly selected the type and number of heterocycles chosen for every molecule. Redundant entries were eliminated using their InChi[54] descriptors.

A ratio of 10:1 of benzene to all of the other types of rings was used, as is evident in Supplementary Fig. 2a, which presents the prevalence of the respective rings in the final dataset. Supplementary Fig. 2a also demonstrates that the prevalence of the rings is equal (with the exception of benzene), which implies that the dataset is not biased toward specific heterocyclic building blocks. In Supplementary Fig. 2b we show the distribution of molecule sizes (as measured by the number of rings in the molecule). The largest subset is that of molecules comprising nine rings, followed by eight- and ten-ring systems. As the number of rings is increased, the number of possible combinations grows. It is thus expected that the number of molecules increases with the number of rings. The ten-ring family is smaller only because we limited the number of entries, due to the higher computational cost of calculating these molecules. Finally, the inherently random nature of our generation procedure can be observed in Supplementary Fig. 2c,d, in which we show the respective distributions of benzene rings and heterocycles present per molecule. As can be seen, the molecules in the dataset span the entire range between 0 and 10 heterocycles/benzenes in a given molecule, which is yet another indication of the diversity of the data.

Following enumeration, we used RDKit[34] to generate 3D structures for the molecules, which were subsequently subjected to Riniker and colleagues' experimental-torsion and basic knowledge distance geometry)[55] and universal force field-based[56] pre-optimization. After pre-optimizing the molecular structures, they underwent a further optimization procedure using GFN2-xTB. Duplicates and molecules that failed to converge to reasonable structures were removed. Furthermore, molecules that failed to converge—or converged to structures inconsistent with the intended Lewis structure—were also removed. The final PAS dataset used in this work comprises 474,174 unique entries (corresponding to 22,414 molecular formulae), with geometries and properties obtained at the GFN2-xTB level of theory.

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 (meaning, graph of atoms). In our previous work[32], we introduced the GOR representation for PBHs. In the GOR (which can be seen in Fig. 1 and in Supplementary Section 3), each node represents a ring (the coordinates of the node are the centroid of the ring). 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, for example, 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 graph of atoms, 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, although 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[32].

We note that the GOR input for GaUDI is a point cloud and no connectivity information is provided. The bonds between neighboring nodes are only used for visualization (for example, Fig. 1) and for translation of the GOR into a full molecular graph. The presence of a bond is determined by the node–node distance, where distances within certain thresholds are considered to be bonded. These thresholds are determined based on the distribution of ring-pair distances extracted from the dataset of optimized geometries.

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[17] are a class of powerful likelihood-based generative models that have recently been shown to outperform generative adversarial networks[57] in image generation tasks[23]. Diffusion models generate samples by gradually removing noise from a signal and their training objective can be expressed as a reweighed variational lower bound[17].

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 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 $\mathbf{z}_0 \approx q_0$. The transition probability $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$ is approximated using a neural network of the form

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

where the vector $\boldsymbol{\theta}$ denotes the learnable parameters of the neural network $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, and $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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 of diffusion models is available in Supplementary Section 1.

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 $E(3)$ EDM[16] employing the $E(n)$ equivariant graph neural network (EGNN)[58] to satisfy the desired properties of $p_{\boldsymbol{\theta}}(\mathbf{z}_{t-1}|\mathbf{z}_t)$ and, consequently, of q_0.

Conditional generation

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.[16] showed that, in practice, such an approach has ample space for improvement[16]. 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[23] and adopted it due to its simplicity[23]. Nevertheless, it is important to note that Song et al.[25] developed a similar approach from a very different perspective[25]. In classifier guidance, 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

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

It is typically intractable to sample from this distribution exactly, but it has been shown that it can be approximated as a perturbed Gaussian distribution[59]: 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 equation (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 the work by Dhariwal and Nichol[23].

Dhariwal and Nichol[23] only considered the scenario in which generation is guided 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 Supplementary 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 toward 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. We therefore 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 of our experiments, we implemented the time-conditioned prediction model using the same EGNN[58] architecture as the network used to approximate the diffusion dynamics, and trained it by minimizing

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

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.

Models and training hyperparameters

Diffusion model hyperparameters. We trained the EDM[16] model with the Adam[60] optimizer with a learning rate of 0.0001 for 1,000 epochs and 1,000 timestamps. The dynamics in EDM are approximate with an EGNN network[58]. We used an EGNN with nine layers, each with 192 features.

Prediction model hyperparameters. For the prediction model, we used an EGNN network[58]. We train it with the Adam[60] optimizer with a learning rate of 0.0006 for 1,000 epochs. We used an EGNN with 12 layers, each with 192 features.

Data availability

All data for cc-PBHs used in this project were obtained from the COMPAS project[33], a freely available data repository at https://gitlab.com/porannegroup/compas. All PAS data are available free of charge at https://doi.org/10.5281/zenodo.7798697 (ref. 61). Source Data are provided with this paper.

Code availability

All codes used to train the models and generate molecules are provided free of charge at https://gitlab.com/porannegroup/gaudi (minted version https://doi.org/10.5281/zenodo.8311764)[62]. The repository also contains an original tutorial for generating GOR representations of PASs and for generating new PASs with user-defined target functions.

References

1. Hwang, J. et al. Perovskites in catalysis and electrocatalysis. *Science* **358**, 751–756 (2017).
2. Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R. & Jensen, K. F. Generative models for molecular discovery: recent advances and challenges. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **12**, e1608 (2022).
3. Fuhr, A. S. & Sumpter, B. G. Deep generative models for materials discovery and machine learning-accelerated innovation. *Front. Mater.* https://doi.org/10.3389/fmats.2022.865270 (2022).
4. Walters, W. P. & Barzilay, R. Applications of deep learning in molecule generation and molecular property prediction. *Acc. Chem. Res.* **54**, 263–270 (2020).
5. Anstine, D. M. & Isayev, O. Generative models as an emerging paradigm in the chemical sciences. *J. Am. Chem. Soc.* **145**, 8736–8750 (2023).
6. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018).
7. Shree Sowndarya, S. V. et al. Multi-objective goal-directed optimization of de novo stable organic radicals for aqueous redox flow batteries. *Nat. Mach. Intell.* **4**, 720–730 (2022).
8. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Sci.* **4**, 268–276 (2018).
9. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
10. Putin, E. et al. Reinforced adversarial neural computer for de novo molecular design. *J. Chem. Inform. Model.* **58**, 1194–1204 (2018).
11. Prykhodko, O. et al. A de novo molecular generation method using latent vector based generative adversarial network. *J. Cheminform.* **11**, 1–13 (2019).
12. Jennings, P. C., Lysgaard, S., Hummelshøj, J. S., Vegge, T. & Bligaard, T. Genetic algorithms for computational materials discovery accelerated by machine learning. *npj Comput. Mater.* **5**, 46 (2019).
13. Henault, E. S., Rasmussen, M. H. & Jensen, J. H. Chemical space exploration: how genetic algorithms find the needle in the haystack. *Peer J. Phys. Chem.* **2**, e11 (2020).
14. Jensen, J. H. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **10**, 3567–3572 (2019).
15. Shi, C. et al. GraphAF: a flow-based autoregressive model for molecular graph generation. Preprint at https://arxiv.org/abs/2001.09382 (2020).
16. Hoogeboom, E., Satorras, V. G., Vignac, C. & Welling, M. Equivariant diffusion for molecule generation in 3D. In *Proc. 39th International Conference on Machine Learning* 8867–8887 (ML Research Press, Cambridge, 2022).
17. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. In *Proc. 34th International Conference on Neural Information Processing Systems* 6840–6851 (Curran Associates Inc., Red Hook, 2020).
18. Ho, J. et al. Video diffusion models. Preprint at https://arxiv.org/abs/2204.03458 (2022).
19. Austin, J., Johnson, D. D., Ho, J., Tarlow, D. & van den Berg, R. Structured denoising diffusion models in discrete state-spaces. In *Proc. 35th Conference on Neural Information Processing Systems* 17981–17993 (Curran Associates Inc., Red Hook, 2021).
20. Xu, M. et al. GeoDiff: a geometric diffusion model for molecular conformation generation. Preprint at https://arxiv.org/abs/2203.02923 (2022).
21. Lyngby, P. & Thygesen, K. S. Data-driven discovery of 2D materials by deep generative models. *npj Comput. Mater.* **8**, 232 (2022).
22. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. DiffDock: diffusion steps, twists, and turns for molecular docking. Preprint at https://arxiv.org/abs/2210.01776 (2022).
23. Dhariwal, P. & Nichol, A. Diffusion models beat GANs on image synthesis. In *Proc. 35th Conference on Neural Information Processing Systems* 8780–8794 (Curran Associates Inc., Red Hook, 2021).
24. Ho, J. & Salimans, T. Classifier-free diffusion guidance. Preprint at https://arxiv.org/abs/2207.12598 (2022).
25. Song, Y. et al. Score-based generative modeling through stochastic differential equations. Preprint at https://arxiv.org/abs/2011.13456 (2021).
26. Balaban, A. T., Oniciu, D. C. & Katritzky, A. R. Aromaticity as a cornerstone of heterocyclic chemistry. *Chem. Rev.* **104**, 2777–2812 (2004).
27. Li, Q. et al. Polycyclic aromatic hydrocarbon-based organic semiconductors: ring-closing synthesis and optoelectronic properties. *J. Mater. Chem. C* **10**, 2411–2430 (2022).

28. Aumaitre, C. & Morin, J.-F. Polycyclic aromatic hydrocarbons as potential building blocks for organic solar cells. *Chem. Rec.* **19**, 1142–1154 (2019).
29. Kilaru, S. et al. Organic materials based on hetero polycyclic aromatic hydrocarbons for organic thin-film transistor applications. *Mater. Sci. Semicond. Process.* **147**, 106730 (2022).
30. Omar, Ö. H., Del Cueto, M., Nematiaram, T. & Troisi, A. High-throughput virtual screening for organic electronics: a comparative study of alternative strategies. *J. Mater. Chem. C* **9**, 13557–13583 (2021).
31. Das, S., Bhauriyal, P. & Pathak, B. Polycyclic aromatic hydrocarbons as prospective cathodes for aluminum organic batteries. *J. Phys. Chem. C* **125**, 49–57 (2020).
32. Weiss, T., Wahab, A., Bronstein, A. M. & Gershoni-Poranne, R. Interpretable deep-learning unveils structure–property relationships in polybenzenoid hydrocarbons. *J. Organic Chem.* https://doi.org/10.1021/acs.joc.2c02381 (2023).
33. 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.* **62**, 3704–3713 (2022).
34. Landrum, G. et al. *RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling* (RDKit, 2013).
35. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9**, 1–14 (2017).
36. Gao, W., Fu, T., Sun, J. & Coley, C. Sample efficiency matters: a benchmark for practical molecular optimization. In *Proc. 36th Conference on Neural Information Processing Systems* 21342–21357 (Curran Associates Inc., Red Hook, 2022).
37. Gebauer, N., Gastegger, M. & Schütt, K. Symmetry-adapted generation of 3D point sets for the targeted discovery of molecules. In *Proc. 33rd Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, 2019).
38. Schilter, O., Vaucher, A., Schwaller, P. & Laino, T. Designing catalysts with deep generative models and computational data. A case study for Suzuki cross coupling reactions. *Digit. Discov.* **2**, 728–735 (2023).
39. Westermayr, J., Gilkes, J., Barrett, R. & Maurer, R. J. High-throughput property-driven generative design of functional organic molecules. *Nat. Comput. Sci.* **3**, 139–148 (2023).
40. Bao, F. et al. Equivariant energy-guided SDE for inverse molecular design. Preprint at https://arxiv.org/abs/2209.15408 (2022),
41. 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.* **36**, e4458 (2022).
42. Gidron, O., Dadvand, A., Sheynin, Y., Bendikov, M. & Perepichka, D. F. Towards 'green' electronic materials. *α*-Oligofurans as semiconductors. *Chem. Commun.* **47**, 1976–1978 (2011).
43. Gidron, O. & Bendikov, M. *α*-Oligofurans: an emerging class of conjugated oligomers for organic electronics. *Angew. Chem. Int. Ed.* **53**, 2546–2555 (2014).
44. Li, X.-H. et al. Narrow-bandgap materials for optoelectronics applications. *Front. Phy.* **17**, 1–33 (2022).
45. Agnoli, S. & Favaro, M. Doping graphene with boron: a review of synthesis methods, physicochemical characterization, and emerging applications. *J. Mater. Chem. A* **4**, 5002–5025 (2016).
46. 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. *Chem. Mater.* **31**, 1891–1898 (2019).
47. Stoycheva, J. et al. Boron-doped polycyclic aromatic hydrocarbons: a molecular set revealing the interplay between topology and singlet fission propensity. *J. Phys. Chem. Lett.* **11**, 1390–1396 (2020).
48. Kothavale, S. S. & Lee, J. Y. Three-and four-coordinate, boron-based, thermally activated delayed fluorescent emitters. *Adv. Optical Mater.* **8**, 2000922 (2020).
49. Brinkmann, G., Grothaus, C. & Gutman, I. Fusenes and benzenoids with perfect matchings. *J. Math. Chem.* **42**, 909–924 (2007).
50. 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). *J. Chem. Theory Comput.* **13**, 1989–2009 (2017).
51. 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. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
52. *SMARTS—A Language for Describing Molecular Patterns* (Daylight Chemical Information Systems, 2007).
53. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Comput. Sci.* **28**, 31–36 (1988).
54. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI—the worldwide chemical structure identifier standard. *J. Cheminform.* **5**, 1–9 (2013).
55. Riniker, S. & Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inform. Model.* **55**, 2562–2574 (2015).
56. Rappé, A. K., Casewit, C. J., Colwell, K., Goddard III, W. A. & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).
57. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
58. Satorras, V. G., Hoogeboom, E. & Welling, M. *E*(n) equivariant graph neural networks. In *Proc. 38th International Conference on Machine Learning* 9323–9332 (ML Research Press, Cambridge, 2021).
59. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. 32nd International Conference on Machine Learning* 2256–2265 (ML Research Press, Cambridge, 2015).
60. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
61. Weiss, T., Mayo-Yanes, E., Chakraborty, S. & Gershoni-Poranne, R. PASs molecular dataset. Zenodo https://doi.org/10.5281/zenodo.7798697 (2023).
62. Weiss, T. GaUDI—2/9/2023. Zenodo https://doi.org/10.5281/zenodo.8311764 (2023).

Acknowledgements

We thank A. Wahab (ETH Zurich) for assistance with implementing the RDKit validity code and for proofreading the paper. We also thank A. Tsybizova (ETH Zurich) for proofreading and for providing helpful comments on the clarity of the text. We gratefully acknowledge P. Chen (ETH Zurich) for his scientific support and mentorship. E.M.Y., S.C. and R.G.P. are grateful for the financial support of the Branco Weiss Fellowship (awarded to R.G.P). R.G.P. is a Branco Weiss Fellow and a Horev Fellow. A.M.B. and T.W. were partially supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 863839) and by the Council For Higher Education - Planning & Budgeting Committee. L.C. is supported by the IRIDE grant from DAIS, Ca' Foscari University of Venice.

Author contributions

R.G.P. and A.M.B. conceived the original idea and designed and supervised the research project. T.W., L.C. and A.M.B. designed the generative and predictive models. T.W. wrote the code and trained the models. E.M.Y. and S.C. performed the quantum chemistry calculations. E.M.Y., S.C. and R.G.P. performed the dataset curation. T.W. and R.G.P. wrote the paper with the help of the other authors. The paper reflects the contributions of all authors.

Competing interests

All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-023-00532-0.

Correspondence and requests for materials should be addressed to Alex M. Bronstein or Renana Gershoni-Poranne.

Peer review information *Nature Computational Science* thanks Ganna Gryn'ova, Rocío Mercado, Rostislav Fedorov and the other, anonymous reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Kaitlin McCardle, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.