nature computational science

Article

https://doi.org/10.1038/s43588-022-00391-1

High-throughput property-driven generative design of functional organic molecules

Received: 27 July 2022

Accepted: 14 December 2022

Published online: 6 February 2023

Check for updates

Julia Westermayr[1,3] ✉, Joe Gilkes[1,2], Rhyan Barrett[1,3] & Reinhard J. Maurer[1] ✉

The design of molecules and materials with tailored properties is challenging, as candidate molecules must satisfy multiple competing requirements that are often difficult to measure or compute. While molecular structures produced through generative deep learning will satisfy these patterns, they often only possess specific target properties by chance and not by design, which makes molecular discovery via this route inefficient. In this work, we predict molecules with (Pareto-)optimal properties by combining a generative deep learning model that predicts three-dimensional conformations of molecules with a supervised deep learning model that takes these as inputs and predicts their electronic structure. Optimization of (multiple) molecular properties is achieved by screening newly generated molecules for desirable electronic properties and reusing hit molecules to retrain the generative model with a bias. The approach is demonstrated to find optimal molecules for organic electronics applications. Our method is generally applicable and eliminates the need for quantum chemical calculations during predictions, making it suitable for high-throughput screening in materials and catalyst design.

The search for new functional molecules and materials is often complicated by several criteria that must be simultaneously satisfied. For example, molecular materials tailored for organic electronics devices must be mechanically flexible, durable and synthetically accessible, while satisfying the relevant described electronic properties that govern the device functionality[1,2]. In addition to these often-competing requirements, it is not always clear how to systematically modify a molecular structure and composition to improve (multiple) properties. Simultaneous multiproperty optimization can be considered the holy grail in molecular and material design[2–4]. A better understanding of how functional groups in a molecule alter its physicochemical properties could, at least in principle, help to facilitate design studies. However, the combinatorial complexity of chemical space consisting of up to 10^{60} organic molecules and the many factors that must be considered often make this problem too complex for traditional optimization and basic heuristic reasoning[2,5]. Candidate identification based on simple structure–property relations and trial-and-error optimization remain the state of the art when it comes to developing new molecules and materials with specific property requirements[2,6].

One area of research where this problem has become apparent is the field of organic optoelectronics, which deals with devices that emit or detect light. Examples in which novel organic electronic materials play a role include sustainable energy sources (solar cells), organic light-emitting diodes, telecommunications, displays in smart devices and optical fibers, to name a few examples. Organic thin-film devices, composed of multiple organic layer components with different tailored properties, have become of particular importance to this research area[7,8]. To deliver new molecular materials for thin-film devices, their

[1]Department of Chemistry, University of Warwick, Coventry, UK. [2]HetSys Centre for Doctoral Training, University of Warwick, Coventry, UK. [3]Present address: Wilhelm-Ostwald-Institut für Physikalische und Theoretische Chemie, Universität Leipzig, Leipzig, Germany. ✉e-mail: julia.westermayr@uni-leipzig.de; r.maurer@warwick.ac.uk

electronic properties, such as the fundamental gap (ΔE), the electron affinity (EA) or the ionization potential (IP), must lie within a narrow window to satisfy the requirements of the device function.

Recently, generative deep learning has emerged as a promising solution for speeding up molecular design[2,9–11]. Generative deep learning is an unsupervised learning technique, in which deep learning models extract knowledge from a dataset of (molecular) geometries and apply the acquired rules to create new molecules with properties similar to those in the original dataset[2]. Several recent works have shown that such methods have the potential to markedly accelerate molecular and material discovery[2,3,9,11–14]; however, there is no guarantee that the generated molecular systems will exhibit properties within a relevant regime.

Unguided search in chemical space is extremely inefficient and fundamentally limits the diversity of structures that can be explored in high-throughput screening, particularly if the molecular generation process requires computationally demanding quantum chemical predictions of electronic properties. Even with hypothetically limitless computational resources, the characterization of generated molecules remains challenging. Several recent works have proposed property targeted generative workflows in the context of drug and molecular design[15–18]. Most generative models predict molecules via fragment-based structural descriptors such as SMILES strings that do not resolve the three-dimensional structure and conformation of molecules. Molecular generation can be guided by recursive workflows that use experimental reference data or quantum chemical calculations. In the latter case, the three-dimensional atomic configuration of the molecular equilibrium structure is required as input for quantum chemical calculations. Generative models that predict three-dimensional conformations of molecules have recently been proposed[3,19–22], yet the requirement of performing quantum chemical calculations introduces a bottleneck that limits the number of molecules that can be screened.

In this work, we propose an approach that delivers high-throughput guided search and design of functional organic molecules with tailored properties. The method achieves this by combining two machine learning algorithms. The first model is an unsupervised, generative autoregressive model that can use chemical rules learned from a structural distribution of molecules to create new, previously unknown three-dimensional equilibrium conformations of molecules. The second model is a supervised physics-inspired deep neural network that, given a three-dimensional structure, can predict the (charged) electronic excitations of functional organic molecules with close to experimental accuracy[23]. The latter eliminates the need for demanding quantum chemical calculations used in previous approaches. The approach presented here provides an automated workflow in which chemical space exploration can be biased towards the generation of molecules that satisfy preset design parameters. We demonstrate the ability to perform high-throughput property-guided molecular design in the context of organic electronics[24]. Key molecular properties relevant in optoelectronic materials targeted here are small ΔE, small IP and large EA[24,25]. Important molecular features that separate the optimal molecules with small ΔE from the rest of the explored molecules can be unveiled using dimensionality reduction techniques and unsupervised clustering algorithms. The trends we find and the rules we discover are verified with quantum chemistry, showing the potential of our method to discover hidden patterns in data. Finally, we provide an outlook for multiproperty optimization by simultaneously biasing the generative model towards systems with low ΔE and low synthetic complexity[26].

Results

Workflow

The proposed approach for automated molecular design is a combination of two deep learning techniques, illustrated in Fig. 1a. The process starts with training of a generative model on a set of molecular structures to learn underlying rules for building molecules that satisfy the same structural distribution and resemble the learned chemical space. The initially trained generative deep learning model is then used to predict a large number (in the range of several thousands to millions) of new molecules. A validity check of molecular structures is carried out and systems are filtered according to their structures: for example, duplicates or disconnected systems are discarded. For the structure generation, we use the generative, autoregressive deep neural network G-SchNet[3]. G-SchNet, in contrast to most generative models[2,27], is able to predict the structural composition and the three-dimensional conformation of molecules, which can serve as an input for electronic structure calculations and deep learning models of electronic structure.

The screening of molecular properties is facilitated with the deep neural network SchNet + H[23] to achieve high computational efficiency. SchNet + H predicts electronic excited states from equilibrium geometries that can be used to compute photoemission spectra with accuracy close to experiment. The high fidelity of the model is achieved by combining a deep learning model for molecular orbital energies obtained from density functional theory (DFT) and a Δ-machine learning model[28], meaning that the difference (Δ) between two levels of theory is learned, to correct these energies to the accuracy of many-body perturbation theory at the level of the *GW* method in the complete basis set limit. The *GW* method acts as a correction to DFT to account for many-body correlation and exchange effects[25]. As Fig. 1a shows, molecules can also be screened on the basis of other properties. In this work, molecules are additionally screened using another deep neural network capable of predicting the synthetic complexity of molecules, namely SCScore[26], which was trained on 12 million reactions from the Reaxys database[29]. The most promising molecules with properties that lie within a predefined target range are then used to bias the generative model, which can subsequently predict new molecules with electronic properties closer to the target[3,12,30]. By iteratively biasing the generative model, the properties of the predicted molecules can be pushed into unexplored regions.

We demonstrate the proposed workflow by training G-SchNet on the OE62 dataset of functional organic molecules. The OE62 data are composed of molecules with high chemical and structural diversity[24]. As can be seen in Fig. 1b, molecules can contain up to 16 different elements. They vary in size from 3 atoms to over 150 atoms. The distributions of ΔE, IP and EA of molecules in the OE62 dataset and of molecules generated by G-SchNet are shown in Fig. 1c. ΔE, IP and EA are important measures to characterize molecules applicable in organic electronic devices; in particular, molecules with small ΔE are interesting and often used in photonics or biomedicine[8], for instance. However, in the OE62 dataset there are not many molecules that exhibit small values of ΔE and IP or large values of EA in regimes that are typically considered relevant for organic electronics applications. Here, we demonstrate that, by iteratively biasing G-SchNet towards the desired property range, molecules can be designed that exhibit values of ΔE, IP and EA that lie outside the property distribution represented by the original training dataset.

Biasing towards desired electronic properties

The results obtained by iteratively biasing G-SchNet towards small ΔE, large EA and small IP are shown in Fig. 2. Panels a, c and e show the distribution of targeted electronic properties for a set of 40,000–90,000 predicted molecules in each iteration (Supplementary Section 1). In the first biasing step of each experiment, a small subset of molecules in the OE62 data with property values below or above a certain threshold (illustrated with shaded areas, corresponding to about 10% of the distribution) are used to retrain G-SchNet with a bias. The molecules are screened using SchNet + H, which is independently validated to accurately predict electronic properties of structures predicted by G-SchNet in Supplementary Section 2. SchNet + H has a mean absolute

https://doi.org/10.1038/s43588-022-00391-1

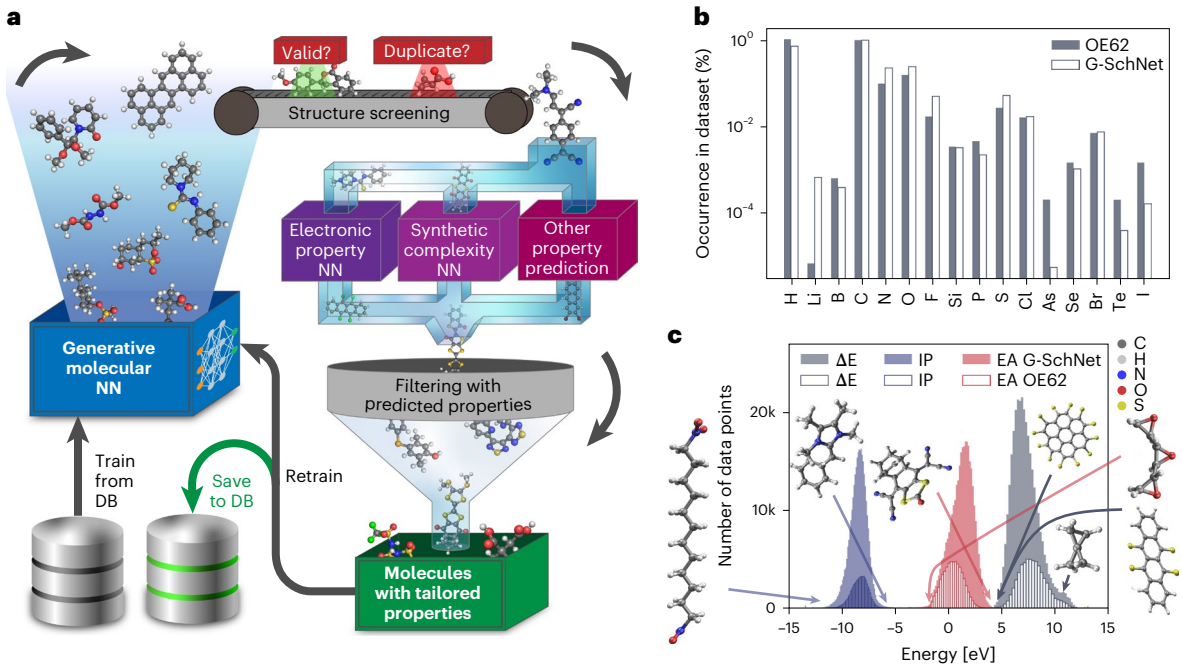

Fig. 1 | Workflow of the proposed method and distribution of molecules in the dataset. a, The proposed method starts by training the generative deep learning model G-SchNet on the OE62 dataset, which can then be applied to build three-dimensional conformations of unseen molecules. These are filtered on the basis of structure (for example, duplicates or disconnected structures are sorted out) and on the basis of electronic properties, synthesizability or other properties. In this work, we use SchNet + H to screen for small ΔE, small IP and large EA. In addition, we apply the SCScore neural network (NN) model to screen for molecules with low complexity in synthesizability. Selected molecules can be used to retrain and bias the generative model. **b**, Elemental distribution of molecules in the OE62 dataset and molecules predicted by G-SchNet. **c**, The distribution of ΔE, IP and EA in the OE62 dataset and in molecules predicted with G-SchNet. Example molecules that highlight the chemical and structural diversity of molecules in the dataset are shown in the plot. DB, database.

error for charged electronic excitations in the range of 0.25 eV with respect to the quantum chemistry reference, which we deem sufficiently accurate for high-throughput screening and identification of candidate molecules. To additionally ensure that G-SchNet is not misled by molecules that are inaccurately predicted with SchNet + H and thus wrongly assumed to fall into the category of molecule properties in the desired range, the variances of the electronic excitations inferred by two SchNet + H models are computed on the fly. Whenever the prediction variances are larger than their average mean absolute errors, we deem SchNet + H to be unreliable and the molecule is discarded; see Supplementary Section 3 for details.

As can be seen from Fig. 2a,c,e, after biasing and retraining of G-SchNet, molecules with a distribution of properties shifted towards the desired energy ranges can be generated. Interestingly, already after the first biasing steps, generated molecules exhibit electronic properties that lie outside of the original dataset. In the subsequent iteration, the molecules with properties at the edges of the distributions are extracted and used to train G-SchNet again. The exact number of molecules and the criteria to select molecules used for biasing are specified in Supplementary Section 4 and Supplementary Data 1. The molecular design process is terminated when the distribution of properties of proposed molecules as predicted by SchNet + H does not overlap anymore with the original distribution. This was after 7, 10 and 11 loops in the cases of ΔE, EA and IP, respectively.

To verify that the distributions outside the original dataset are not artifacts due to molecules outside of the training regime of SchNet + H, we recalculated ΔE, EA and IP at the G0W0 level of theory for 66, 79 and 33 molecules, respectively, that are extracted randomly from the last three iterations of each experiment. Indeed, we found that the molecules consistently had electronic properties that were not present in the original dataset. The smallest ΔE value of the extracted molecules is 3.2 eV, which is 1.6 eV smaller than the smallest value reported in the original dataset. The largest EA is 6.6 eV, which is 2.4 eV larger than the largest EA reported in the original dataset, and the smallest IP is 4.2 eV, 0.8 eV smaller than the smallest IP reported. The reference calculations and distribution of molecular properties are discussed in more detail in Supplementary Section 2 (Supplementary Fig. 3).

The fact that the generative model can produce molecules with ΔE, EA and IP values that are not reported in the original training set may seem surprising, but this can be explained by taking a look at the chemical space spanned by the molecules in the OE62 dataset and the structures predicted by G-SchNet (Fig. 2b,d,f). The chemical space represented and formed by the OE62 dataset is shown by two representative collective variables obtained from dimensionality reduction via principal component analysis (PCA) based on the Smooth Overlap of Atomic Positions structural descriptor[31] (see Dimensionality reduction of generated molecules for details). The molecules generated in each consecutive biasing loop are shown in the same chemical space and indicated by different colors. As can be seen, the generated molecules are contained within the structural space spanned by the original OE62 dataset. Interestingly, similar regions of chemical space are identified to be important for molecules that feature a small ΔE (panel b) and large EA (panel d). In contrast, a different region of chemical space is identified for molecules that predominantly exhibit a small IP (panel f).

Identification of bonding patterns

To correlate key bonding patterns in molecules with trends in electronic properties, we combined dimensionality reduction with clustering techniques and analyze which molecules feature small ΔE. For dimensionality reduction, PCA is applied to two types of descriptor, one that encodes bonding patterns and one that encodes structural distributions of molecules in the OE62 dataset and for the collection of all molecules generated during consecutive biasing iterations. Five principal components obtained from each descriptor type were used as

https://doi.org/10.1038/s43588-022-00391-1

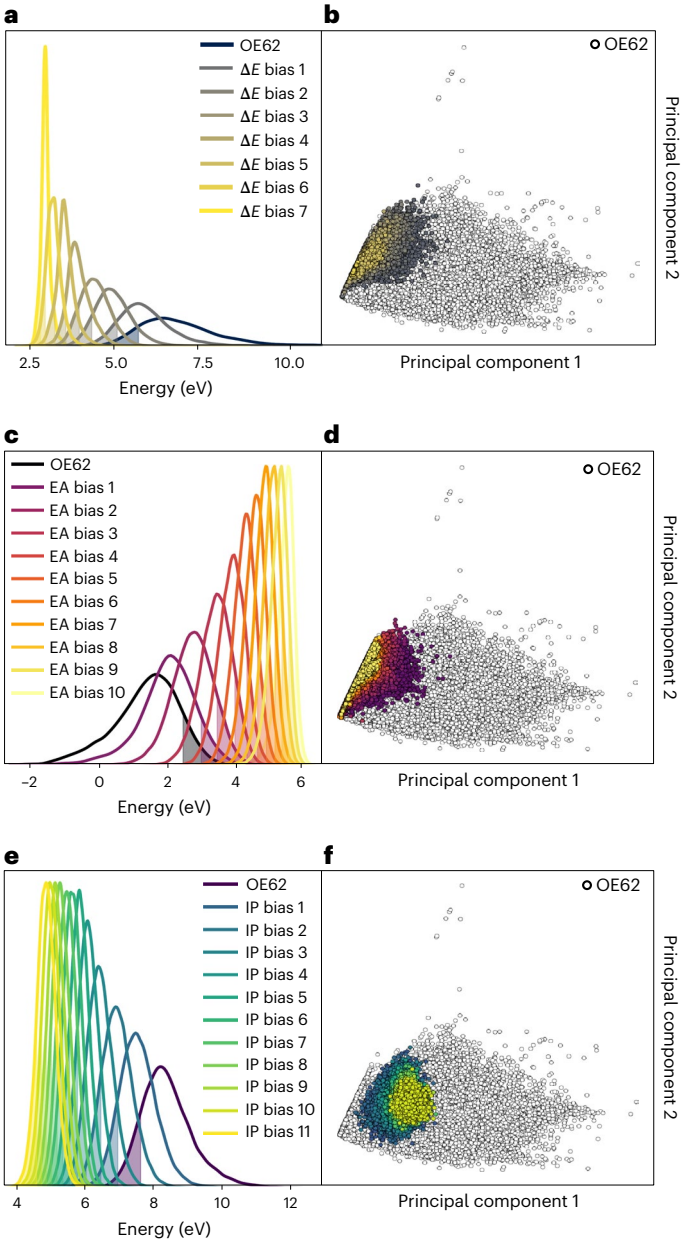

Fig. 2 | Distribution of electronic properties and structures of generated molecules. a,**c**,**e**, Distribution of Δ*E* (**a**), EA (**c**) and IP (**e**) after biasing towards small Δ*E* (**a**), large EA (**c**) and small IP (**e**). **b**,**d**,**f**, Distribution of data points in chemical space spanned by principal components obtained from OE62 data using structural descriptors. The color code indicates the biasing step.

an input for clustering analysis, which contained over 98% of the variance in the data. Further principal components were not required as they would make negligible contribution to the analysis. For clustering, we used BIRCH[32] to find cluster centroids coupled with agglomerative clustering[33]. Details of descriptors, PCA and clustering analysis can be found in Methods. Data points plotted along the first principal components obtained from the structural descriptor against Δ*E* are shown in Fig. 3, where colors in panel a indicate iterations and colors in panel b indicate subclusters found across iterations.

Manual inspection of the centroids of the subclusters indicated that an increased number of cyano groups (−C≡N) is present in molecules with small Δ*E*. This trend can also be observed for representative molecules plotted next to Fig. 3b and is quantified in Fig. 3c. While in the original dataset mostly C−N single bonds are present and only a few molecules have C≡N triple bonds, molecules generated during the last loops mainly contain C≡N triple bonds. To analyze whether this trend is sensitive to the original training dataset, that is, to find out whether G-SchNet still predicts a high content of cyano groups in molecules optimized for small Δ*E* even if they are not contained in the original data, we eliminated all C≡N triple bonds from the original training dataset and performed a knockout study. The modified OE62 dataset was used to train another G-SchNet model, which was then applied in a separate experiment to generate molecules with iteratively smaller Δ*E*. Already after the first loop, G-SchNet based on the knockout dataset generated molecules with an increased number of C≡N triple bonds (Supplementary Fig. 9). The reason that G-SchNet can recover some functional groups not contained in the original dataset lies in the nature of the SchNet descriptor, which is represented by a set of continuous atom-centered filter functions trained to optimally represent the data. These functions encode the probability of finding atoms at a certain distance and constellation around each atom. The G-SchNet model thus has some likelihood of also generating shorter CN bonds with different coordination than in the training set, which are then enhanced in the distribution via the biasing approach with SchNet + H.

Further, quantitative analysis of elemental composition of molecules generated by later loops revealed a notable increase of sulfur and selenium content in molecules with small Δ*E*. The respective percentages are depicted in Fig. 3d–g. As can be seen from panel g, when sulfur and selenium content rises the oxygen content falls, which indicates replacement of oxygen by sulfur or selenium. To investigate whether this result is an artifact of our models or a real trend that leads to small Δ*E*, we carried out 144 quantum chemical calculations (see Quantum chemistry calculations for details) of molecules with oxygen atoms replaced by sulfur and selenium and compared their HOMO (highest occupied molecular orbital)–LUMO (lowest unoccupied molecular orbital) gaps as approximate analogs of Δ*E* (refs. [23,24]). Our results clearly indicate that replacing one or all oxygen atoms with sulfur reduces the HOMO–LUMO gap on average by 0.5 eV and 1.1 eV, respectively. Further replacement of sulfur by selenium additionally decreases the HOMO–LUMO gap by 0.2 eV in both cases, hence on average a decrease in HOMO–LUMO gap by 0.7 eV and 1.3 eV can be found when replacing one or all oxygen atoms with selenium atoms. The effect of selenium to promote photoconducting properties was reported as early as 1873 (ref. [34]).

Molecules predicted by G-SchNet contain unusually high concentrations of selenium and sulfur atoms as well as cyano groups, considering that they were generated from a base distribution of known crystal-forming organic molecules. To find out if such molecules are used in real applications, a literature search with SciFinder (https://scifinder-n.cas.org) was conducted (Supplementary Fig. 8). In addition to the literature search, we parsed all molecules from the final three loops and compared them with approximately 250,000 small aromatic molecules considered applicable to organic electronics[35] by using the Tanimoto similarity measure[36] (see Similarity analysis of molecules for details). The findings suggest that the identified molecules contain structural motifs, such as tetrathiafulvalenes[37] and (selenium-enriched) trithiapentacene derivatives[38] shown in bold in panel b, that are frequently mentioned in literature relevant to organic molecular electronics[39], especially in the context of (dye-sensitized) solar cells[40,41], for synthesis of organic electronic materials, electroluminescent materials[42,43] or single-molecule switches.

As it is evident from the results above and Fig. 3c–g, G-SchNet changes the relative distribution of elements and bonding patterns to shift the electronic properties into the desired range of small Δ*E*. In doing so, molecules are generated that feature known structural motifs that are already in use in organic electronics.

However, the property-based biasing approach also comes with downsides. While the biased generative method successfully creates molecules with desirable properties, as iterations progress the method

https://doi.org/10.1038/s43588-022-00391-1

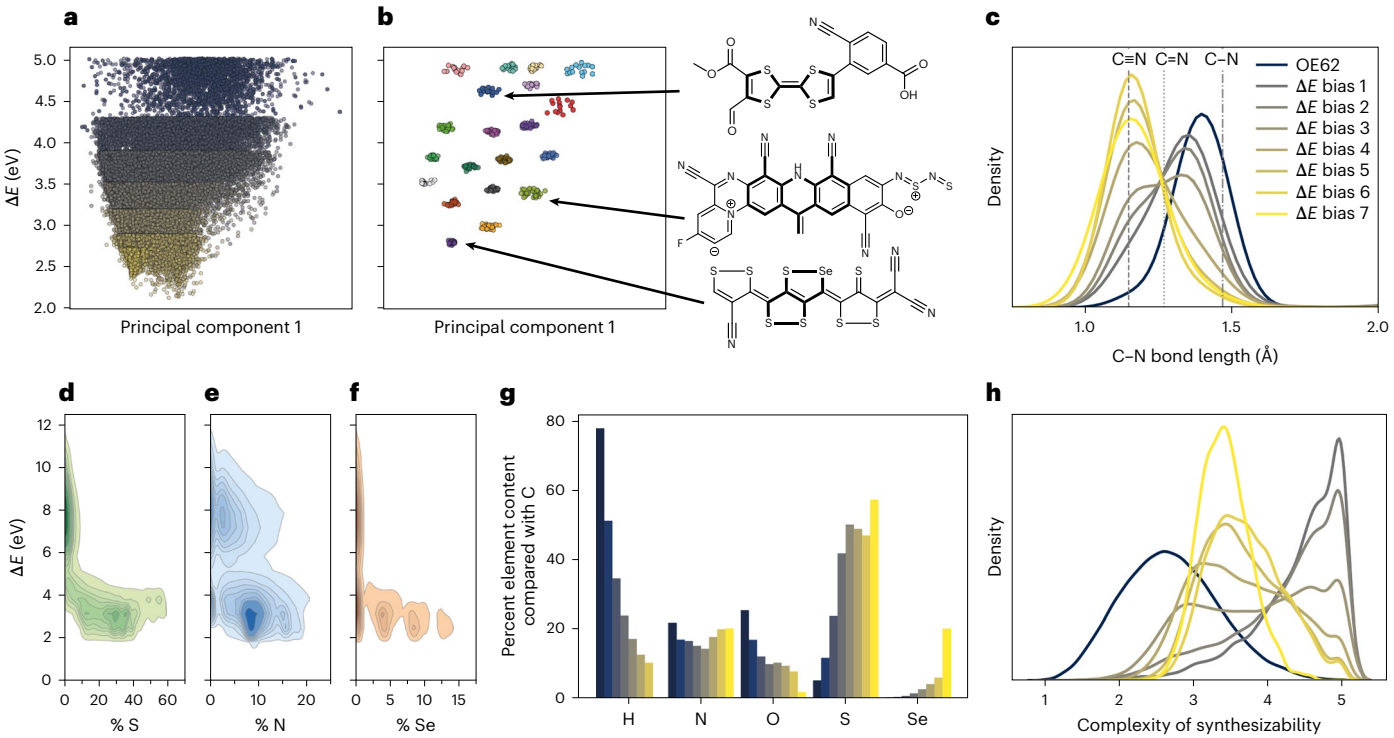

Fig. 3 | Cluster analysis for molecules with small Δ*E*. **a**, Molecules obtained with G-SchNet after biasing towards small Δ*E* are represented using the first principal component using structural (Smooth Overlap of Atomic Positions) descriptors of all molecules (OE62 and G-SchNet-generated molecules). The color gradient corresponds to the different loops. The same legend as in **c** applies. **b**, Subclusters found with unsupervised learning obtained from data of **a** and representative molecules illustrated next to it. **c**, C–N bond length distribution. **d**–**f**, Relative elemental content of sulfur (S) (**d**), nitrogen (N) (**e**) and selenium (Se) (**f**) in molecules obtained with G-SchNet and in the original dataset. **g**,**h**, Elemental composition (**g**) and distribution of SCScore (**h**) of molecules of the OE62 dataset and obtained from G-SchNet (the same legend as in **c** applies).

also creates molecules with narrow structural distributions and increasing synthetic complexity, and in many cases with highly improbable structural arrangements. The complexity of synthesizability of the generated molecules is shown in Fig. 3h as obtained from a neural network for the synthetic complexity score (SCScore) metrics by Coley et al.[26] The SCScore ranges from 1 (low synthetic complexity) to 5 (high synthetic complexity), and as can be seen the minimization of Δ*E* emerges to the detriment of the synthetic complexity of the molecules. Ideally, molecules should be designed that feature electronic properties in an optimal range while still being synthetically accessible.

Targeting multiple properties

To generate molecules that exhibit both low synthetic complexity and small Δ*E*, we selected 2,670 molecules with small Δ*E* and small SCScore out of an initial dataset created by merging the OE62 dataset with an additional set of 340,000 molecules generated with G-SchNet trained on OE62. These data points were used to bias G-SchNet; in each consecutive loop, molecules were selected that satisfy selection criteria for both properties. The distributions of Δ*E* and SCScore for each iteration are shown in Fig. 4a,b, respectively. As can be seen, after each biasing step, G-SchNet successfully predicts molecules with iteratively smaller Δ*E* and smaller SCScore. Analysis of the elemental distributions of the generated molecules (Supplementary Fig. 11) reveals that the overall structural trends observed in single-property biasing of molecules that lead to small Δ*E* are retained. However, selenium is effectively eliminated from the distribution due to the additional criterion of achieving small SCScore. This trend is encouraging as selenium is a trace element and less abundant than sulfur. In addition, it is considered a contaminant of concern in water systems[44]. This is especially problematic as selenium has one of the narrowest windows between concentrations where it serves as a vital trace mineral and concentrations where it is toxic, hence industrially caused accumulation in the environment poses a risk[44,45].

Discussion

The presented method constitutes an efficient workflow for the (multi) property-driven design of previously unseen molecules. One of the limitations of the model is that it requires the prediction and screening of several hundred thousand molecules in each loop to obtain a large enough number of molecules with which the generative model can be biased after screening. This process is limiting, especially when the chemical diversity of generated structures is small, and can become a computational bottleneck if molecules are screened towards more than two properties. This limitation can be tackled with conditional generative models, such as conditional G-SchNet[12], which enable the conditioning of the generative model towards predicting molecules with certain properties by including these properties of interest as labels during training.

The ability to generate viable molecules is not unique to G-SchNet, and previously proposed generative models have been shown to achieve similar results. The novelty of our approach lies in its high-throughput capability. For example, previously reported approaches, such as the one by Sumita et al.[16], perform generative search based on SMILES strings, which are translated into three-dimensional structures with RDKit[46] and then screened using quantum chemistry calculations. This has several downsides. First, the conversion with RDKit of the generated structures does not necessarily yield equilibrium structures, whereas G-SchNet is only trained on relaxed equilibrium structures and has been shown to predict structures close to structural equilibrium (see Supplementary Fig. 1)[3]. A prediction based on SMILES would also not have allowed us to predict cyano groups from an original training

https://doi.org/10.1038/s43588-022-00391-1

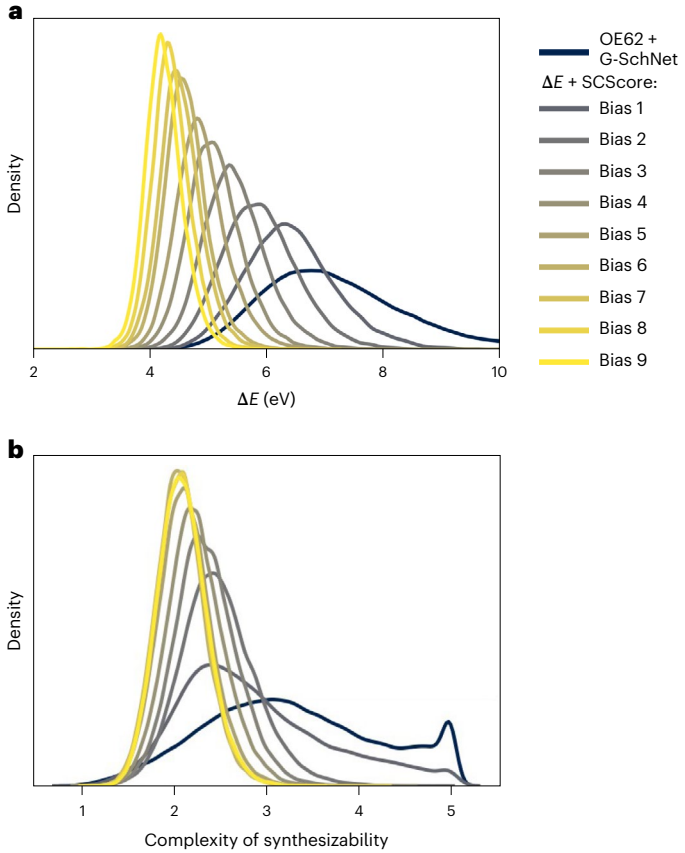

Fig. 4 | Multiproperty biasing. a,**b**, Distribution of ΔE (**a**) and SCScore (**b**) of molecules in the OE62 dataset and generated with G-SchNet biased against both properties.

database that does not contain such functional groups. Furthermore, the screening of 1,000 generated molecules with quantum chemistry calculations at the accuracy that we require would have taken over 500,000 computing hours or roughly 20,000 days. In contrast, in this work, we have screened many hundreds of thousands of molecules in a few days. The combination of the machine learning models applied here is thus a clear advantage that provides true high-throughput molecular design capabilities.

The ability of the method to predict molecules with electronic properties beyond the initial training dataset will be useful for a range of applications from high-throughput drug discovery to molecular design for organic electronics. Future work will explore how the performance of the method can be further improved by using different neural network architectures. By coupling this approach with a generative model of condensed phase structures, the property-driven design of crystalline solids may be possible.

Methods

Quantum chemistry calculations

Quantum chemistry calculations to verify results were carried out using the same procedure as in ref. [24] that was used to generate the dataset. All calculations were carried out using FHI-aims[47]. Every molecule was first relaxed using DFT with the Perdew–Burke–Ernzerhof (PBE) functional[48] and the standard default 'light' basis set as defined in FHI-aims. We augment the PBE functional with the Tkatchenko–Scheffler (PBE + vdW) correction to account for long-range dispersion corrections[49]. Afterwards, structure relaxations using the same settings, but with a standard default 'tight' basis set, were carried out. PBE0[50,51] orbital energies were calculated on the basis of the PBE + vdW optimized structures.

Using the PBE + vdW optimized structures, additional G0W0@PBE0 calculations were carried out as implemented in FHI-aims with analytic continuation[52]. To extract quasiparticle energies in the complete basis set limit, two calculations were conducted: one with the triple-zeta basis set def2-TZVP and one with the quadruple-zeta basis set def2-QZVP[53]. The extrapolated values were calculated by a linear regression against the inverse of the total number of basis functions[24,54].

To analyze the effect of sulfur and selenium content in molecules, we carried out DFT calculations for 144 randomly selected molecules generated with G-SchNet that contained no sulfur and no selenium, but oxygen atoms. We then carried out five calculations: one with the original molecule, two with a molecule in which a single oxygen atom is once replaced with a selenium atom and once with a sulfur atom and two with a molecule in which all oxygen atoms are replaced with either sulfur or selenium. The HOMO–LUMO gaps were compared as approximations to ΔE, because, despite them being underestimated with DFT, the trends are similar to those found with G0W0[23,24].

G-SchNet for OE62

G-SchNet was originally developed for small organic molecules made up of carbon, hydrogen, oxygen, nitrogen and fluorine (QM9 dataset[55,56]). We adapted G-SchNet to train on molecules that are part of the OE62 dataset[24], which features large chemical and structural diversity (Fig. 1b) and contains 62,000 molecular structures that are extracted from experimentally discovered organic crystals.

To generate molecules, the autoregressive, generative model learns from atomic positions, $\boldsymbol{r}_i$, and corresponding atom types, Z_i, $\mathbf{R}_{\leq n} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ with $\mathbf{r}_i \in \mathbb{R}^3$ and $\mathbf{Z}_{\leq n} = (Z_1, \ldots, Z_n)$ with $Z_i \in \mathbb{N}$, respectively. Thus, n point sets of atom types and positions are considered.

Rotationally and translationally invariant feature vectors are created using SchNet[57,58], a continuous-filter convolutional neural network that was originally developed to map molecular structures to properties such as energies or polarizabilities. The atomic features obtained from SchNet are multiplied element-wise with outputs of an embedding layer obtained from atom types and two additional auxiliary tokens. The number of tokens can be generalized using the variable t. The resulting feature vectors are then processed using dense atom-wise layers to obtain the probabilities of the next atom types and positions. The probability of the next atom type, $P\left(Z_{t+i}|\mathbf{R}^t_{\leq i-1}, \mathbf{Z}^t_{\leq i-1}\right)$, is obtained via

$$P\left(Z_{t+i}|\mathbf{R}^t_{\leq i-1}, \mathbf{Z}^t_{\leq i-1}\right) = \frac{1}{\beta}\prod_{j=i}^{t+i-1} P\left(Z_{t+i}|x_j\right) \quad (1)$$

Probabilities for atomic positions of the next atom are obtained in a similar way. Note that because of the t auxiliary tokens, which can be seen as auxiliary atom types that do not belong to the final generated molecule, indices run from 1 to $t + n$. One token marks the origin of the structure generation process and is fixed. The use of this token was found to improve training and lead to generated structures closer to the original distribution. In addition, another token, the focus point, breaks the symmetry of molecules and reduces artifacts. Each additional atom is always placed such that it is a neighbor of the focus point. β is a normalization constant[3].

Note that to generate rotationally invariant probabilities the three-dimensional information is obtained from pairwise distances, $d_{(t+i)j} = ||r_{t+i} - r_j||_2$, rather than absolute positions, with α being a normalization constant:

$$P\left(r_{t+i}|\mathbf{R}^t_{\leq i-1}, \mathbf{Z}^t_{\leq i}\right) = \frac{1}{\alpha}\prod_{j=i}^{t+i-1} P\left(d_{(t+i)j}|\mathbf{R}^t_{\leq i-1}, \mathbf{Z}^t_{\leq i}\right) \quad (2)$$

To train G-SchNet on molecules of the OE62 dataset, the original code was adapted. Importantly, the additional atom types that are

present in the OE62 dataset compared with the QM9 dataset had to be added and minimum and maximum bonding distances and orders had to be defined. In addition, since only molecules with even numbers of electrons were available in the OE62 dataset, we added a filtering function that excluded all molecules with unpaired electrons.

G-SchNet has the advantage of generating molecules in three dimensions. We validate that the generated molecules are close to their equilibrium structures according to the reference method, DFT, in Supplementary Section 1.

G-SchNet for OE62 was trained using a batch size of 2, a cutoff of 10 Å, 128 features (size of atom-wise representation), nine regular SchNet interaction blocks and 25 Gaussian functions to expand distances between atoms. In G-SchNet, the batch size depends on the number of samples per batch, but also on the size of the molecules within a batch. This is because a molecule is generated one atom at a time. By default, the whole trajectory to create a molecule is sampled, which can lead to large memory consumption, especially when molecules in a batch are large. Since molecules in the OE62 dataset can contain up to 200 atoms, we drew five random atom placements per molecule per batch instead of the complete trajectory. To still sample from the whole trajectory during training, a high number of epochs was chosen. Besides these, default parameters were used to train G-SchNet: an initial learning rate of 0.0001 and a decay of the learning rate by 0.5 after 10 epochs without improvement of the model during training.

SchNet + H for quasiparticle energies

The unsupervised, autoregressive generative deep neural network G-SchNet[3] is combined with a supervised, physically inspired deep neural network to design molecules with decreasing IP and increasing EA, as well as decreasing ΔE. Compared with recent studies that aimed to optimize the HOMO–LUMO gap as a theoretical proxy of ΔE, we optimize the gap as obtained from charged electronic excitations[59]. We used the already trained SchNet + H models from ref. [23] for this study.

As illustrated at the bottom of Fig. 2, the ionization potential of a given state i, IP_i, describes the energy of a bound state, which can be reconstructed experimentally in photoelectron spectroscopy by ejection of electrons with kinetic energy E_{kin} from a sample with work function Φ after irradiation with ultraviolet/visible light or X-rays with energy $h\nu$:

$$\mathrm{IP}_i = h\nu - E_{\mathrm{kin}} - \Phi = -\varepsilon_i \text{ for } \varepsilon_i < E_{\mathrm{Fermi}} \tag{3}$$

E_{Fermi} indicates the Fermi level and ε_i the electron removal energy or quasiparticle energy of ionization. In contrast, the electron affinity of a state i, EA_i, is equal to the negative energy of unoccupied states or the quasiparticle energy of electron addition and can be measured by measuring emitted Bremsstrahlung of electrons scattered in a sample:

$$-\mathrm{EA}_i = E_{\mathrm{kin}} - h\nu + \Phi = -\varepsilon_i \text{ for } \varepsilon_i \geq E_{\mathrm{Fermi}} \tag{4}$$

ΔE is the energy difference between IP and EA. The HOMO and LUMO energy levels according to DFT are often used to approximate the IP and EA, respectively, because they are computationally cheaper to calculate, but they are less accurate compared with many-body perturbation theory at the *GW* level of theory. Consequently, the HOMO–LUMO gap is often used as an approximation of ΔE but is known to underestimate energies[24].

In this work, the *GW* quasiparticle energies are obtained from SchNet + H, a physically inspired deep neural network trained on orbital energies from DFT/PBE0 of molecules in the OE62 dataset. As in G-SchNet, SchNet + H uses the SchNet descriptor[57,58] to represent molecules. In contrast to G-SchNet or the conventional SchNet model for molecular properties, however, SchNet + H predicts multiple energy levels, $\varepsilon_i^{\mathrm{ML(DFT)}}$, by inferring a latent Hamiltonian, $H^{\mathrm{ML(DFT)}}$, which is diagonalized using a transformation matrix, U:

$$\mathrm{diag}\left(\left\{\varepsilon_i^{\mathrm{ML(DFT)}}\right\}\right) = U^{\mathrm{T}} H^{\mathrm{ML(DFT)}} U \tag{5}$$

In this way, a transferable representation of molecular energies for molecules of arbitrary sizes is created. The energy levels obtained after diagonalization of the machine learning-inferred Hamiltonian can be corrected to *GW* accuracy at the complete basis set limit by another model trained on the difference between $\varepsilon_i^{\mathrm{ML(DFT)}}$ and $\varepsilon_i^{\mathrm{GW}}$, meaning quasiparticle energies at the *GW* level of theory. Adding corrections to the energy levels, $\varepsilon_i^{\mathrm{ML(DFT)}}$, results in energy levels at the *GW* level of theory:

$$\varepsilon_i^{\mathrm{ML(GW)}} = \varepsilon_i^{\mathrm{ML(DFT)}} + \varepsilon_i^{\mathrm{ML(GW\text{-}DFT)}} \tag{6}$$

This model has been shown to be accurate to predict photoemission spectra of molecules in the OE62 dataset and functional organic molecules outside of this dataset[23]. In this work, this model is applied to screen G-SchNet-predicted structures on the basis of their ΔE, EA and IP. The applicability of SchNet + H for this purpose is validated in Supplementary Section 2.

Computational details of the workflow for targeted design

The generation of molecules with desired electronic properties was conducted by biasing G-SchNet, meaning retraining it, with a subset of molecules that exhibit specific electronic properties. In this work, G-SchNet was biased independently three times: towards small ΔE, large EA and small IP.

In each loop, we generated 200,000 molecules for biasing towards small ΔE, large EA and multiple properties, and 100,000 for biasing towards small IP and during the knockout study. The number for IP biasing was reduced to maintain the balance between computational effort and accuracy, as molecules generated during this loop were on average larger and required about twice the computational resources. As 100,000 molecules yielded satisfactory results, while reducing computation time, this number was selected for the knockout study too. One loop in all studies took approximately 2 days. This time duration includes the molecule generation and the screening of these molecules with SchNet + H (computational costs of SchNet + H are specified in Supplementary Section 5). These molecules were then sorted on the basis of their electronic properties. Those molecules with electronic properties below their mean minus s.d. (IP and ΔE) or above their mean plus s.d. (EA) were selected for retraining of G-SchNet.

When biasing towards multiple properties, SCScore[26] and ΔE, we found that out of the OE62 dataset only 47 of the predicted molecules had lower ΔE and lower SCScore than their respective means minus s.d. To increase the initial dataset for biasing, another 340,000 molecules were generated with G-SchNet and molecules with values for SCScore and ΔE lower than their mean minus 0.5 times s.d. were selected, which resulted in an initial biasing dataset of about 2,670 molecules. During every biasing step, molecules that had ΔE smaller than their mean minus 0.5 times s.d. and SCScore smaller than their mean minus 0.5 times s.d. or SCScore ≤ 2 were selected for biasing G-SchNet in the next iteration.

For biasing towards small IP, the process terminated after two loops due to the generation of very large molecules, which made the structure generation and filtering process extremely computationally costly and finally infeasible with the existing computational resources at the time. As stated in ref. [12], where a conditional G-SchNet model was trained on drug-like molecules with about 50 atoms at most, further adaptations, such as a cutoff or long-range interactions, are needed to allow for scalability to larger systems. The problem was circumvented in this work by restricting the IP-biasing experiment to the prediction of molecules with up to 70 atoms. With this adaptation, the biased retraining could be conducted straightforwardly.

We terminated each experiment by continuously checking the electronic properties of a predicted dataset and the chemical diversity. As soon as there was no meaningful change in the distribution of properties for the predicted molecules between iterations, the workflow was terminated. All loops until then were used for analysis. We ended up with 7, 11 and 10 iterations for biasing towards small ΔE, small IP and large EA, respectively.

SCScore to predict the complexity of synthesizability of molecules

To estimate the complexity of the synthesis of a molecule, the SCScore is used as obtained from a deep neural network trained on 12 million reactions from the Reaxys database[29]. This score correlates with the number of steps used for synthesis. As inputs, this model uses canonical SMILES strings[60] that are generated using Open Babel[61].

The SCScore runs from 1 to 5, where molecules that have an SCScore of 5 are expected to be highly complex to synthesize and molecules with an SCScore of 1 are expected to be easily synthesizable. The SCScore defines synthesizability according to the number of reaction steps that are needed to synthesize from reasonable starting materials. Information on what starting materials are useful is learned and thus included in the model implicitly[26].

Dimensionality reduction of generated molecules

To visualize the chemical space that is spanned by molecules generated with G-SchNet compared with molecules in the original OE62 dataset and to create inputs for subsequent cluster analysis, we applied dimensionality reduction, for which we used PCA as implemented in scikit-learn[62].

The inputs for PCA were one of two applied molecular descriptors that we refer to as bonding and structural descriptors. The structural descriptors are obtained using the Smooth Overlap of Atomic Positions descriptor[31], which leads to a 57,792-dimensional description of molecules with the aim of accounting for the whole molecule. To obtain bonding descriptors, we take raw molecular geometries and apply Open Babel[61] and RDKit[46] to extract as many interesting features relating to the bonding of the molecules as possible. Features can be as simple as the number of atom types in a molecule, but also more complex, such as the number of rings of certain sizes, the aromaticity of a molecule and targeted electronic properties. The final dimension of the bonding descriptor was 732.

Each defined descriptor obtained from molecules of the OE62 dataset was used as input for PCA. To visualize the chemical space spanned by the OE62 dataset in comparison with the space spanned by the G-SchNet-generated molecules, we produced the descriptor for G-SchNet-generated structures and represented them using the same principal components as obtained from the OE62 data. The first two principal components cover 94% and 90% of the variance in the OE62 dataset for the bonding and structural descriptor, respectively (Supplementary Fig. 5). Results obtained from bonding descriptors are shown in Supplementary Fig. 6.

Clustering analysis

For clustering, we use a mixture of BIRCH[32] and agglomerative clustering[33] to allow for uneven cluster sizes as implemented in scikit-learn[62], which was chosen due to its memory efficiency. As an input, we used five principal components obtained after PCA of all molecules using both previously defined descriptors. The inputs were normalized such that features obtained from the different descriptors are equally weighted. An exception to this is that clustering was weighted towards changes in energy to better resolve energetic trends in subclusters. Clustering was conducted for molecules pooled from all biasing iterations. For each of the 13 clusters found, we extracted 10 molecules around the centroid. This procedure provides us with a condensed view of what makes up each cluster and reduces the task of analyzing clusters that contain too many molecules for manual inspection. The clusters found are illustrated using structural principal components and small ΔE in Supplementary Fig. 7a,b and the subclusters are shown in Supplementary Fig. 7c.

Similarity analysis of molecules

To measure similarity of molecules generated with G-SchNet and found in literature, we used SMILES strings and computed the Tanimoto score[36]. The mean of the Tanimoto score between molecules obtained after biasing against small ΔE is 0.54. Note that a similarity of 0.5 is often considered large enough to consider molecules to be similar[63]. The maximum similarity found was 0.72. We used key groups of molecules that exhibited the highest similarities and searched for hits using SciFinder (Supplementary Fig. 8).

Data availability

The OE62 dataset is available in ref. [24] and the OE62 + 340k G-SchNet molecule dataset is uploaded on https://figshare.com/articles/dataset/G-SchNet_for_OE62/20146943 (ref. [64]). Quantum chemistry calculations carried out in this study are uploaded to NOMAD under DOI 10.17172/NOMAD/2022.07.02-1 (ref. [65]). A supplementary data file showing the number of molecules predicted and used for training in each experiment and each loop is included as Supplementary Data 1.

Code availability

The modified G-SchNet version is available on GitHub (https://github.com/rhyan10/G-SchNetOE62) and tagged as version v0.1 (minted version under DOI 10.5281/zenodo.7430248)[66]. The GitHub repository includes scripts to analyze the data and carry out PCA. SchNet + H is published in ref. [23] and available on http://www.github.com/schnarc (minted version under DOI 10.5281/zenodo.7424017)[67]. We include a tutorial for using SchNet + H and G-SchNet models for OE62 on figshare (https://figshare.com/articles/dataset/G-SchNet_for_OE62/20146943), including instructions for installation[64]. Original tutorials for training and using G-SchNet and SchNet + H are available on GitHub with the original code of G-SchNet (https://github.com/atomistic-machine-learning/G-SchNet)[3] and SchNarc (https://github.com/schnarc/SchNarc/tree/develop)[68], respectively.

References

1. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
2. Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R. & Jensen, K. F. Generative models for molecular discovery: recent advances and challenges. *WIRES Comput. Mol. Sci.* **12**, e1608 (2022).
3. Gebauer, N. W. A., Gastegger, M. & Schütt, K. T. Symmetry-adapted generation of 3D point sets for the targeted discovery of molecules. *Adv. Neural Inf. Process. Syst.* **32** (2019).
4. Tkatchenko, A. Machine learning for chemical discovery. *Nat. Commun.* **11**, 4125 (2020).
5. Coley, C. W. Defining and exploring chemical spaces. *Trends Chem.* **3**, 133–145 (2021).
6. Wu, T. C. et al. A materials acceleration platform for organic laser discovery. *Adv. Mater.* https://doi.org/10.1002/adma.202207070 (2022).
7. Gryn'ova, G., Lin, K.-H. & Corminboeuf, C. Read between the molecules: computational insights into organic semiconductors. *J. Am. Chem. Soc.* **140**, 16370–16386 (2018).
8. Li, X.-H. et al. Narrow-bandgap materials for optoelectronics applications. *Front. Phys.* **17**, 13304 (2022).
9. Xue, D. et al. Advances and challenges in deep generative models for de novo molecule generation. *WIRES Comput. Mol. Sci.* **9**, e1395 (2019).

10. Meyers, J., Fabian, B. & Brown, N. De novo molecular design and generative models. *Drug Discov. Today* **26**, 2707–2715 (2021).
11. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
12. Gebauer, N. W. A., Gastegger, M., Hessmann, S. S. P., Müller, K.-R. & Schütt, K. T. Inverse design of 3D molecular structures with conditional generative neural networks. *Nat. Commun.* **13**, 973 (2022).
13. Li, Y., Pei, J. & Lai, L. Structure-based de novo drug design using 3D deep generative models. *Chem. Sci.* **12**, 13664–13675 (2021).
14. Elton, D. C., Boukouvalas, Z., Fuge, M. D. & Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **4**, 828–849 (2019).
15. Tan, X. et al. Automated design and optimization of multitarget schizophrenia drug candidates by deep learning. *Eur. J. Med. Chem.* **204**, 112572 (2020).
16. Sumita, M., Yang, X., Ishihara, S., Tamura, R. & Tsuda, K. Hunting for organic molecules with artificial intelligence: molecules optimized for desired excitation energies. *ACS Cent. Sci.* **4**, 1126–1133 (2018).
17. Bilodeau, C. et al. Generating molecules with optimized aqueous solubility using iterative graph translation. *React. Chem. Eng.* **7**, 297–309 (2022).
18. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
19. Simm, G. N. & Hernández-Lobato, J. M. A generative model for molecular distance geometry. In *Proc. 37th International Conference on Machine Learning* 8949–8958 (JMLR.org, 2020).
20. Xu, M., Luo, S., Bengio, Y., Peng, J. & Tang, J. Learning neural generative dynamics for molecular conformation generation. Preprint at https://arxiv.org/abs/2102.10240 (2021).
21. Axelrod, S. & Gómez-Bombarelli, R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Sci. Data* **9**, 185 (2022).
22. Ganea, O. et al. GeoMol: torsional geometric generation of molecular 3D conformer ensembles. *Adv. Neural Inf. Process. Syst.* **34** (2021).
23. Westermayr, J. & Maurer, R. J. Physically inspired deep learning of molecular excitations and photoemission spectra. *Chem. Sci.* **12**, 10755–10764 (2021).
24. Stuke, A. et al. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data* **7**, 58 (2020).
25. Golze, D., Dvorak, M. & Rinke, P. The GW compendium: a practical guide to theoretical photoemission spectroscopy. *Front. Chem* **7**, 377 (2019).
26. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).
27. Brown, N., Fiscato, M., Segler, M. H. S. & Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).
28. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the Δ-machine learning approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).
29. Lawson, A. J., Swienty-Busch, J., Géoui, T. & Evans, D. in *The Future of the History of Chemical Information ACS Symposium Series* Vol. 1164, 127–148 (American Chemical Society, 2014).
30. Joshi, R. P. et al. 3D-Scaffold: a deep learning framework to generate 3D coordinates of drug-like molecules with desired scaffolds. *J. Phys. Chem. B* **125**, 12166–12176 (2021).
31. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
32. Zhang, T., Ramakrishnan, R. & Livny, M. BIRCH: a new data clustering algorithm and its applications. *Data Min. Knowl. Discov.* **1**, 141–182 (1997).
33. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **42**, 19 (2017).
34. Liotta, D. & Monahan, R. Selenium in organic synthesis. *Science* **231**, 356–361 (1986).
35. Wilbraham, L., Smajli, D., Heath-Apostolopoulos, I. & Zwijnenburg, M. A. Mapping the optoelectronic property space of small aromatic molecules. *Commun. Chem.* **3**, 14 (2020).
36. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, 20 (2015).
37. Bendikov, M., Wudl, F. & Perepichka, D. F. Tetrathiafulvalenes, oligoacenenes, and their buckminsterfullerene derivatives: the brick and mortar of organic electronics. *Chem. Rev.* **104**, 4891–4946 (2004).
38. Hu, Y., Chaitanya, K., Yin, J. & Ju, X.-H. Theoretical investigation on the crystal structures and electron transfer properties of cyanated TTPO and their selenium analogs. *J. Mater. Sci.* **51**, 6235–6248 (2016).
39. Ferri, N. et al. Hemilabile ligands as mechanosensitive electrode contacts for molecular electronics. *Ang. Chem. Int. Ed.* **58**, 16583–16589 (2019).
40. Manzoor, F. et al. Theoretical calculations of the optical and electronic properties of dithienosilole- and dithiophene-based donor materials for organic solar cells. *Chem. Sel.* **3**, 1593–1601 (2018).
41. Li, Y., Liu, J., Liu, D., Li, X. & Xu, Y. D–A–π–A based organic dyes for efficient DSSCs: a theoretical study on the role of π-spacer. *Comput. Mater. Sci.* **161**, 163–176 (2019).
42. Kim, T. H. & Kim, K. S. Acridine derivative and organic electroluminescence device comprising the same. South Korea patent KR101120892B1 (2009).
43. Seifermann, S. & Choné, R. Organic molecules, in particular for use in optoelectronic devices. Europe patent EP3916072 (2018).
44. Sharma, V. K., Sohn, M. & McDonald, T. J. in *Advances in Water Purification Techniques* (ed. Ahuja, S.) 203–218 (Elsevier, 2019).
45. Fordyce, F. M. in *Essentials of Medical Geology: Revised Edition* (ed. Selinus, O.) 375–416 (Springer, 2013).
46. Landrum, G. *RDKit: Open-Source Cheminformatics* (2006); https://www.rdkit.org/
47. Blum, V. et al. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175–2196 (2009).
48. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
49. Tkatchenko, A. & Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **102**, 073005 (2009).
50. Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: the PBE0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).
51. Perdew, J. P., Ernzerhof, M. & Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **105**, 9982–9985 (1996).
52. Ren, X. et al. Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions. *New J. Phys.* **14**, 053020 (2012).
53. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297–3305 (2005).

54. van Setten, M. J. et al. GW100: benchmarking G0W0 for molecular systems. *J. Chem. Theory Comput.* **11**, 5665–5687 (2015).
55. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).
56. Ruddigkeit, L., van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **52**, 2864–2875 (2012).
57. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet—a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
58. Schütt, K. T. et al. SchNetPack: a deep learning toolbox for atomistic systems. *J. Chem. Theory Comput.* **15**, 448–455 (2019).
59. Reining, L. The GW approximation: content, successes and limitations. *WIRES Comput. Mol. Sci.* **8**, e1344 (2018).
60. 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).
61. O'Boyle, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminform* **3**, 33 (2011).
62. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
63. Baldi, P. & Nasr, R. When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. *J. Chem. Inf. Model.* **50**, 1205–1222 (2010).
64. Westermayr, J., Barrett, R., Gilkes, J. & Maurer, R. J. G-SchNet for OE62. *Figshare* https://doi.org/10.6084/m9.figshare.20146943.v2 (2022).
65. Westermayr, J. & Maurer, R. J. Organic molecules from generative autoregressive models. *NOMAD* https://doi.org/10.17172/NOMAD/2022.07.02-1 (2022).
66. Westermayr, J. & Barrett, R. G-Schnet for OE62 dataset (v0.1). *Zenodo* https://doi.org/10.5281/zenodo.7430248 (2022).
67. Westermayr, J. SchNarc for SchNet+H. *Zenodo* https://doi.org/10.5281/zenodo.7424017 (2021).
68. Westermayr, J., Gastegger, M. & Marquetand, P. Combining SchNet and SHARC: the SchNarc machine learning approach for excited-state dynamics. *J. Phys. Chem. Lett.* **11**, 3828–3834 (2020).

Acknowledgements

This work was funded by the Austrian Science Fund (FWF; J 4522-N) (J.W.), the EPSRC Centre for Doctoral Training in Modelling of Heterogeneous Systems (EP/S022848/1) (R.J.M.), the EPSRC-funded Network+ on Artificial and Augmented Intelligence for Automated Scientific Discovery (EP/S000356/10) (R.J.M.) and the UKRI Future Leaders Fellowship program (MR/S016023/1) (R.J.M.). Computational resources have been provided by the Scientific Computing Research Technology Platform of the University of Warwick, the EPSRC-funded Northern Ireland High Performance Computing service (EP/T022175/1) via access to Kelvin2, the EPSRC-funded HPC Midlands+ computing service (EP/P020232/1) via access to Athena and Sulis and the EPSRC-funded High End Computing Materials Chemistry Consortium (EP/R029431/1) for access to the ARCHER2 UK National Supercomputing Service (https://www.archer2.ac.uk). We thank N. Gebauer (TU Berlin) for fruitful discussions on the G-SchNet model. For the purpose of open access, we have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

Author contributions

R.J.M. conceived the original idea and supervised the research project. R.J.M. and J.W. designed the research project. R.B. and J.W. trained the deep learning models and created the property-guided design workflow. J.G. and J.W. performed the dataset curation, predictions, model validation and data analysis. J.W. performed the quantum chemistry calculations. J.W. and R.J.M. wrote the manuscript with the help of the other authors. The manuscript reflects the contributions of all authors.

Competing interests

R.J.M. is an editorial board member of the journal *Communications Materials*. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-022-00391-1.

Correspondence and requests for materials should be addressed to Julia Westermayr or Reinhard J. Maurer.

Peer review information *Nature Computational Science* thanks Camille Bilodeau 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.