nature computational science

Article

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

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

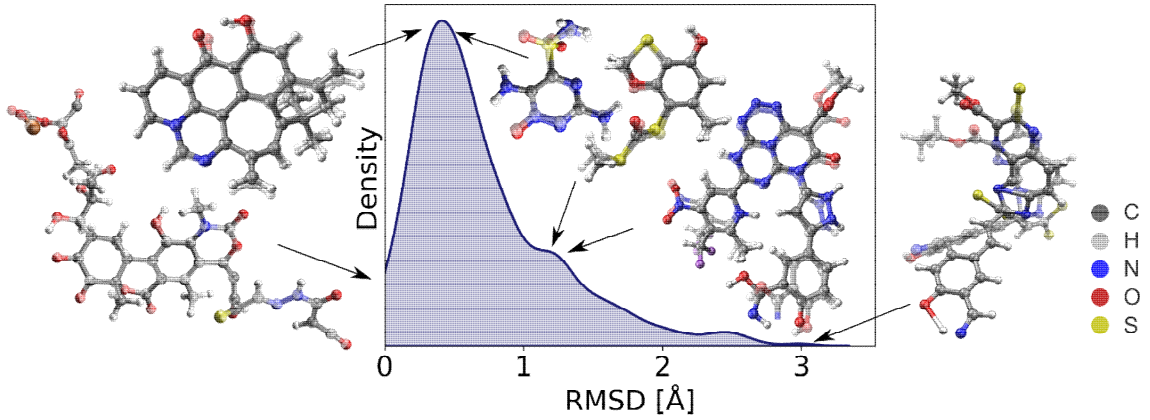
In the format provided by the authors and unedited

Table of Contents

Supplementary Section 1 Validation of G-SchNet for OE62

To ensure that the trained autoregressive, generative deep neural network, G-SchNet,[1] predicts sensible structures that resemble the molecules in the original data set (OE62),[2] we carried out a two-fold analysis. First, we generated a data set of 100k molecules with G-SchNet with molecules that contain up to 100 atoms. We then randomly selected 400 data points and optimized them with PBE+vdW[3-5] and tight basis set settings using FHI-aims.[6] For structure relaxations, the same protocol reported for the OE62 data set was used (see also Methods section on quantum chemistry calculations).

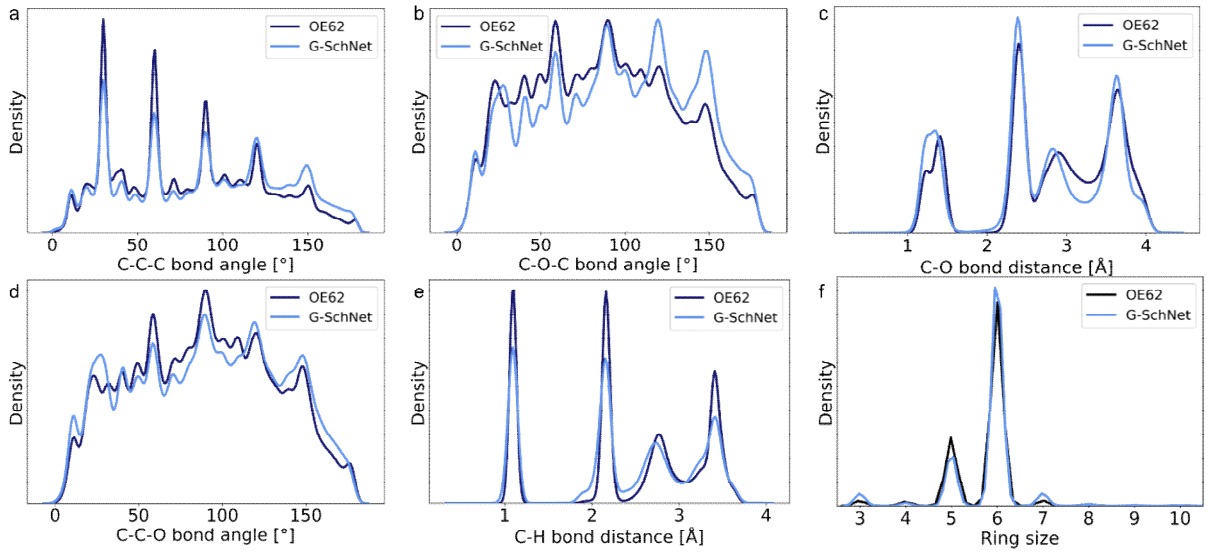
The optimized molecular geometries were then aligned with the G-SchNet predicted molecules and the root mean squared deviations (RMSD) were computed. The distribution of RMSD values is shown in **Supplementary Figure 1**. In addition to the RMSDs, sample molecules are shown. The G-SchNet predicted structures are solid, while the density functional theory (DFT)-optimized structures are shown slightly transparent. The left-most molecule shows the structure with lowest RMSD of 0.022 Å, where both structures are almost identical. The second pair of two structures illustrates deviations of around 0.5 Å (i.e., 0.46 Å and 0.48 Å) and deviations are representatives for most of the predicted molecules with G-SchNet. The next pair of two structures to the right have an RMSD at around 1.09 Å und 1.18Å. At RMSD > 1Å, deviations between G-SchNet-predicted structures and DFT-optimized structures become clearly visible but can be deemed minor. The molecule with the largest deviation of 3.00 Å is shown on the right and shows a geometry that GSchNet predicts to be more strongly distorted than the DFT reference result.



Supplementary Figure 1: Validation of G-SchNet predicted structures. *The root mean squared deviations (RMSD) of molecules predicted with G-SchNet were compared to structures obtained after structure relaxation with the reference density functional theory method. Exemplary molecules are shown, where the G-SchNet predicted structure (solid colors) is overlaid with the DFT-optimized structure (transparent). Examples for molecules that have very low RMSD, RMSD at around 0.5 Å, 1.1-1.2 Å, and >3 Å* are illustrated.

In addition to the RMSD, we compared distributions of some the most common bond lengths and bond angles. This analysis is based on the validation of G-SchNet that was carried out for the QM9 data set in Ref. [1]. The distributions for C-C-C bond angles, C-O-C bond angles, C-O bond distances, C-C-O bond angles, C-H bond distances, and ring sizes of molecules in the OE62 data set and molecules predicted by G-SchNet can be seen in **Supplementary Figure 2**. As can be seen, the distributions are very similar and indicate that, at least for the illustrated bonds and angles, G-SchNet structures resemble the molecular structures of the OE62 data set. The similarity of G-SchNet structures

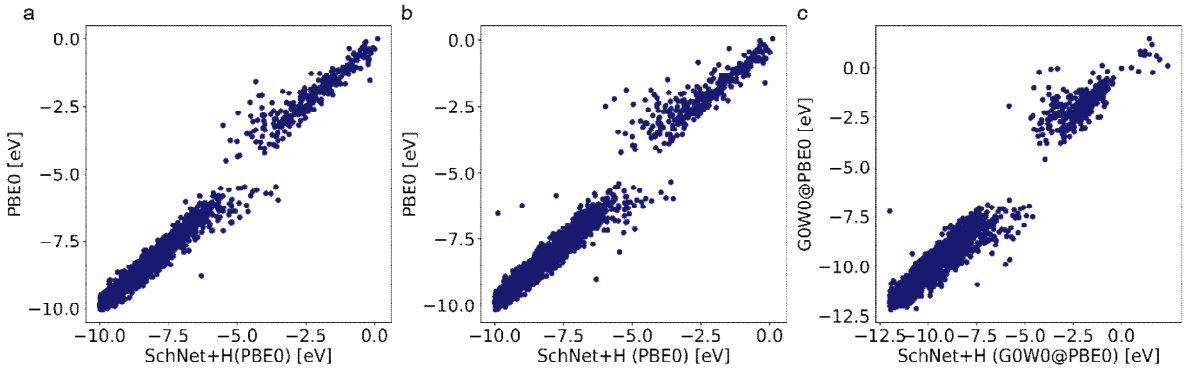
compared with molecules of the OE62 data set can be further assessed from Figure 1b, which shows the elemental composition of molecules with respect to the amount of carbon. Besides the amount of lithium and arsenic, which appear to differ strongly in the plot but in reality deviate only minorly due to the log-scale used for better visibility of elements with negligible amounts, the molecular compositions are very similar.



Supplementary Figure 2: Comparison of structures predicted with G-SchNet with structures of the OE62 data set. a) *Probability distribution of C-C-C bond angles,* ***b)*** *C-O-C bond angles,* ***c)*** *C-O bond distances,* ***d)*** *C-C-O bond angles,* ***e)*** *C-H bond distances, and* ***f)*** *ring sizes of molecules in the OE62 data set and G-SchNet-predicted molecules.*

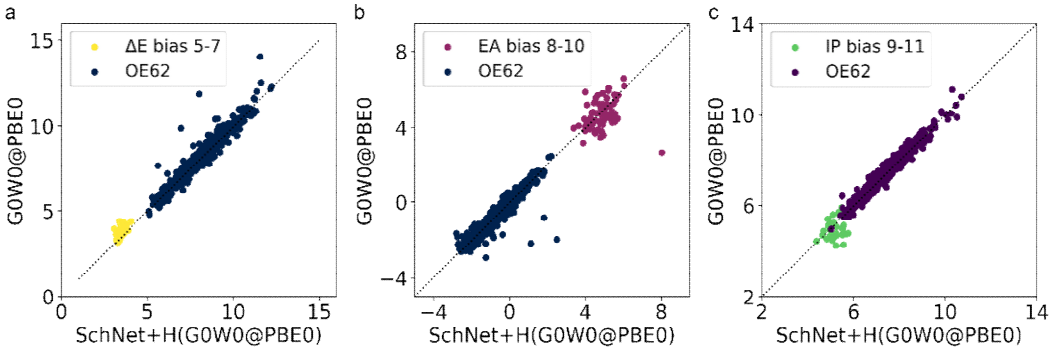
Supplementary Section 2 Validation of SchNet+H for G-SchNet-predicted structures

To assess the influence of structural differences on the electronic properties of molecules, i.e., orbital energies and quasiparticle energies, we predicted orbital energies of the 400 molecules used for validation of G-SchNet in **Supplementary Figure 3**, as obtained from G-SchNet and after structure optimization with DFT. The orbital energies of DFT-optimized structures predicted with SchNet+H are plotted against orbital energies obtained from DFT (PBE0[7,8] and tight basis set settings) using the DFT-optimized molecules as inputs. The mean absolute error (MAE) is about 0.24 eV (**Supplementary Figure 3**a). For comparison, the error of SchNet+H orbital energies for G-SchNet predicted structures compared to orbital energies obtained with DFT using DFT-optimized structures are only slightly larger, i.e., 0.26 eV (**Supplementary Figure 3**b). The same test was executed with SchNet+H for quasiparticle energies. The MAE error obtained using DFT-optimized and G-SchNet predicted structures is about 0.25 eV and 0.28 eV, respectively. Scatter plots of quasiparticle energies using G-SchNet-predicted structures for SchNet+H predictions can be seen in **Supplementary Figure 3**c. For comparison, the error of SchNet+H for molecules in the test data of the OE62 data set is about 0.13 eV. The method can be deemed sufficiently accurate for the purpose of high-throughput targeted design of functional organic molecules.



Supplementary Figure 3: Validation of SchNet+H for G-SchNet generated structures. a) *Scatter plots of SchNet+H predicted orbital energies and PBE0 orbital energies for structures obtained from G-SchNet and for* ***b)*** *optimized structures with PBE+vdW and the tight basis set settings. The same procedure as in the original data set was carried out to relax molecules.* ***c)*** *Scatter plots of SchNet+H predicted quasiparticle energies and reference G0W0@PBE0 quasiparticle energies for molecules obtained from G-SchNet without additional DFT optimization.*

Supplementary Section 3 Validation of molecules at the edges of of the distributions



Supplementary Figure 4: Validation of electronic properties of generated molecules*. a) Fundamental gaps, ΔE,* ***b)*** *electron affinities, EA, and* ***c)*** *ionization potentials, IP, for molecules of the original data set and G-SchNet generated structures of the last 3 biasing steps predicted with SchNet+H and computed with G0W0@PBE0.*

To assess the reliability of SchNet+H predictions, two SchNet+H models were employed that were trained on different random train/test splits of the same dataset. By computing the deviation of ΔE, EA, and IP values between the two models, which should be well below the MAE of the individual models, it is possible to identify structures for which quasiparticle energies are predicted with high uncertainty. The threshold was set to the MAE of the models that was determined for a given data set. This approach is known as query by committee. [9,10]

To further validate the predictions of SchNet+H for molecules obtained in the last biasing steps of the fundamental gap, ΔE, the electron affinity, EA, and the ionization potential, IP, G0W0@PBE0 calculations were carried for 66, 79, and 33 data points, respectively. These data points were obtained by taking every 50th data point of molecules in the last 3 loops that had a ΔE or EA small than their mean minus standard deviation and an IP larger than their mean plus standard deviation of the model. In this way, 99, 86, 71 geometries were obtained for ΔE, EA, and IP, respectively. These were, as done in the original data set, optimized with PBE+vdW using first light and later tight settings of the basis set. The relaxed geometry was then used to compute G0W0@PBE0 values at the complete basis set limit. Therefore, two calculations were carried out, once with the QZVP basis set and once with the TZVP basis set. G0W0@PBE0 values at the complete basis set limit were extrapolated from TZVP and QZVP quasiparticle energies by a linear fit using the procedure employed for the GW100 benchmark set[11] with the script obtained from NOMAD of ref. [2]. Out of all calculations, 66, 79, and 33 converged

for ΔE, EA, and IP, respectively. The reference values are plotted against the SchNet+H predictions in **Supplementary Figure 4**. In addition, the G0W0@PBE0 values of the original data set are shown. It is clearly visible that molecules predicted in the last iterations of the biasing process exhibit properties at the edges or outside of the training set.

As can be seen, SchNet+H accurately predicts the trends of almost all molecules correctly. There is one data point for the EA, which is predicted with a large error. The mean absolute error for ΔE, EA, and IP of molecules of the last biasing steps are 0.4 eV, 0.6 eV, and 0.4 eV, respectively. Given the fact that these molecules are at the edge of the originally learned distribution exhibiting electronic properties outside the training set and the use case of computationally efficient high-throughput screening, the accuracy can be deemed sufficient.

The smallest ΔE value computed with G0W0@PBE0 was 3.2 eV, while the smallest ΔE value of the OE62 data set is 4.8 eV, which is 1.6 eV larger. The mean ΔE value of the molecules recomputed is 3.9 eV, which is still smaller than the smallest value found in the OE62 data set. The mean ΔE value of the OE62 data set is 8.1 eV.

The largest EA value computed with G0W0@PBE0 was 6.6 eV, while the largest EA value of the OE62 data set is 2.4 eV, which is 4.2 eV larger. The mean EA of the molecules recomputed is 4.6 eV, which is still much larger than the largest value found in the OE62 data set. The mean EA of the OE62 data set is -0.7 eV.

The smallest IP computed with G0W0@PBE0 was 4.2 eV, while the smallest IP of the OE62 data set is 5.0 eV, which is 0.8 eV larger. The mean IP of the molecules recomputed is 5.0 eV, while the mean IP of the OE62 data set is 7.4 eV.

Supplementary Section 4 Iterative biasing

For biasing of G-SchNet towards large EA, we selected all molecules with a target property, P, that was larger than the mean of each property, $\bar{P}$, plus the corresponding standard deviation, σ_P: $P = \bar{P} + x \cdot \sigma_P$. For biasing of G-SchNet towards small IP, ΔE, and SCScore, we selected all molecules with a target property, P, that was larger than the mean of each property, $\bar{P}$, minus the corresponding standard deviation, σ_P: $P = \bar{P} - x \cdot \sigma_P$. For single property biasing we set x to 1. In case of biasing towards two properties, x was set to 0.5. The number of valid molecules generated in each loop and the number of molecules selected for biasing G-SchNet are shown in **Supplementary Datafile 1**.

Supplementary Section 5 Computational costs of quantum chemistry calculations and machine learning training and predictions

The computational costs for G0W0@PBE0 and SchNet+H quasiparticle energies are compared in **Supplementary Table 1**. As can be seen, the computational costs of G0W0@PBE0 calculations are extremely large with several 1000 CPUhs for molecules larger than 80 atoms. The computational costs for SchNet+H predictions are almost independent of atom size and are averaged from predictions made for over 10k molecules. Dell PowerEdge C6420 compute nodes each with 2 x Intel Xeon Platinum 8268 (Cascade Lake) were used for molecules with up to about 45 atoms and Dell PowerEdge R640 nodes each with 2 x Intel Xeon Platinum 8268 (Cascade Lake) were used for larger molecules. SchNet+H predictions were carried out on Dell PowerEdge R740 nodes each with 3 x NVIDIA RTX 6000 24 GB RAM GPUs.

As can be seen in **Supplementary Table 1** the screening of several hundred thousand molecules is computationally extremely costly and can be regarded as infeasible, especially because high memory nodes are necessary for molecules larger than about 45 atoms. In contrast, SchNet+H is computationally efficient enough to predict several hundred thousand molecules within less than a day. Note that the costs of obtaining G0W0@PBE0 calculations are more expensive than PBE0 calculations, because two calculations are carried out: The first step is the prediction of orbital energies at PBE0 level of theory and the second step is the correction of these energy levels with a Δ-ML model for G0W0@PBE0. Since two slightly differently trained SchNet+H models were executed each time G-SchNet generated structures were screened, one loop took approximately 2 days on a GPU. G-SchNet training on OE62 data took approximately 1 week, while biasing took less than 1 day on a GPU.

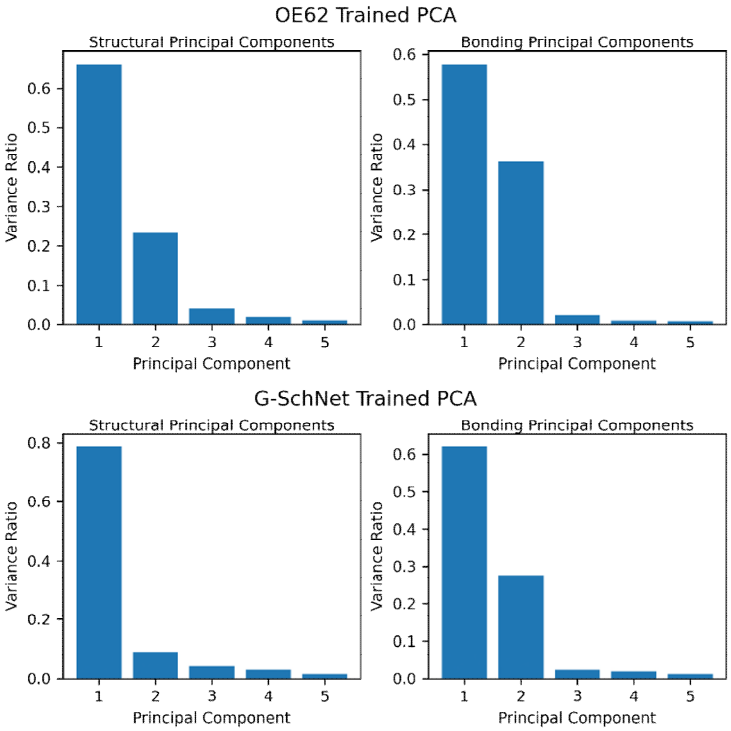
Supplementary Table 1: Computational costs of quantum chemical calculations and machine learning predictions*. The computational costs of calculating PBE0 orbital energies and G0W0@PBE0 quasiparticle energies at the complete basis set (CBS) limit with density functional theory and SchNet+H are compared for two molecules of different sizes. Dell PowerEdge C6420 compute nodes each with 2 x Intel Xeon Platinum 8268 (Cascade Lake) were used for molecules with up to about 45 atoms and Dell PowerEdge R640 nodes each with 2 x Intel Xeon Platinum 8268 (Cascade Lake) were used for larger molecules. SchNet+H predictions were carried out on Dell PowerEdge R740 nodes each with 3 x NVIDIA RTX 6000 24 GB RAM GPUs.*

Type of calculation	Molecule size	QC [CPUh]	SchNet+H [GPUh]
PBE0	42	7.1	$4.4 \cdot 10^{-5}$
G0W0@PBE0 CBS	42	502	$1.8 \cdot 10^{-4}$
PBE0	85	47.3	$4.4 \cdot 10^{-5}$
G0W0@PBE0 CBS	85	4,126	$1.8 \cdot 10^{-4}$

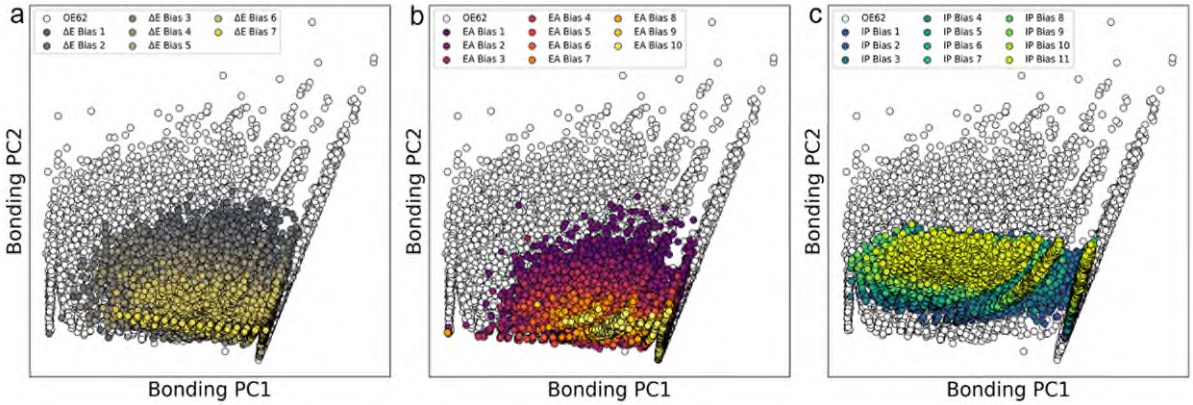
Supplementary Section 6 Clustering and principal component analysis (PCA)

The variance covered by the first 5 principal components using descriptors of molecules of the OE62 data and of all molecules as input are shown in **Supplementary Figure 6**.

In addition to the representation of the chemical space spanned by principal components obtained from the OE62 data set and the structural descriptors, we carried out PCA using bonding descriptors of the OE62 data set. The chemical space spanned by the OE62 data represented by the first two principal components of the bonding descriptors can be seen in **Supplementary Figure 5**. The plots verify results found by using structural descriptors (Figure 2b, d, and f) and suggest similar relevant regions in chemical space for small fundamental gaps and large electron affinities and different important regions in chemical space that make up small ionization potentials. Also here, we can see that generated molecules are within the regions covered by molecules in the OE62 data set.

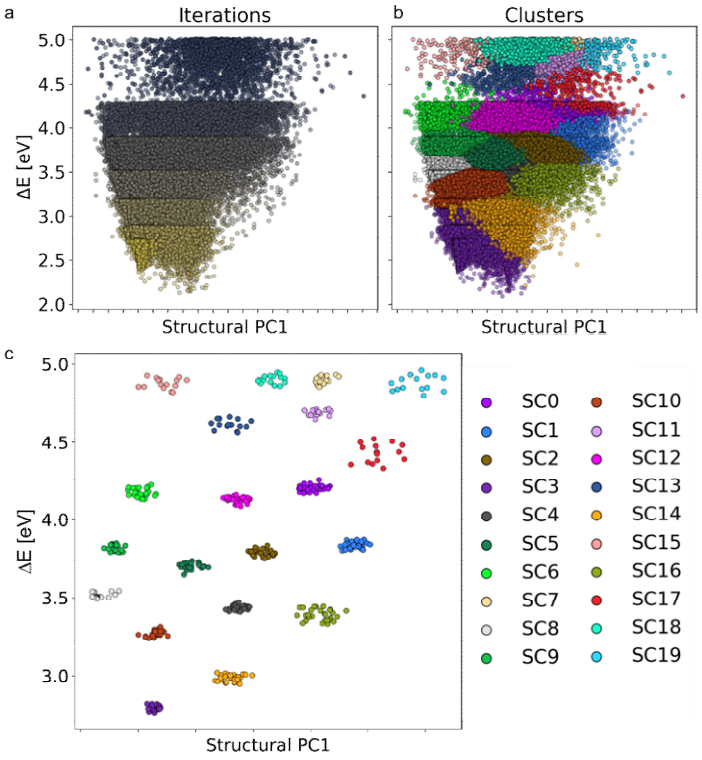


Supplementary Figure 6: Variance in principal components. a) *Variance of the first 5 principal components (PCs) obtained for the structural descriptor, i.e., SOAP,* ***b)*** *and the bonding descriptor, for molecules of the OE62 data set.* ***c)*** *Variance of the first 5 principal components (PCs) obtained for the structural descriptor, i.e., SOAP,* ***d)*** *and the bonding descriptor, for molecules of the OE62 data set and the generated molecules used for biasing towards small fundamental gaps.*



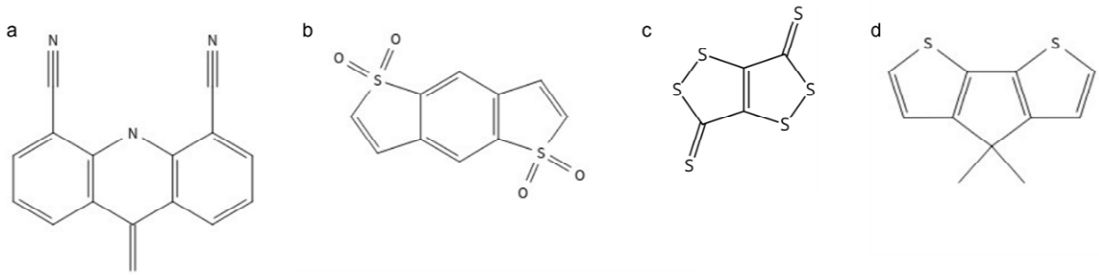
Supplementary Figure 5: Chemical space spanned by OE62 data. *Distribution of data points in chemical space made up by principal components obtained from OE62 data using bonding descriptors and results from biasing towards* ***a)*** *small fundamental gaps, ΔE,* ***b)*** *large electron affinities, EA, and* ***c)*** *small ionization potentials, IP. The color code indicates the biasing step. The plots are complementary to Figure 2 in the main text panels b, d, and f.*

Supplementary Figure 7 shows the clusters plotted against the first principal components (PCs) obtained from structural descriptors and ΔE colored with respect to the loops (panel a) and clusters found (panel b). The subclusters are shown in panel c.



Supplementary Figure 7: Clustering analysis for biasing G-SchNet towards small fundamental gaps, ΔE. A) *Data points obtained from OE62 and G-SchNet colored according to iterations and* ***b)*** *colored according to clusters found.* ***C)*** *10 representatives of each cluster obtained with subclustering using centroids of b) as inputs SC indicates sub cluster.*

Supplementary Section 7 Molecular features

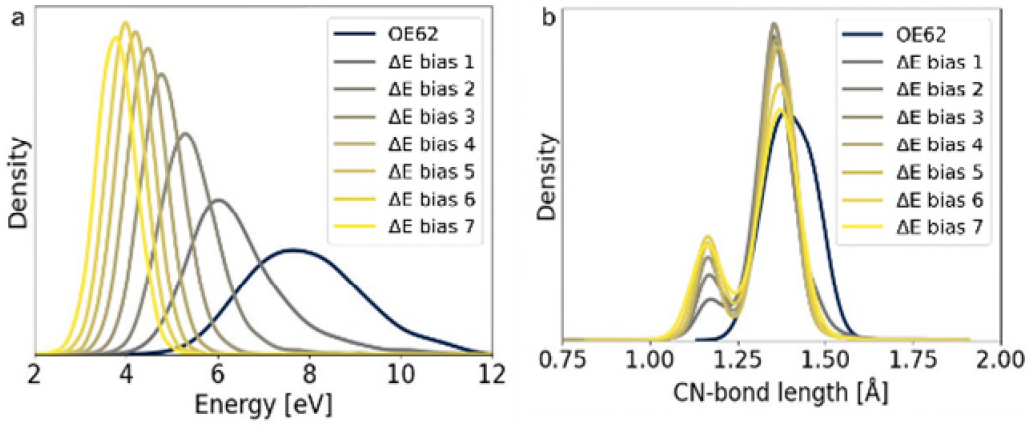


***Supplementary Figure 8*: Functional groups represented in molecules with small fundamental gaps.** *Molecules strongly represented in the data set biased towards small fundamental gaps and generated with G-SchNet that are also found in the data set in Ref.44.*

Supplementary Figure 8 shows functional groups that are represented frequently in molecules that have a small ΔE value. These molecular groups are parsed in SciFinder and are found in applications and are discussed in the main text.[12-14]

Supplementary Section 8 Knock-out study

To analyze whether G-SchNet can predict bonding patterns that are not present in the original data set, we eliminate all molecules containing cyano groups of the OE62 data set. These are molecules that have a C-N bond length of less than 1.25 Å, as C-N triple bonds are usually in the range of 1.15 Å. The modified OE62 data set is used to train a new G-SchNet model, which is then used to predict new molecules and is biased against small ΔE. As can be seen in **Supplementary Figure 9**a, the ΔE values iteratively decrease, when biased against them, which is expected. Supplementary Figure 9b shows that already after the first biasing step, G-SchNet predicts molecules with increased number of cyano groups. The trend of increased number of cyano groups in molecules with small ΔE values is thus retained.

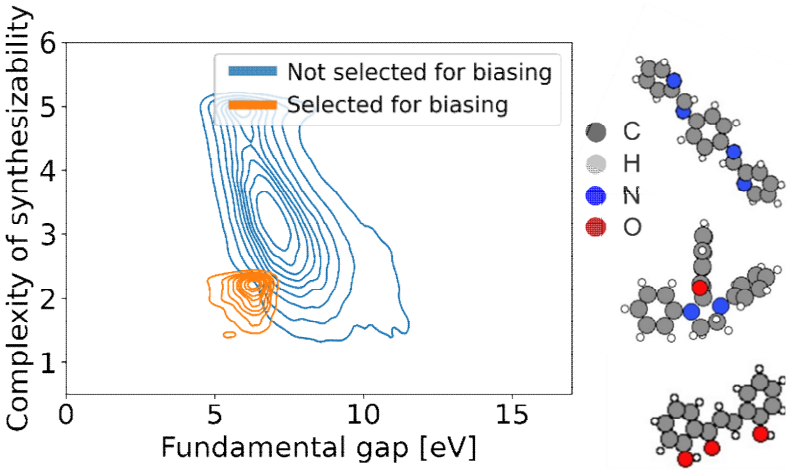


Supplementary Figure 9: Knock-out study. A) *Distribution of fundamental gaps, ΔE, and* ***b)*** *C-N bond lengths of molecules in the OE62 data set excluding molecules with a C-N bond length < 1.25 Å and of molecules generated with G-SchNet biased against ΔE.*

Supplementary Section 9 Multi-property biasing

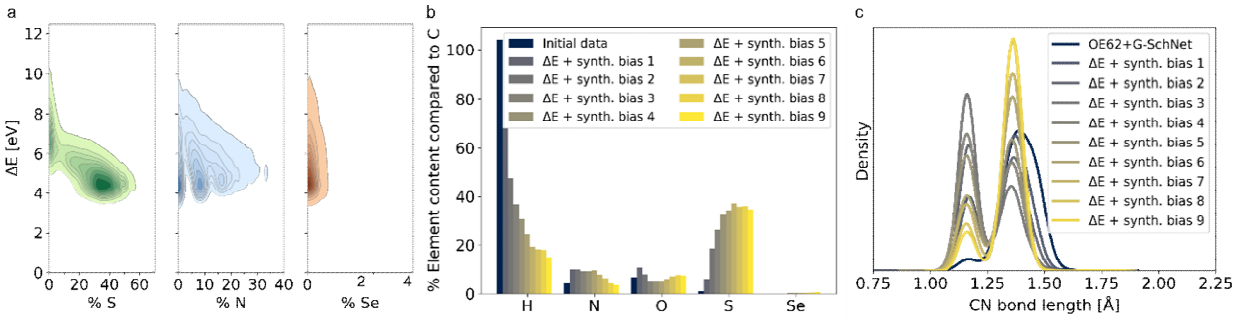
As discussed in the main text in section 3.2 and 3.3, the synthetic complexity of molecules increases when minimizing the fundamental gap (see Figure 3h). This effect seems to revert after the third loop, when the complexity of synthesizability drops and becomes more favorable towards the end of the biasing process. However, it does not return to its original, lower distribution. This lowering of the complexity of synthesizability is possibly due to the fact that molecules become smaller with iterations, which generally reduces synthetic complexity.[15] The conclusion that our method is successful in finding rules in molecules that could be potentially relevant to optoelectronics, but that the molecules we generate are possibly too complex to synthesize, is not very encouraging. Therefore, we further sought to investigate the potential of the method to simultaneously optimize multiple properties, i.e., small fundamental gaps and low synthetic complexity of molecules.

The molecules selected for biasing G-SchNet initially are shown in **Supplementary Figure 10**. This image shows the fundamental gap against the SCScore of 340k molecules obtained from the OE62 data set and predicted with G-SchNet. The orange distribution is used for biasing G-SchNet initially.



***Supplementary Figure 10: Molecules selected for multi-property biasing**. Fundamental gap of molecules plotted against synthetic complexity score (SCScore) of molecules of the OE62 data set and generated with G-SchNet trained on the OE62 data set (blue distribution). The distribution of molecules selected for biasing towards small fundamental gaps are shown in orange. Some example molecules with small fundamental gaps and synthetic complexity (orange area) are shown right to the plot.*

The results, i.e., the sulfur nitrogen and selenium content (panel a), the elemental distribution in molecules (panel b), and the C-N bond lengths (panel c) are shown in **Supplementary Figure 11**. In addition to **Figure 4** in the main text. The plots are complementary to **Figure 4** in the main text, but contain results obtained by multi-property biasing, i.e., biasing towards small fundamental gaps and small SCScore, instead of results obtained only from biasing towards a single property, i.e., small fundamental gaps



***Supplementary Figure 11: Cluster analysis for molecules with small fundamental gaps and small SCScore.** a) Distribution of sulfur (S), nitrogen (N), and selenium (Se), b) elemental distribution and c) distribution of C-N bond lengths of molecules generated during biasing towards small fundamental gaps, ΔE, and small synthetic complexity score (SCScore).*

References:

1 Gebauer, N. W., Gastegger, M. & Schütt, K. T. Generating equilibrium molecules with deep neural networks. *arXiv* **1810.11347** (2018).

2 Stuke, A. *et al.* Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data* **7**, 58 (2020). https://doi.org:10.1038/s41597-020-0385-y

3 Tkatchenko, A., DiStasio, R. A., Car, R. & Scheffler, M. Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.* **108**, 236402 (2012). https://doi.org:10.1103/PhysRevLett.108.236402

4 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). https://doi.org:10.1103/PhysRevLett.102.073005

5 Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **77**, 3865-3868 (1996). https://doi.org:10.1103/PhysRevLett.77.3865

6 Blum, V. *et al.* Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175-2196 (2009). https://doi.org:https://doi.org/10.1016/j.cpc.2009.06.022

7 Perdew, J. P., Ernzerhof, M. & Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **105**, 9982-9985 (1996). https://doi.org:10.1063/1.472933

8 Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110**, 6158-6170 (1999). https://doi.org:10.1063/1.478522

9 Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.* **115**, 1032-1050 (2015). https://doi.org:https://doi.org/10.1002/qua.24890

10 Freund, Y., Seung, H. S., Shamir, E. & Tishby, N. Selective Sampling Using the Query by Committee Algorithm. *Machine Learning* **28**, 133-168 (1997). https://doi.org:10.1023/A:1007330508534

11 van Setten, M. J. *et al.* GW100: Benchmarking G0W0 for Molecular Systems. *J. Chem. Theory Comput.* **11**, 5665-5687 (2015). https://doi.org:10.1021/acs.jctc.5b00453

12 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). https://doi.org:10.1021/cr030666m

13 Ferri, N. *et al.* Hemilabile Ligands as Mechanosensitive Electrode Contacts for Molecular Electronics. *Ang. Chem. Int. Ed.* **58**, 16583-16589 (2019). https://doi.org:https://doi.org/10.1002/anie.201906400

14 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. Mat. Sci.* **51**, 6235-6248 (2016).

15 Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inform. Model.* **58**, 252-261 (2018). https://doi.org:10.1021/acs.jcim.7b00622