nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-22611-4 OPEN

Active discovery of organic semiconductors

Christian Kunkel[1], Johannes T. Margraf[1], Ke Chen[1], Harald Oberhofer[1] & Karsten Reuter[1,2]✉

The versatility of organic molecules generates a rich design space for organic semiconductors (OSCs) considered for electronics applications. Offering unparalleled promise for materials discovery, the vastness of this design space also dictates efficient search strategies. Here, we present an active machine learning (AML) approach that explores an unlimited search space through consecutive application of molecular morphing operations. Evaluating the suitability of OSC candidates on the basis of charge injection and mobility descriptors, the approach successively queries predictive-quality first-principles calculations to build a refining surrogate model. The AML approach is optimized in a truncated test space, providing deep methodological insight by visualizing it as a chemical space network. Significantly outperforming a conventional computational funnel, the optimized AML approach rapidly identifies well-known and hitherto unknown molecular OSC candidates with superior charge conduction properties. Most importantly, it constantly finds further candidates with highest efficiency while continuing its exploration of the endless design space.

[1] Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, Garching, Germany. [2] Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany. ✉email: reuter@fhi-berlin.mpg.de

The sheer vastness of chemical spaces[1] has long motivated prior-to-synthesis virtual discovery. In corresponding work, promising candidate molecules or materials for refined study are often searched and identified on the basis of a small number of quantities that are deemed representative for the targeted application[2–4]. Prevalent for first-principles computational screening approaches is to calculate such descriptors at predictive quality through electronic structure theory for every candidate in a somehow enumerated chemical space or otherwise given database. Initially performed for small focused libraries, the screening is now extended to search spaces of ever increasing size and—since discovery is limited to the explicitly considered molecules or materials—to ever more systematic and exhaustive enumerations within these spaces.

Unfortunately, the combinatorial explosion characteristic for chemical versatility quickly leads to intractable numbers of candidates for such exhaustive first-principles screenings, even if based on computationally comparably undemanding descriptors. A common strategy to tackle this problem is a computational funnel[5]. Here, the exhaustive screening is only performed for computationally least-demanding descriptors or even less demanding estimates thereof. Subsequently, the large candidate set is narrowed in staged filtering and the calculation of other descriptors is only performed for smaller and smaller subsets which appear promising in terms of the previously calculated descriptors. Unfortunately, chemical diversity suggests the multi-objective (descriptor) landscape spanned over the search space to be quite rugged[6], with molecular or materials sub-classes likely constituting separate funnels and related analogs leading to multiple local minima. This raises concerns whether the true optimum candidates can reliably be identified through such computational funneling.

An ever more appealing alternative is therefore to completely abandon the original idea to exhaustively screen a once defined chemical space or database. Instead, the explicit first-principles computation of the descriptors is restricted to candidates emerging in an iteratively refining search[7–9]. In the context of data science, this is afforded by several learning concepts, which additionally allow to even avoid predefining or a priori enumerating the search space itself. Examples include (semi-)supervised learning, meta-, transfer-, or few-shot learning and generative models[10,11]. For drug-discovery tasks[12,13], such concepts have already been successfully employed to further accelerate molecular de novo design[14] and drive autonomous discovery[15]. For materials discovery based on first-principles descriptors, in particular active machine learning (AML)[16] has been explored as a most data-efficient method[17–22].

In AML, the acquired knowledge in form of explicitly calculated descriptors is used to successively establish a surrogate model of larger and larger regions of the rugged descriptor landscape. In an iterative procedure, the predictive-quality calculations for new candidates can then also be balanced between exploitation and exploration. In exploitation, the global insight provided by the current surrogate model is used for a targeted identification of new promising candidates. In exploration, descriptors for new candidates are specifically calculated to refine and extend the surrogate model. For this, we here employ Gaussian Process Regression (GPR) and use high values of its inherent Bayesian uncertainty estimate to flag candidates (or regions in chemical space) for which an explicit descriptor calculation will maximally contribute new information.

We pursue this concept for the efficient virtual discovery of organic semiconductors (OSCs) for electronic applications. Used in organic field effect transistors (OFETs),[23] photovoltaics (OPVs),[24] or light emitting diodes (OLEDs),[25] OSCs offer great versatility and novel materials' properties, paired with a low ecologic and economic footprint. Typical OSC-constituting molecules are, however, of considerable size (e.g., 22 or 42 non-hydrogen atoms in the classic examples pentacene or rubrene, respectively) and the spanned electronic property landscapes are known to be highly sensitive even to small molecular substitutions.[26–28] A vast number of ~10^{33} similar-sized molecules is estimated to be synthesizable[1], raising the suspicion that presently known well-performing OSC molecular materials are not even the tip of the iceberg. This has motivated a number of preceding exhaustive screening or virtual discovery studies in more or less restricted closed subspaces.[3,5,29–34].

In this work we first analyze a diverse set of OSC molecules to derive clear molecular-construction rules that allow to generate an in principle unlimited OSC chemical space. This space is then successively explored by the AML discovery strategy, rapidly identifying molecular candidates that are superior to well-known OSC materials in terms of their molecular electronic descriptors assessing efficient charge injection and charge mobility. Deep methodological insight is gained by analyzing and visualizing the AML exploration inside a chemical space network (CSN) containing only a subset of the design space, limited to allow its full enumeration. Even inside this truncated chemical space the AML-discovery clearly outperforms a conventional funnel approach.

Results

Morphing based generation of an unlimited OSC search space. The basis for our efficient AML exploration of an a priori unlimited molecular search space is the development of a concise set of molecular construction rules that allow to generate this space by iterative application. To establish a diverse, but problem-specific chemical space, we resort to existing domain knowledge and analyze the building blocks and motives contained in molecules constituting a number of well-performing crystalline OSC molecular materials. For this analysis, we exploit the fact that most functionalized organic molecules can be unambiguously fragmented into a molecular backbone (of one or more cores), linkers (that connect cores) and side groups (attached to cores) as illustrated in Fig. 1. Without loss of generality, we correspondingly fragment 30 prominent π-conjugated molecules that belong to a variety of important molecular families[23] (Acenes, Thienoacenes, TTF-derivatives, Carbazoles, Triphenylamines, Diimides, Quinacridones and Azaacenes) and consist of the most common organic elements C, H, N, O and S. Figure 1 highlights some of these peer molecules and the full set is given in the SI in Supplementary Fig. 1. Intriguingly, the richness of chemical building blocks identified in this way can be exhaustively generated by a set of only 22 simple molecular morphing operations starting from the smallest aromatic building block benzene. As illustrated in Fig. 1 these morphing operations each act on a molecule's individual atomic sites or fragments, each time adding, modifying or removing fragments. These morphing operations should be seen as alchemical transformations to navigate between molecules, while applying organic synthesis steps could be a viable alternative.[35] Even though at a first glance rather unintuitive for the generation of successively larger or complex molecules, we also note that the inclusion of every morphing operation in a backwards step, i.e., resubstituting a fragment substructure, is crucial to increase the interconnectivity of the forming chemical space, see Supplementary Fig. 3.

The generic nature of the morphing operations identified through the fragmentation ansatz is not only a stepping stone for the efficient AML exploration. It also provides a blueprint for future variations of the present search space or the generation of different search spaces for other applications. Additional morphing operations will lead to more general search spaces and could be automatically extracted from a diverse chemical database[36],

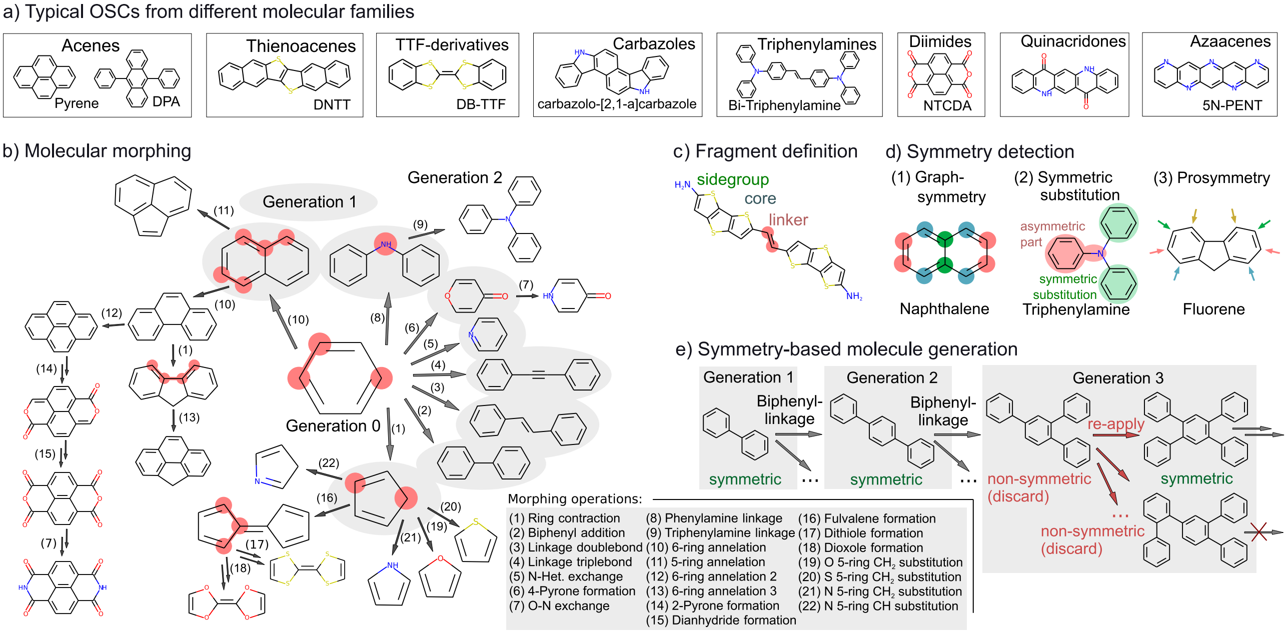

Fig. 1 Molecular construction approach to generate an unlimited OSC chemical space. a Important π-conjugated molecular families and examples of well-performing OSC-molecules therein. Molecular morphing operations are designed such that the generated OSC space includes these families. **b** Schematic overview of the molecular generation process. Starting from benzene, diverse molecules are created by iterative application of up to 22 morphing operations. The first generation resulting from the 8 morphing operations applicable to benzene is fully shown. Molecules in further generations are only shown as examples, but every operation type is depicted at least once, see also Supplementary Fig. 2 for an extended depiction. **c** Fragment-definitions used throughout the text exemplified for the molecule BDTTE. Connected aromatic ring structures are cores. Linkers and sidegroups both branch from a core structure with a single bond, but are either connecting to at least two core structures or only bonded to one core fragment. **d** Concepts for symmetry detection used throughout the molecular generation process. **e** Modified molecular morphing step, adapted to the symmetry constraints imposed on candidate molecules.

while deliberate suppression of morphing operations can be used to focus on molecular sub-classes. Ring-annelation type morphing operations as well as biphenylic addition are for example essential for the iterative construction of core Acene fragments, such as in Pyrene or DPA. To build structures like Thienoacenes, Azaacenes or Carbazoles, ring contractions that lead to 5-membered rings are included as intermediates for heteroaromatic ring construction. This, though, comes at the cost of potentially yielding pericyclically reactive molecules, as discussed further below. Similarly, two types of linker operations are included to access the family of Triphenylamines. Further examples together with a detailed description of every morphing operation are provided in Supplementary Note 1. Considering their known OSC tuning potential,[28,37,38] we note that in particular the augmentation of the present backbone-oriented set of construction rules by specific morphing operations for side groups or additional functional groups is expected to lead to an important extension of the here showcased search space.

The construction rules may also be modified to incorporate further prior knowledge about the OSC design problem. Here, we notably include constraints on molecular symmetry. Molecular symmetry may be beneficial for synthetic accessibility. Furthermore, it can mitigate mobility reducing charge localization[27] and in particular in monomolecular crystals often favors charge percolation pathways[3,39,40] (albeit its role can be intricate[41]). We correspondingly prune the construction rules for the present OSC context to enforce 2D graph symmetries expected to provide a prosymmetry for the 3D case. Specifically, generated molecules are only considered for further morphing, if they fall into three types of symmetry classes as explained in Fig. 1d, e: They (1) exhibit a full graph-symmetry, with all atomic environments appearing at least twice. (2) An asymmetric part in the molecule made of one or more fragments is symmetrically substituted by an even number of similar fragments, or (3) a molecule is prosymmetric such that it has atomic sites on which a single substitution operation could lead to a molecule of class (1) or (2). Further details on symmetry detection are provided in Supplementary Note 2. As always, incorporation of any such domain-specific heuristics like symmetry is thereby a double-edged sword, possibly generating more meaningful search spaces as much as introducing a limiting bias. AML is particularly appealing in this respect. Any such rules can readily be added or dropped without incurring excessive computational costs as in exhaustive screenings of predefined search spaces.

Charge-conduction based fitness. In the spanned search space, we assess the suitability of candidate molecules for OSC applications by two descriptors known to probe two important and complementary aspects related to the conduction of charge. One concerns the efficient injection of charge from a contacting electrode into the OSC material. The other assesses the required high charge mobility inside the OSC bulk. For predominantly *p*-type OSC materials[23] a detrimentally high barrier for a corresponding hole injection from a standard gold electrode is readily probed by a level-alignment descriptor $\epsilon_{\mathrm{align}} = |\epsilon_{\mathrm{HOMO}} - \Phi_{\mathrm{Au}}|$,[42] which evaluates the energetic mismatch between the Au work function $\Phi_{\mathrm{Au}} = -5.1$ eV[43] and the energetic position of the highest occupied molecular orbital (HOMO) ϵ_{HOMO} as a common approximation of the material's ionization potential.[44,45] Adapting this descriptor to other electrode materials or to *n*-type OSC materials (then involving the energetic position of the lowest unoccupied molecular orbital, LUMO) is straightforward. As an equally established descriptor for the bulk charge mobility we employ the intra-molecular (hole) reorganization energy λ_h, which measures the cost of accommodating a new charge state after the carrier has moved to the next molecular site.[46,47] As molecular properties, both ϵ_HOMO and λ_h can be determined by efficient first-principles calculations as detailed in Supplementary Note 2, where the density-functional theory (DFT) B3LYP[48–50] level of theory constitutes a well established accuracy standard[27,31,39,40,51], matching experimental data[44,52]. We emphasize though that using the lowest-energy gas-phase conformer for the descriptor calculation disregards packing-effects in the molecular crystal[53–55] and we further discuss the influence of conformers on descriptor values in Supplementary Note 3.

To evaluate molecular fitness and prioritize candidates during AML discovery, both objectives are combined in a scalarized fitness function

$$F = -\left\| \begin{pmatrix} \lambda_\mathrm{h} \\ \epsilon_\mathrm{align} \end{pmatrix} \cdot \mathbf{w} \right\|_2 , \qquad (1)$$

which an ideal candidate molecule will maximize.[56] Here, the weight vector $\mathbf{w} = (1.0, 0.7)^\top$ accommodates the generally different absolute scales of the two descriptors, with the value of 0.7 chosen to yield an essentially Ohmic alignment with the electrode of $|\epsilon_\mathrm{align}| < 0.3$ eV if λ_h falls into the range of commonly known OSCs. We note, though, that the exact choice of weights is rather unimportant for the performance of the AML search, as it only linearly biases F towards either of the descriptors, as further detailed below. With the currently chosen weight and at the DFT-B3LYP level of theory, pentacene and rubrene – materials that have been contacted by gold electrodes before[57,58] – will feature F values of -0.16 and -0.2, respectively. A threshold $F \geq -0.2$ will therefore later on be used to measure discovery success of the AML.

AML: design and search strategy. By successively querying the explicit first-principles calculation of the descriptors for identified candidate molecules, the AML algorithm establishes an ever improving surrogate model of the fitness function F over the search space. Out of a manifold of in principle possible surrogate models, we found GPR to already achieve outstanding performance at very moderate amounts of data. In brief, the employed model uses circular Morgan fingerprints[59] to compare the structural similarity of not yet explicitly calculated molecules with the hitherto acquired ones. Specifically, counts of substructures that can be extracted by moving up to two bonds away from each central atom are generated. The similarity between two molecules is then measured with a substructure count kernel. A full account of the GPR learning through log-marginal likelihood maximization is provided in Supplementary Note 2. A central advantage of GPR for the AML context is that it not only provides a prediction for the targeted fitness function F, but also the corresponding predictive uncertainty σ from the Gaussian variance. Balancing between exploitation and exploration, the AML algorithm can thus query new candidate molecules either because they are highly promising in terms of a maximum predicted fitness F or because they exhibit a high uncertainty σ such that their explicit calculation will maximally improve the surrogate model. Practically, molecules are thereby chosen according to an upper confidence bound acquisition function

$$F_\mathrm{acq} = F + \kappa\sigma. \qquad (2)$$

This represents a simple, well-tested strategy in Bayesian optimization[60–62] or active-search[63,64] with GPRs, which contains only one hyperparameter κ to balance exploration and exploitation.

Multiple possibilities arise how to actually execute the iterative AML process. After initializing the surrogate model by training on a defined number N_initial of molecules, central questions concern the acquisition of new data before the surrogate model is retrained. Compatible with super-computing resources that encourage a parallel first-principles evaluation of the descriptors for multiple molecules, we opt for a batch-based learning where N_batch molecules with maximum F_acq are queried and the model is then retrained on the basis of the accumulated new descriptor data. Future improvements could include an additional enforcement of diversity in the prioritized batch.[18,21,65,66] In an in principle infinite chemical space, another central AML design choice regards the extent over which new molecules are practically assessed with the established, conceptually global surrogate model. Aiming for high-performance OSC molecules of tractable size and complexity, we here opt for a single tree expansion that limits the candidates to those in the vicinity of already sampled ones[67].

In a most straightforward realization and if all molecules for which first-principles descriptors have already been computed define the current population at step n of the AML search, then the N_batch molecules for the next step $n+1$ are identified in the search space formed by all molecules that can be generated by one-time application of any of the morphing operations to every molecule in the current population. While this nicely exploits the evolutionary pressure contained in the current population of size $N_\mathrm{pop} = N_\mathrm{initial} + n \times N_\mathrm{batch}$, the search space for step $n+1$ could also be systematically increased by exhaustive multiple-time application of the morphing operations. As illustrated below by comparing a corresponding search depth of one- or two-time application, this may help to overcome local funnels and navigate more efficiently through chemical space. On the other hand and regardless of the actual search depth d_search, the continuously growing population size will at later learning steps n inevitably lead to a combinatorial explosion of new candidates for any such exhaustive enumeration. Eventually, this requires to decrease the resolution in the ever increasing search space. Note that precisely this combinatorial explosion also precludes popular supervised machine learning approaches that exhaustively learn molecular properties in a closed chemical space, possibly followed by some form of data mining[3].

A decreasing resolution in the AML search space can for instance be achieved by imposing additional heuristic selection criteria, e.g., selectively suppressing certain morphing operations for increasing search depths, or other more sophisticated tree-search policies[68] also employed in reinforcement learning[35,69]. Here, we realize deeper partial expansions of the search tree up to a search depth d_search by applying the molecular morphing operations only to a fixed number of N_deep molecules selected first from the current population and then subsequently from those molecules that were created by the previous morphing operations. By each time selecting the N_deep molecules through fitness-rank based roulette-wheel selection, i.e., by assigning higher selection probabilities to molecules with high F_acq values, the search tree is thus preferentially expanded into regions of the OSC space that the surrogate model anticipates to be rewarding (either in terms of exploitation or exploration).

Hyperparameter optimization. The thus defined AML approach contains a number of hyperparameters that may critically affect its performance. Most notably, these are κ that balances exploration and exploitation in the acquisition function, N_batch the size of the prioritized batch in each learning step, as well as d_search the depth of the search space in terms of the number of applied morphing operations. The decreased resolution strategy additionally requires the specification of the fixed subset size of N_deep molecules to which morphing operations are applied. Less

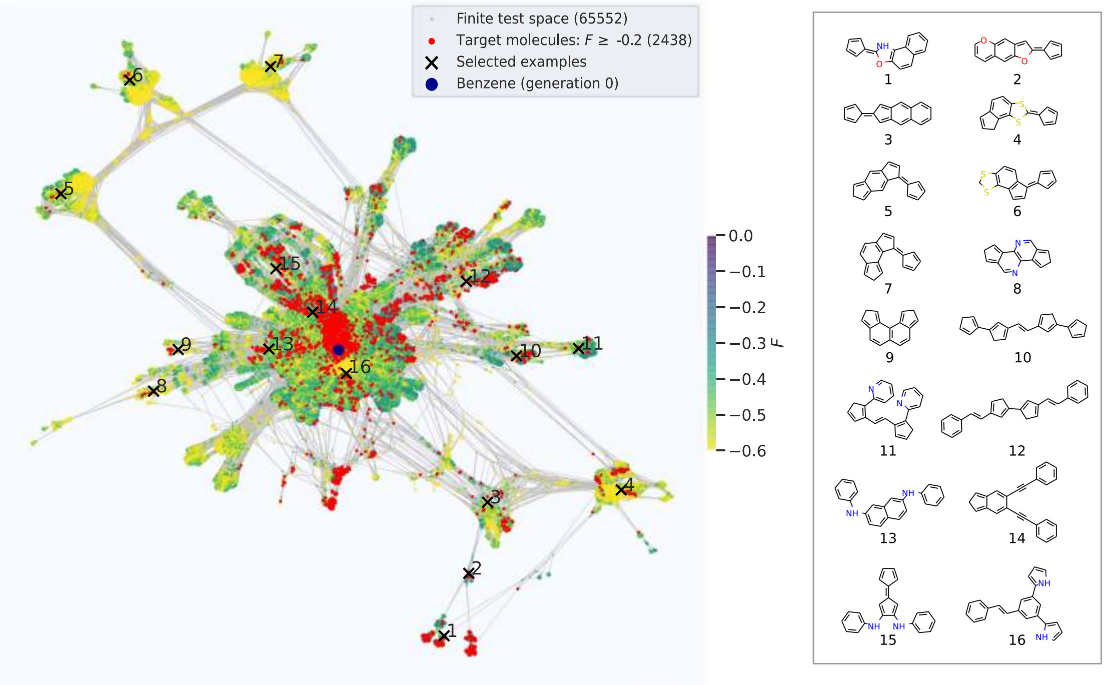

Fig. 2 Finite OSC test space. Left panel: Chemical space network (CSN) representation of the finite OSC test space of 65.552 unique molecules generated by exhaustive application of all morphing operations up to 14 times. Each molecule is surrounded by morphing-related analogs (see text). Benzene as the smallest base molecule is colored in blue. All other molecular nodes are colored according to their fitness function F as calculated at the semi-empirical density-functional tight-binding level. 2438 red nodes form the target discovery group of top-performing molecules with high fitness $F \geq -0.2$. Right panel: Example molecules from the top-performing group, chosen randomly from different areas of the CSN to illustrate the structural diversity contained in the test space.

decisive is the initial number of molecules N_{initial} used for the first training of the surrogate model, which defines only an insignifiant part of the total executed first-principles calculations and which should only be large enough to somehow kick-start the AML process. Here, we suitably set N_{initial} to the 179 unique molecules that result in the first two generations when applying all morphing operations up to two times starting from the simplest building block benzene, cf. Fig. 1.

In order to explore the effect of the other hyperparameters and optimize them for first-principles OSC discovery, we consider the finite subspace formed of all molecules up to a maximum size of 4 rings, 4 heteroatoms and 2 linkers that are generated by exhaustive application of all morphing operations up to 14 times, see Supplementary Note 2. With 65.552 unique molecules this subspace is already representative for the design problem and contains many and diversely structured high-performing molecules as illustrated in Fig. 2. At the same time, the still tractable size of the finite test space allows for the exhaustive calculation of all molecular descriptors with van der Waals (vdW) corrected density functional tight-binding (DFTB).[70] While this semi-empirical level of theory is not fully quantitative, it provides a sufficiently realistic account of the descriptor landscape for the intended method testing as analyzed in detail in Supplementary Fig. 4. Further details on molecular test space generation and descriptor calculation are provided in the Supplementary Note 2.

The finite test space contains a total of 2438 top-performing molecules with a high fitness $F \geq -0.2$. As a quantitative benchmark, we thus measure the discovery success $S(N)$ as the fraction of these molecules that are identified after the descriptors of N molecules have been queried. With 179 queries used for the initialization, see above, the final measure $S(5179)$ thus evaluates the discovery success after $n = 50$ learning steps when using $N_{\text{batch}} = 100$. Supplementary Fig. 6 compiles the corresponding success curves $S(N)$, when systematically combining $N_{\text{batch}} = 50$, 100, or 200 with κ values in half-integer steps between 0 and 5, as well as for a search depth of one- or two-time exhaustive application of all morphing operations. Fortunately, we find the AML search to be highly robust with respect to the choice of N_{batch} and κ. Only a small variation of $0.71 < S(N = 5179) < 0.80$ is obtained over all tested combinations for a search depth of one, meaning that 70–80% of the top-performing molecules are consistently found after descriptors for less than 8% of the entire test space have actually been computed. For a search depth of two, this success rate becomes slightly higher, reaching up to 85% as compiled in Supplementary Fig. 7. Generally, larger batch sizes seem to implicitly increase the explorative behavior, such that an almost indistinguishably optimum performance is obtained for larger N_{batch} in combination with successively smaller exploration weights κ in the acquisition function, cf. Eq. (2). For too small κ, the success curves become stepped though, indicating that temporarily the mainly exploitative algorithm then only meanders through identified sub-pockets of the test space. Too large κ, on the other hand, diminish the initial success of a then too explorative algorithm in the first learning steps. Overall, an intermediate value pair $(N_{\text{batch}}, \kappa) = (100, 2.5)$ thus provides a robust setting and is henceforth employed in all AML runs. For these values of $(N_{\text{batch}}, \kappa)$, we also performed a sensitivity analysis with regard to the employed weight vector $\mathbf{w}$ in Eq. (1) and the bond radius in the Morgan fingerprints used to assess molecular similarity. The results are summarized in Supplementary Figs. 8 and 9, respectively, and again demonstrate a high robustness with respect to these parameters.

The higher success rate for $d_{\text{search}} = 2$ indicates that it is generally advantageous to further expand the search space away from the known topologies of the current population. Assessing the dependence of the decreased resolution AML algorithm on its two additional hyperparameters, Supplementary Table 1 summarizes the corresponding discovery successes when systematically combining a varying subset size $N_{\mathrm{deep}} = 100, 250, 500$ and 1000 with search depths $d_{\mathrm{search}} = 1, 2, 3, 4, 5$ and 10. Again, we find the algorithm to be quite robust, with higher d_{search} compensating smaller N_{deep}. Within the finite test space, many combinations thus saturate at success rates around 82–83%. This is essentially as good as the best performance of the previous exhaustive enumerations, but comes at the advantage of a controlled growth of the search space at later learning steps. For the first-principles AML discovery in the virtually unlimited OSC space below we correspondingly employ this decreased resolution search strategy with a top-performing hyperparameter combination (d_{search}, N_{deep}) = (3, 500).

Visualizing AML at work. The finite test space can also be viewed as a chemical space network (CSN), in which the morphing operations establish a total of 315.451 directed connections between the constituting molecules. This allows us to visualize the space in form of a 2D graph structure, in which the molecules are mutually repelling nodes, while morphing relationships between them lead to attractive edges[71], see Supplementary Note 1 for details. In such a representation each molecule is thus spatially surrounded by morphing-related analogs. Figure 2 shows the resulting graph, in which the individual nodes are colored according to their DFTB calculated fitness. As expected, the target group for discovery in form of the 2438 top-performing molecules is widely scattered over disjoint parts of chemical space, with ensembles of related molecules often clustered in sub-pockets.

Apart from providing a bird's eye view of the design problem, the CSN representation also affords a direct visual access to the AML process. Plotting the evolving population N over subsequent learning steps n reveals how much a chosen AML strategy is able to focus its exploration onto the interesting regions of chemical space and how efficiently it prioritizes OSC molecules with desired properties. Figure 3 illustrates this for the determined optimum hyperparameters and contrasts the learning for exhaustive searches with depths of one or two, with the decreased resolution strategy where the searches partially expand subsets of

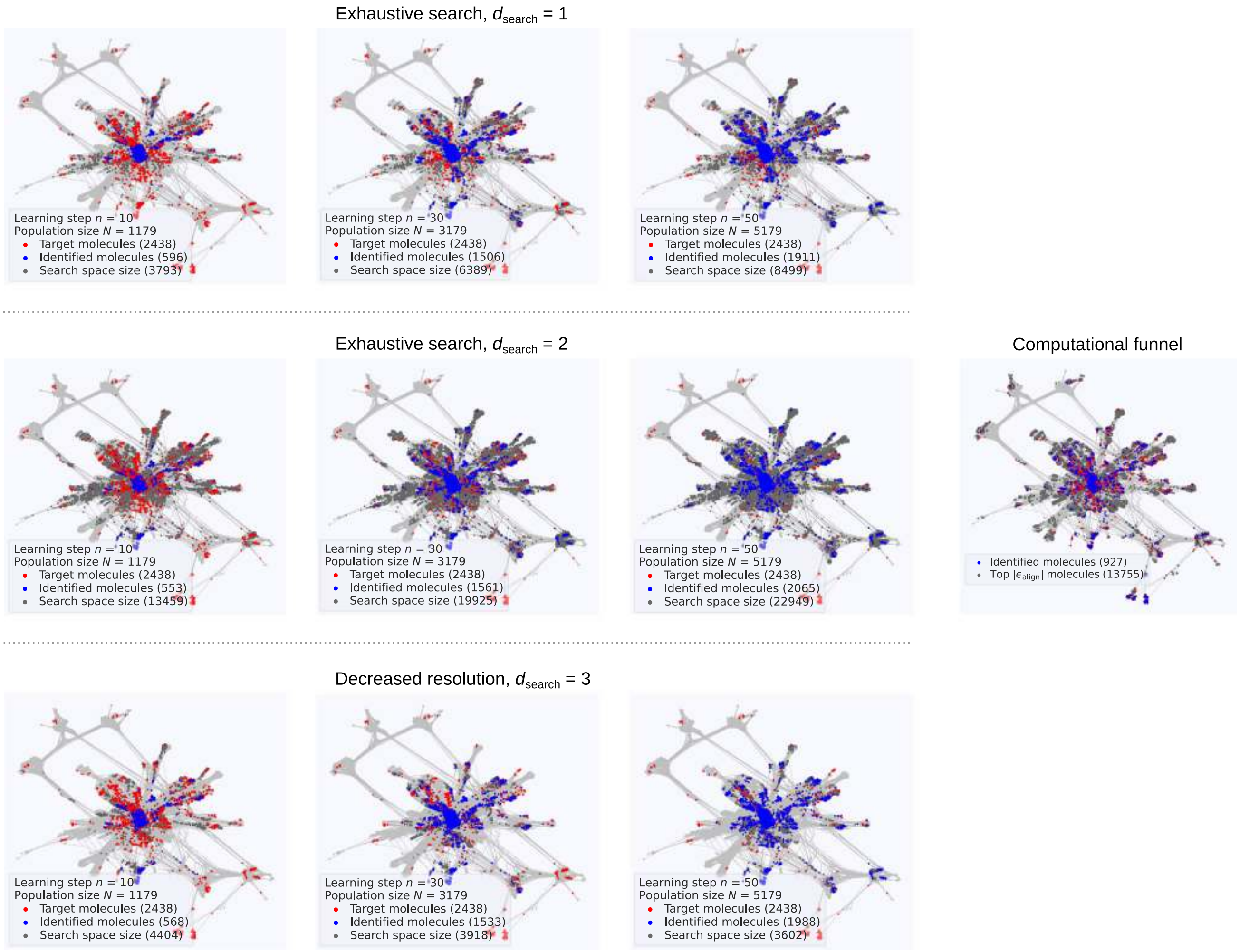

Fig. 3 AML exploration of the finite test space. The same CSN representation of the OSC test space as in Fig. 2 is shown in gray. Superimposed are the target group of 2438 top-performing molecules in red. Each panel shows the discovery success after n learning steps with the color of all identified top-performing molecules changed to blue and the search space for the next learning step $n+1$ colored in dark gray. Left upper panels: Steps $n = 10, 30, 50$ for an exhaustive search with search depth of one. Left middle panels: Steps $n = 10, 30, 50$ for an exhaustive search with search depth of two. Left lower panels: Steps $n = 10, 30, 50$ for a decreased resolution search ($N_{\mathrm{deep}} = 500$) with search depth of three (see text). Supplementary Movies 1–3 provide the detailed, full trajectory of all three AML discovery runs over learning steps 1–50. Right centered panel: Discovery success of a conventional computational funnel after computing an equal number of descriptors (5179) as after 50 learning steps, and anticipating that knowledge of 13.755 molecules with optimum $|\epsilon_{\mathrm{align}}| < 0.3$ eV is present (see text).

$N_{\mathrm{deep}} = 500$ molecules at search depth three. For the exhaustive search with $d_{\mathrm{search}} = 1$, the discovery is centered to more morphing-related top-performing molecules all more or less located in the core region of the CSN. In contrast, for the deeper exhaustive search, the algorithm also successfully identifies top-performing molecules in the periphery of the network that are topologically quite disconnected from the initial population. The downside is a rapidly increasing size of the search space that in the present case is only bounded by the finiteness of the considered test space. This is largely mitigated by the decreased resolution search, which nevertheless equally successfully identifies top-performing molecules at the CSN periphery.

To put this performance of the AML searches into perspective, we also contrast them in Fig. 3 with the result of a conventional computational funnel. For the latter we pretend that the calculation of ϵ_{HOMO} has a negligible computational cost and the value of this descriptor is known for every molecule in the test space. This allows to identify a subset of 13.755 promising molecules for which $|\epsilon_{\mathrm{align}}| < 0.3\,\mathrm{eV}$ and which contains all previously considered 2438 top-performing molecules. The computational funnel approach would then focus the explicit calculation of the more demanding λ_{h} descriptor to molecules in this subset. To enable a direct comparison with the preceding AML assessment, a random selection of 5179 molecules out of this subset would then lead to a success rate of $S(5179) \approx 0.4$. Even in this finite test space, where the AML algorithm can not even unfold its real strength, less than half of the top-performing molecules are thus found by this prevalent computational screening strategy after spending the same amount of CPU time (assuming that the exhaustive calculation of 65.552 ϵ_{HOMO} descriptors for the entire test space would constitute an insignificant computational effort).

First-principles AML discovery in a virtually unlimited OSC chemical space. Based on the gathered methodological understanding and optimized algorithmic settings ($N_{\mathrm{batch}} = 100, \kappa = 2.5, d_{\mathrm{search}} = 3, N_{\mathrm{deep}} = 500$) we now proceed to first-principles AML discovery at the vdW-corrected DFT-B3LYP level of theory. This is a truly challenging endeavor, considering the vastness of the OSC design space. While the space of molecules that can be generated through the morphing operations is in principle unbounded, we here restrict it to the realm of “small molecules” containing a maximum of 100 atoms (including H atoms). This realm appears as a first, more practical target for synthesis and crystallization, also considering that essentially all known top-performing OSC molecules to date fall into this size range. Estimated to surpass a size of 10^{30} molecules, see Supplementary Note 2, the corresponding chemical space is nevertheless virtually unlimited for all practical purposes and would defy any conventional exhaustive computational screening. While an iterative search as with AML is thus the only tractable means to explore this space at predictive quality, an additional technical aspect emerges that did not yet play a role in the analysis of the finite test space at the semi-empirical level before. It concerns the typically massively parallel processing on the required high-performance computing (HPC) infrastructure. As a result of queuing or downtimes, as well as convergence behavior of the first-principles calculations, the results for the N_{batch} descriptor calculations can become available at quite different times (or in rare cases of failed convergence or system instabilities may not become available at all). A practical way to avoid long waiting times before the last calculations are ready is to initially select a larger batch size for descriptor calculation and then continue with the forthcoming learning steps whenever the desired number of N_{batch} molecules has been processed (successfully or unsuccessfully). We found this strategy to afford an efficient and continuous HPC workflow, here initially submitting the 200 molecules with highest F_{acq} values for descriptor calculations. These are continuously processed on the HPC system by 40–100 parallel worker processes, to reach the targeted batch size $N_{\mathrm{batch}} = 100$, while for a retraining of the surrogate model only successfully processed cases are included. In this respect, the above determined robustness of the AML performance with regard to the exact batch size also constitutes an important asset for such HPC operation.

Figure 4 summarizes the results of the AML discovery run over its first 15 learning steps. Gratifyingly, the algorithm quickly stabilizes into a highly efficient mode of operation while simultaneously meandering deep into unknown chemical space. Already after five learning steps even the median fitness of the entire prioritized batch exceeds the threshold value $F \geq -0.2$ for the first time, reflecting top-performing molecules. However, as clearly seen from the violin plots of the F distribution over the batches in Fig. 4b, this high efficiency does not simply result from the algorithm just exploiting its established knowledge. Even at later learning steps, the algorithm steadily queries quite unfavorable molecules with a fitness worse than $F < -0.3$. While such exploratory queries can either be based on high model uncertainty or induced by model prediction errors, they serve to continuously improve the surrogate model also outside the already considered search space. As a result, at each later learning step, the algorithm keeps on identifying top-performing molecules at a stable, high rate.

After 15 learning steps and a corresponding calculation of first-principles descriptors for 1680 molecules (and only 35 unsuccessfully terminated calculations), a total of 900 molecules with molecular fitness $F \geq -0.2$ have been found. A relative success rate of 54%, i.e., essentially every second first-principles calculation yields a promising molecule and this without any a priori knowledge of the vast OSC space. A second AML discovery run described in Supplementary Note 4 confirms the robustness of this high performance. Notably, due to the random nature in our search strategy, significantly different, but equally favorable molecules are identified in this run. This performance becomes even more impressive from the viewpoint that these molecules are true discoveries, as essentially none of them are contained in existing focused libraries assembled in previous screening studies[3,31–34]. With typically ~$10^5 - 10^6$ entries, these data sets reflect the wealth of our existing knowledge and synthesis efforts, but simply do not even scratch the surface of the true OSC design possibilities. To this end, the negligible overlap with the top-performing molecules identified in these previous studies also has to do with molecular size. Within the first learning steps, the average size in the prioritized batch quickly rises to around 90 atoms, which is at the edge of the limit currently imposed on our search and in a size regime that could barely be addressed by the previous exhaustive enumeration studies. At the same time, even archetypical and acclaimed molecular OSC materials like DNTT ($C_{22}H_{12}S$) or rubrene ($C_{42}H_{28}$) approach this size regime, with many other experimentally tested candidates falling right into it[23]. The preferred prioritization of such larger molecules is thereby to some extent likely simply a result of the combinatorially exploding phase space. On the other hand, another physical factor could be that the AML algorithm learns and exploits the tendency of λ_{h} to decrease with increasing molecular size[3] as a consequence of a larger hole delocalization (which even at the hybrid DFT-B3LYP level of theory may be slightly overestimated[72]). The inclusion of molecular coupling-sensitive descriptors into the fitness function is therefore certainly a promising topic for future studies.

The discovered molecules exhibit a diverse set of structures, incorporating distinct core fragments and the full set of allowed

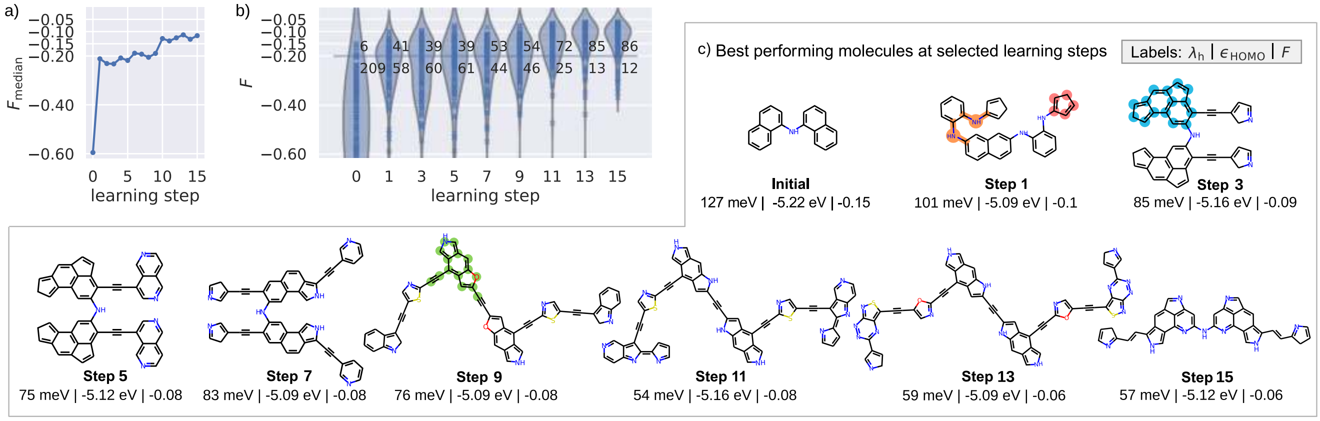

Fig. 4 First-principles AML discovery in a virtually unlimited space. a Median values of molecular fitness F over the prioritized N_{batch} molecules at the different learning steps (step 0 shows the median of the initial population $N_{initial}$). **b** Corresponding violin plot showing the (kernel-density estimated) distribution of molecular fitness F over the batch. These smooth kernel-density estimated distributions can slightly extend beyond the true range of F values as indicated by the explicit values marked by blue crosses. The number of queries leading to favorable and unfavorable molecules is indicated next to each violin. Due to descriptor calculation failures (see text) these numbers do not always add up to $N_{batch} = 100$. **c** Examples of top-performing molecules identified at various learning steps (see text for an explanation of the different color-highlighted geometric motifs). An extended list of the 4 top-performing molecules of each learning step is shown in Supplementary Fig. 10.

heteroatoms and linkers. Figure 4c illustrates this with the best-performing molecules identified at selected learning steps, and an extended list being compiled in Supplementary Fig. 11. This diversity indicates that the AML algorithm successfully explored topologically widely differing areas of the OSC space and did not get stuck in one or a few subpockets. Nevertheless, some commonalities can be spotted, like the recurrent presence of phenylamine linker motifs (marked in orange in the best-performing molecule of learning step 1 in Fig. 4c). Similarly, more complex ring systems emerged at later learning stages (marked in blue and green in the most favorable molecule of step 3 and 9, respectively) and are from thereon quite pronounced among well-performing molecules. While a diverse molecular space is searched, the AML discovery thus automatically identifies and prioritizes privileged design motifs. After harvesting a larger number of molecules in further learning steps, an exciting prospect for future studies is therefore to mine the accumulating data set and systematically extract this implicit knowledge for rational design. To this end, the trained surrogate model can also be used to quickly assess the suitability of such manually constructed molecules or of deliberate modifications of the here identified ones. The latter could be particularly appealing in view of long-term device-stability or synthetic accessibility. We note that certainly not all identified molecules are suitable in this regard. For instance, the 5-membered unsaturated rings of the displayed compound of learning step 1 (marked in red) in Fig. 4c could be problematic as they might undergo Diels-Alder type reactions, and we attribute the appearance of such ring motives as the algorithm's intent to provide intermediates on the way to the later explored, more stable 5-membered heterocycles. Nonetheless, multiple of the favorable molecules are symmetric and composed of standard building blocks that should be easily accessible through short and reliable synthesis routes, with the surrogate model furthermore available to gauge the effect of stabilizing modifications.

Discussion

In our view, active machine learning based on first-principles descriptors constitutes a most promising route to prior-to-synthesis virtual discovery. Its iterative refinement allows to most efficiently focus the data-generating calculations and meaningfully explore the vastness of chemical spaces at predictive quality and without a priori specifications, enumeration or reliance on empirical descriptors with limited validity range. In this work we have established such an AML discovery approach for molecular OSC materials through versatile molecular morphing operations and based on charge injection and conduction querying descriptors. Fortunately and with a view on explainable ML models, our systematic assessment within a finite test space suggests the approach to be quite robust with respect to the algorithmic hyperparameters. Most promising to further increase its already high efficiency and prevent an over-exploitation of particular structural motifs, is likely to additionally enforce structural diversity among the N_{batch} molecules selected at each learning step, instead of the present purely fitness-ranked roulette-wheel selection.

Central to assess this performance and enable an unbiased and systematic comparability of different AML approaches will be the establishment of well-designed, balanced and freely available benchmark platforms for unlimited search spaces. As clear from the present work, already within the here pursued single-tree expansion there are multiple design strategies and concomitant algorithmic parameters. While we have explored these in a truncated test space, AML only unfolds its full potential in the exploration of unlimited spaces. Representative and standardized benchmark platforms as already available for drug-design tasks[13] will therefore be pivotal to truly compare various learning concepts that work without a priori enumeration or pre-definition of the search problem.

Further challenges and advancements in the physico-chemical domain comprise the adaption and extension of the molecular morphing operations to tailor the OSC search space. The present set derived from literature domain knowledge spans a design space geared towards flexible, π-conjugated molecules. Ultimately, a generic, but chemically-valid creation of morphing operations could drive discovery of many novel structural motives. Heavier requirements on the surrogate GPR-model in such cases could then be tackled with improved covariance functions for 2D molecular graphs[73] or conformer-specific 3D coordinates[74], while alleviating the limited scaling by sparse approximations[75], or application of alternative models[76–79].

Another major area for development concerns the first-principles descriptors entering the employed multi-objective fitness function. Devising such suitable descriptors has evolved into an important research area of its own[80–83], independent of the present AML and OSC context. With the presently employed level-alignment descriptor ϵ_{align} and the hole reorganization energy λ_{h} our search readily identified a diverse range of hitherto unknown molecular candidates. Just as in conventional computational screening, there are numerous possibilities to refine the underlying candidate evaluation through additional (or alternative) descriptors. In the exemplified OSC context, obvious avenues could be to explicitly consider synthetic accessibility[84], electronic coupling and charge-transport networks in the molecular solid[46,51,85,86] or electron-phonon coupling[87]. In view of the high data efficiency of the AML approach, one may also drop the present focus on computationally least-demanding descriptors, originally dictated by the excessive queries in conventional exhaustive screening work. More elaborate descriptors like structural interfacing with electrode materials[88] could therefore routinely (or at least occasionally) be requested. Eventually, one could even think of incorporating experimental feedback from self-driving laboratories[89]. The prospects are thus as manifold as exciting. Regardless of the specific road chosen, it is conceptually clear that autonomously operating workflows like the present AML approach offer an unparalleled means to accelerate the discovery and design of viable future materials like the high-mobility organic semiconductors featured in this work.

Data availability

The source data necessary to reproduce the main figures of the manuscript is provided in the supplementary materials of this article. Source data are provided with this paper.

Code availability

The code used to run AML discovery is available at https://doi.org/10.5281/zenodo.4554331

Received: 29 September 2020; Accepted: 15 March 2021;
Published online: 23 April 2021

References

1. Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. Estimation of the size of drug-like chemical space based on gdb-17 data. *J. Comput. Aided Mol. Des.* **27**, 675–679 (2013).
2. Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **303**, 1813–1818 (2004).
3. Schober, C., Reuter, K. & Oberhofer, H. Virtual screening for high carrier mobility in organic semiconductors. *J. Phys. Chem. Lett.* **7**, 3973–3977 (2016).
4. Pulido, A. et al. Functional materials discovery using energy–structure–function maps. *Nature* **543**, 657–664 (2017).
5. Pyzer-Knapp, E. O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J. & Aspuru-Guzik, A. What is high-throughput virtual screening? a perspective from organic materials discovery. *Annu. Rev. Mater. Res.* **45**, 195–216 (2015).
6. Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D. & Varnek, A. Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. *J. Chem. Inf. Model.* **55**, 84–94 (2015).
7. Reymond, J.-L., van Deursen, R., Blum, L. C. & Ruddigkeit, L. Chemical space as a source for new drugs. *Med. Chem. Commun.* **1**, 30–38 (2010).
8. Devi, R. V., Sathya, S. S. & Coumar, M. S. Evolutionary algorithms for de novo drug design – a survey. *Appl. Soft Comput.* **27**, 543–552 (2015).
9. Le, T. C. & Winkler, D. A. Discovery and optimization of materials using evolutionary approaches. *Chem. Rev.* **116**, 6107–6132 (2016).
10. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
11. 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).
12. Reker, D. & Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discov. Today* **20**, 458–465 (2015).
13. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. Guacamol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).
14. Schneider, G. *De novo Molecular Design* (Wiley, 2013) https://books.google.de/books?id=ZxlrmwEACAAJ.
15. Jensen, K. F., Coley, C. W. & Eyke, N. S. Autonomous discovery in the chemical sciences part i: progress. *Angew. Chem. Int. Ed.* **59**, 22858–22893 (2019).
16. Settles, B. *Active learning literature survey* (University of Wisconsin, Madison, 2010).
17. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *Npj Comput. Mater.* **5**, 21 (2019).
18. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
19. Häse, F., Roch, L. M., Kreisbeck, C. & Aspuru-Guzik, A. Phoenics: a bayesian optimizer for chemistry. *ACS Cent. Sci.* **4**, 1134–1145 (2018).
20. Vandermause, J. et al. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *Npj Comput. Mater.* **6**, 20 (2020).
21. Janet, J. P., Ramesh, S., Duan, C. & Kulik, H. J. Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization. *ACS Cent. Sci.* **6**, 513–524 (2020).
22. Bisbo, M. K. & Hammer, B. Efficient global structure optimization with a machine-learned surrogate model. *Phys. Rev. Lett.* **124**, 086102 (2020).
23. Wang, C., Dong, H., Hu, W., Liu, Y. & Zhu, D. Semiconducting π-conjugated systems in field-effect transistors: a material odyssey of organic electronics. *Chem. Rev.* **112**, 2208–2267 (2012).
24. Lin, Y., Li, Y. & Zhan, X. Small molecule semiconductors for high-efficiency organic photovoltaics. *Chem. Soc. Rev.* **41**, 4245–4272 (2012).
25. Xu, R.-P., Li, Y.-Q. & Tang, J.-X. Recent advances in flexible organic light-emitting diodes. *J. Mater. Chem. C* **4**, 9116–9142 (2016).
26. Geng, H. et al. Theoretical study of substitution effects on molecular reorganization energy in organic semiconductors. *J. Chem. Phys.* **135**, 104703 (2011).
27. Uejima, M., Sato, T., Tanaka, K. & Kaji, H. Vibronic coupling density analysis for the chain-length dependence of reorganization energies in oligofluorenes: a comparative study with oligothiophenes. *Phys. Chem. Chem. Phys.* **15**, 14006–14016 (2013).
28. Wilbraham, L., Smajli, D., Heath-Apostolopoulos, I. & Zwijnenburg, M. A. Mapping the optoelectronic property space of small aromatic molecules. *Commun. Chem.* **3**, 14 (2020).
29. 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).
30. Saeki, A. & Kranthiraja, K. A high throughput molecular screening for organic electronics via machine learning: present status and perspective. *Jpn. J. Appl. Phys.* **59**, SD0801 (2019).
31. Matsuzawa, N. N. et al. Massive theoretical screen of hole conducting organic materials in the heteroacene family by using a cloud-computing environment. *J. Phys. Chem. A* **124**, 1981–1992 (2020).
32. Nematiaram, T., Padula, D., Landi, A. and Troisi, A. On the largest possible mobility of molecular semiconductors and how to achieve it. *Adv. Funct. Mater.* **30**, 2001906 (2020).
33. Atahan-Evrenk, S. & Atalay, F. B. Prediction of intramolecular reorganization energy using machine learning. *J. Phys. Chem. A* **123**, 7855–7863 (2019).
34. Cheng, C. Y., Campbell, J. E. & Day, G. M. Evolutionary chemical space exploration for functional materials: computational organic semiconductor discovery. *Chem. Sci.* **11**, 4922–4933 (2020).
35. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
36. Segler, M. H. S. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. Eur. J.* **23**, 5966–5971 (2017).
37. Yao, Z.-F., Wang, J.-Y. & Pei, J. Control of π–π stacking via crystal engineering in organic conjugated small molecule crystals. *Cryst. Growth Des.* **18**, 7–15 (2018).
38. Kunkel, C., Schober, C., Margraf, J. T., Reuter, K. & Oberhofer, H. Finding the right bricks for molecular legos: a data mining approach to organic semiconductor design. *Chem. Mater.* **31**, 969–978 (2019a).
39. Stehr, V., Pfister, J., Fink, R. F., Engels, B. & Deibel, C. First-principles calculations of anisotropic charge-carrier mobilities in organic semiconductor crystals. *Phys. Rev. B* **83**, 155208 (2011).
40. Li, P., Cui, Y., Song, C. & Zhang, H. Electronic and charge transport properties of dimers of dithienothiophenes: effect of structural symmetry and linking mode. *RSC Adv.* **5**, 50212–50222 (2015).
41. Ren, L. et al. Critical role of molecular symmetry for charge transport properties: a paradigm learned from quinoidal bithieno[3,4-b]thiophenes. *Chem. Mater.* **29**, 4999–5008 (2017).

42. Ishii, H., Sugiyama, K., Ito, E. & Seki, K. Energy level alignment and interfacial electronic structures at organic/metal and organic/organic interfaces. *Adv. Mater.* **11**, 605–625 (1999).
43. Michaelson, H. B. The work function of the elements and its periodicity. *J. Appl. Phys.* **48**, 4729–4733 (1977).
44. Schwenn, P., Burn, P. & Powell, B. Calculation of solid state molecular ionisation energies and electron affinities for organic semiconductors. *Org Electron.* **12**, 394 – 403 (2011).
45. Bhandari, S., Cheung, M. S., Geva, E., Kronik, L. & Dunietz, B. D. Fundamental gaps of condensed-phase organic semiconductors from single-molecule calculations using polarization-consistent optimally tuned screened range-separated hybrid functionals. *J. Chem. Theory Comput.* **14**, 6287–6294 (2018).
46. Oberhofer, H., Reuter, K. & Blumberger, J. Charge transport in molecular materials: an assessment of computational methods. *Chem. Rev.* **117**, 10319–10357 (2017).
47. Nelsen, S. F., Blackstock, S. C. & Kim, Y. Estimation of inner shell marcus terms for amino nitrogen compounds by molecular orbital calculations. *J. Am. Chem. Soc.* **109**, 677–682 (1987).
48. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098–3100 (1988).
49. Lee, C., Yang, W. & Parr, R. G. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).
50. Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **98**, 11623–11627 (1994).
51. Moral, M., Garzón-Ruiz, A., Castro, M., Canales-Vázquez, J. & Sancho-García, J. C. Virtual design in organic electronics: screening of a large set of 1,4-bis (phenylethynyl)benzene derivatives as molecular semiconductors. *J. Phys. Chem. C* **121**, 28249–28261 (2017).
52. Kera, S. et al. Experimental reorganization energies of pentacene and perfluoropentacene: effects of perfluorination. *J. Phys. Chem. C* **117**, 22428–22437 (2013).
53. Stuke, A. et al. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data* **7**, 58 (2020).
54. Mitzel, N. W. & Rankin, D. W. H. Saracen – molecular structures from theory and experiment: the best of both worlds. *Dalton Trans.* 3650–3662 (2003).
55. Blomeyer, S. et al. Intramolecular π–π interactions in flexibly linked partially fluorinated bisarenes in the gas phase. *Angew. Chem. Int. Ed.* **56**, 13259–13263 (2017).
56. Besnard, J. et al. Automated design of ligands to polypharmacological profiles. *Nature* **492**, 215–220 (2012).
57. Takeya, J. et al. Very high-mobility organic single-crystal transistors with in-crystal conduction channels. *Appl. Phys. Lett.* **90**, 102120 (2007).
58. Jurchescu, O. D., Baas, J. & Palstra, T. T. M. Effect of impurities on the mobility of single crystal pentacene. *Appl. Phys. Lett.* **84**, 3061–3063 (2004).
59. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
60. Srinivas, N., Krause, A., Kakade, S. M. & Seeger, M. W. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Trans. Inf. Theory* **58**, 3250–3265 (2012).
61. Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* **3**, 397–422 (2002).
62. Srinivas, N., Krause, A., Kakade, S. & Seeger, M. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10,* 1015–1022 (Omnipress, Madison, WI, USA, 2010).
63. Vanchinathan, H. P., Marfurt, A., Robelin, C.-A., Kossmann, D. & Krause, A. Discovering valuable items from massive data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15*, 1195–1204 (Association for Computing Machinery, New York, NY, USA, 2015) https://doi.org/10.1145/2783258.2783360.
64. Ma, Y., Huang, T.-K. & Schneider, J. Active search and bandits on graphs using sigma-optimality. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence, UAI 2015*, 542–551 (2015).
65. Pinsler, R., Gordon, J., Nalisnick, E. & Hernández-Lobato, J. M. Bayesian batch active learning as sparse subset approximation. In *Advances in Neural Information Processing Systems 32*, (eds Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. & Garnett, R.) 6359–6370 (Curran Associates, Inc., 2019).
66. Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J. & Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations* (2020).
67. Madhawa, K. & Murata, T. A multi-armed bandit approach for exploring partially observed networks. *Appl. Netw. Sci.* **4**, 26 (2019).
68. Browne, C. et al. A survey of monte carlo tree search methods. *IEEE Trans. Comput. Intell. AI Games* **4**, 1–43 (2012).
69. Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **9**, 10752 (2019).
70. Grimme, S., Bannwarth, C. & Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (z = 1–86). *J. Chem. Theory Comput.* **13**, 1989–2009 (2017).
71. Kunkel, C., Schober, C., Oberhofer, H. & Reuter, K. Knowledge discovery through chemical space networks: the case of organic electronics. *J. Mol. Model.* **25**, 87 (2019b).
72. Brückner, C. & Engels, B. A theoretical description of charge reorganization energies in molecular organic p-type semiconductors. *J. Comput. Chem.* **37**, 1335–1344 (2016).
73. Ralaivola, L., Swamidass, S. J., Saigo, H. & Baldi, P. Graph kernels for chemical informatics. *Neural Netw.* **18**, 1093 – 1110 (2005).
74. Himanen, L. et al. Dscribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
75. Quinonero-Candela, J. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.* **6**, 1939–1959 (2005).
76. Beluch, W. H., Genewein, T., Nürnberger, A. & Köhler, J. M. The power of ensembles for active learning in image classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018) pp. 9368–9377.
77. Kendall, A. & Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems* (2017).
78. Zhang, Y. & Lee, A. A. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.* **10**, 8154–8163 (2019).
79. Wenzel, F. et al. How good is the bayes posterior in deep neural networks really? In *International Conference on Machine Learning* (2020).
80. Curtarolo, S., Hart, G. L. W., Nardelli, M. B., Mingo, N., Sanvito, S. & Levy, O. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
81. Pracht, P., Bauer, C. A. & Grimme, S. Automated and efficient quantum chemical determination and energetic ranking of molecular protonation sites. *J. Comput. Chem.* **38**, 2618–2631 (2017).
82. Andersen, M., Levchenko, S. V., Scheffler, M. & Reuter, K. Beyond scaling relations for the description of catalytic materials. *ACS Catal.* **9**, 2752–2759 (2019).
83. Cubuk, E. D., Sendek, A. D. & Reed, E. J. Screening billions of candidates for solid lithium-ion conductors: a transfer learning approach for small data. *J. Chem. Phys.* **150**, 214701 (2019).
84. 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).
85. Ishii, H. et al. Charge mobility calculation of organic semiconductors without use of experimental single-crystal data. *Sci. Rep.* **10**, 2524 (2020).
86. Friederich, P. et al. Molecular origin of the charge carrier mobility in small molecule organic semiconductors. *Adv. Funct. Mater.* **26**, 5757–5763 (2016).
87. Landi, A. & Troisi, A. Rapid evaluation of dynamic electronic disorder in molecular semiconductors. *J. Phys. Chem. C* **122**, 18336–18345 (2018).
88. Egger, A. T. et al. Charge transfer into organic thin films: a deeper insight through machine-learning-assisted structure search. *Adv. Sci.* **7**, 2000992 (2020).
89. MacLeod, B. P. et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020).

Acknowledgements

C.K. and J.T.M are grateful for support by Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81. C.K., H.O. and K.R. gratefully acknowledge support from the Solar Technologies Go Hybrid initiative of the State of Bavaria. K.C. acknowledges funding from the China Scholarship Council. We also thankfully acknowledge computational resources provided by the Leibniz Supercomputing Centre. J.T.M. would like to acknowledge illuminating discussions on synthesizability and chemical stability with Tobias Schaub.

Author contributions

C.K, H.O., J.T.M and K.R. conceived the idea. C.K. implemented the algorithms in code and carried out the calculations. Methodological details were thereby worked out by C.K., K.C. and J.T.M. C.K, H.O., J.T.M and K.R. wrote the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-22611-4.

Correspondence and requests for materials should be addressed to K.R.

Peer review information *Nature Communications* thanks Graeme Day and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.