PAPER

Check for updates

Generative design of functional organic molecules for terahertz radiation detection

Zsuzsanna Koczor-Benda, [ID] *[a] Shayantan Chaudhuri, [ID] [ab] Joe Gilkes, [ID] [ac] Francesco Bartucca, [ID] [a] Liming Li [ID] [a] and Reinhard J. Maurer [ID] *[ad]

Plasmonic nanocavities are molecule–nanoparticle junctions that offer a promising approach to upconvert terahertz radiation into visible or near-infrared light, enabling nanoscale detection at room temperature. However, the identification of molecules with strong terahertz-to-visible frequency upconversion efficiency is limited by the availability of suitable compounds in commercial databases. Here, we employ the generative autoregressive deep neural network, G-SchNet, to perform property-driven design of novel monothiolated molecules tailored for terahertz radiation detection. To design functional organic molecules, we iteratively bias G-SchNet to drive molecular generation towards highly active and synthesizable molecules based on machine learning-based property predictors, including molecular fingerprints and state-of-the-art neural networks. We study the reliability of these property predictors for generated molecules and analyze the chemical space and properties of generated molecules to identify trends in activity. Finally, we filter generated molecules and plan retrosynthetic routes from commercially available reactants to identify promising novel compounds and their most active vibrational modes in terahertz-to-visible upconversion.

1 Introduction

Terahertz (THz) radiation has applications in numerous fields, including medical diagnostics, security screening, communications, and astronomy.[1,2] Historically, the development of both powerful and affordable light sources, and efficient THz detectors, has been technologically challenging.

Nanoscale, room-temperature detection of terahertz and mid-infrared radiation is enabled by molecular optomechanical devices utilizing the enhancement of electric fields in plasmonic nanocavities to convert terahertz radiation into visible or near-infrared light.[3,4] These nanocavities can be assembled on silicon-based photonic integrated circuits,[5] opening possibilities for low-cost fabrication and multiplexed detection. To enhance the light–matter interaction, molecules are typically placed between two metallic nanoantennas.[4,6,7] One of the two antennas focuses terahertz radiation at the design frequency over the molecular sample volume to enhance the absorption of terahertz radiation *via* the surface-enhanced infrared absorption[8] mechanism. The second optical antenna confines visible or near-infrared light to volumes below 100 nm^3,[3] which induces surface-enhanced Raman scattering[9] of molecules within the

plasmonic nanocavity. Absorption of THz radiation by molecules within the nanocavity results in the vibrational excitation of a specific normal mode, which leads to an increase in the measured Raman anti-Stokes intensity of the same normal mode, similar to resonant sum-frequency generation spectroscopy.[10] For centrosymmetric molecules, simultaneous activity in absorption and Raman scattering is not possible. Even in asymmetric molecules, it is rare to have vibrational modes that can efficiently upconvert the THz radiation signal, as this requires a large change in both electronic dipole moment and in polarizability along the vibrational mode. Vibrational modes of organic molecules in the THz frequency range are often delocalized across several functional groups or across molecules, which makes it challenging to use chemical intuition to suggest promising candidates or define molecular design rules. This makes it necessary to use quantum chemical calculations in connection with computational screening or design approaches to identify good candidate molecules and their active vibrational modes.[11,12] Such computational predictions motivate more detailed experimental investigations for the fabrication and application of new molecular optomechanical devices.[5]

Machine learning (ML) methods can facilitate the design and discovery of new functional materials by enabling the fast computational screening of large structural databases.[13–15] ML-based screening has previously been used to identify promising candidates for THz radiation detection from commercially available compound databases.[11] However, a drawback of this

[a]Department of Chemistry, University of Warwick, Coventry, CV4 7AL, UK. E-mail: zsuzsanna.koczor-benda@warwick.ac.uk; r.maurer@warwick.ac.uk

[b]School of Chemistry, University of Nottingham, Nottingham, NG7 2RD, UK

[c]Centre for Doctoral Training in Modelling of Heterogeneous Systems, University of Warwick, Coventry, CV4 7AL, UK

[d]Department of Physics, University of Warwick, Coventry, CV4 7AL, UK

approach was that there was a limited search pool of molecules that have an affinity to the gold surfaces of the nanoantennas used in detector prototypes. Self-assembled monolayers of thiol-containing molecules have been shown to have high stability and reproducibility on gold surfaces,[16] which are often used in plasmonic devices. It is therefore prudent to focus on thiol-containing molecules that are commercially available or easily synthesizable. These requirements pose a challenge for high-throughput screening methods as the number of thiol compounds within large commercial databases is relatively low, with only around 150 000 out of more than 20 million compounds in the eMolecules database and 32 000 out of 8 million compounds from the MolPort database identified in Koczor-Benda et al.[11] being monothiols, respectively.

An alternative solution for accelerating the discovery of promising molecules is generative deep learning, which in the past has been used for the property-driven design of functional organic molecules.[17–22] Most proposed generative deep learning models use text-based or two-dimensional (2D) molecular representations.[23,24] G-SchNet is a generative autoregressive deep neural network that has the advantage of being able to generate molecules in three-dimensional (3D) space.[25] Previous studies have shown that G-SchNet can be iteratively biased to generate molecules satisfying certain target properties. Westermayr et al.[17] coupled G-SchNet with a neural network that predicts molecular quasiparticle energies[26] to bias molecular generation towards small fundamental gaps, low ionization potentials, or high electron affinities, while conserving low synthetic complexity of the molecules. Gebauer et al.[18] developed conditional G-SchNet, which, in addition to structures, trains on electronic property and structural motif labels to condition molecular generation.

In this paper, we perform property-driven generative design of functional organic molecules for THz radiation detection using G-SchNet by driving the generative model to create novel molecules with high frequency-upconversion efficiency, affinity to gold surfaces, and synthetic accessibility. To predict the upconversion properties of molecules, we use the target property P introduced in Koczor-Benda et al.,[11] which is based on the total spectral intensity in a wide frequency window (30–1000 cm^{-1}) relevant for THz and mid-infrared applications. Due to the challenges and high cost associated with experimental preparation and characterization, the quantity P is not yet experimentally validated as an established surrogate. We therefore only use it as a semi-quantitative guide in the generative design. To increase the pool of candidates for this application, we train G-SchNet models on a dataset of around 30 000 thiol-containing molecules and generate hundreds of thousands of monothiolated molecules by iterative biasing. We analyze chemical trends in the generated databases and identify functional groups that correlate with high upconversion intensity. Previously used ML predictors of the frequency upconversion efficiency based on molecular fingerprints[11] become unreliable as the property-driven generative biasing workflow explores novel molecules beyond the training dataset. We replace them with more transferable equivariant graph neural network (GNN) models that make use of the 3D

molecular conformations that G-SchNet generates. To train these models, we use calculations based on density functional theory (DFT) for P values contained in Molecular Vibration Explorer,[12] which are available for around 2800 gold-thiolate molecules, and extend this database with new DFT calculations on generated molecules. Finally, highly spectroscopically active compounds are identified by generative design and further validated with quantum chemistry calculations and retrosynthetic route planning to identify promising, novel compounds for THz radiation detection.

2 Methods

2.1 Generative machine learning

2.1.1 Training dataset. A training dataset of 29246 monothiolated molecules was compiled from the eMolecules[27] commercial molecular database, that was previously used by Koczor-Benda et al.[11]. This database contains over 20 million readily available or custom-synthesized compounds from over 15 suppliers, aimed mainly at drug discovery applications.[27] This training dataset was selected to ensure that the generative model creates molecules that are chemically similar to known synthesizable compounds, thus facilitating the search for viable candidates. The eMolecules database was first filtered for monothiols based on the corresponding SMARTS pattern. Charged molecules and duplicates were removed, resulting in an initial pool of 147 623 molecules containing the following elements: hydrogen (H), boron (B), carbon (C), nitrogen (N), oxygen (O), fluorine (F), silicon (Si), phosphorus (P), sulfur (S), chlorine (Cl), selenium (Se), bromine (Br), tin (Sn), and iodine (I). In contrast to Koczor-Benda et al.,[11] molecular size and number of rotatable bonds were not restricted, resulting in a larger pool of molecules. Initial 3D structures for the unique monothiolated molecules were created from Simplified Molecular Input Line Entry System (SMILES) strings[28] and relaxed with the MMFF94 Merck molecular force field[29] using the RDKit package.[30] To maximize chemical diversity, a Smooth Overlap of Atomic Positions (SOAP)[31] descriptor with a local region cut-off of 4.0 Å, 4 radial basis functions, and a maximal degree of spherical harmonics of 3 was calculated for each molecule (resulting in 6384 features), using the DScribe package.[32] After singular value decomposition with 500 components, 30 000 clusters were identified with k-means clustering using the sci-kit-learn[33] library. For each cluster, the molecule closest to the cluster center was selected. Molecules that had already been calculated in the THz database were removed (604 duplicates), resulting in the final training set of 29246 molecules. Structure optimization was performed with the xTB software package using the GFN2-xTB parametrization,[34] based on which the final database of 3D geometries for the generative model was constructed. For a discussion on using the computationally less expensive xTB method instead of commonly used DFT for structure optimisation, and a comparison of generated unrelaxed and relaxed structures see Section S10 in the SI.

2.1.2 Training workflow. The schnetpack-gschnet[35,36] package was used to train G-SchNet models on the aforementioned training database. Each G-SchNet model was trained

using a SchNet[37] neural network with 128 features, 9 interaction blocks, a cut-off of 10 Å and 25 centers for the radial basis expansion of distances. A learning rate of 0.0001 was used and 5 random atom placements per molecule per batch were drawn. For all trained G-SchNet models, data were randomly split (as implemented within schnetpack-gschnet) 80%/10%/10% for training, validation and testing, respectively. Approximately 100 000 molecules were generated with each trained model, with a maximum molecular size of 60 atoms. Non-unique, disconnected, or invalid (incorrect valency) generated molecules were discarded. Molecules were filtered to only contain one thiol group, which can act as the linker to the gold nano-antenna in a THz radiation detector device. The number of molecules generated and remaining after filtering are summarized in the SI (Table SI).

2.1.3 Iterative biasing of G-SchNet. The generation of molecules with desired properties was achieved by an iterative workflow similar to the one proposed by Westermayr et al.[17] Herein, in each iteration, the G-SchNet model is trained, molecules are generated, molecules are filtered with a property prediction model, and a new training dataset is built that contains the original and a subset of the novel generated molecules with selected properties above or below a certain threshold value. As a result, molecular generation is iteratively biased towards molecules with desired properties. In each iteration, G-SchNet was trained (from scratch) with the modified dataset. The sizes of the training databases for each of the six biasing iterations are detailed in Table SII in the SI.

In each iteration, molecules were selected according to two properties: THz upconversion efficiency, predicted with a previously trained Kernel Ridge Regression (KRR) model,[11] and the SCScore metric of synthetic complexity.[38] The upconversion efficiency figure of merit, P, is defined as the logarithm of the orientation-averaged upconversion intensity (I_m^c) summed over all M vibrational frequencies in the 1–30 THz frequency window (30–1000 cm^{-1}):[11]

$$P = \log\left(\sum_{m \in M} \langle I_m^c \rangle\right) \quad (1)$$

Higher P values correspond to greater total frequency upconversion capability of vibrations in the selected frequency range, providing a semi-quantitative measure to guide the design process. I_m^c is based on the absorption and Raman scattering intensities of vibrational mode m (a full definition of I_m^c can be found in Section S2 of the SI). I_m^c was calculated using DFT for a simplified model of the molecule–metal interface in Koczor-Benda et al.[11] resulting in the Gold database of Molecular Vibration Explorer,[12] and used as training data for the KRR model.[11] We also discuss the details of these DFT calculations in the next Section.

The SCScore neural network by Coley et al.[38] was trained on 12 million reactions from the Reaxys[39] database. The SCScore correlates with the number of reaction steps required to synthesize the molecule from reasonable starting materials and ranges between 1 and 5, where higher numbers indicate reduced synthesizability.[38] Canonical SMILES[28] representations of molecules

generated using Open Babel[40] were used as input for the KRR predictor and the SCScore calculator. To simultaneously bias molecular generation towards large P (high THz upconversion efficiency) and low SCScore (S, low synthetic complexity) values, molecules with properties satisfying both $P \geq \bar{P}+0.5\sigma_P$ and $S \leq \bar{S}-0.5\sigma_S$ were appended to the training dataset for the subsequent training iteration, where $\bar{X}$ and σ_X are the mean average and standard deviation, respectively, of property X.

2.1.4 Reference calculations and property predictors. As reference data for the ML models, a database of about 2800 gold-thiolate molecules, available from Molecular Vibration Explorer,[12] was used, henceforth referred to as the 'THz database'. This database was originally compiled in Koczor-Benda et al.[11] and contains P values calculated with Kohn–Sham DFT,[41,42] using the B3LYP[43,44] hybrid generalized gradient approximation, the DFT-D3 (ref. 45) dispersion correction, the Karlsruhe basis set with split valence polarization (def2-SVP),[46] and a tight energy convergence threshold. The molecules were modeled as gold-thiolates to consider the most immediate chemical effects of the metal-molecule interface. This choice of modeling was also validated against surface-enhanced Raman spectroscopy measurements in Griffiths et al.,[47] Boehmke Amoruso et al.,[48] and Wright et al.[49]. Koczor-Benda et al.[11] validated the computational approach in detail against Raman and infrared measurements for powder, solution and nanoparticle-on-mirror constructs of a set of test molecules and found that individual spectral features as well as surface-enhanced Raman spectroscopy intensities integrated over a wide spectral window correlate well with measurements. However, for an accurate modeling of low-frequency vibrational features (below 200 cm^{-1}), considering the metal facets as well as molecule–molecule interactions becomes necessary,[48] which increases computational costs. To enable a fast computational assessment of a large number of molecules, and benefit from the existing, openly available database, we follow the approach of Koczor-Benda et al.[11] To assess the accuracy of ML property predictors along the biasing iterations, additional reference calculations at the same level of theory were performed whereby the thiol group in each molecule was modified to a gold-thiolate group. The Gaussian16 (ref. 50) software package was used to run DFT calculations and analysis tools from Molecular Vibration Explorer[12] were used to calculate P values. The pretrained KRR model from Koczor-Benda et al.[11] was used to predict P values; additionally, PaiNN[51] and MACE[52] equivariant GNN models were trained on the P values of the DFT-optimized structures of the THz database. Full details of training and hyperparameter optimization, as well as learning curves, are provided in the SI (Tables SIII, SIV and Fig. S1–S3). The PaiNN and MACE predictions of P values are based on the unrelaxed 3D structures of generated molecules, following Westermayr et al.[17]. Section 10 of the SI discusses the effect of using unrelaxed structures on the predicted P values for a subset of generated molecules.

2.2 Dimensionality reduction and clustering

To visualize the chemical space spanned by molecules within various datasets and to create inputs for subsequent cluster

analysis, dimensionality reduction *via* principal component analysis (PCA) was applied. The inputs for PCA were one of two applied molecular descriptors, henceforth referred to as bonding and structural descriptors. Structural descriptors were averaged SOAP[31] descriptors, obtained using the DScribe[53] package, which results in a 50820-dimensional description of molecules that encodes the average atomic environment around each atom. To obtain bonding descriptors from molecules, the Open Babel[40] and RDKit[30] software packages were used to extract as many interesting features as possible relating to molecular bonding. These ranged from simple quantities, such as the number of different elements within the molecule, to complex quantities such as the molecular aromaticity, resulting in a 403-dimensional bonding descriptor. Descriptor vectors were calculated for each molecule of the training database and used as inputs for PCA. To visualize the chemical space spanned by the training database in comparison with the spaces spanned by the generated molecules, the descriptor for generated molecules was represented using the same principal components as obtained from the training database. For clustering, a mixture of the balanced iterative reducing and clustering using hierarchies (BIRCH)[54] data mining algorithm and agglomerative clustering[55] was used to allow for uneven cluster sizes. Clustering was performed across the first three principal components of the bonding and structural descriptors, in addition to the PaiNN-predicted P values, weighted to achieve an approximately equal contribution of the first principal components of each descriptor and the predicted P value across all clusters.

2.3 Retrosynthetic planning

The AiZynthFinder[56] software was used for the retrosynthetic planning of selected molecules. The retrosynthesis algorithm is based on a Monte Carlo tree search that recursively breaks down a molecule to existing precursor molecules[56] based on a stock from compounds available within the ZINC[57] database. The tree search itself is guided by a policy that suggests possible precursors by utilizing a neural network trained on a library of known reaction templates. The employed policy[58] was trained on US patent office data,[59] as available within AiZynthFinder. The SMILES strings of molecules with successful retrosynthetic routes were cross-referenced against the PubChem[60,61] database using the PubChemPy[62] package.

3 Results and discussion

3.1 Analysis of generated molecules

The G-SchNet generative model is initially trained on the original dataset and used to generate novel and 'unbiased' molecules. A subset of the generated molecules is selected according to their predicted THz upconversion efficiency (high P value) and synthetic complexity (low SCScore) and added to the dataset. This process is repeated in six successive iterations during which properties of the generated molecules are driven towards the desired ranges (Fig. 1a and b). Iterative biased generation of molecules successfully leads to molecules with higher P and

lower SCScore in later iterations when compared to the training dataset ('Train') and the unbiased initial generation ('Unbiased'). Further shifts in property values after iteration 5 were not significant and biasing was stopped after Iteration 6.

The composition of generated compounds differs significantly from the training set, as shown by the elemental composition of molecules in Fig. 1d. The differences are largest between the training set and the unbiased generated molecules, which highlights the fact that G-SchNet, without biasing or conditioning, does not fully reproduce the chemical features of its training set. This shortcoming has been previously observed by Westermayr *et al.*[17] and Gebauer.[63] This effect is more significant for models trained on diverse datasets featuring many elements and molecular sizes than for models trained on small and simple molecules (such as QM9 (ref. 64 and 65)). The unbiased generated molecules feature a significantly reduced proportion of hydrogen atoms compared to the training dataset, which suggests increased numbers of unsaturated bonds and heteroatomic groups. The proportion of hydrogen atoms slightly increases through the subsequent biased iterations. Nitrogen atoms also become more prevalent in generated sets, while the proportion of carbon and fluorine atoms decreases. There is a shift of the size distribution of molecules to smaller values, as shown in Fig. S4a. While unbiased generation creates significant numbers of molecules with 30–60 atoms, generated molecules in later iterations have, on average, about 20 atoms. A significant number of molecules generated by the unbiased model have an SCScore above 4 (Fig. 1b), which was also observed by Westermayr *et al.*[17]. We note that all training molecules are commercially available so the SCScore metric does not fully reflect their accessibility but rather was used as an indicative metric by which we filter out generated molecules that are overly complex. For the most promising generated candidate molecules, we perform comprehensive retrosynthetic planning analysis to assess their synthesizability more accurately (*vide infra*).

As the training database only contained monothiols, the proportion of thiols in generated molecules is high, around 65% in the unbiased case, which increases in subsequent iterations to around 85%, as shown in Fig. S5 in the SI. It is interesting to see that the frequency of certain functional groups is significantly increased throughout the biasing iterations. An example of this is the aromatic amine group, which is present in only 0.5% of training molecules, but found in 9.8% of molecules generated by the unbiased G-SchNet model (Fig. 1c). By Iteration 6, 58.7% of generated molecules contain one or more aromatic amine groups. Simultaneously, the number of instances of this functional group per molecule also increases with iterations, as shown in Fig. 1c, with some of the generated molecules having as much as five aromatic amine groups. This functional group was identified by Koczor-Benda *et al.*[11] to correlate with high P values according to the ML predictor and as shown in Fig. 1e, the presence of this functional group also correlates with significantly higher predicted P values. We note that the sudden increase in the presence of this and other functional groups between the training and the unbiased
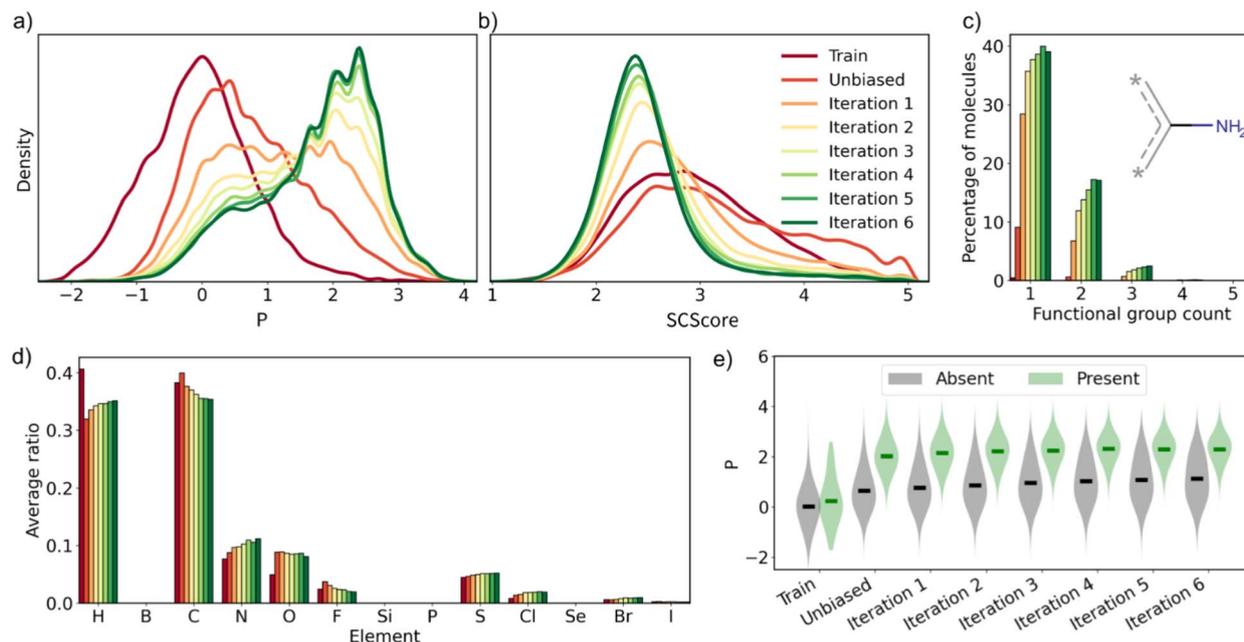
Fig. 1 Distribution of (a) predicted P values and (b) SCScore for molecules used for training G-SchNet (thiol database) and molecules generated in the biasing iterations. (c) Increase in relative occurrence and number of aromatic amine groups in molecules through the biasing iterations, (d) the average elemental composition of training and generated molecules; and (e) the distribution and mean average of P values predicted by the KRR model for molecules in which an aromatic amine is absent or present.

generated molecules could explain the significant shift in the predicted P value distribution between the two sets in Fig. 1a.

3.2 Evaluation and improvement of property predictors

As shown above, generated molecules significantly differ in chemical composition from the training molecules. This raises the question of whether the KRR predictor of the THz upconversion efficiency metric, P, provides transferable prediction accuracy for the novel, generated molecules – a crucial prerequisite for targeted property-driven molecular design. To assess this, DFT structure optimizations and vibrational spectrum calculations were performed on randomly selected molecules from the thiol database that was used to train the G-SchNet model and from the dataset generated in Iteration 6. Table 1 shows the performance of the KRR predictor on these molecules. The mean absolute error (MAE) on the Thiol database is

similar to the MAE on the test set of the THz database, while the MAE increases significantly for molecules generated in Iteration 6. In particular, the KRR model severely underestimates the P values of high-P molecules, as shown in the SI (Fig. S6), which suggests that the true P values of molecules generated in the biasing workflow reach much higher values than what is predicted in Fig. 1a.

As the KRR predictor uses SMILES strings as input and is based on 2D Morgan fingerprints, it does not benefit from the information contained in the 3D structures generated by G-SchNet. As the THz upconversion efficiency sensitively depends on the molecular conformation and vibrational frequencies, this limits the expressiveness and prediction accuracy of the model. We therefore trained two equivariant GNN models with 3D atom-wise embeddings on the same THz dataset, namely the MACE and PaiNN models. Table 1 compares the MAE of the different ML models for the reference DFT-calculated P values, determined for the DFT-optimized structures of test molecules from the THz dataset. Both MACE and PaiNN provide improved predictions compared to the EN and KRR models of Koczor-Benda et al.,[11] with PaiNN providing the best prediction. PaiNN also learns faster than MACE from less data, as shown by the learning curves in the SI (Fig. S3); for this reason, the PaiNN predictor was used for all subsequent analyses. When testing the PaiNN model on the molecules generated in Iteration 6, the MAE is larger with 0.73 (Table 1). PaiNN also underestimates the P values of high-P value molecules, as shown in the SI (Fig. S6), though this is slightly less pronounced than with KRR. Therefore, all tested models show reduced prediction accuracy when applied to the iteratively

Table 1 Performance of different ML models for P prediction, reported as mean absolute error for test molecules from the THz database, thiol database, and molecules generated in Iteration 6. EN and KRR models are taken from Koczor-Benda et al.[11] with predictions based on SMILES strings of molecules. In the case of MACE and PaiNN, predictions are based on DFT-optimized molecular structures

	Dataset		
Model	THz	Thiol	Iteration 6
EN[11]	0.60	—	—
KRR[11]	0.59	0.62	0.89
MACE	0.46	—	—
PaiNN	0.41	0.53	0.73

biased datasets, suggesting that the models are forced to predict outside of the chemical space spanned by the training data. This severely limits their ability to act as a transferable property predictor that drives molecule generation. The deterioration of the model accuracy for the THz upconversion efficiency is more significant than what was observed by Westermayr *et al.*[17] for electronic property prediction. We hypothesize that this is due to the integrated nature of the THz upconversion metric P and its sensitive dependence on collective low-frequency molecular vibrations and the molecular polarizability.

To alleviate the problem of underestimated high P values and the lack of transferability of the PaiNN predictor across the biased generation runs, the PaiNN predictor was retrained on a random subset of DFT-calculated P values from molecules generated in Iteration 6 and molecules from the thiol database. A committee of 5 PaiNN models was trained on different train/validation splits, and the mean average and standard deviation of their predictions were analyzed (SI, Fig. S7). The standard deviation of predictions was found to not correlate strongly with the absolute error of the prediction, indicating that the uncertainty of predictions cannot be used in an active learning-type workflow for augmenting the training set in a data-efficient way. After retraining, the mean average of the prediction becomes significantly more accurate for high P values, as shown in the SI (Fig. S8). The retrained PaiNN model achieves an MAE of 0.43 in P prediction on the Iteration 6 dataset which is consistent with the MAE previously achieved on the validation set when training on only the THz dataset (Table 1).

Equipped with a robust and transferable P predictor, new P values were predicted using the committee of 5 PaiNN models for all molecules in the training and generated molecule datasets (Fig. 2). Compared to the KRR predictions, the distribution of PaiNN-predicted P values for the generated molecules shifts to significantly higher values, with the highest predicted P value reaching 7.30. The presence of specific functional groups can be analyzed alongside the PaiNN predictions for P values. This analysis (SI, Fig. S9), indicates that some of the promising features identified by Koczor-Benda *et al.*,[11] such as the aromatic

Fig. 2 Distribution of PaiNN predictions (full lines) and original KRR predictions (dotted lines) for P values on all training and generated molecules. In the case of PaiNN, the distributions show the mean predicted P value by a committee of 5 PaiNN models that were trained on the original THz database augmented by randomly selected molecules from the G-SchNet training database (thiols) and molecules generated in Iteration 6.

amine group (Fig. 1d), correlate with higher P values in the generated molecules as well as in the training set of commercial thiols.

3.3 Analysis of the chemical space of generated molecules

Structural and bonding descriptors were calculated for all generated molecules. Principal components of these descriptors span a latent representation of the chemical space covered by the molecules. A heat map of the distribution of molecules in this latent space is projected into the basal plane of Fig. 3a, where it is clear that molecular generation is prioritized in a specific region of latent space. Previous efforts at biasing G-SchNet have shown significant localization in such latent chemical spaces as biasing iterations proceed.[17] This can be visualized by separating out the contributions of each iteration, as shown in the SI (Fig. S10). However, unlike in Westermayr *et al.*,[17] in this work, we did not find a clear correlation between the progression of biasing iterations and the occupied chemical space decreasing in size; while there was an initial decrease in the covered area for the molecules of the unbiased generation, the molecules in successive iterations did not localize any further to one particular area of chemical space. This is because we retain original molecules in each biasing iteration, but will likely relate to the P value biasing target being less related to specific changes in functional groups and chemical composition. The P value is likely more closely related to several features that can appear across a diverse range of molecules.

To better resolve the types of molecules that were being generated in different areas of the latent space, the heat map in Fig. 3a was expanded through the inclusion of the PaiNN-predicted P values and was clustered as previously described. These clusters are also shown in Fig. 3a, with data points corresponding to their counterparts in the heat map. Many of the clusters span a wide range of P values and a large area of latent space, indicating that there is little correlation between the latent space and the THz radiation sensitivity of each molecule, again signifying that the P value is a complex biasing target. This leads to inefficiency in the biasing procedure, as structurally similar molecules can result in dramatically different P value predictions. The high-density region of the heat map results in many closely packed clusters, while the lower-density regions are inhabited by fewer large clusters. We note that while the sheer number of data points makes it difficult to see all the clusters, it is clear that some generated molecules with high P values, clustered near the top of Fig. 3a, have the potential to perform very well for THz radiation detection.

To perform further analysis, each cluster was subsampled to find the twenty closest molecules to the centroid of each cluster (Fig. 3b). While the subsampling omits molecules at the edges of the respective clusters, it allows for analysis of the nature of the molecules that exist in each cluster. The densely packed region of the latent space is now more visible, with over half of the clusters localized in a narrow slice of the bonding/structural principal component space on the right of the plot.

Five subclusters (labeled C1–C5 and indicated in Fig. 3b) were chosen for detailed analysis, to establish trends in the
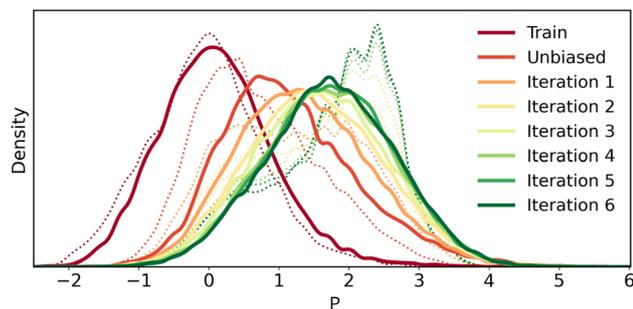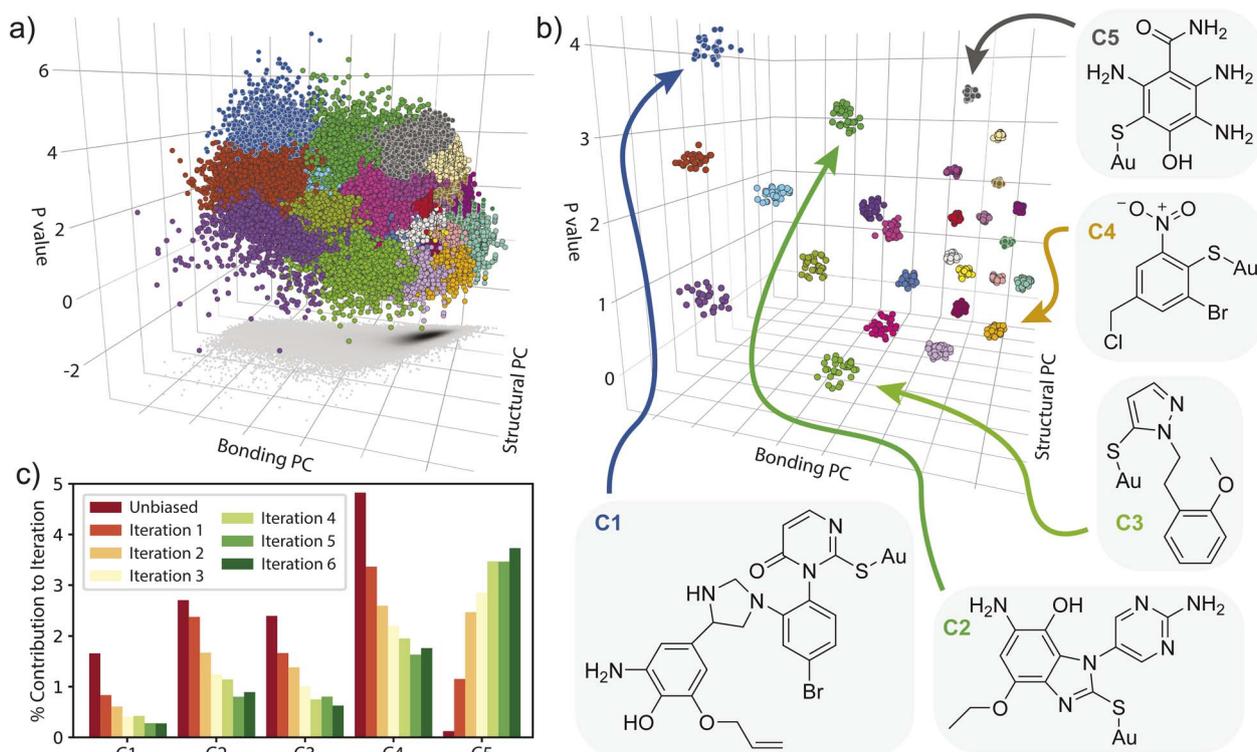
Fig. 3 Latent chemical space clustering results for all generated molecules. Shown are: (a) generated molecules in the latent space formed by the first principal components (PCs) of the bonding and structural descriptors, separated vertically by their predicted P values and clustered with respect to these axes. The bottom plane depicts the density of points within the principal component space, with darker areas indicating regions of high density; (b) subsamples of clusters around their centroids to reveal the 20 most representative molecules for each cluster, with illustrative examples from five such subclusters (C1–C5) shown; (c) separation of molecules in their respective clusters from (a) into contributions from each biasing iteration to reveal trends in the types of molecules that are prioritized and penalized during iterative biasing.

types of molecules that were being predicted and the features that increase or reduce the predicted P value. Statistics for the molecules in these subclusters are shown in Table 2. Subclusters C1 and C2 show high average P values. They are both composed of highly conjugated molecules with numerous aromatic rings. These contained a variety of heteroatomic functional groups, including alcohols and aromatic amines, as previously noted in Fig. 1c, and both subclusters contained very few molecules with halogen substituents. The main difference between molecules in these subclusters was their overall size – molecules in C1 were generally larger and contained more aromatic rings.

Subcluster C5 also exhibits a large average P value, although it differed from subclusters C1 and C2 due to all of its molecules

being much smaller and centred around a single highly substituted benzene ring. Molecules in this subcluster contain a high proportion of aromatic amine groups, in addition to other oxygen- and nitrogen-containing groups. Again, there were very few halogenated molecules present. This is in direct contrast to the molecules of subcluster C4, which were also based around a single benzene ring but were predicted to have a very low P value. These rings were characterized by being less heavily substituted than those in C5 and contained a comparatively high proportion of halogens and nitro groups, the latter of which were not found in any high-P value clusters. It is notable that these subclusters, and indeed all of those in the previously noted high-density region of the latent space heat map, were based around substituted benzene molecules.

Finally, molecules within subclusters C3 and C2 are structurally very similar when judged from their vicinity in the principal component latent space. However, molecules in subcluster C3 exhibit much lower P values than molecules in C2. While C3 molecules contain aromatic rings, all molecules lacked conjugation between these rings due to aliphatic joining chains. Compared to the other high-P value subclusters, their rings were also significantly less substituted, and molecules were less heteroatomic overall.

We can conclude that molecules with high predicted P values fall into one of two categories: either they are large, conjugated

Table 2 Statistics for the generated molecules in the chosen subclusters shown in Fig. 3, including PaiNN-predicted P values

Subcluster	Average P value	SCScore	Number of atoms
C1	4.1	3.9–4.9	50–59
C2	3.2	3.3–3.4	35–40
C3	0.1	2.7–3.8	28–33
C4	−0.2	1.6–2.9	17–0
C5	3.4	2.2–3.0	21–25

aromatic systems, or they are smaller, highly substituted benzene rings. In both cases, the presence of oxygen and nitrogen-based substituents (particularly amines) was desired, while halogenation and nitro groups lead to lower P values.

To establish how the presence of each of these types of molecules varied over the biasing iterations, each analyzed subcluster's respective full cluster was separated out into a percentage contribution to each iteration, as shown in Fig. 3c. While C1, C2, C3 and C4 all contributed less to each iteration as biasing proceeded, C5 contributed significantly more, indicating that G-SchNet was consistently biased towards molecules similar to those in subcluster C5. This is sensible when the multi-property biasing task that was undertaken is considered, as the molecules in subcluster C5 were smaller and chemically simpler than those in subclusters C1 and C2, thereby receiving a lower SCScore since they would be simpler to synthesize. Since molecules in subcluster C5 have a relatively high P value and a relatively low SCScore, they were prioritized; molecules in subclusters C1 and C2 were too complex, yielding a higher SCScore, while molecules in subclusters C3 and C4 were simpler but had a low predicted P value, so molecules from these clusters did not fulfil the multi-property biasing criteria.

3.4 Identification of candidate molecules

We selected generated molecules with $P \geq 4.25$ (based on predictions by the retrained PaiNN predictor) and employed AiZynthFinder to perform retrosynthetic planning. From the 1011 molecules satisfying this selection criterion, only 34 were predicted to have retrosynthetic routes from purchasable precursors[56] based on a stock from compounds available within the ZINC[57] database; retrosynthetic paths for these molecules can be found in Fig. S11–S17. Notably, all 34 molecules belong to clusters from which subclusters C2 and C5 were drawn (SI, Table SVI).

To confirm the suitability of these molecules for THz radiation detection, their absorption, Raman scattering and frequency upconversion spectra were calculated, and their P values were determined using DFT. Fig. 4 shows the relevant properties and vibrational spectra of the top candidate, while vibrational spectra and properties of other candidate molecules with DFT-calculated P values above 5.20 are shown in the SI (Fig. S18–S21). The top candidate, 2-amino-5-(4-aminophenylamino)pyridine-4-thiol, has a DFT-calculated P value of 7.88. Considering that the P value is a logarithmic quantity (eqn (1)), this is significantly higher than any of the molecules previously identified within commercial databases in Koczor-Benda et al.,[11] where the top 5 candidates had P values between 5.30 and 6.18. For the fifth top molecule from Koczor-Benda et al.,[11] 5-amino-2-mercaptobenzimidazole, Redolat et al.[5] developed a functionalization technique to prepare self-assembled molecular monolayers in gold-based plasmonic nanocavities and successfully integrated these nanocavities on a silicon-based photonic chip. While frequency upconversion measurements are not yet available for this compound, our DFT simulations suggest about 14 times higher upconversion capability in the THz/mid-infrared range for the most active mode
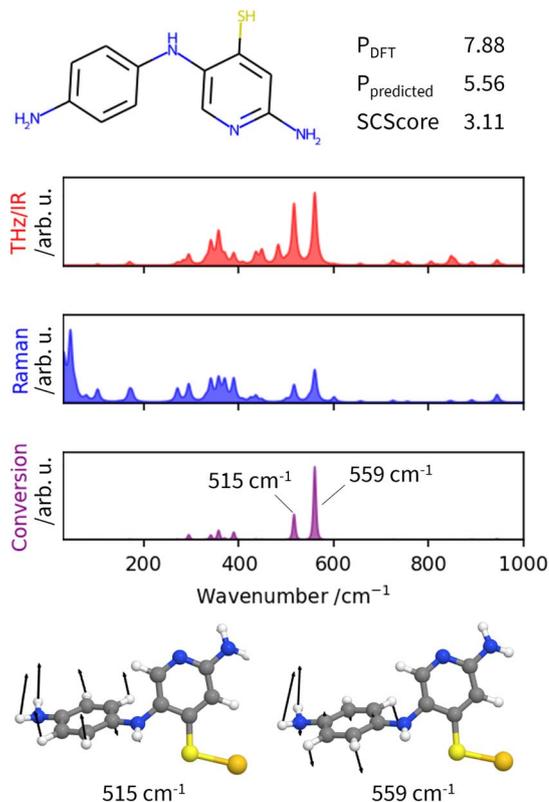

Fig. 4 Properties of the top candidate molecule, 2-amino-5-(4-aminophenylamino)pyridine-4-thiol, generated by G-SchNet. Density functional theory (DFT)-calculated (P_{DFT}) and PaiNN-predicted ($P_{predicted}$) P values, predicted SCScore, as well as DFT-calculated terahertz (THz)/infrared (IR) radiation absorption, Raman scattering and frequency upconversion spectra are shown. The two most intensive vibrational modes for frequency upconversion are also depicted.

(559 cm^{-1}) of our top candidate compared to the most active mode (458 cm^{-1}) of 5-amino-2-mercaptobenzimidazole (see Fig. S22 in SI).

The top molecule has two vibrational modes that are highly active in frequency upconversion, which are located at 515 cm^{-1} and 559 cm^{-1}. Both modes involve an out-of-plane (umbrella) motion of one of the amino groups that is coupled to out-of-plane motion of hydrogen atoms of the neighboring ring. This out-of-plane motion of the amino group is also responsible for the highest intensity peaks of other top candidates, as shown in the SI (Fig. S18–S21). This provides evidence that the aromatic amine functional group not only correlates with high P values, but is also directly involved in the upconversion process. The highly active mode appears in the $515–832 \text{ cm}^{-1}$ spectral range for the top candidates, showing that the chemical environment and the coupling of the out-of-plane motion of the amino group with other vibrations of the molecule have a significant effect on the position of the peak. This can be advantageous for the tuning of narrowband THz radiation detectors operating at different frequencies. We note that this does not mean that all molecules that contain amino groups are necessarily good

candidates for frequency upconversion: the spectral intensities are heavily influenced by other functional groups within the molecule, such as the thiol group, and the top candidates rely on the intricate interplay of atomic motion from the whole molecule to achieve outstanding frequency upconversion properties. We also note that within the top candidates, molecules with the same SMILES string were generated multiple times with different 3D structures in the different biasing iterations. As the SCScore and KRR-predicted P values depend only on 2D information, they remain the same for different conformers. However, the PaiNN-predicted P values for raw generated structures and DFT-calculated P values for structures that have undergone geometry optimization can differ, as shown in Section S10 and Fig. S24 of the SI. This further highlights the benefits of working with property predictors that are based on 3D descriptors.

Of the 34 molecules listed in Table SVI, only one compound (generated three times as different conformers, all sharing the same SMILES string) was identified in the PubChem[60,61] database, Nc1cc(S)c(cc1N)N, which corresponds to 2,4,5-triaminobenzenethiol (Compound Identifier 67981805 (ref. 66)). The remaining 31 molecules were not found in PubChem, likely representing novel candidate structures for THz upconversion applications.

4 Conclusions and outlook

Generative design of functional organic molecules can be biased towards certain properties by iteratively adapting the underlying training dataset. Here, we do this to design candidate molecules for THz radiation detection by mixing molecules from an existing database with selected molecules created by the autoregressive generative deep learning model G-SchNet. This enables us to perform property-driven design of novel and synthesizable monothiolated molecules with high THz-to-visible upconversion efficiencies. By performing a comprehensive structural analysis on the dataset of generated molecules, we have revealed key chemical trends among generated molecules and identified functional groups that contribute to enhanced upconversion, such as aromatic amines. From the novel, generated molecules, we were able to select several candidates and provide potential retrosynthetic pathways from commercially available reactants. The top candidate molecule has a DFT-calculated THz upconversion efficiency of 7.88, which is significantly higher than any of the molecules previously identified from commercial databases.

This work also revealed several practical challenges associated with property-driven generative design that require careful consideration when designing such workflows. First of all, we have seen that even unbiased molecular generation in G-SchNet creates a distribution of molecules that significantly differs from the training dataset in terms of elemental and functional group composition. If the model cannot capture the chemical space spanned by the data, this means that the ability of the property-driven design workflow to drive the generation in a directed way is limited. The performance of G-SchNet and other generative algorithms in this regard needs to be analysed

in greater detail in the future. Secondly, during sequential iterations of biasing with a changing training dataset, the ML-based property predictor that selects suitable molecules must continue to provide accurate predictions. We showed that GNN-based ML predictors, based on MACE and PaiNN models and 3D input structures, gave more accurate P values than predictors based on 2D molecular fingerprints. The figure of merit of THz upconversion efficiency, P, was shown to be a highly integrated quantity that is challenging to learn due to its dependence on low-lying vibrational modes. Careful validation revealed that contrary to previous work on the property-driven generative design of fundamental electronic gaps[17] none of the P predictors trained on the original data set were transferable to the newly generated molecules. Their prediction accuracy deteriorated during the iterative biasing workflow. Therefore, the PaiNN predictor had to be retrained based on new DFT training data. Uncertainty-based active learning during biasing iterations would not have been a robust strategy due to the lack of correlation between prediction accuracy and uncertainty in highly regularized GNNs. Therefore, active learning based on structural diversity sampling is likely a more robust choice to retain ML predictor performance throughout the iterative biasing procedure.

Significant future work will be needed to make property-driven generative design workflows more efficient and robust. To this end, constrained generation with (semi-)supervised generative models such as constrained G-SchNet[18] that can constrain specific functional groups or diffusion models able to perform inpainting tasks will likely be beneficial. This would reduce the portion of generated molecules that are discarded during the workflow due to the absence of a thiol group. The question of whether generative models faithfully represent the structural and functional group distribution of the underlying training dataset requires further attention. Commonly, generative models are only assessed on their ability to generate valid and unique molecules, which is insufficient when aiming to employ models for directed exploration of chemical space.

Both the property-driven design workflow and the novel candidate molecules we have identified in this study will contribute to advancing the discovery of functional organic materials for nanosensor applications such as THz radiation detection. Our results highlight the potential of generative models to not only expand the chemical space of viable molecules but also to guide future experimental and computational efforts in the molecular design of plasmonic nanocavities.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data for this article, including molecular databases in ASE database format, DFT-optimized best candidate molecules, and ASE databases for xTB calculations are available online: **https://doi.org/10.6084/m9.figshare.28539995.v3**.[67] The repository also contains the trained ML models, and Jupyter Notebooks and

scripts associated with the generation, prediction, and analysis workflows described in this work. Code for the extraction of bonding features from molecular databases and obtaining the principal components of the structural/bonding descriptors has been released in our GSchNetTools package, available at **https://github.com/maurergroup/GSchNetTools**. The SCScore model used in this work is publicly available at **https://github.com/connorcoley/scscore**, and files pertaining to retrosynthetic planning with AiZynthFinder are publicly available at **https://figshare.com/articles/dataset/ AiZynthFinder_a_fast_robust_and_flexible_open-source_software_for_retrosynthetic_planning/12334577**.

Supplementary data including numerical convergence results and additional data is available. See DOI: **https://doi.org/10.1039/d5dd00106d**.

References

1 M. Tonouchi, *Nature Photon.*, 2007, **1**, 97–105.

2 S. S. Dhillon, M. S. Vitiello, E. H. Linfield, A. G. Davies, M. C. Hoffmann, J. Booske, C. Paoloni, M. Gensch, P. Weightman, G. P. Williams, E. Castro-Camus, D. R. S. Cumming, F. Simoens, I. Escorcia-Carranza, J. Grant, S. Lucyszyn, M. Kuwata-Gonokami, K. Konishi, M. Koch, C. A. Schmuttenmaer, T. L. Cocker, R. Huber, A. G. Markelz, Z. D. Taylor, V. P. Wallace, J. Axel Zeitler, J. Sibik, T. M. Korter, B. Ellison, S. Rea, P. Goldsmith, K. B. Cooper, R. Appleby, D. Pardo, P. G. Huggard, V. Krozer, H. Shams, M. Fice, C. Renaud, A. Seeds, A. Stöhr, M. Naftaly, N. Ridler, R. Clarke, J. E. Cunningham and M. B. Johnston, *J. Phys. D: Appl. Phys.*, 2017, **50**, 043001.

3 P. Roelli, C. Galland, N. Piro and T. J. Kippenberg, *Nature Nanotech.*, 2016, **11**, 164–169.

4 P. Roelli, D. Martin-Cano, T. J. Kippenberg and C. Galland, *Phys. Rev. X*, 2020, **10**, 031057.

5 J. Redolat, M. Camarena-Pérez, A. Griol, M. S. Lozano, M. I. Gómez-Gómez, J. E. Vázquez-Lozano, E. Miele, J. J. Baumberg, A. Martínez and E. Pinilla-Cienfuegos, *Nano Lett.*, 2024, **24**, 3670–3677.

6 A. Xomalis, X. Zheng, R. Chikkaraddy, Z. Koczor-Benda, E. Miele, E. Rosta, G. A. E. Vandenbosch, A. Martínez and J. J. Baumberg, *Science*, 2021, **374**, 1268–1271.

7 W. Chen, P. Roelli, H. Hu, S. Verlekar, S. P. Amirtharaj, A. I. Barreda, T. J. Kippenberg, M. Kovylina, E. Verhagen, A. Martínez and C. Galland, *Science*, 2021, **374**, 1264–1267.

8 F. Neubrech, C. Huck, K. Weber, A. Pucci and H. Giessen, *Chem. Rev.*, 2017, **117**, 5110–5145.

9 P. L. Stiles, J. A. Dieringer, N. C. Shah and R. P. Van Duyne, *Annu. Rev. Anal. Chem.*, 2008, **1**, 601–626.

10 C. Humbert, T. Noblet, L. Dalstein, B. Busson and G. Barbillon, *Materials*, 2019, **12**, 836.

11 Z. Koczor-Benda, A. L. Boehmke, A. Xomalis, R. Arul, C. Readman, J. J. Baumberg and E. Rosta, *Phys. Rev. X*, 2021, **11**, 041035.

12 Z. Koczor-Benda, P. Roelli, C. Galland and E. Rosta, *J. Phys. Chem. A*, 2022, **126**, 4657–4663.

13 R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. Sik Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. Ik Hong, M. Baldo, R. P. Adams and A. Aspuru-Guzik, *Nature Mater.*, 2016, **15**, 1120–1127.

14 H. Sahu, F. Yang, X. Ye, J. Ma, W. Fang and H. Ma, *J. Mater. Chem. A*, 2019, **7**, 17480–17488.

15 A. Saeki and K. Kranthiraja, *Jpn. J. Appl. Phys.*, 2020, **59**, SD0801.

16 V. Chechik and C. J. M. Stirling, Gold–Thiol Self-Assembled Monolayers, in *Patai's Chemistry of Functional Groups*, ed. Z. Rappoport, Wiley, 1999.

17 J. Westermayr, J. Gilkes, R. Barrett and R. J. Maurer, *Nat. Comput. Sci.*, 2023, **3**, 139–148.

18 N. W. A. Gebauer, M. Gastegger, S. S. P. Hessmann, K.-R. Müller and K. T. Schütt, *Nat. Commun.*, 2022, **13**, 973.

19 R. P. Joshi, N. W. A. Gebauer, M. Bontha, M. Khazaieli, R. M. James, J. B. Brown and N. Kumar, *J. Phys. Chem. B*, 2021, **125**, 12166–12176.

20 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.

21 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. Miguel Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.

22 J. Meyers, B. Fabian and N. Brown, *Drug Discov. Today*, 2021, **26**, 2707–2715.

23 J. Arús-Pous, A. Patronov, E. J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen and O. Engkvist, *J. Cheminform.*, 2020, **12**, 38.

24 W. Kong, Y. Hu, J. Zhang and Q. Tin, *Front. Pharmacol.*, 2022, **13**, 1046524.

25 N. W. A. Gebauer, M. Gastegger and K. T. Schütt, NeurIPS Proceedings, Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules, in *Advances in Neural Information Processing Systems 32*, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox and R. Garnett, 2019.

26 J. Westermayr and P. Marquetand, *Chem. Rev.*, 2021, **121**, 9873–9926.

27 *eMolecules database*, **https://www.emolecules.com**, accessed 01 March 2020.

28 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.

29 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519.

30 G. Landrum, *RDKit: Open-source cheminformatics*, **http://www.rdkit.org/**, accessed November 13, 2024.

31 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **87**, 184115.

32 L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Comput. Phys. Commun.*, 2020, **247**, 106949.

33 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **23**, 2825–2830.

34 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.

35 N. Gebauer and K. T. Schütt, *Conditional G-SchNet extension for SchNetPack 2.0 – A generative neural network for 3d molecules*, **https://github.com/atomistic-machine-learning/schnetpack-gschnet**, accessed November 13, 2024.

36 K. T. Schütt, S. S. P. Hessmann, N. W. A. Gebauer, J. Lederer and M. Gastegger, *J. Chem. Phys.*, 2023, **158**, 144801.

37 K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko and K.-R. Müller, NeurIPS Proceedings SchNet: a continuous-filter convolutional neural network for modeling quantum interactions, in *Advances in Neural Information Processing Systems 30*, ed. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, 2017.

38 C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, *J. Chem. Inf. Model.*, 2018, **58**, 252–261.

39 A. J. Lawson, J. Swienty-Busch, T. Géoui and D. Evans, *The Making of Reaxys–Towards Unobstructed Access to Relevant Chemistry Information*, American Chemical Society, 2014, pp. 127–148.

40 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminform.*, 2011, **3**, 33.

41 P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864–B871.

42 W. Kohn and L. Sham, *Phys. Rev.*, 1965, **140**, A1133–A1138.

43 A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098.

44 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785.

45 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**.

46 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.

47 J. Griffiths, T. Földes, B. de Nijs, R. Chikkaraddy, D. Wright, W. M. Deacon, D. Berta, C. Readman, D.-B. Grys, E. Rosta, et al., *Nat. Commun.*, 2021, **12**, 6759.

48 A. Boehmke Amoruso, R. A. Boto, E. Elliot, B. de Nijs, R. Esteban, T. Földes, F. Aguilar-Galindo, E. Rosta, J. Aizpurua and J. J. Baumberg, *Nat. Commun.*, 2024, **15**, 6733.

49 D. Wright, Q. Lin, D. Berta, T. Földes, A. Wagner, J. Griffiths, C. Readman, E. Rosta, E. Reisner and J. J. Baumberg, *Nat. Catal.*, 2021, **4**, 157–163.

50 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussiañ16 Revision C.01*, Gaussian Inc. Wallingford CT, 2016.

51 K. Schütt, O. Unke and M. Gastegger, Proceedings of Machine Learning Research, Equivariant message passing for the prediction of tensorial properties and molecular spectra, in *Proceedings of the 38th International Conference on Machine Learning*, ed. M. Meila and T. Zhang, 2021.

52 I. Batatia, D. P. Kovacs, G. Simm, C. Ortner and G. Csanyi, NeurIPS Proceedings MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields, in *Advances in Neural Information Processing Systems 35*, ed. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh, 2022.

53 L. Himanen, M. O. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Comput. Phys. Commun.*, 2020, **247**, 106949.

54 T. Zhang, R. Ramakrishnan and M. Livny, *Data Min. Knowl. Discov.*, 1997, **1**, 141–182.

55 E. Schubert, J. Sander, M. Ester, H. P. Kriegel and X. Xu, *ACM Trans. Database Syst.*, 2017, **42**, 1–21.

56 S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, *J. Cheminform.*, 2020, **12**, 70.

57 T. Sterling and J. J. Irwin, *J. Chem. Inf. Model.*, 2015, **55**, 2324–2337.

58 A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist and E. J. Bjerrum, *Chem. Sci.*, 2020, **11**, 154–168.

59 D. Lowe, *Chemical reactions from US patents (1976–Sep 2016)*, 2017, DOI: **10.6084/m9.figshare.5104873.v1**, accessed November 13, 2024.

60 *PubChem*, **https://pubchem.ncbi.nlm.nih.gov/**, accessed February 12, 2025.

61 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton, *Nucleic Acids Res.*, 2025, **53**, D1516–D1525.

62 *PubChemPy documentation*, https://pubchempy.readthedocs.io/en/latest/, accessed February 12, 2025.

63 N. W. A. Gebauer, PhD thesis, Technische Universität Berlin, 2024.

64 L. Ruddigkeit, R. van Deursen, L. C. Blum and J.-L. Reymond, *J. Chem. Inf. Model.*, 2012, **52**, 2864–2875.

65 R. Ramakrishnan, P. O. Dral, M. Rupp and O. Anatole von Lilienfeld, *Science*, 2014, **1**, 140022.

66 *PubChem Compound Summary for CID 67981805, 2,4,5-Triaminobenzenethiol*, https://pubchem.ncbi.nlm.nih.gov/compound/2_4_5-Triaminobenzenethiol, accessed February 12, 2025.

67 Z. Koczor-Benda, S. Chaudhuri, J. Gilkes, F. Bartucca, L. Li and R. J. Maurer, *G-SchNet for THz Radiation Detection*, 2025, DOI: 10.6084/m9.figshare.28539995.v1, accessed March 10, 2025.