RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2022, **12**, 30962

Combining machine learning and quantum chemical calculations for high-throughput virtual screening of thermally activated delayed fluorescence molecular materials: the impact of selection strategy and structural mutations†

Chunyun Tu,[a] Weijiang Huang,[a] Sheng Liang,[b] Kui Wang,[a] Qin Tian[a] and Wei Yan *[a]

In view of the theoretical importance and huge application potential of Thermally Activated Delayed Fluorescence (TADF) materials, it is of great significance to conduct High-Throughput Virtual Screening (HTVS) on compound libraries to find TADF candidate molecules. This research focuses on the computational design of pure organic TADF molecules. By combining machine learning and quantum chemical calculations, using cheminformatics tools, and introducing the concept of selection and mutation from evolutionary theory, we have designed a computational program for HTVS of TADF molecular materials, especially the impact of selection strategy and structural mutations on the results of HTVS was explored. An initial compound library (size = 10^3) constructed by enumeration of typical donors and acceptors was used to evolve by successively applying selection and 10 different structural mutations. And a group fingerprint similarity (Δ_{MSPR}) index was proposed to account for the similarity between two compound libraries with comparable sizes. Based on the computed data, we have found that the mix of selection and mutations into the evolution map does have great impact on the HTVS results: (a) except the fast mutation **Sub2**, all the rest of the mutations can effectively concentrate 'good' molecules in a compound library, and hence give large material abundance (typically >0.8) for high mutation generations ($n_g \geq 6$). (b) The mean energy gap can exhibit a fast convergent trend toward very low values, hence the studied mutations (except **Sub2**) can cooperate very well with the studied DA substrates to generate optimal molecules, and the group fingerprint similarity can retain high enough values for large n_g, which can be associated with the apparent convergence in molecular skeletons as n_g increases. (c) The distribution of skeleton frequencies for a specific mutation is generally uneven with one dominant skeleton. The overall numbers of common and generic cores for all mutations are 11 and 7 as $n_g = 9$. Hence, in a sense, the 'optimal' skeletons seem unique and useful in realizing low energy gaps. With these observations and the development of related HTVS software, we expect to provide insight and tools to the research community of HTVS of molecular (TADF) materials.

Received 7th September 2022
Accepted 9th October 2022

DOI: 10.1039/d2ra05643g

rsc.li/rsc-advances

1 Introduction

Since Tang and VanSlyke's first big breakthrough in 1987,[1] organic light-emitting diodes (OLEDs) have been profoundly improved in materials, device structures, and luminous efficiency.[2,3] In recent years, OLEDs have been widely used in the manufacture of display devices (such as TV screens, computer monitors, smart phone screens, flexible display panels, *etc.*), and are considered to have great potential in the field of lighting.[4] A modern OLED typically has a three-layer architecture, [anode|hole transport layer|light-emitting layer|electron transport layer|cathode]. The light-emitting material is dispersed in the light-emitting layer by doping or non-doping manner, and emits light in response to the current generated by the potential applied across the electrodes, which is so-called electroluminescence.[4]

So far, the luminescent materials as core OLED materials have undergone profound improvements, starting from the first generation of fluorescent materials (*e.g.*, aluminum octahydroxyquinoline), through the second generation of

[a]*School of Chemistry and Materials Engineering, Guiyang University, Guiyang, 550005, P. R. China. E-mail: lrasyw@163.com; Tel: +86-180-9605-0905*

[b]*School of Mathematics and Information Science, Guiyang University, Guiyang, 550005, P. R. China*

† Electronic supplementary information (ESI) available: Structure of donors and acceptors; details for machine learning; definition of group molecular similarity; *etc.* See DOI: https://doi.org/10.1039/d2ra05643g

phosphorescent materials represented by heavy transition noble metal organic complexes (*e.g.*, bipyridine complexes of Ir(III)) until the third generation of TADF materials (*e.g.*, organic donor–π-bridge–acceptor molecules).

Upon electric excitation, TADF materials (compounds characterized by very low first excited singlet-triplet energy gaps (ΔE_{ST})) get thermally activated to induce efficient reverse intersystem crossing (rISC) where the triplet excitons get converted into singlet excitons, so as to emit light dominantly from the emissive singlet excited state. In Fig. 1, the electroluminescence process of TADF material is schematically shown. Compared with noble metal–organic complex phosphorescent materials, TADF materials have the advantages of larger material space, low price, easy preparation and synthesis, easy fabrication of flexible screens, and more stable blue light emission. Therefore, in the last decade, as the most promising electroluminescent material for modern OLEDs, they have been experimentally,[2,5–9], theoretically[10–23] and theory-experiment jointly[15,24,25] studied in depth.

Basically, there are two classes of TADF materials that have been carefully explored.[4] The first type is pure organic D–A or D–π–A systems whose electron donor (D) or acceptor (A) are mainly constructed by nitrogen-containing aromatic heterocycles. The lowest excitation states typically possess significant intramolecular charge transfer (CT) transition character. After reasonable design and optimization, the external quantum efficiency (EQE) of OLED devices based on such TADF materials can even be as high as 30%. From the perspective of structural characteristics, the best luminous efficiency usually corresponds to the twisted D–A (or D–π–A) compounds due to enough steric hindrance between the donor and acceptor parts. Another type is transition metal (Cu(I), Ag(I), Zn(II), *etc.*) complexes with electronic configuration of d^{10}, and their lowest excited states usually have significant metal–ligand Charge transfer (MLCT) transition character. The saturated d^{10} electronic configuration of the central metal is very beneficial to reduce the possible quenching of the dπ–dπ* transitions in the complex and achieve deep blue emission.

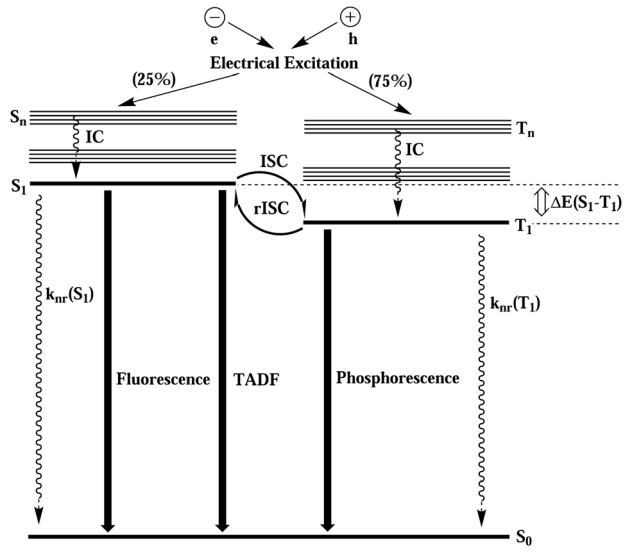

Fig. 1 Schematic diagram of the electroluminescence process of thermally activated delayed fluorescent materials.

The experimental breakthroughs came mainly from Adachi and collaborators, who focused on designing organic molecules with D–π–A (and other) frameworks, and tuning the frameworks to achieve a small enough ΔE_{ST} while maintaining a suitable fluorescence radiation rate, so that efficient TADF becomes possible. Recently developed blue TADF OLED devices have an EQE approaching 37%, which is rather impressive considering the EQE of Tang and VanSlyke's 1987 version of fluorescent OLEDs is about 1%.[1]

In a review on molecular design patterns of organic TADF materials,[3] Im *et al.* suggested that high-efficiency TADF materials should have at least a small ΔE_{ST} and a high photoluminescence quantum yield (PLQY). ΔE_{ST} is associated with upconverting triplet excitons to singlet excitons, while PLQY is closely related to the radiative transition probability. To obtain a small ΔE_{ST}, a strong donor/acceptor should be used and the molecular backbone should be twisted. The acquisition of high PLQY should have: a phenyl bridge as a connecting unit, delocalized and dispersed highest occupied molecular orbital (HOMO), and a double luminescent core. These strategies will undoubtedly provide useful guidance for further molecular design of TADF materials.

Contemporary electronic structure theory methods (*e.g.*, density functional theory, DFT) have been able to predict the optoelectronic properties of molecules (or materials) with relatively high accuracy.[26,27] Theoretical research is playing an increasingly important role in the in-depth understanding of the structure–property relationship and luminescence mechanism of TADF materials, and has a significant impact on the molecular design of such materials. As pointed out by Olivier and collaborators, theoretical research on this type of materials requires careful consideration.[19] Designing new molecules with efficient TADF emission is a difficult task, as they must exhibit a strong transition between singlet and triplet states without using heavy elements to enhance spin–orbit coupling fast conversion (large k_{rISC}). They should also show a large fluorescence rate (large k_F), but at the same time a small energy difference between excited singlet and triplet states (small ΔE_{ST}). In a feature article, Penfold *et al.* reviewed recent advances in theoretical and computational chemistry to understand TADF materials and mechanisms.[20] For luminescence dynamics, simply assume $k_{rISC} \gg k_F$, and apply eqn (1)

$$k_{TADF} = \frac{1}{3}k_F \exp(-\Delta E_{ST}/k_B T) \quad (1)$$

for rate estimation is considered inappropriate, and a new way to uniformly deal with the relevant quantities is needed. For electronic structure calculations, the standard time-dependent density functional method (TDDFT) may fail, in which case a tuned range-separated hybrid DFT or multi-reference configuration interaction (MRCI) method needs to be introduced. To understand the TADF mechanism, it is not enough to calculate the reverse intersystem crossover rate (k_{rISC}) between singlet (S_1) and triplet (T_1) only by using first-order perturbation theory and

Fermi's golden rule. Rather, the second-order perturbation theory including the spin-vibronic mechanism needs to be taken into account. In addition, conformational, regioisomerization, as well as environmental effects, are also crucial in determining the properties of TADF materials, and should therefore also be taken into account.

Commonly viewed as a branch discipline of theoretical chemistry, the rise of cheminformatics in recent years is deemed to make great impact on chemical science. With the continuous development of the theoretical system,[28–33] additionally, the open-sourceization of many high-quality cheminformatics tools (*e.g.*, RDKit, Mordred, stk *etc.*),[34–37] those make it possible (even for non-experts) to efficiently manage large amounts of chemical information. The efficient management of virtual molecules as well as molecular libraries *in silico* by using cheminformatics tools is crucial for large-scale computational design of (molecular) materials. On the other hand, in the field of computational design of (organic) molecules and (solid-state) materials, exploration of chemical compound space (CCS) using high-throughput virtual screening (HTVS) methods is being accepted as a routine procedure for molecular or material lead discovery. The important material categories involved include photovoltaic materials, optoelectronic materials, organic matrix flow battery materials, *etc.*[38–40] By designing computational funnels to efficiently deploy computational programs, the HTVS approach allows researchers to make data-driven discoveries by observing trends in the data.

As one of the branches of artificial intelligence (AI), machine learning (ML) can efficiently extract hidden relationships from large amounts of complex data. With advances in algorithmic models and open-source tools (general purpose: Scikit-learn, TensorFlow, Pytorch *etc.*;[41–44] chemistry or materials orientation: DeepChem, MLatom, MAST-ML *etc.*[45–48]), ML has profoundly changed the research paradigm of computational chemistry (or materials) science in the last decade.[49] Classical algorithm developments and applications include: predicting molecular atomization energies;[50] finding density functionals for model systems;[51] improving high-level electron correlation methods, learning universal molecular force fields; predicting molecular thermochemical properties, chemical reaction active sites, molecular excited state properties, molecular crystallization behavior, *etc.*[39,52–55] On the other hand, the establishment of open-source molecular databases has also promoted the development and calibration of models and algorithms which combine quantum chemistry with machine learning.[45,56–59]

Considering the rarity and high price of heavy metal transition metal complex phosphorescent materials, as well as the difficulty in achieving high-performance blue light emission, it is undoubtedly very attractive to design and develop stable and efficient TADF blue light materials as an alternative.[4] A pioneering attempt at high-throughput virtual screening of organic TADF materials was first made by Aspuru-Guzik and collaborators. By utilizing machine learning and time-dependent density functional theory methods, the screening procedure is rationally set to screen thousands of promising candidate TADF molecules from a search space of 1.6 million molecules, among which the best candidate molecules can be used to prepare OLED devices. The achieved external quantum efficiencies can be as high as 22%.[15] In another distinguished study, the same authors designed a deep neural network incorporating a variational autoencoder (VAE),[60] by accepting hundreds of thousands of existing chemical structures to build three coupled functions: encoder, decoder and predictor.[61] This model can convert discrete molecular representations to and from multidimensional continuous ones. Notably, the continuous representation allows the use of powerful gradient-based optimization to efficiently guide the search for optimal functional compounds.

This study focuses on the computational design of pure organic TADF molecules, by examining the effects of structural mutations as well as selection strategy on the results of high-throughput virtual screening of TADF materials, we expect to provide theoretical basis and guidance for the optimization of organic (or metallic complex type) TADF materials (lead) for larger-scale chemical space exploration in the future.

2 Theoretical methods

Sieving materials effectively within a large chemical compound space is as difficult as finding a needle in a haystack. Organic TADF molecules are typically electron donor–acceptor systems (DA, D–π–A, *etc.*) with N-containing heteroaromatic rings as building blocks. By introducing the concept of mutation and selection from genetic algorithm[62] (GA) and combining machine learning algorithm with quantum chemical computations in the calculation steps, a relatively simplified calculation program is proposed. The main aim is to efficiently explore the chemical compound space to obtain organic TADF material candidates.

Structural mutations can play an important role in tuning the electronic properties of molecular systems. Suppose our starting molecule is Biphenyl with 10 aromatic C–H bonds (aC–H) in the structure, if we allow two types of simple structural mutations:

(1) The whole is replaced by an aromatic N (aN),

$$\mathrm{aC{-}H} \rightarrow \mathrm{aN};$$

(2) The terminal H is substituted by a group G (G is a common simple electron donor or acceptor),

$$\mathrm{aC{-}H} \rightarrow \mathrm{aC{-}G}$$

If we further set substitution group G to be F (Fluorine group), for this molecule, we would virtually have 2^{10} mutant offspring (assuming all positions are distinguishable), and the real size would be 210 after removing the duplicates. This only considers the consequences of a single mutation. If there are more than one possible mutations at a single substitutable position, the number of combinations will expand dramatically beyond the calculable extent, given typically available computation resources owned by a computational research group. Obviously, the size of our initial molecular library G_0 will not be

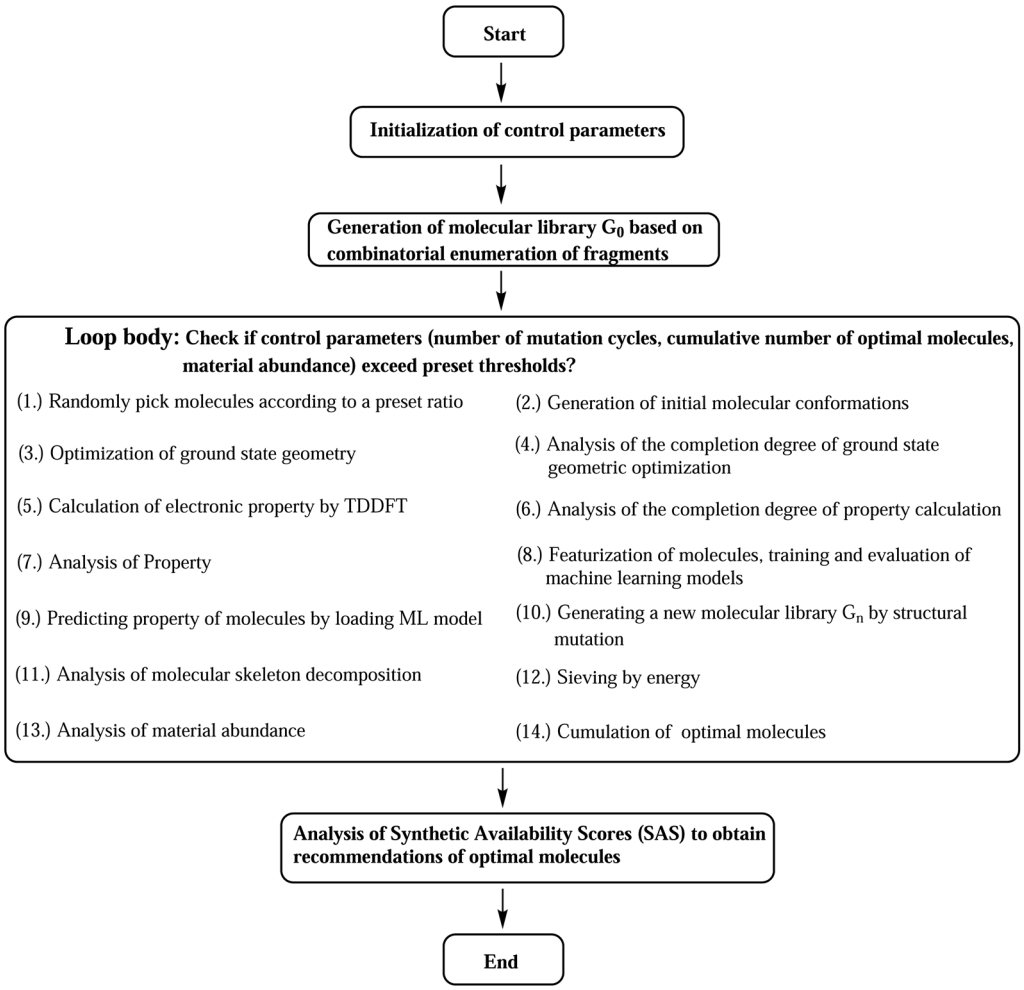

Fig. 2 Computational road map for HTVS of TADF molecules.

1 (typically greater than 10^3). Therefore, designing computational funnels based on a core property (or several core properties) of a material is crucial for efficient exploration of chemical compound space. For TADF material in current case, this property was chosen to be the energy difference between the first singlet excited state and the first triplet excited state (ΔE_{ST}).

Both single and mixed mutations have been taken into account. They are: N(slow), N(fast), F, CN, OMe, and NMe_2; F or OMe, F or NMe_2, CN or OMe, CN or NMe_2. N(slow) and N(fast) denote different mutation speeds, where N(slow) restricts only one position to be substituted, and N(fast) allows at most two. These mutations are denoted symbolically as **Sub1**, **Sub2**, **Sub3**, …, **Sub10**, respectively.

The designed computational framework for high-throughput virtual screening for TADF materials in this study is schematically presented in Fig. 2. The brief process is as follows:

(a) The control parameters get initialized. The convergence criteria for the loop are set as a combination of three: number of generation of mutations (n_g), number of accumulated optimal molecules ($n_{acc\_opt\_mols}$), and material abundance (ω_{MA}).

(b) Through rational selection of donor and acceptor fragments (30 donors and 43 acceptors, see Fig. S1 and S2 in the ESI†), under the donor–acceptor (DA) structural framework, using the open source cheminformatics package RDKit, and based on the Simplified Molecular Input Line Entry System (SMILES),[29] an initial molecular library G_0 (limit its size to 10^3) was obtained by combinatorial enumeration of fragments.

(c) Starting from this library, some molecules are randomly selected (the selection ratio is set to 10%), and their initial molecular conformations are generated by the RDKit package, where the ETKGD algorithm[35] is adopted.

(d) The core properties of the selected molecules are quickly and accurately calculated by quantum chemical calculations. The geometry optimization of ground state is performed by semi-empirical quantum chemical methond PM6-D3.[63,64] Based on the optimized geometry, the vertical energy gap (ΔE_{ST}) is calculated by TD-ωB97XD/6-31G(d) method.[65]

The differences between ground state geometries computed by B3LYP/6-31G(d) and PM6-D3 levels of theory is mesured by the root-mean-square deviation (RMSD) of the computed molecules for **Sub3** ($n_g = 0$ only). The RMSDs is calculated by the Python code rmsd with adoption of the Kabsch algorithm to

align molecules.[66,67] The distribution of frequency of RMSDs is given in Fig. S3.† The mean of the RMSDs is 0.59, and 75% of them are smaller than 0.69, which indicates the size of difference in geometries might be acceptable. Hence, the PM6-D3 method is adopt. In addition, the effect of varied ground state geometry optimization methods on the HTVS results have been briefly tested (see ESI†). Moreover, tuned range-separated hybrid functional methods (*e.g.*, LC-ω*PBE, ω*B97XD and CAM-B3LYP) are typically chosen to accurately compute the related electronic properties of TADF molecules. In this study, owing to the limit on available computational resources, the TD-ωB97XD/6-31G(d) method is chosen with the range-separation parameter not tuned, with the hope that the tuned range-separation parameters of molecules could not deviate considerably from the default values or if the deviations are considerable, they could induce the same direction changes on the distribution of the computed property of the compound library.

(e) The molecular structures get featurized by molecular fingerprint method, and are introduced into the machine learning algorithm to train and learn a model. The chosen fingerprint is the ECFP method, and the computation is assisted by the DeepChem package. And the Random Forest (RF) Regressor[68] of the machine learning package Scikit-learn is used. (For more details, refer to the related section in the ESI.†)

(f) By using the learned ML model, we predict the property of entire molecular library so as to obtain the optimal molecules within the library.

(g) A certain proportion (10%) of the top-ranked molecules are taken out to generate a new generation of molecular libraries (named G_n, $n = 1, 2, 3, \ldots$) by means of structural mutations. Here, selection and mutation get incorporated into the computational paths.

(h) Analysis of molecular skeleton decomposition is performed to access the corresponding evolution of skeleton of molecules in library. There the Murcko Skeleton Decomposition[69] method in the RDKit package is adopted.

(i) An energy sieve is then applied to divide the molecules into regions of different colors. Molecules with predicted vertical first excited energies (E_{S1}) larger than 2.80 eV, between 2.50 to 2.80 eV, and smaller than 2.50 eV are partitioned into the blue, green, and red regions of colors, respectively.

(j) Compute material abundance (ω_{MA}) and accumulate optimal molecules to get number of accumulated optimal molecules ($n_{acc\_opt\_mols}$). The threshold for 'good' material is the predicted $\Delta E_{ST} < 0.15$ eV. The material abundance is computed by eqn (2)

$$\omega_{MA} = \frac{\text{number of molecules with } \Delta E_{ST} \text{ lower than 0.15 eV}}{\text{number of all molecules}} \quad (2)$$

(k) The accumulated optimal molecules are finally ranked based on Synthetic Accessibility Scores (SAS) to obtain the best TADF material candidates. Low SASs imply relative ease of synthesis of molecules. Since the perpetual mutations on the molecular framework would profoundly disturb the structure, even could make the synthesis impossible, a final control of SAS is certainly necessary to sieve out bad structures from the good ones.

Repeat the above steps from (c) to (j) using the newly formed library G_n to generate a next library G_{n+1}, until we have reached the preset loop convergence criterion. The definition of calculation completeness for geometry optimization and property evaluation is meant to assist the automation of related calculation routines. Unavoidably, the geometry or property of some molecules (and their mutation offspring) may not converge under the chosen computational methods, therefore, only a preset completeness ratio is required to escape the steps. For geometry optimization and property calculation, the ratios are 0.80 and 0.90, respectively. Inside the loop, the interconversion of chemical files between different formats is facilitated by the Open Babel cheminformatics tool.[70] Gratefully, the analysis of data is assisted by the Anaconda3 (ref. 71) scientific computing platform and the Spyder[72] integrate development environment, where several numeric Python packages have been used, including NumPy, Pandas, SciPy and Matplotlib.[73–76] The above computational program has been packaged and distributed as an open source Python code (SALAM).[77]

All quantum chemical computations is done by the Gaussian 16 package.[78]

3 Results and discussion

This work is mainly designed to examine the effect of different structural mutations as well as selection strategy on the result of HTVS of TADF materials. The initial compound library G_0 is generated by combinatorial enumeration of 30 donors and 43

Table 1 The evolution of material abundance (ω_{MA}) with increase of mutation generation (n_g) for different mutations

n_g	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	Sub10
0	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
1	0.133	0.076	0.200	0.025	0.121	0.093	0.186	0.133	0.067	0.152
2	0.460	0.033	0.697	0.351	0.337	0.027	0.967	0.648	0.901	0.799
3	0.766	0.210	0.825	0.634	0.953	0.179	0.968	0.967	0.865	0.708
4	0.592	0.481	0.798	0.762	0.805	0.352	0.975	0.649	0.720	0.567
5	0.789	0.330	0.837	0.857	0.926	0.591	0.879	0.423	0.916	0.845
6	0.760	0.376	0.834	0.782	0.959	0.785	0.929	0.941	0.982	0.894
7	0.747	0.368	0.833	0.814	0.949	0.927	0.958	0.979	0.812	1.000
8	0.760	0.197	0.860	0.867	0.877	0.876	1.000	0.823	0.841	0.777
9	0.642	0.306	0.852	0.946	0.895	0.716	0.995	0.915	0.982	1.000

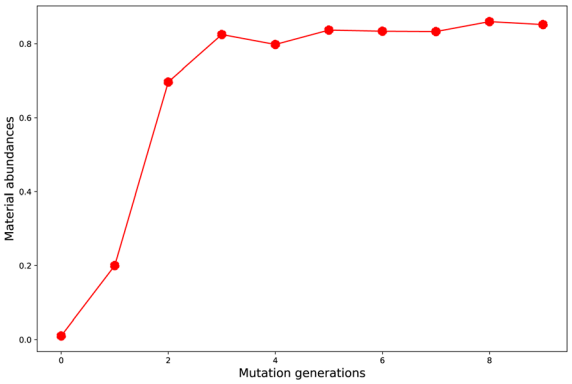

Fig. 3 The material abundance *versus* mutation generation for **Sub3**.

acceptors under the donor–acceptor (DA) molecular frameworks (details in ESI†). The kernel property adopted for optimization is the vertical energy gap ΔE_{ST}.

3.1 Material abundance

Starting from the common baseline library (G_0), the evolution of material abundance (ω_{MA}) with the increase of mutation generation (n_g) for different mutations has been listed in Table 1. For any of the studied mutations, the initial ω_{MA} is 0.010 ($n_g = 0$), as n_g increases, the ω_{MA} typically can display a sharp increase at low n_g values, and give slow increase at meadian n_g values, and finally trend to approach (or oscillate around) their respective limits. For an illustration of this behavior, the material abundance *versus* mutation generation for mutation = F (**Sub3**) is depicted in Fig. 3. From the rather low 0.010 ($n_g = 0$) increases to relatively high 0.825 ($n_g = 3$) and finally ceases at 0.852 ($n_g = 9$). For succeeded generations, the corresponding ω_{MA} may experience significant decrease, however, the overall trend to stay around a limit is rather obvious.

A comparison of **Sub1** and **Sub2** (correspond to different mutation speeds: slow *versus* fast), tells that for this type of mutations (aC–H → aN) a slow mutation speed is favorable than a fast one in achieving high ω_{MA} at large n_g. As $n_g = 9$, the ω_{MA} for **Sub1** and **Sub2** are 0.642 and 0.306, correspondingly. For the second type of mutations (aC–H → aC–G), the evolution of ω_{MA} seem relatively small as n_g increases to high values. To sum up, except the fast mutation **Sub2**, all of the rest mutations can effectively concentrate 'good' molecules in compound library, hence give large ω_{MA} values.

3.2 Average number of aromatic aCH bonds (n_{aCH})

Turn back to the setup of the computational road map, it's natural to expect that as the two types mutations occupy positions (say, the mutation generation increases) the average number of aromatic aCH bonds of a compound library would experience significant drop in magnitude across the process. The evolution of average number of aromatic aCH bonds (n_{aCH}) with increase of mutation generation (n_g) for different mutations is listed in Table 2. It's noted that the overall evolution trend for n_{aCH} (uniformly decreases in value as n_g increases) is

Table 2 The evolution of average number of aromatic aCH bonds (n_{aCH}) with increase of mutation generation (n_g) for different mutations

n_g	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	Sub10
0	18.3	18.3	18.3	18.3	18.3	18.3	18.3	18.3	18.3	18.3
1	15.7	14.8	14.9	14.9	14.9	14.9	15.8	15.9	15.8	15.7
2	13.7	11.1	11.3	10.0	12.9	10.7	13.1	12.6	14.7	12.8
3	11.8	7.8	10.1	11.1	9.4	8.9	17.8	10.5	13.8	14.5
4	11.9	6.5	6.3	12.5	8.0	15.6	11.1	9.3	16.6	15.1
5	11.1	5.5	6.2	10.9	6.9	19.2	10.7	13.4	19.2	20.1
6	11.1	5.7	5.4	7.9	7.2	18.1	11.3	6.1	14.0	21.1
7	11.2	5.8	5.3	8.0	5.9	20.2	10.5	4.8	10.5	20.4
8	10.9	5.5	5.0	7.8	6.3	18.2	10.6	4.4	7.8	4.0
9	10.8	5.4	4.9	8.2	5.4	16.0	9.7	3.6	7.5	2.9

as we expected, though exceptions (alternative increase and decrease) do exist. Starting from the baseline 18.3 ($n_g = 0$), the n_{aCH} can drop down to a smaller value (generally lower than 10.0 as $n_g = 9$).

Take **Sub3** as a case, whose n_{aCH} starts from a large value 18.3 ($n_g = 0$), sharply decreases to 6.3 ($n_g = 4$), finally drops to 4.9 ($n_g = 9$). The related diagram for **Sub3** has been depicted in Fig. 4. The starting large n_{aCH} should be attributed to large amount of unsubstituted molecules with multi-cyclic aromatic structures in the library. A small ending n_{aCH} should be attributed to large amount of oversubstituted molecules in the library, while a large ending n_{aCH} should be attributed to concentration of very large multi-cyclic aromatic molecules with low substitutions in the library. Thus, the seemingly anomalous phenomena of alternative increase and decrease in n_{aCH} can be understood by tracing the evolution of molecular skeletons of the compound library.

As compared with the slow mutation (**Sub1**), the n_{aCH} of the fast mutation (**Sub2**) exhibits a very rapid drop in value. However, this rapid drop in n_{aCH} is not sufficient to guarantee a meaningful increase in ω_{MA} (compare Tables 2 and 1). For some mutations (**Sub6** and **Sub10**), the anomalous alternative increase and decrease in n_{aCH} is a sign of violent transformation of dominant molecular skeletons under selection and mutation process. Therefore, it can be used as an indicator to differentiate

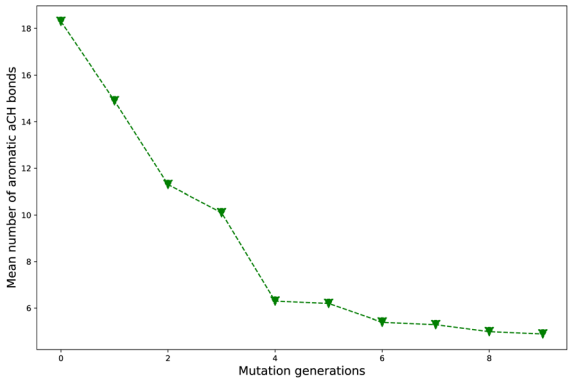

Fig. 4 The mean number of aromatic aCH bonds *versus* mutation generation for **Sub3**.

Table 3 The evolution of number of accumulated optimal molecules ($n_{acc\_opt\_mols}$) with increase of mutation generation (n_g) for different mutations

n_g	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	Sub10
0	10	10	10	10	10	10	10	10	10	10
1	83	83	202	34	125	97	191	137	75	156
2	288	108	791	341	418	122	1048	695	910	870
3	556	275	1429	807	1195	285	1894	1500	1679	1463
4	781	533	1808	1464	1607	576	2728	2043	2329	1947
5	1052	615	2041	2168	1841	1078	3455	2399	3143	2666
6	1345	696	2144	2667	2191	1768	4265	3104	4012	3453
7	1626	732	2182	3074	2380	2655	5083	3807	4730	4318
8	1896	741	2236	3433	2596	3423	5961	4328	5376	4645
9	2063	753	2273	3970	2730	4057	6833	4803	6172	4902

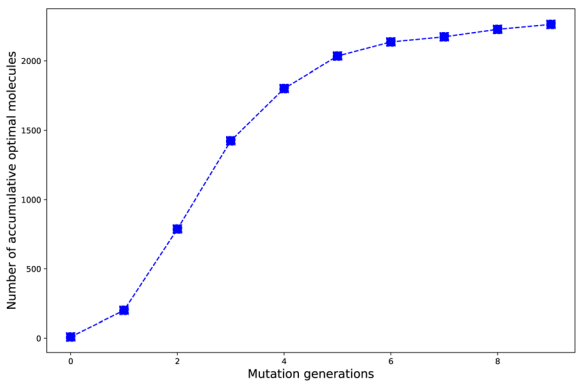

Fig. 5 The number of accumulated optimal molecules *versus* mutation generation for **Sub3**.

the skeleton transformation effect of different mutations on the same DA substrates. If n_{aCH} retains large values as n_g turns large, there would be great amount of relatively 'big' molecules accumulated in library. Otherwise, if n_{aCH} exhibits rapid drop as n_g increases, there would be great amount of relatively 'small' molecules accumulated in library. Generally, from a point of view of synthetic chemistry, the 'small' molecules is more favorable than the 'big' ones.

To sum up, analysis of the n_{aCH} of different mutations tells us that the uniform drop in n_{aCH} with increase of n_g is not a sufficient condition to guarantee a meaningful increase in ω_{MA}, rather it can be used a an indicator to differentiate different mutations on skeleton transformation effect.

3.3 Accumulated optimal molecules

Harvesting optimal molecules as more as possible by applying the designed HTVS program is one of the most important aims to be achieved. In other words, the best mutation should be the one which can induce the molecule's property in the right direction to fulfill the requirement as material. Since in the design of the selection and mutation computational step, the parent molecules are intrinsically added as part to form the new compound library, we think a number of accumulated optimal molecules along the mutation generations should be more suitable to account for the concentrating ability of the different mutations.

The evolution of number of accumulated optimal molecules ($n_{acc\_opt\_mols}$) with increase of mutation generation (n_g) for different mutations has been listed in Table 3. The $n_{acc\_opt\_mols}$ for any of mutations can exhibit a sharp increase in low to middle n_g values ($0 < n_g \leq 5$), follow by slower growth for middle to high n_g ($6 \leq n_g \leq 9$), and may finally trend to flatten out. The probability of finding identical molecules between adjacent libraries will increase as n_g becomes large. This behavior is best demonstrated by the data of **Sub3**, as depicted in Fig. 5.

As compared with the fast mutation **Sub2**, the slow mutation **Sub1** can give approximately 2.7 times increase in $n_{acc\_opt\_mols}$. Thus, **Sub1** is more favorable than **Sub2** in producing optimal molecules. Taking $n_g = 9$ as base, for the 4 terminal single mutations (**Sub3** to **Sub6**), the precedence order is: **Sub3** < **Sub5** < **Sub4** ≈ **Sub6**, a strong donor (or acceptor) is superior to a weak one; for the rest 4 mixed mutations (**Sub7** to **Sub10**), the precedence order is: **Sub8** < **Sub10** < **Sub9** < **Sub7**, the weak–weak pair exhibits superiority among others. The mixed mutations would produce more optimal molecules as expected since they correspond to larger chemical compound spaces, however, the price is the significant increase in molecular complexity, which might eventually prohibit them as material due to difficulty from experimental synthesis.

3.4 Mean energy gap $\overline{\Delta E_{ST}}$

A successful HTVS program ought to effectively drive the optimized property to the optimal direction. For the current case,

Table 4 The evolution of mean energy gap ($\overline{\Delta E_{ST}}$) with increase of mutation generation (n_g) for different mutations

n_g	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	Sub10
0	0.669	0.669	0.669	0.669	0.669	0.669	0.669	0.669	0.669	0.669
1	0.292	0.369	0.246	0.332	0.280	0.310	0.262	0.298	0.253	0.264
2	0.218	0.316	0.140	0.218	0.275	0.236	0.065	0.150	0.096	0.115
3	0.088	0.276	0.098	0.139	0.086	0.331	0.046	0.134	0.103	0.116
4	0.187	0.207	0.095	0.116	0.106	0.178	0.051	0.144	0.121	0.151
5	0.104	0.268	0.078	0.101	0.086	0.165	0.082	0.256	0.098	0.085
6	0.120	0.233	0.075	0.108	0.085	0.124	0.058	0.052	0.057	0.078
7	0.110	0.249	0.076	0.102	0.084	0.076	0.050	0.066	0.108	0.100
8	0.114	0.274	0.074	0.077	0.089	0.104	0.042	0.092	0.090	0.096
9	0.155	0.229	0.076	0.059	0.085	0.128	0.043	0.073	0.063	0.043

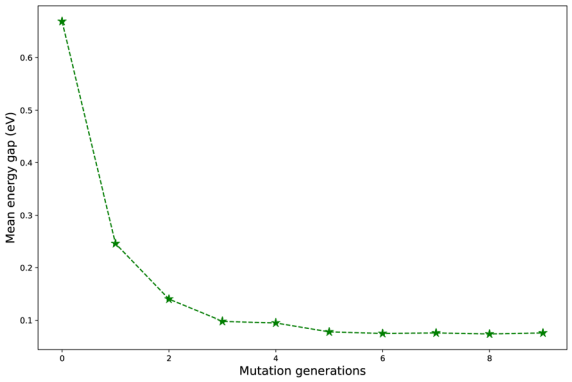

Fig. 6 The evolution of mean energy gaps *versus* mutation generation for **Sub3**.

the evolution of mean energy gap $\overline{\Delta E_{ST}}$ of library along with the mutation generations has been listed in Table 4. It's observed that: the $\overline{\Delta E_{ST}}$ for any of all mutations can exhibit a sharp decrease as n_g moves from low to middle values ($0 < n_g \leq 3$), and trend to flatten out from middle to high values ($3 < n_g \leq 9$). A large proportion of molecules with low ΔE_{ST} is enough to maintain the mean quantity $\overline{\Delta E_{ST}}$ of library at low value.

Regardless of the types of mutations, the fast convergent trend of $\overline{\Delta E_{ST}}$ is rather impressive. The evolution of mean energy gaps *versus* mutation generation for **Sub3** has been depicted in Fig. 6. The $\overline{\Delta E_{ST}}$ starts from a value of 0.669 eV ($n_g = 0$), experiences a sharp drop to 0.098 eV ($n_g = 3$), finally trends to flatten out to 0.076 eV ($n_g = 9$). There should have a clear correlation between the evolutionary behaviors of $\overline{\Delta E_{ST}}$ and ω_{MA}, since both of them are group quantities based on the ΔE_{ST} of molecules.

To give more details on the impact of the mutation along with mutation generations, the evolution of energy gaps frequency distribution *versus* mutation generation for **Sub3** has been depicted in Fig. 7. The fast shift to low ΔE_{ST} is apparent (n_g moves from 0 to 3), then the distribution retains a large proportion in the very low value range and tails in the low to medium range.

To sum up, regardless of the types of mutations, the mean energy gap can exhibit a fast convergent trend toward very low values, hence the studied mutations (except **Sub2**) can cooperate very well with the DA substrates to generate optimal molecules.

3.5 Skeleton decomposition

Decomposition of a closely connected compound library into molecular skeletons is an effective way to examine it's main skeleton compositions. Considering the loop nature of the designed HTVS program, it's interesting to observe how the dominant molecular skeletons can evolve with the increase of mutation generation. By applying the Murcko Skeleton Decomposition method, the related data has been computed. Two types of molecular skeletons (common and generic cores) are adopted. The common cores identify both types of elements and bonds, while the generic cores neglect this information.

The evolution of skeletons (common cores) *versus* mutation generation for **Sub3** has been depicted in Fig. 8, and that of

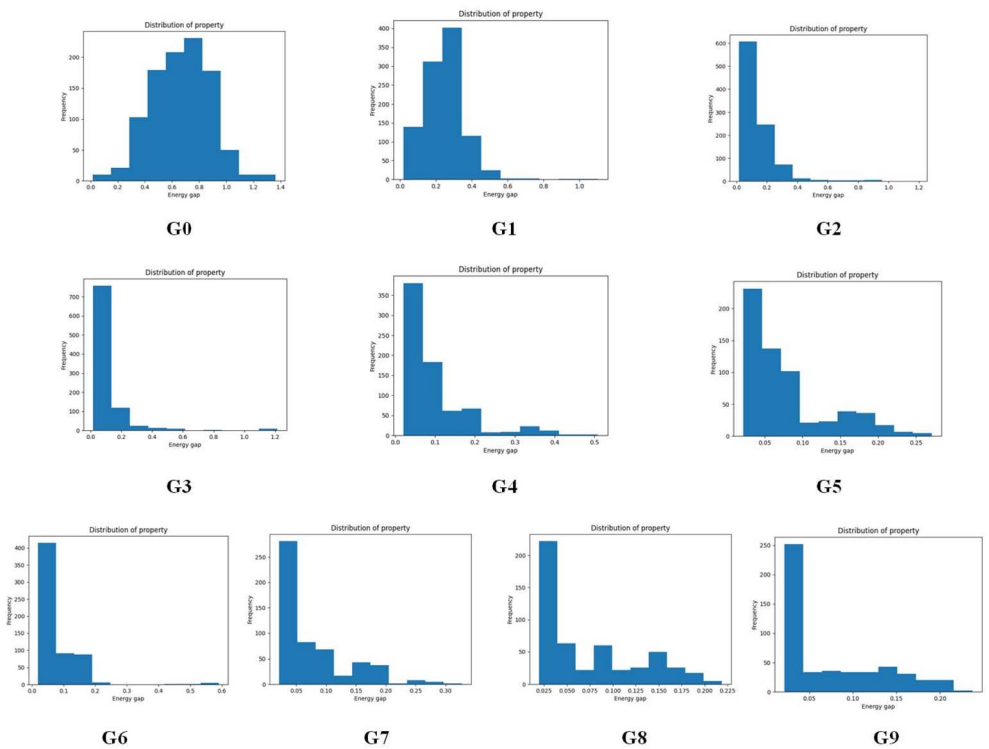

Fig. 7 The evolution of energy gaps frequency distribution *versus* mutation generation for **Sub3**.

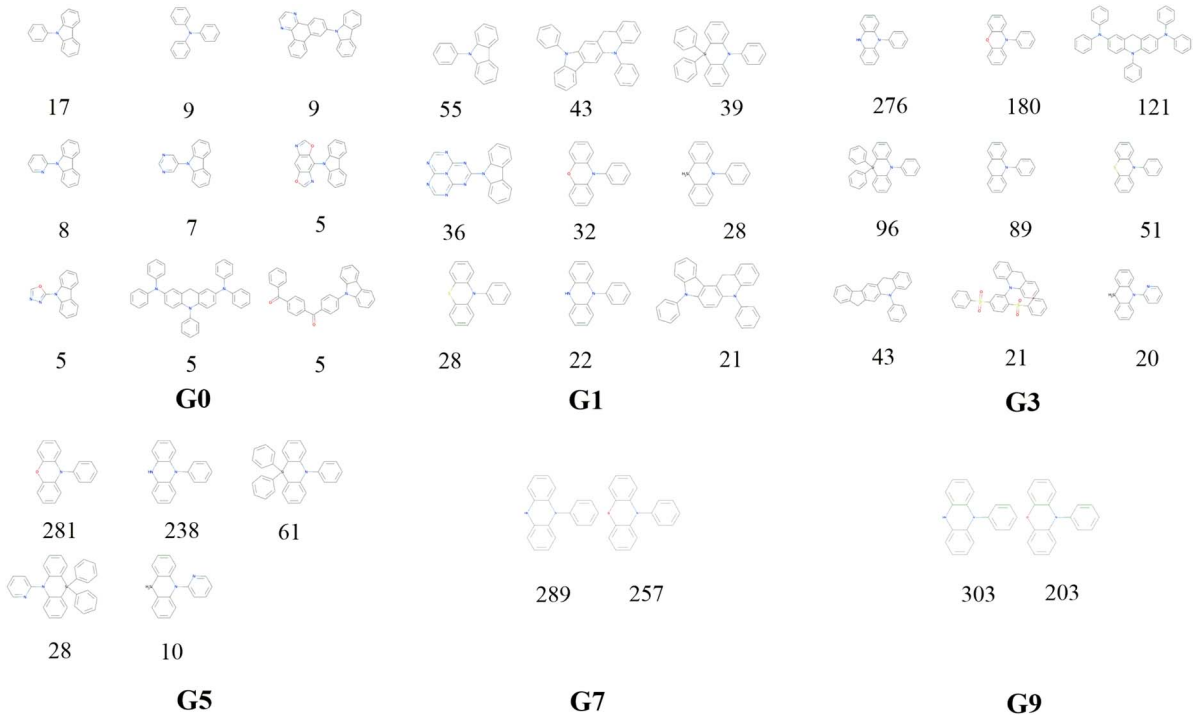

Fig. 8 The evolution of skeleton (core) with mutation generation for **Sub3** (the numbers below structures denote the corresponding frequencies).

skeletons (generic cores) has been given in Fig. S4 in the ESI.† For simplicity, at most 9 dominant high-frequency skeletons and selected mutation generations have been shown. For both types of cores, there exists explicit quantitative shrinkage of dominant high-frequency skeletons. The common cores starts from the relatively uniform distribution of frequencies of 9 cores ($n_g = 0$), then collapses to very uneven distribution of frequencies of 5 cores ($n_g = 5$), and finally collapses further to a distribution of frequencies of only 2 cores ($n_g = 9$) (Fig. 8). The generic cores can exhibit more profound collapse in number of cores (Fig. S4 in the ESI†). This collapse is a sign of efficient convergence of the structures around 'excellent' molecules. Hence is beneficial for obtaining optimal molecules.

Table 5 The evolution of the number of molecules in library (n_{tot}), the number of intersection molecules (n_{inter}), and the group fingerprint similarity between two libraries (Δ_{MSPR}) with increase of mutation generation (n_g) for **Sub1**, **Sub3** and **Sub7**

	Sub1			**Sub3**			**Sub7**		
n_g	n_{tot}	n_{inter}	Δ_{MSPR}	n_{tot}	n_{inter}	Δ_{MSPR}	n_{tot}	n_{inter}	Δ_{MSPR}
0	1000	—	—	1000	—	—	1000	—	—
1	588	100[a]	0.660[a]	1000	100	0.542	1000	98	0.553
2	531	125	0.738	960	111	0.638	1000	127	0.666
3	487	127	0.778	929	140	0.787	1000	122	0.737
4	585	179	0.900	747	227	0.893	1000	135	0.735
5	541	161	0.834	618	282	0.949	1000	171	0.834
6	570	151	0.872	608	354	0.908	1000	126	0.779
7	588	180	0.898	546	350	0.932	1000	138	0.910
8	574	173	0.930	514	339	0.980	1000	120	0.891
9	565	212	0.928	506	364	0.992	1000	117	0.959

[a] Calculated with respect to the corresponding precedent n_g.

3.6 Similarity analysis between mutation generations

Similarity is an important concept used to describe the degree of difference in the structure of two molecules. Using this concept, and introducing a valid fingerprint representation for molecule and a similarity metric, the similarity can be easily calculated for any pair of two molecules.

Following the concept of similarity for two molecules and based on molecular fingerprint representation, we propose a numerical method to calculate group fingerprint similarity (Δ_{MSPR}) between two compound libraries. Here, the molecular fingerprint representation method is ECFP, and the similarity is measured by the Tanimoto metric.

The calculation of the group fingerprint similarity (Δ_{MSPR}) is based a algorithm, which we name it the Maximum Similarity Pairing Rule (MSPR) (refer to ESI†). The evolution of the number of molecules in library (n_{tot}), the number of intersection molecules (n_{inter}), and the group fingerprint similarity between two libraries (Δ_{MSPR}) with increase of mutation generation (n_g) for **Sub1**, **Sub3** and **Sub7** is listed in Table 5.

For **Sub1**, the n_{tot} keeps a size of about 500 for n_g in range from 1 to 9, and the n_{inter} exhibits a slowly increase trend in that range, hence the Δ_{MSPR} can change from a low value of 0.660 (for $n_g = 1$) to a high value 0.928 (for $n_g = 9$). For **Sub3**, the rather high values of Δ_{MSPR} for high mutation generations ($n_g \geq 5$) can

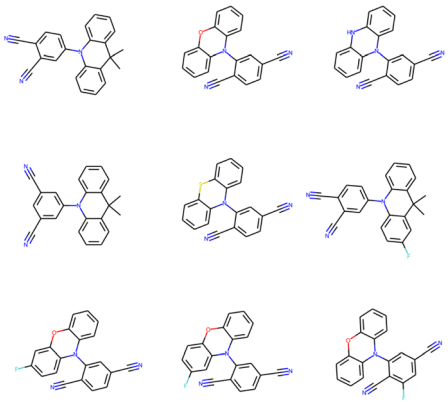

Fig. 9 The structures of 9 optimal molecules with lowest SAS for **Sub3**.

be ascribed by the ordered descending of n_{tot} and incrementing of n_{inter}. For **Sub7**, both of n_{tot} and n_{inter} keep their sizes as n_g increases. The high values of Δ_{MSPR} for high mutation generations ($n_g \geq 7$) can be ascribed by the convergence in molecular skeletons, which can retain the similarity between pairs of molecules at high values of range.

To sum up, regardless of types of mutations, the group fingerprint similarity (Δ_{MSPR}) at high mutation generations can retain high enough values (typically larger than 0.90), which can be associated with the apparent convergence in molecular skeletons at high mutation generations.

3.7 Optimal TADF molecules (low ΔE_{ST} and SAS)

The (energy gap) optimal molecules for all mutations have been sorted by synthetic accessibility scores from low to high, so as to give recommendation for TADF materials candidates. The structures of 9 molecules with lowest SAS for **Sub3** have been depicted in Fig. 9. For other mutations, the related structures are given by Fig. S5 in the ESI.†

In principle, molecules with simpler and more symmetric structures are favored by the SAS sorting routine. Within the studied compound space, those compounds constructed by typical tri-cyclic donors connecting with (polyacetonitrile substituted) benzenes acceptors can possess the lowest SAS. In addition, they can exhibit low enough energy gaps. Therefore, from the point of view of synthetic chemistry, they are recommended as optimal TADF molecules (low ΔE_{ST} and SAS), although only a small energy gap might not be enough to guarantee the occurrence of TADF emission.

Notably, possessing a low enough ΔE_{ST} as a necessary condition, the real occurrence of TADF emission for a compound should at least be accompanied with an acceptable radioactive fluorescent rate. Since the number of accumulated optimal molecules for high mutation generation are typically larger than 2000, we expect the extra fulfillment of the radioactive fluorescent rate may have great chance to occur possibly by further sieving the already obtained compound library of accumulated optimal molecules.

3.8 Recommendation for combinations of DA substrates with mutations

3.8.1 Optimal molecular skeletons for realizing low energy gaps. Regardless of the mutations, the high material abundances (typically >0.8) at large mutation generations ($n_g \geq 6$) naturally indicate the existence of 'optimal' molecular skeletons for realizing low energy gaps as substituted by suitable mutations. The SMILES of common and generic cores, and their frequencies for studied mutations as $n_g = 9$ have been listed in Table 6.

The distribution of frequencies of different skeletons for a specific mutation is generally uneven with one dominant skeleton. The common core with SMILES = "c1ccc(N2c3ccccc3Nc3ccccc32)cc1" exists for several mutations (**Sub3**, **Sub5**, **Sub8** and **Sub10**) with associated frequencies (303, 243, 662 and 555). The common core with SMILES = "c1ccc(N2c3ccccc3Oc3ccccc32)cc1" exists for several mutations (**Sub3**, **Sub4** and **Sub8**) with associated frequencies (203, 29 and 134). The common core with SMILES = "c1ccc(N2c3ccccc3Cc3cc4c(cc32)c2ccccc2n4-c2ccccc2)cc1" exists for two mutations (**Sub7** and **Sub9**) with associated frequencies (984 and 688). Similarly, the common core with SMILES = "O=S(=O)(c1ccccc1)c1ccc(S(=O)(=O)c2ccccc2)c(N2c3ccccc3Cc3ccccc32)c1" exists only for **Sub4** with frequencies = 640. Different skeletons can exhibit distinguishable preference to associate with different mutations. In short, the mutation can select the optimal skeleton(s) out from thousands of original DA substrates to realize low energy gaps.

After removing duplicates, the structures of optimal skeletons (common cores) for mutations from **Sub3** to **Sub10** as $n_g = 9$ have been depicted in Fig. 10. The related diagram for generic cores is given by Fig. S6 in ESI.† By definitions of common and generic cores, the common cores for **Sub1** and **Sub2** cannot collapse, however, their generic cores do belong to 1 or 2 skeletons (Table 6). It's interesting to note that the overall numbers of common and generic cores for all mutations are 11 and 7. Hence, in a sense, the 'optimal' skeletons seem unique and useful in realizing low energy gaps.

3.8.2 Emitting colors. The optimization of structures is designed for one kernel property (the energy gap), hence the accompanied emitting color (energy) should not be optimized (as mutation generation increases). Conceptually, clean (with limited substitution) DA compounds should mainly exhibit blue emission owing to the intrinsic $\pi \rightarrow \pi^*$ transition pattern. As hetero-atoms doping in the structure, there would be considerable red shift to occur. In addition, the connection of strong electronic donor to conjugate aromatic system would have effect to red-shift the emission. For the chosen DA substrates and mutations, analysis of data from the energy sieve step shows that:

(1) To access red color, **Sub10** (CN or NMe_2) is the best mutation groups, which can produce considerable proportion red molecules when the mutation generation equals to 9.

(2) To access green color, **Sub1** (N(slow)), **Sub2** (N(fast)), **Sub4** (CN), **Sub9** (CN or OMe), and **Sub10** (CN or NMe_2) are favored groups.

Table 6 The SMILES of common and generic cores, and their frequencies for studied mutations as $n_g = 9$

Mutations		Freq
SMILES of common cores		
Sub3	c1ccc(N2c3ccccc3Nc3ccccc32)cc1	303
	c1ccc(N2c3ccccc3Oc3ccccc32)cc1	203
Sub4	O=S(=O)(c1ccccc1)c1ccc(S(=O)(=O)c2ccccc2)c(N2c3ccccc3Cc3ccccc32)c1	640
	O=S1(=O)c2ccccc2S(=O)(=O)c2cc(N3c4ccccc4Cc4ccccc43)ccc21	122
	O=S1(=O)c2ccccc2S(=O)(=O)c2cc(N3c4ccccc4Oc4ccccc43)ccc21	80
	c1ccc(N2c3ccccc3Oc3ccccc32)cc1	29
Sub5	c1ccc(N2c3ccccc3Nc3ccccc32)cc1	243
	c1ccc(N2c3ccccc3Cc3c2ccc2c3c3ccccc3n2-c2ccccc2)cc1	233
Sub6	c1ccc(N(c2ccccc2)c2ccc3c(c2)Cc2cc(N(c4ccccc4)c4ccccc4)ccc2N3c2ccccc2)cc1	1000
Sub7	c1ccc(N2c3ccccc3Cc3cc4c(cc32)c2ccccc2n4-c2ccccc2)cc1	984
	c1ccc(N2c3ccccc3Sc3ccccc32)cc1	16
Sub8	c1ccc(N2c3ccccc3Nc3ccccc32)cc1	662
	c1ccc(N2c3ccccc3Oc3ccccc32)cc1	134
Sub9	c1ccc(N2c3ccccc3Cc3cc4c(cc32)c2ccccc2n4-c2ccccc2)cc1	688
	c1ccc(N2c3ccccc3Cc3ccccc32)cc1	176
	c1ccc(N2c3ccccc3Cc3cc4c(cc32)Cc2ccccc2-4)cc1	136
Sub10	c1ccc(N2c3ccccc3Nc3ccccc32)cc1	555
SMILES of generic cores		
Sub1	C1CCC(C2C3CCCCC3CC3CC4C(CC32)C2CCCCC2C4C2CCCCC2)CC1	565
Sub2	C1CCC(C2C3CCCCC3CC3CCCCC32)CC1	506
	C1CCC(C2C3CCCCC3CC3CC4C(CC5CCCCC54)CC32)CC1	170
Sub3	C1CCC(C2C3CCCCC3CC3CCCCC32)CC1	506
Sub4	CC(C)(C1CCCCC1)C1CCC(C(C)(C)C2CCCCC2)C(C2C3CCCCC3CC3CCCCC32)C1	640
	CC1(C)C2CCCCC2C(C)(C)C2CC(C3C4CCCCC4CC4CCCCC43)CCC21	202
	C1CCC(C2C3CCCCC3CC3CCCCC32)CC1	29
Sub5	C1CCC(C2C3CCCCC3CC3CCCCC32)CC1	243
	C1CCC(C2C3CCCCC3CC3C2CCC2C(C4CCCCC4)C4CCCCC4C32)CC1	233
Sub6	C1CCC(C(C2CCCCC2)C2CCC3C(CC4CC(C(C5CCCCC5)C5CCCCC5)CCC4C3C3CCCCC3)C2)CC1	1000
Sub7	C1CCC(C2C3CCCCC3CC3CC4C(CC32)C2CCCCC2C4C2CCCCC2)CC1	984
	C1CCC(C2C3CCCCC3CC3CCCCC32)CC1	16
Sub8	C1CCC(C2C3CCCCC3CC3CCCCC32)CC1	796
Sub9	C1CCC(C2C3CCCCC3CC3CC4C(CC32)C2CCCCC2C4C2CCCCC2)CC1	688
	C1CCC(C2C3CCCCC3CC3CCCCC32)CC1	176
	C1CCC(C2C3CCCCC3CC3CC4C(CC5CCCCC54)CC32)CC1	136
Sub10	C1CCC(C2C3CCCCC3CC3CCCCC32)CC1	555

(3) To access blue color, all mutations seem valid. **Sub3** (F) and **Sub5** (OMe) are the recommended groups.

3.9 Perspective on the applicability of the designed HTVS program

By now, we have demonstrated the ability of the designed HTVS program to sieve organic TADF molecules (and skeletons) out by providing a set of preset mutations and starting from a medium size of compound library (10^3 DA molecules) constructed by enumeration of donors and acceptors. By the design of the mutations (only conjugated aromatic units containing molecular systems are valid), once the mutations could occur within a new compound library, the electronic properties could effectively be modulated by the preset mutations, then the designed HTVS program would have great chance to be applicable to accumulate 'good molecules' for specific kind of properties. Accordingly, upon certain modifications of the code (*e.g.,* adding more property analysis functions), we expect the designed HTVS program might have the following possible fields of applications:

* Systems: organic molecules within DA, D–π–A, D–A–D, A–D–A and D_3–A frameworks; organometallic complexes with conjugate organic aromatic ligands.

* Properties: electronic and electric properties based on ground state (and possibly excited state) geometry of molecule.

* Materials: organic and organometallic TADF, nonlinear optical and two-photon absorption materials; and possibly organic conductive and photovoltaic materials.

Obviously, there are some areas for further improvement. Whether other types of mutations are possible for conjugated aromatic systems, and whether it is possible to design a valid crossover operator to combine two parent molecules to give offsprings molecules, have not been explored yet. Moreover, the program is driven by only one core property, sometimes there would be several properties to be optimized simultaneously, therefore, more design efforts should be devoted to support this kind of requirement. And, there should be more supporting on (artificial neural network based) deep learning methods to improve the models' accuracy for property prediction. Additionally, more quantum chemical packages as computing

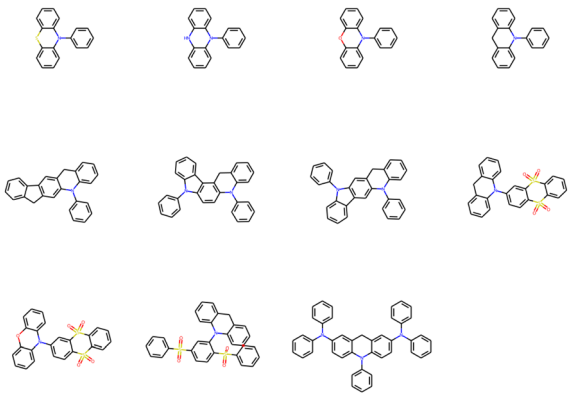

Fig. 10 The structures of optimal skeletons (common cores) for mutations from **Sub3** to **Sub10** as $n_g = 9$.

engines for electronic structure should be supported. Accompanied with the above-mentioned areas for improvement, we still hope that the designed HTVS programs could provide some valuable insights into related fields.

4 Conclusion

By combining machine learning and quantum chemical calculations, using cheminformatics tools, and introducing the concept of selection and mutation from evolutionary theory, we designed a computational program for high-throughput virtual screening of thermally activated delayed fluorescence molecular materials, especially the impact of selection strategy and structural mutations on the results of HTVS was explored. The energy gap was chosen as the kernel property to be optimized. The initial compounds library (DA substrates) was generated by combinatorial enumeration of fragments; 10 mutations was used; the Random Forest Regressor was adopt as the ML algorithm to be learned; a 10% ratio was set to randomly pick molecules from library to form the training set, and their geometries and electronic properties was computed by Gaussian; the molecular structure was featurized by the ECFP fingerprint method; by searching hyper-parameter space with 5-fold cross validation, along with the computed property, the training data was fed to the ML algorithm to obtain the best ML model; then the best ML model was used to predict unseen molecules in library. We have found that the mix of selection and mutations into the evolution map can have great impact on the HTVS results:

(1) Except the fast mutation **Sub2**, all of the rest mutations can effectively concentrate 'good' molecules in compound library, hence give large ω_{MA} (typically >0.8) for mutation generation at high values ($n_g \geq 6$).

(2) Analysis of the n_{aCH} of different mutations tells us that the uniform drop in n_{aCH} with increase of n_g is not a sufficient condition to guarantee a meaningful increase in ω_{MA}, rather it can be used a an indicator to differentiate different mutations on skeleton transformation effect.

(3) The $n_{acc\_opt\_mols}$ for any of mutations can exhibit a sharp increase in low to middle n_g values ($0 < n_g \leq 5$), follow by slower growth for middle to high n_g ($6 \leq n_g \leq 9$), and may finally trend to flatten out. **Sub1** is more favorable than **Sub2** in producing optimal molecules. For the 4 terminal single mutations, the precedence order is: **Sub3** < **Sub5** < **Sub4** ≈ **Sub6**; for the rest 4 mixed mutations, the precedence order is: **Sub8** < **Sub10** < **Sub9** < **Sub7**. The mixed mutations would produce more optimal molecules as expected in price of significant increase in molecular complexity, which might eventually prohibit them as material due to difficulty from experimental synthesis.

(4) The $\overline{\Delta E_{ST}}$ can exhibit a fast convergent trend toward very low values, hence the studied mutations (except **Sub2**) can cooperate very well with the DA substrates to generate optimal molecules.

(5) A group fingerprint similarity (Δ_{MSPR}) index was proposed to account for the similarity between two compound libraries with comparable sizes. The Δ_{MSPR} can retain high enough values (typically larger than 0.90) for large n_g, which can be associated with the apparent convergence in molecular skeletons at high mutation generations.

(6) The distribution of frequencies of different skeletons for a specific mutation is generally uneven with one dominant skeleton. The overall numbers of common and generic cores for all mutations are 11 and 7 as $n_g = 9$. Hence, in a sense, the 'optimal' skeletons seem unique and useful in realizing low energy gaps.

With above observations and the development of HTVS software, we expect to provide insight and tool to the research community of HTVS of molecular (TADF) materials.

Author contributions

C. Tu contributed to the conception of the study, designed the computational code, conducted the theoretical studies and data analysis, and drafted the manuscript; W. Huang contributed to the design of compound library; S. Liang contributed to the programming of the Python code; K. Wang and Q. Tian contributed to the conceptual and functional design of the computational code; W. Yan contributed to the conception of the study, revised the manuscript and supervised the whole work.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

C. Tu thanks the financial support of Guizhou Youth Science and Technology Talents Project (KY[2020]086), and the fund of Guiyang University (GYU-KY-2022); W. Huang thanks the financial support of Guizhou Youth Science and Technology Talents Project (KY[2020]085); and W. Yan thanks the financial support of Natural Science Foundation of China (52063005), Science and Technology Support Project of Guizhou Province (2021488), Outstanding Young Science and Technology Talent Project of Guizhou Province (20215622), Innovative talent

project of Guizhou Province (202004) and the Innovation Group Project of Guizhou Provincial Department of Education (2020024).

Notes and references

1 C. W. Tang and S. A. VanSlyke, *Appl. Phys. Lett.*, 1987, **51**, 913–915.
2 C. Adachi, *Jpn. J. Appl. Phys.*, 2014, **53**, 060101.
3 Y. Im, M. Kim, Y. J. Cho, J.-A. Seo, K. S. Yook and J. Y. Lee, *Chem. Mater.*, 2017, **29**, 1946–1963.
4 *Highly Efficient OLEDs: Materials Based on Thermally Activated Delayed Fluorescence*, ed. H. Yersin, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2018.
5 H. Uoyama, K. Goushi, K. Shizu, H. Nomura and C. Adachi, *Nature*, 2012, **492**, 234–238.
6 Q. Zhang, B. Li, S. Huang, H. Nomura, H. Tanaka and C. Adachi, *Nat. Photonics*, 2014, **8**, 326–332.
7 S. Hirata, Y. Sakai, K. Masui, H. Tanaka, S. Y. Lee, H. Nomura, N. Nakamura, M. Yasumatsu, H. Nakanotani, Q. Zhang, K. Shizu, H. Miyazaki and C. Adachi, *Nat. Mater.*, 2015, **14**, 330–336.
8 Q. Zhang, D. Tsang, H. Kuwabara, Y. Hatae, B. Li, T. Takahashi, S. Y. Lee, T. Yasuda and C. Adachi, *Adv. Mater.*, 2015, **27**, 2096–2100.
9 L.-S. Cui, H. Nomura, Y. Geng, J. U. Kim, H. Nakanotani and C. Adachi, *Angew. Chem., Int. Ed.*, 2017, **56**, 1571–1575.
10 T. J. Penfold, *J. Phys. Chem. C*, 2015, **119**, 13535–13544.
11 M. K. Etherington, J. Gibson, H. F. Higginbotham, T. J. Penfold and A. P. Monkman, *Nat. Commun.*, 2016, **7**, 1–7.
12 M. K. Etherington, F. Franchello, J. Gibson, T. Northey, J. Santos, J. S. Ward, H. F. Higginbotham, P. Data, A. Kurowska, P. L. Dos Santos, D. R. Graves, A. S. Batsanov, F. B. Dias, M. R. Bryce, T. J. Penfold and A. P. Monkman, *Nat. Commun.*, 2017, **8**, 1–11.
13 J. Föller, M. Kleinschmidt and C. M. Marian, *Inorg. Chem.*, 2016, **55**, 7508–7516.
14 J. Gibson, A. P. Monkman and T. J. Penfold, *ChemPhysChem*, 2016, **17**, 2956–2961.
15 R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams and A. Aspuru-Guzik, *Nat. Mater.*, 2016, **15**, 1120–1127.
16 Q. Peng, D. Fan, R. Duan, Y. Yi, Y. Niu, D. Wang and Z. Shuai, *J. Phys. Chem. C*, 2017, **121**, 13448–13456.
17 P. K. Samanta, D. Kim, V. Coropceanu and J.-L. Brédas, *J. Am. Chem. Soc.*, 2017, **139**, 4042–4051.
18 J.-M. Mewes, *Phys. Chem. Chem. Phys.*, 2018, **20**, 12454–12469.
19 Y. Olivier, J.-C. Sancho-Garcia, L. Muccioli, G. D'Avino and D. Beljonne, *J. Phys. Chem. Lett.*, 2018, **9**, 6149–6163.
20 T. J. Penfold, F. Dias and A. P. Monkman, *Chem. Commun.*, 2018, **54**, 3926–3935.
21 Y.-J. Gao, W.-K. Chen, Z.-R. Wang, W.-H. Fang and G. Cui, *Phys. Chem. Chem. Phys.*, 2018, **20**, 24955–24967.
22 P. de Silva, C. A. Kim, T. Zhu and T. Van Voorhis, *Chem. Mater.*, 2019, **31**, 6995–7006.
23 I. Kim, S. O. Jeon, D. Jeong, H. Choi, W.-J. Son, D. Kim, Y. M. Rhee and H. S. Lee, *J. Chem. Theory Comput.*, 2020, **16**, 621–632.
24 M. Z. Shafikov, A. F. Suleymanova, R. Czerwieniec and H. Yersin, *Chem. Mater.*, 2017, **29**, 1708–1715.
25 H. Yersin, L. Mataranga-Popa, R. Czerwieniec and Y. Dovbii, *Chem. Mater.*, 2019, **31**, 6110–6116.
26 R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, 1989.
27 E. Runge and E. K. U. Gross, *Phys. Rev. Lett.*, 1984, **52**, 997–1000.
28 H. L. Morgan, *J. Chem. Doc.*, 1965, **5**, 107–113.
29 D. Weininger, *J. Chem. Inf. Model.*, 1988, **28**, 31–36.
30 G. W. Bemis and M. A. Murcko, *J. Med. Chem.*, 1996, **39**, 2887–2893.
31 X. Q. Lewell, D. B. Judd, S. P. Watson and M. M. Hann, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 511–522.
32 J. Degen, C. Wegscheid-Gerlach, A. Zaliani and M. Rarey, *ChemMedChem*, 2008, **3**, 1503–1507.
33 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
34 G. Landrum, *The RDKit Documentation—The RDKit 2020.09.1 documentation*, 2020, **http://www.rdkit.org/docs/index.html**.
35 S. Riniker and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 2562–2574.
36 H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *J. Cheminf.*, 2018, **10**, 4.
37 L. Turcani, E. Berardo and K. E. Jelfs, *J. Comput. Chem.*, 2018, **39**, 1931–1942.
38 E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre and A. Aspuru-Guzik, *Annu. Rev. Mater. Res.*, 2015, **45**, 195–216.
39 L. Wilbraham, E. Berardo, L. Turcani, K. E. Jelfs and M. A. Zwijnenburg, *J. Chem. Inf. Model.*, 2018, **58**, 2450–2459.
40 X.-Y. Ma, J. P. Lewis, Q.-B. Yan and G. Su, *J. Phys. Chem. Lett.*, 2019, **10**, 6734–6740.
41 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos and D. Cournapeau, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
42 T. Amr, *Hands-On Machine Learning with Scikit-learn and Scientific Python Toolkits*, Packt Publishing Ltd, Birmingham, 2020.
43 M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu and X. Zheng, *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, Savannah, GA, 2016, pp. 265–283.
44 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison,

A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, *Adv. Neural Inf. Process. Syst.*, 2019, 8024–8035.
45 Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, *Chem. Sci.*, 2018, **9**, 513–530.
46 B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing and Z. Wu, *Deep Learning for the Life Sciences*, O'Reilly Media, 2019.
47 P. O. Dral, *J. Comput. Chem.*, 2019, **40**, 2339–2347.
48 R. Jacobs, T. Mayeshiba, B. Afflerbach, L. Miles, M. Williams, M. Turner, R. Finkel and D. Morgan, *Comput. Mater. Sci.*, 2020, **176**, 109544.
49 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.
50 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.
51 J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller and K. Burke, *Phys. Rev. Lett.*, 2012, **108**, 253002.
52 R. Ramakrishnan and O. A. von Lilienfeld, *Reviews in Computational Chemistry*, John Wiley & Sons, Inc., 2017, pp. 225–256.
53 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2015, **11**, 2087–2096.
54 O. A. v. Lilienfeld, K.-R. Müller and A. Tkatchenko, *Nat. Rev. Chem.*, 2020, **4**, 347–358.
55 B. Huang and O. A. von Lilienfeld, *Nat. Chem.*, 2020, **12**, 945–951.
56 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.
57 S. Ye, W. Hu, X. Li, J. Zhang, K. Zhong, G. Zhang, Y. Luo, S. Mukamel and J. Jiang, *Proc. Natl. Acad. Sci.*, 2019, **116**, 11612–11617.
58 S. Ma and Z.-P. Liu, *ACS Catal.*, 2020, **10**, 13213–13226.
59 W.-K. Chen, X.-Y. Liu, W.-H. Fang, P. O. Dral and G. Cui, *J. Phys. Chem. Lett.*, 2018, **9**, 6702–6708.
60 D. P. Kingma and M. Welling, Auto-Encoding Variational Bayes, *arXiv*, 2014, preprint, arXiv:1312.6114.
61 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. 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.
62 J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*, MIT press, 1992.
63 J. J. Stewart, *J. Mol. Model.*, 2007, **13**, 1173–1213.
64 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
65 J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615–6620.
66 *Calculate Root-mean-square deviation (RMSD) of Two Molecules Using Rotation*, version 1.4, **https://github.com/charnley/rmsd**.
67 W. Kabsch, *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.*, 1976, **32**, 922–923.
68 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
69 W. P. Walters and M. A. Murcko, *Adv. Drug Delivery Rev.*, 2002, **54**, 255–271.
70 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 1–14.
71 *Anaconda Software Distribution*, 2022, **https://docs.anaconda.com/**.
72 P. Raybaut, *Spyder-documentation*, 2009, Available online at: **https://pythonhosted.org**.
73 T. E. Oliphant, *A guide to NumPy*, Trelgol Publishing, USA, 2006, vol. 1.
74 W. McKinney, *Python for high performance and scientific computing*, 2011, 14, pp. 1–9.
75 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, *Nat. Methods*, 2020, **17**, 261–272.
76 J. D. Hunter, *Comput. Sci. Eng.*, 2007, **9**, 90–95.
77 C. Tu, *SALAM: an HTVS tool for organic materials*, 2022, **https://github.com/yidapa/salam**.
78 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, *Gaussian 16 Revision C.01*, Gaussian Inc, 2019.