Supplementary Information: Machine Learning-enabled Globally Guaranteed Evolutionary Computation

Bin Li[1,3], Ziping Wei[1], Jingjing Wu[1], Shuai Yu[2], Tian Zhang[2], Chunli Zhu[3]
Dezhi Zheng[3], Weisi Guo[4,5], Chenglin Zhao[1], Jun Zhang[3]

I. Structured Random Sampling

We take the 2D problem space to explain how to perform the random sampling in the first stage; the extension to d-dimensional problems is straightforward. In the case of the optimal sampling, s rows and columns should be selected according to the leveraging scores [1]–[4]. In other words, the probabilities of sampling one of M rows and N columns are given by:

$$\Pr\{r_j = m\} = ||\mathbf{U}(m,:)||_F^2 \big/ ||\mathbf{U}||_F^2, \quad \text{for } r_j \in \mathcal{R};$$
$$\Pr\{c_i = n\} = ||\mathbf{V}(:,n)||_F^2 \big/ ||\mathbf{V}||_F^2, \quad \text{for } r_i \in \mathcal{C},$$

where $\mathbf{F} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ is the singular value decomposition (SVD) of the original problem space; $\mathbf{U}(m,:)$ denotes the m-th row of a left singular matrix $\mathbf{U}$ and $\mathbf{V}(:,n)$ denotes the n-th column of a right singular matrix $\mathbf{V}$. Although this leveraging score-based sampling strategy could acquire the best reconstruction performance, however, with only the sampled versions of $\mathbf{F}$ (i.e. $\mathbf{C}$ and $\mathbf{R}$), we are not able to compute the leveraging scores.

1. School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, 100876, China.
2. State Key Laboratory of Information Photonics and Optical Communications, BUPT, Beijing 100876, China.
3. School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081, China.
4. Centre for Autonomous and Cyberphysical Systems, Cranfield University, MK43 0AL, UK
5. The Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK

For this reason, we adopt a refined sampling strategy. To be specific, we firstly sample s columns according to the uniform distribution. On this basis, we approximately compute the sampling score of M rows, which is defined as the row variance of a sampled sub-matrix $\mathbf{C}$. Then, we can sample s rows according to the probabilities

$$\Pr\{c_i = n\} = 1/N, \quad \Pr\{r_j = m\} = p_j \Big/ \sum\nolimits_j p_j,$$

whereby

$$p_j = \sum\nolimits_{c_i \in \mathcal{C}} \left\{ f(j, c_i) - \frac{1}{s} \sum\nolimits_{c_i} f(j, c_i) \right\}^2.$$

In practice, the computation of this sampling scores can be iteratively implemented. For example, we may firstly sample few columns and then compute the scores of each rows. In the next round, we compute the sampling scores of each columns, relying on the already sampled rows; and thus update the scores of remaining rows. As demonstrated by empirical results, this method obtains the high reconstruction accuracy.

A. Online Estimation of the Rank

Despite the general low rank structures (for example, see SI Fig. E1 for benchmark functions), the *a prior* knowledge on the rank of discrete problem space $\mathbf{F}$ may be unavailable. Fortunately, during the structured random sampling stage, we can estimate the rank value $r = \text{rank}(\mathbf{F})$ in an online mode. As noted, the singular values of a reconstructed representation matrix $\hat{\mathbf{F}}$ would accurately approximate the singular values of the original space $\mathbf{F}$, when the sampling length s is sufficiently large. Inspired by this, we propose one sample-efficient method for the online estimation of rank, which is critical to the low-rank representation learning stage (see SI Fig. E2).

At the beginning, we set a small sampling length s_0, and obtain the representation space $\hat{\mathbf{F}} \in \mathbb{R}^{M \times N}$. Subsequently, in each iteration we increase the sampling length to $s_{i+1} = s_i + 1$; and accordingly, we obtain the new singular values, i.e., $\sigma^2_{s_{i+1}} \leq \sigma^2_{s_i}$. For the strictly rank-restricted problems, we obtain the estimation of unknown rank, i.e., $\hat{r} = s_i$, once the condition $\sigma^2_{s_{i+1}} \ll \sigma^2_{s_i}$ was satisfied. For the approximate rank-restriction problems, we estimate the rank via the residual error, which is defined as $\sum_{s_i+1}^{N} \sigma^2_{s_i} < (N - s_i)\sigma^2_{s_i}$. Once the condition $\sum_{j=1}^{s_i} \sigma_j^2 / [\sum_{j=1}^{s_i} \sigma_j^2 + (N - s_i)\sigma^2_{s_{i+1}}] >$0.95 was satisfied, we obtain $\hat{r} = s_i$; see an illustration in SI Fig. E3.

II. Theoretical Probability of Finding the Global Optimum

In this section, we provide the theoretical analysis on the lower bound probability of EVOLER in finding the globally optimum. To sum up, our following analysis consists of 4 steps. Here, we take the 2D problems for example; see the extension to d-dimensional case in SI section III.

Firstly, we show that the residual error, between the original discrete problem space $\mathbf{F} \in \mathbb{R}^{M\times N}$ and its low rank representation $\hat{\mathbf{F}} \in \mathbb{R}^{M\times N}$, would follow a non-Gaussian distribution which is well described by the long-tail α-stable distribution.

Secondly, we study the probability of one critical case, whereby the next local solution $\mathbf{x}^{\text{local}} \in \mathbf{X}$ is mistaken as the initial solution $\hat{\mathbf{x}}^{\text{opt}} = \arg\min_{\mathbf{x}\in\mathbf{X}} \hat{f}(\mathbf{x})$ in a representation space $\hat{\mathbf{F}}$; $\mathbf{X} \in \mathbb{R}^{M\times N}$ is the discrete grid of continuous variable $\mathbf{x} \in \mathbb{R}^2$; $f(\mathbf{x}) : \mathbb{R}^2 \mapsto \mathbb{R}$ denotes a fitness function.

Thirdly, we give the probability of an initial solution $\hat{\mathbf{x}}^{\text{opt}}$ falls into the basin of the global optimum $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x})$.

Finally, we show that, after the limited T generations, EVOLER converges to the global optimum with probability $\Pr\{\mathbf{x}^T \rightarrow \mathbf{x}^*\}$. In particular, the probability of finding the global optimum would approach 1, i.e., $\Pr\{\mathbf{x}^T \rightarrow \mathbf{x}^*\} \rightarrow 1$, when the problem has the general low-rank structure, as is often the case of real-world applications.

In this following, we give the detail analysis on each step.

A. *Statistical Distribution of Residual Error*

For the general rank-restricted problems, we can firstly reconstruct one low rank representation space $\hat{\mathbf{F}} \in \mathbb{R}^{M\times N}$; M and N are the grid sizes. On this basis, we show that the residual error, i.e., $e_{m,n} = \{\hat{\mathbf{F}} - \mathbf{F}\}$, will theoretically follow one non-Gaussian probability density; see the following Theorem 1.

Theorem 1. *Denote the rank-restricted original problem space as $\boldsymbol{F} \in \mathbb{R}^{M\times N}$. We sample s rows and columns to abstract two submatrics $\boldsymbol{C} = \boldsymbol{F}(\mathcal{C}, :) \in \mathbb{R}^{M\times s}$ and $\boldsymbol{R} = \boldsymbol{F}(:, \mathcal{R}) \in \mathbb{R}^{s\times N}$, $\mathcal{C}$ and $\mathcal{R}$ are two sampling index sets; we thus reconstruct a representation space $\hat{\boldsymbol{F}} = \boldsymbol{CWR} \in \mathbb{R}^{M\times N}$, whereby $\boldsymbol{W} = [\boldsymbol{F}(\mathcal{C}, \mathcal{R})]^\dagger$. The residual error matrix is denoted as $\boldsymbol{\Xi} = \boldsymbol{F} - \hat{\boldsymbol{F}} \in \mathbb{R}^{M\times N}$. Then, the element $e_{m,n}$ of $\boldsymbol{\Xi}$ statistically follows the long-tail α-stable distribution.*

Proof: Before going a further step, we firstly give an important result on the multivariate t-distribution [5], [6], which is closely related with our following analysis.

Lemma 1. *Denote one covariance matrix* $\boldsymbol{S} \in \mathcal{R}^{n\times n}$*, which follows the Wishart distribution. Provided one* $n \times 1$ *random vector* $\boldsymbol{r} \sim \mathcal{N}(0, Q\boldsymbol{I}_n)$ *independent of* $\boldsymbol{S}$*, then the multi-variate variable* $\boldsymbol{t} \triangleq (\boldsymbol{S}^{1/2^{'}})^{-1}\boldsymbol{r}$ *follows the* t*-distribution.*

Based on the Lamma 1, we will show that the elements in each vector $\mathbf{y}_j \triangleq \mathbf{WR}(:, j) \in \mathbb{R}^{s\times 1}$ $(j = 1, 2, \cdots, N)$ would be the multivariate t-distributed variables.

As noted, after randomly sampling s rows and columns from $\mathbf{F}$, the intersection matrix $\mathbf{F}_s \triangleq \mathbf{F}(\mathcal{C}, \mathcal{R}) \in \mathbb{R}^{s\times s}$ consists of s independent $s \times 1$ vectors. For simplicity, we assume the elements of $\mathbf{F}_s$ follow the Gaussian distribution (as seen latterly, our following analysis holds also for the non-Gaussian cases). So, we can define one covariance matrix as:

$$\mathbf{S} \triangleq \mathbf{F}_s^T\mathbf{F}_s \in \mathbb{R}^{s\times s},$$

which is a semi-positive symmetric definite (SPSD) matrix, or in the literature refereed as the Wishart matrix. $\mathbf{X}^T$ denotes the transpose matrix of $\mathbf{X}$.

Similarly, we assume each row of the sampled sub-matrix $\mathbf{R}$, i.e., $\mathbf{r} \triangleq \mathbf{F}(j, :) \in \mathbb{R}^{s\times 1}$, to be Gaussian distributed. Furthermore, since the involved s samples are randomly selected and we have $s \ll \min\{M, N\}$, we reasonably assume $\mathbf{r}$ is independent and identically distributed (i.i.d). That is to say, its covariance matrix is given by $\mathbf{r}\mathbf{r}^T = Q\mathbf{I}_s$; $\mathbf{I}_s$ is the $s \times s$ dimensional identity matrix. In addition, it is easily noted that the weight matrix $\mathbf{W} \triangleq \mathbf{F}_s^{-1} \in \mathbb{R}^{s\times s}$ [1]–[3], [7], after the pseudo-inverse operation, would be independent of $\mathbf{r} \in \mathbb{R}^{s\times 1}$.

As a consequence, we conclude that the multi-variate variable

$$\mathbf{y}_j \triangleq \mathbf{WR}(:, j) = \left(\mathbf{S}^{1/2^{'}}\right)^{-1}\mathbf{r} \in \mathbb{R}^{s\times 1},$$

would follow the t-distribution, by applying the above Lemma 1. As noted, there are two necessary conditions for this result, i.e., (i) the independence between $\mathbf{W}$ and $\mathbf{r}$, and (ii) the i.i.d Gaussian distributed $\mathbf{r}$. In some practical situations (e.g. benchmark functions), the random samples of $\mathbf{r}$ may not be Gaussian distributed. Fortunately, our empirical results show that the above t-distribution still holds; see SI Fig. E10 for the results on 14 benchmark functions.

On this basis, the residual error can be reformatted as:

$$e_{m,n} = \mathbf{C}(m,:)\mathbf{WR}(:,n) = \sum_{j=1}^{s} \mathbf{C}(m,j)\mathbf{y}_n(j),$$
$$\triangleq \sum_{j=1}^{s} z_j, \quad j = 1, 2, \cdots, s.$$

For clarity, here we define a new variable as $z_j \triangleq \mathbf{C}(m,j)\mathbf{y}_n(j)$. As the linear combination of multivariate t-distributed variables $\mathbf{y}_n(j)$, the new variable z_j also follows the t-distribution [8].

It is noteworthy that the variance of t-distribution variables, i.e., z_j, would become *infinite*; so that the celebrated Central Limit Theorem (CLT) becomes also inapplicable. Still, we may resort to the Generalized Central Limit Theorem (GCLT) to derive the probability density function (PDF) of the residual errors $\{e_{m,n}\}$.

Before applying the GCLT, we need to show the residual errors $e_{m,n} = \sum_{j=1}^{s} z_j$ meet the following two necessary properties. (1) *Stable* property: As discussed, the linear combination of multivariate t-distributed variables $\mathbf{y}_n$ still follows the t-distribution [9], i.e., $\sum_{j=1}^{s} z_j = \sum_{j=1}^{s} \mathbf{C}(m,j)\mathbf{y}_n(j)$ is also t-distributed. (2) *Limit* property: We easily note that, for $y \to \infty$, the PDF of the t-distributed variables $\mathbf{y}_n(j)$ should satisfy:

$$y^{\alpha} \times \Pr\{Y > y\} \to c^{+}, \ |y|^{\alpha} \times \Pr\{Y < -y\} \to c^{-},$$
$$c^{+} + c^{-} > 0.$$

Based on the above two properties, we are allowed to apply the GCLT [10]; and therefore, we conclude the residual error $z \triangleq e_{m,n}$ follows the α-stable distribution, i.e., $p(z) = S_{\alpha}(\beta, \gamma, \delta)$, whose characteristic function (i.e., the Fourier transform of its PDF) is given by [11]:

$$\mathbb{E}\{\exp(juZ)\} = \int_z \exp(juz)p(z)dz$$
$$= \exp\left\{-\gamma^{a}|u|^{a}\left[1 - \beta\tan\left(\frac{\pi\alpha}{2}\right)\text{sign}(u)(|\gamma u|^{1-\alpha} - 1)\right] + j\delta u\right\},$$

for $\alpha \neq 1$; and

$$\mathbb{E}\{\exp(juZ)\} = \exp\left\{-\gamma|u|\left[1 - \frac{2\beta}{\pi}\text{sign}(u)\log(\gamma|u|)\right] + j\delta u\right\},$$

for $\alpha = 1$. In the above, $\alpha \in (0, 2]$ is the *characteristic exponent*; $\beta \in [-1, 1]$ is the *skewness* parameter; γ is the *scale* parameter and δ is the *location* parameter [11]. Note that, in most cases the PDF of α-stable variables, i.e., $p(z) = \int_u \cos(uz)\mathbb{E}\{\exp(juz)\}du$ (the inverse Fourier transform of the above characteristic function), has no closed-form expression. Only in some special cases, can the closed-form of this PDF be derived. For example, (i) when $\alpha = 2$ and $\beta = 0$, the α-stable density becomes the Gaussian distribution; (ii) when $\alpha = 1$ and $\beta = 0$, the α-stable density becomes the Cauchy distribution.

Despite the non-analytic complex expression, in practice the α-stable PDF can be easily computed, which is shown to well fit the empirical data of residual errors; see SI Fig. E11 for the strictly low-rank problem (e.g. Schwefel function, $d = 2$) and the approximately low-rank problem (e.g. Rotated Ackley function, $d = 2$). For more details on the parameters estimation of the α-stable PDF, see ref. [12], [13].

B. Probability of mistaking the global optimum as the next local optimum

After finding an initial solution $\hat{\mathbf{x}}^{\text{opt}}$ in the low-rank representation space $\hat{\mathbf{F}}$, the particles are placed around the identified attention subspace $\mathcal{A} \triangleq \{|\mathbf{x} - \hat{\mathbf{x}}^{\text{opt}}| < \delta|\}$. If there was no residual error (e.g., $e_{m,n} = 0$), the initial solution $\hat{\mathbf{x}}^{\text{opt}} = \arg\min_{\mathbf{x}\in\mathbf{X}} \hat{f}(\mathbf{x})$ should equal to the optimal solution $\mathbf{x}^{\text{opt}} = \arg\min_{\mathbf{x}\in\mathbf{X}} f(\mathbf{x})$ in the original space $\mathbf{F}$. Given the non-zero residual errors, we need to characterize the probability of mistaking the next local solution $\mathbf{x}^{\text{local}} \in \mathbf{X}$ as an initial solution $\hat{\mathbf{x}}^{\text{opt}} \in \mathbf{X}$, which is critical to the following analysis. That is because, once we have obtained an deviated estimation $\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}$, then the attention subspace $\mathcal{A}$ would likely reject the global optimum $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, finally leading to the non-global solutions.

Theorem 2. *The problem space $\boldsymbol{F} \in \mathbb{R}^{N\times N}$ is rank-restricted (i.e., $M = N$). We sample $s \sim \mathcal{O}(rlog(\frac{r}{\epsilon}))$ rows / columns to abstract two submarines $\boldsymbol{C} = \boldsymbol{F}(\mathcal{C}, :) \in \mathbb{R}^{M\times s}$ and $\boldsymbol{R} = \boldsymbol{F}(:, \mathcal{R}) \in \mathbb{R}^{s\times N}$; we thus reconstruct a representation space $\hat{\boldsymbol{F}} = \boldsymbol{CWR} \in \mathbb{R}^{M\times N}$, whereby $\boldsymbol{W} = [\boldsymbol{F}(\mathcal{C}, \mathcal{R})]^{\dagger}$. The initial solution in an representation space $\hat{\boldsymbol{F}}$ is $\hat{\boldsymbol{x}}^{opt} = \arg\min_{\boldsymbol{x}\in\boldsymbol{X}} \hat{f}(\boldsymbol{x})$, while the next local solution of $\boldsymbol{F}$ is $\boldsymbol{x}^{local} \in \boldsymbol{X}$. Then, the probability $Pr\{\hat{\boldsymbol{x}}^{opt} = \boldsymbol{x}^{local}\}$ is given by:*

$$\Pr\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}\} < 1 - \frac{2}{\pi}\arctan\frac{\Delta f}{\frac{1}{N}\sqrt{\| \mathbf{F} - \hat{\mathbf{F}} \|_F^2}}. \tag{1}$$

where $\Delta f \triangleq |f(\boldsymbol{x}^{global}) - f(\boldsymbol{x}^{local})|$ is the distance between the global optimum and the next local

optimum; $\frac{1}{N^2} \parallel \boldsymbol{F} - \hat{\boldsymbol{F}} \parallel_F^2$ *is the averaged residual error of each element in* $\hat{\boldsymbol{F}}$*;* $\parallel \cdot \parallel_F^2$ *denotes the Frobenius norm.*

Proof: As noted, the statistical analysis on α-stable distribution would become intractable, as it involves the complex non-analytic PDF and many unknown parameters. In order to attain the simple analytical result, we firstly adopt one bi-parameter Cauchy-Gaussian mixture (CGM) distribution to approximate the α-stable density [14], [15], i.e.,

$$\begin{aligned} p(z) &\simeq (1-\lambda) \times p_G(z) + \lambda \times p_C(z) \\ &= (1-\lambda)\frac{1}{2\sigma\sqrt{\pi}} \exp\left(-\frac{z^2}{4\sigma^2}\right) + \lambda \frac{\sigma}{\pi(z^2+\sigma^2)}, \end{aligned} \tag{2}$$

where $\lambda \in (0, 1]$ is a weighting coefficient; $\sigma > 0$ denotes an equivalent variance of the Gaussian distribution and the Cauchy distribution. As demonstrated, this CGM distribution is capable of approximating the α-stable density with arbitrary accuracy [16], [17]; see SI Fig. E12 for the approximation result, taking a rotated Ackley function for example. On this basis, we are able to derive the probability of mistaking the next local solution as the initial solution, in which case the optimal solution in $\mathbf{F}$ may be initially rejected from an attention subspace.

Due to the non-zero residual error between a representation space $\hat{\mathbf{F}}$ and the original space $\mathbf{F}$, the initial solution $\hat{\mathbf{x}}^{\text{opt}}$ in $\hat{\mathbf{F}}$ may be different from the optimal solution $\mathbf{x}^{\text{opt}}$ in $\mathbf{F}$. In the following, we consider the probability of mistaking the next local optimum as the initial solution, i.e., $\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}$. In this situation, the estimated next local optimum $\hat{f}(\mathbf{x}^{\text{local}})$ would be smaller than the optimum $f(\mathbf{x}^{\text{opt}})$, or the estimated optimum $\hat{f}(\mathbf{x}^{\text{opt}})$ would be larger than the next local optimum $f(\mathbf{x}^{\text{local}})$. Provided the statistical PDF of the residual errors, i.e., $p(z)$, the probability $\Pr\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}\}$ is given by:

$$\begin{aligned} \Pr\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}\} &= \Pr\{\hat{f}(\mathbf{x}^{\text{local}}) < f(\mathbf{x}^{\text{opt}})\} + \Pr\{\hat{f}(\mathbf{x}^{\text{opt}}) > f(\mathbf{x}^{\text{local}})\} \\ &= 2 \times \Pr\{z > \Delta f\} = 2 \times \int_{\Delta f}^{\infty} p(z) dz. \end{aligned}$$

So, the probability of rejecting the local optimum when estimating the initial solution is:

$$\begin{aligned} \Pr\{\hat{\mathbf{x}}^{\text{opt}} \neq \mathbf{x}^{\text{local}}\} &= 1 - \Pr\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}\} \\ &= 1 - 2 \times \int_{\Delta f}^{\infty} p(z) dz. \end{aligned}$$

This is also illustrated by SI Fig. E13. I.e., the estimation error may occur when: (1) the estimated next local optimum is smaller than the original global optimum, or (2) the estimated global optimum is larger than the original next local optimum. After substituting the PDF of a Gaussian-Cauchy mixture, the following result is directly obtained

$$\begin{aligned}\Pr\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}\} &= 2\left\{\frac{1-\lambda}{2}\text{erfc}\left(\frac{\Delta f}{2\sigma}\right) + \lambda\left[\frac{1}{2} - \frac{1}{\pi}\arctan\frac{\Delta f}{\sigma}\right]\right\}, \\ &\overset{\text{(a)}}{\simeq} \lambda\left[1 - \frac{2}{\pi}\arctan\frac{\Delta f}{\sigma}\right],\end{aligned}$$

where $\text{erfc}(x) \triangleq \int_x^{\infty} \frac{2}{\sqrt{\pi}}\exp(-t^2)dt$ denotes the error complementary function (ECF). Note that, the above relation (a) holds for the rank-restricted problems. I.e., provided a large sampling length s, we may have a small variance σ, and thus $\text{erfc}\left(\frac{\Delta f}{2\sigma}\right) \to 0$, for a relatively large Δf. Another consideration is that, as an approximation of the long-tail α-stable distribution, the coefficient of the rapidly decaying Gaussian component would be very small, i.e., $(1-\lambda) \to 0$ (see SI Fig. E12, for example $\lambda = 0.999$).

Furthermore, we note that the parameter σ in Eq. (2) represents a variance of the Gaussian and the Cauchy distribution; so it is supposed σ should be smaller than the whole variance of residual errors, i.e.,

$$\begin{aligned}\sigma^2 &\leq \mathbb{E}\left\{\left[f - \hat{f} - \mathbb{E}(f - \hat{f})\right]^2\right\} \\ &\simeq \frac{1}{N^2}||\mathbf{F} - \hat{\mathbf{F}}||_F^2.\end{aligned}$$

Here, we also assume $M = N$ for clarity. Combined the above relation, then the probability $\Pr\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}\}$ would become

$$\Pr\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}\} < \lambda\left[1 - \frac{2}{\pi}\arctan\frac{\Delta f}{\frac{1}{N}\sqrt{\| \mathbf{F} - \hat{\mathbf{F}} \|_F^2}}\right].$$

Given the weight coefficient $0 < \lambda \leq 1$, the above probability is finally simplified to:

$$\Pr\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}\} < 1 - \frac{2}{\pi}\arctan\frac{\Delta f}{\frac{1}{N}\sqrt{\| \mathbf{F} - \hat{\mathbf{F}} \|_F^2}}.$$

C. *Probability of the initial solution falling into the basin of global solution*

Before proceeding, we define the attraction basin of the global solution $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x})$, which is denoted as $\mathcal{B}^{\text{global}}$:

$$\mathcal{B}^{\text{global}} \triangleq \{\mathbf{x} \in \mathbb{R}^2 : \ |\mathbf{x} - \mathbf{x}^*| < \theta, \ f(\mathbf{x}) \text{ is unimodal}\}. \tag{3}$$

Here, $\theta \in \mathbb{R}^+$ is the radius of the global basin $\mathcal{B}^{\text{global}}$. It is easily seen that there exist a small basin radius θ, so that the fitness function $f(\mathbf{x})$ is unimodal in $\mathcal{B}^{\text{global}}$; see also the illustration in SI Fig. E13.

Once the estimated initial solution in a representation space $\hat{\mathbf{F}}$ falls into the attraction basin of the global optimum, i.e., $\hat{\mathbf{x}}^{\text{opt}} \in \mathcal{B}^{\text{global}}$, then the global optimum can be reliably identified, by applying classical evolutionary computation methods, as seen lately. That is to say, by initializing the particles around the basin of global solution, the local optimum can be reliably rejected while the global optimum will become achievable. In the following, we will give the probability of the initial solution falls into the basin of the global solution, see the following Theorem 3.

Theorem 3. *The rank-restricted original space is denoted as $\boldsymbol{F} \in \mathbb{R}^{N\times N}$. We sample $s \sim \mathcal{O}(rlog(\frac{r}{\epsilon}))$ rows and columns to abstract two submarines $\boldsymbol{C} = \boldsymbol{F}(\mathcal{C}, :) \in \mathbb{R}^{M\times s}$ and $\boldsymbol{R} = \boldsymbol{F}(:, \mathcal{R}) \in \mathbb{R}^{s\times N}$; we thus reconstruct a representation space $\hat{\boldsymbol{F}} = \boldsymbol{CWR} \in \mathbb{R}^{M\times N}$, whereby $\boldsymbol{W} = [\boldsymbol{F}(\mathcal{C}, \mathcal{R})]^\dagger$. The global basin is defined as $\mathcal{B}^{global} = \{\boldsymbol{x} \in \mathbb{R}^2 : |\boldsymbol{x} - \boldsymbol{x}^*| \le \theta\}$ (θ is the basin radius). The initial solution in the representation space is $\hat{\boldsymbol{x}}^{opt} = \arg\min_{\boldsymbol{x}\in\boldsymbol{X}} \hat{f}(\boldsymbol{x})$. Then, the probability of the global basin containing this initial solution is:*

$$\Pr\left\{\hat{\mathbf{x}}^{\text{opt}} \in \mathcal{B}^{\text{global}}\right\} > (1-\epsilon)\cdot\frac{2}{\pi}\arctan\left(\frac{\Delta f}{\Delta E(s)}\right),$$

where Δf is the distance between the global optimum and the next local optimum; $\Delta E(s) \triangleq \frac{1}{N}\sqrt{(1+\eta)\sum_{j=s+1}^{N}\sigma_j^2}$ is the upper bound residual error which is related to the sampling length s. $\epsilon \in (0, 1]$ is a small constant also related to s.

The proof of Theorem 3 is straightforward. We firstly have the following relation, i.e.,

$$\begin{aligned}\Pr\left\{\hat{\mathbf{x}}^{\text{opt}} \in \mathcal{B}^{\text{global}}\right\} &= 1 - \Pr\left\{\hat{\mathbf{x}}^{\text{opt}} \notin \mathcal{B}^{\text{global}}\right\}, \\ &= 1 - \sum\nolimits_{\mathbf{x}_j \notin \mathcal{B}^{\text{global}}} \Pr\{\mathbf{x}_j\} \times \Pr\left\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}_j\right\},\end{aligned}$$

where $0 < \Pr\{\mathbf{x}_j\} < 1$ denotes the occurrence probability of $\mathbf{x}_j \notin \mathcal{B}^{\text{global}}$; and for simplicity, we

denote $p_j = \Pr\{\mathbf{x}_j\}$ and have $\sum_j p_j = 1$. $\Pr\left\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}_j\right\}$ denotes the probability of mistaking the solution $\mathbf{x}_j$ as the initial solution $\hat{\mathbf{x}}^{\text{opt}}$, and we denote $q_j = \Pr\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}_j\}$. Denote the fitness distance between the global optimum $\mathbf{x}^*$ and one feasible solution $\mathbf{x}_j$ as $\Delta f_j = |f(\mathbf{x}_j) - f(\mathbf{x}^{\text{opt}})|$. As illustrated by SI Fig. E13, any feasible solution $\mathbf{x}_j$ outside $\mathcal{B}^{\text{global}}$ should satisfy

$$\Delta f_j \geq \Delta f = |f(\mathbf{x}^{\text{local}}) - f(\mathbf{x}^{\text{opt}})|, \quad \text{for } \mathbf{x}_j \notin \mathcal{B}^{\text{global}}.$$

From Eq. (1), we note $\Pr\left\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}_j\right\}$ is decreased with an increase of Δf_j. So, we have

$$q_j \leq \Pr\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}\}.$$

Given the fact $\sum_{\mathbf{x}_j \notin \mathcal{B}^{\text{global}}} \Pr\{\mathbf{x}_j\} = 1$, we would obtain the following result:

$$\begin{aligned}
&\Pr\left\{\hat{\mathbf{x}}^{\text{opt}} \in \mathcal{B}^{\text{global}}\right\} = 1 - \sum_{\mathbf{x}_j \notin \mathcal{B}^{\text{global}}} p_j \times q_j, \\
&> 1 - \sum_{\mathbf{x}_j \notin \mathcal{B}^{\text{global}}} p_j \times \Pr\left\{\hat{\mathbf{x}}^{\text{opt}} = \mathbf{x}^{\text{local}}\right\} > 1 - \left[1 - \frac{2}{\pi}\arctan\frac{\Delta f}{\frac{1}{N}\sqrt{\| \mathbf{F} - \hat{\mathbf{F}} \|_F^2}}\right] \sum_{\mathbf{x}_j \notin \mathcal{B}^{\text{global}}} p_j, \\
&= \frac{2}{\pi}\arctan\frac{\Delta f}{\frac{1}{N}\sqrt{\| \mathbf{F} - \hat{\mathbf{F}} \|_F^2}}.
\end{aligned} \tag{4}$$

Note that, the above theoretical result can be generalized to the case of $d > 2$; see SI section III. In the case of $d = 2$, we can simplify the above result based on the following fact:

$$\frac{1}{N}\sqrt{||\mathbf{F} - \hat{\mathbf{F}}||_F^2} \overset{\text{(b)}}{\leq} \frac{1}{N}\sqrt{(1+\eta)\sum\nolimits_{s+1}^{N} \sigma_j^2} \triangleq \Delta E(s).$$

Since the above inequality (b) holds with probability $1 - \epsilon$, we finally obtain the probability of the initial solution falling into the basin of global solution, i.e.,

$$\Pr\left\{\hat{\mathbf{x}}^{\text{opt}} \in \mathcal{B}^{\text{global}}\right\} > (1 - \epsilon) \times \frac{2}{\pi}\arctan\left(\frac{\Delta f}{\Delta E(s)}\right). \tag{5}$$

D. Probability of finding the globally optimal solution

Based on the above analysis, we finally obtain the probability of finding the globally optimal solution, see the following Theorem 4.

Theorem 4. *The rank-restricted original problem space is denoted as* $\boldsymbol{F} \in \mathbb{R}^{N \times N}$ *(i.e.* $M =$

N). We randomly sample $s \sim \mathcal{O}(rlog(\frac{r}{\epsilon}))$ rows and columns to abstract two submarines $\boldsymbol{C} = \boldsymbol{F}(\mathcal{C},:) \in \mathbb{R}^{M\times s}$ and $\boldsymbol{R} = \boldsymbol{F}(:,\mathcal{R}) \in \mathbb{R}^{s\times N}$; we thus reconstruct a representation space $\hat{\boldsymbol{F}} = \boldsymbol{CWR} \in \mathbb{R}^{M\times N}$; $\boldsymbol{W} = [\boldsymbol{F}(\mathcal{C},\mathcal{R})]^{\dagger}$. The particles are initialized around the attention subspace $\mathcal{A} \triangleq \{\boldsymbol{x} \in \mathbb{R}^2 : |\boldsymbol{x} - \hat{\boldsymbol{x}}^{opt}| < \delta|\}$. After the limited T generations, the obtained final solution is denoted as $\hat{\boldsymbol{x}}^T$. Then, the probability of this final solution $\hat{\boldsymbol{x}}^T$ converging to the globally optimal solution $\boldsymbol{x}^*$ is lower bounded, i.e.,

$$\Pr\left\{\mathbf{x}^T \to \mathbf{x}^*\right\} > (1-\epsilon)\frac{2}{\pi}\arctan\left(\frac{\Delta f}{\Delta E(s)}\right). \tag{6}$$

The proof of the above Theorem 4 is straightforward. We firstly have the following relation, i.e.,

$$\begin{aligned}\Pr\left\{\mathbf{x}^T \to \mathbf{x}^*\right\} = \Pr\left\{\hat{\mathbf{x}}^{\text{opt}} \in \mathcal{B}^{\text{global}}\right\} \times \Pr\left\{\mathbf{x}^T \to \mathbf{x}^*; \mathcal{A} \cap \mathcal{B}^{\text{global}} \neq \varnothing\right\} + \\ \Pr\left\{\hat{\mathbf{x}}^{\text{opt}} \notin \mathcal{B}^{\text{global}}\right\} \times \Pr\left\{\mathbf{x}^T \to \mathbf{x}^*; \mathcal{A} \cap \mathcal{B}^{\text{global}} = \varnothing\right\}.\end{aligned}$$

Here, $\cap$ gives the intersection set and $\varnothing$ denotes the empty set. The case of $\mathcal{A} \cap \mathcal{B}^{\text{global}} \neq \varnothing$ indicates the intersection set between the identified attention space $\mathcal{A}$ and the global basin $\mathcal{B}^{\text{global}}$ is not empty; in other words, the global optimum is directly achievable from the initialized solutions. Similarly, the other case $\mathcal{A} \cap \mathcal{B}^{\text{global}} = \varnothing$ indicates the intersection set between $\mathcal{A}$ and $\mathcal{B}^{\text{global}}$ is empty; i.e., the global optimum is unachievable from the initialized solutions.

Further considering the fact $\Pr\left\{\hat{\mathbf{x}}^{\text{opt}} \notin \mathcal{B}^{\text{global}}\right\} \geq 0$ and $\Pr\left\{\mathbf{x}^T \to \mathbf{x}^*; \mathcal{A} \cap \mathcal{B}^{\text{global}} = \varnothing\right\} \geq 0$, we may obtain:

$$\Pr\left\{\mathbf{x}^T \to \mathbf{x}^*\right\} \geq \Pr\left\{\hat{\mathbf{x}}^{\text{opt}} \in \mathcal{B}^{\text{global}}\right\} \times \Pr\left\{\mathbf{x}^T \to \mathbf{x}^*; \mathcal{A} \cap \mathcal{B}^{\text{global}} \neq \varnothing\right\}.$$

According to the previous theoretical results on the algorithm convergence, after initialized around $\mathcal{B}^{\text{global}}$ and the global optimum becomes achievable, PSO algorithm would converge to the optimum of this local subspace with probability 1 [18], i.e., $\Pr\left\{\mathbf{x}^T \to \mathbf{x}^*; \mathcal{A} \cap \mathcal{B}^{\text{global}} \neq \varnothing\right\} \to 1$. By applying this result, the lower bound probability of EVOLER in finding the global solution is derived.

Note that, in the current framework of evolutionary computation, there is still a lack of theoretical analysis on the *precision* of the returned global solution. Even so, as validated in 20 representative benchmark functions of various geometric properties, the attained global solution

could be highly accurate. For example, we may practically have $|f(\mathbf{x}^T) - f(\mathbf{x}^*)| < \varepsilon$ with $\varepsilon < 10^{-10}$; see Fig. 3 in the main text, SI Table I and SI Fig. E5[1]. The theoretical analysis on the precision of our global solution is left for the future study.

For the case of $d = 2$, furthermore, there are two remarks for the lower bound probability of EVOLER in finding the global solution.

Remark 1: When the distance Δf is very small, we may have a small ratio $\frac{\Delta f}{\Delta E(s)}$. By applying Taylor series expansion when $\frac{\Delta f}{\Delta E(s)} < 1/2$ for example, i.e., $\arctan(x) = x + o(x^3)$, then the above probability becomes:

$$\Pr\left\{\mathbf{x}^T \to \mathbf{x}^*\right\} > (1-\epsilon) \times \frac{2\Delta f}{\pi \Delta E(s)}. \tag{7}$$

That is to say, the lower bound probability of finding the global optimum is proportional to the ratio between the distance Δf and an upper bound residual error $\Delta E(s)$.

Remark 2: If we configure a relatively large sampling length s, the residual error can be very small, especially for widespread low-rank problems. When we have $s \sim \mathcal{O}(r\log(\frac{r}{\epsilon}))$, the ration between the local distance Δf and the upper bound residual error $\Delta E(s)$ could be very large. Then, we have:

$$\begin{aligned} &\Pr\left\{\mathbf{x}^T \to \mathbf{x}^*\right\} > (1-\epsilon), \\ &\Rightarrow \ \Pr\left\{\mathbf{x}^T \to \mathbf{x}^*\right\} \to 1, \quad \text{for } \Delta f > C_0 \times \Delta E(s). \end{aligned} \tag{8}$$

That means we would have $\Pr\left\{\mathbf{x}^T \to \mathbf{x}^*\right\} \to 1$, when configuring a relatively large s which leads to an ignorable ϵ. For example, we obtain $\arctan\left(\frac{\Delta f}{\Delta E(s)}\right) \to \frac{\pi}{2}$ when $C_0 \geq 30$.

III. Extension of Theoretical Results to $d > 2$

For the high dimension case of $d > 2$, the above theoretical analysis can be directly generalized. For one thing, from the reconstruction process in Eq. (6) in Method E of the main text, we observe that the similar form (i.e. $\mathbf{C}_1\mathbf{U}_1^\dagger$) appears in the reconstructed tensor; so the residual error would also follow a non-Gaussian α-stable distribution; see also the empirical data in SI Fig. E14 (e.g. $d = 5$). For another, the probability of the initial solution $\hat{\mathbf{x}}^{\text{opt}} \in \mathbf{X} \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_d}$

[1]Note that, for the benchmark Schwefel function, the globally optimum fitness is about 2.54×10^{-5}, due to a truncated constant. For example, in the numerical experiments the constant is truncated to 418.9829.

falls into the global basin $\mathcal{B}^{\text{global}} \triangleq \{\mathbf{x} \in \mathbb{R}^d : \ |\mathbf{x} - \mathbf{x}^*| < \theta, \ f(\mathbf{x}) \text{ is unimodal}\}$, as well as the probability of finding the global solution, is similarly given by:

$$\Pr\left\{\mathbf{x}^T \to \mathbf{x}^*\right\} > \frac{2}{\pi} \arctan\left(\frac{\Delta f}{\frac{1}{N^{d/2}}\sqrt{\| \boldsymbol{\mathcal{F}} - \hat{\boldsymbol{\mathcal{F}}} \|^2}}\right). \tag{9}$$

Note that, the residual error $\| \boldsymbol{\mathcal{F}} - \hat{\boldsymbol{\mathcal{F}}} \|^2$ is also upper bounded, which yet may have more complex expressions [19]. As validated, the global optimum solution will be attained by our method (see Fig. E7 in SI, $d = 5$), taking the inverse design of nano-photonics devices for example. For the diverse designing targets, our method attains the best performance when compared to classical evolutionary computation methods (see Figs. E8 and E9 in SI).

IV. Hierarchy Structured Exploration

EVOLER firstly applies the structured exploration / sampling to identify one small attention subspace, by finding an initial solution in the reconstructed space. In this process, one practical consideration is that the required sampling length s or the grid size N may be large, especially when the decay rate of singular values is relatively slow (e.g. in the rotated benchmark functions) or the variable ranges are very large (e.g. in the Griewank function). In order to make our method more sample-efficient, we further suggest one *hierarchy* structured sampling mechanism, which, in combination with the hierarchy reconstruction strategy in SI Figs. E4∼E6, can maximize the efficiency of exploration and minimize the number of samples.

The main concept is that we iteratively apply the structured exploration for multiple times; see an illustration in SI Fig. E15. Firstly, we apply the standard EVOLER to identify an attention subspace $\mathcal{A}$ and an initial solution $\hat{\mathbf{x}}^{\text{opt}}$. Rather than directly initializing the particles, we immediately perform a successive structured exploration around this small subspace $\mathcal{A}$, which further shrinks an attention subspace of exploitation and enhances the resolution of initial solution. To be specific, we further explore a shrink problem space, i.e., $\mathbf{F}^1 = \{f(\mathbf{X}^1_{m,n}), \ m = 1, 2, \cdots, M_1; \ n = 1, 2, \cdots, N_1\} \in \mathbb{R}^{M_1 \times N_1}$, where $\mathbf{X}^1 \in \mathbb{R}^{M_1 \times N_1}$ is a discrete grid of bounded variable $\mathbf{x} \in [\hat{\mathbf{x}}^{\text{opt}} - \delta, \hat{\mathbf{x}}^{\text{opt}} + \delta]^2$. Subsequently, the randomly sampling is used to reconstruct this shrink space $\mathbf{F}^1$, and another smaller attention subspace $\mathcal{A}^1$ is identified. After multiple times (e.g. T_0) of the hierarchy structured explorations, we finally obtain a very small attention subspace $\mathcal{A}^{T_0-1}$ and the corresponding initial solution $\hat{\mathbf{x}}^{\text{opt}}$, which steers the exploitation towards

this shrink subspace. In practice, $T_0 = 2$ times would be sufficient in most cases. For example, in the more challenging rotated benchmarks we may use this mechanism to effectively reduce the required samples.

REFERENCES

[1] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Relative-error CUR matrix decompositions," *SIAM Journal on Matrix Analysis and Applications*, 2008.

[2] S. Wang and Z. Zhang, "Improving CUR matrix decomposition and the nyström approximation via adaptive sampling," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2729–2769, 2013.

[3] B. Li, P. Chen, H. Liu, W. Guo, X. Cao, J. Du, C. Zhao, and J. Zhang, "Random sketch learning for deep neural networks in edge computing," *Nature Computational Science*, vol. 1, no. 3, pp. 221–228, 2021.

[4] H. Liu, Z. Wei, H. Zhang, B. Li, and C. Zhao, "Tiny machine learning (tiny-ml) for efficient channel estimation and signal detection," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6795–6800, 2022.

[5] E. A. Cornish, "The multivariate t-distribution associated with a set of normal sample deviates," *Australian Journal of Physics*, vol. 7, p. 531, 1954.

[6] J. M. Dickey, "Matricvariate generalizations of the multivariate t-distribution and the inverted multivariate t-distribution," *The Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 511–518, 1967.

[7] B. Li, S. Wang, J. Zhang, X. Cao, and C. Zhao, "Ultra-Fast Accurate AoA Estimation via Automotive Massive-MIMO Radar," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 1172–1186, 2021.

[8] G. A. Walker and J. G. Saw, "The distribution of linear combinations of t-variables," *Journal of the American Statistical Association*, vol. 73, no. 364, pp. 876–878, 1978.

[9] S. Kotz and S. Nadarajah, *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

[10] B. Gnedenko, A. Kolmogorov, B. Gnedenko, and A. Kolmogorov, "Limit distributions for sums of independent," *Am. J. Math*, vol. 105, 1954.

[11] J. Nolan, *Stable distributions: models for heavy-tailed data*. Birkhauser New York, 2003.

[12] L. I. Li, "The study of parameter estimation of alpha-stable distribution based on truncated extreme value and classifying skewness," *Signal Processing*, vol. 24, no. 4, pp. 574–577, 2008.

[13] M. H. Bibalan, H. Amindavar, and M. Amirmazlaghani, "Characteristic function based parameter estimation of skewed alpha-stable distribution: An analytical approach," *Signal Processing*, vol. 130, no. 1, pp. 323–336, 2017.

[14] X. Li, J. Sun, L. Jin, and M. Liu, "Bi-parameter cgm model for approximation of α-stable PDF," *Electronics Letters*, vol. 44, no. 18, pp. 1096–1097, 2008.

[15] J. Park, G. Shevlyakov, and K. Kim, "Maximin distributed detection in the presence of impulsive alpha-stable noise," *IEEE Transactions on Wireless Communications*, vol. 10, no. 6, pp. 1687–1691, 2011.

[16] A. Swami, "Non-gaussian mixture models for detection and estimation in heavy-tailed noise," in *In Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 6, pp. 3802–3805, IEEE, 2000.

[17] Y. Xia, Z. Deng, L. Li, and X. Geng, "A new continuous-discrete particle filter for continuous-discrete nonlinear systems," *Information Sciences*, vol. 242, pp. 64–75, 2013.

[18] F. Van den Bergh and A. P. Engelbrecht, "A convergence proof for the particle swarm optimiser," *Fundamenta Informaticae*, vol. 105, no. 4, pp. 341–374, 2010.

[19] Z. Song, D. P. Woodruff, and P. Zhong, "Relative error tensor low rank approximation," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2772–2789, SIAM, 2019.

[20] F. Van den Bergh and A. P. Engelbrecht, "A cooperative approach to particle swarm optimization," *IEEE transactions on evolutionary computation*, vol. 8, no. 3, pp. 225–239, 2004.

[21] B. Niu, Y. Zhu, X. He, and H. Wu, "Mcpso: A multi-swarm cooperative particle swarm optimizer," *Applied Mathematics and computation*, vol. 185, no. 2, pp. 1050–1062, 2007.

[22] W. Li, X. Meng, Y. Huang, and Z.-H. Fu, "Multipopulation cooperative particle swarm optimization with a mixed mutation strategy," *Information Sciences*, vol. 529, pp. 179–196, 2020.

[23] J. J. Liang, A. K. Qin, P. N. Suganthan, and S. Baskar, "Comprehensive learning particle swarm optimizer for global optimization of multimodal functions," *IEEE transactions on evolutionary computation*, vol. 10, no. 3, pp. 281–295, 2006.

[24] C. Yang and D. Simon, "A new particle swarm optimization technique," in *18th International Conference on Systems Engineering (ICSEng'05)*, pp. 164–169, IEEE, 2005.

[25] N. Zeng, Z. Wang, W. Liu, H. Zhang, K. Hone, and X. Liu, "A dynamic neighborhood-based switching particle swarm optimization algorithm," *IEEE Transactions on Cybernetics*, 2020.

[26] F. Van den Bergh and A. P. Engelbrecht, "A new locally convergent particle swarm optimiser," in *IEEE International conference on systems, man and cybernetics*, vol. 3, pp. 6–pp, IEEE, 2002.

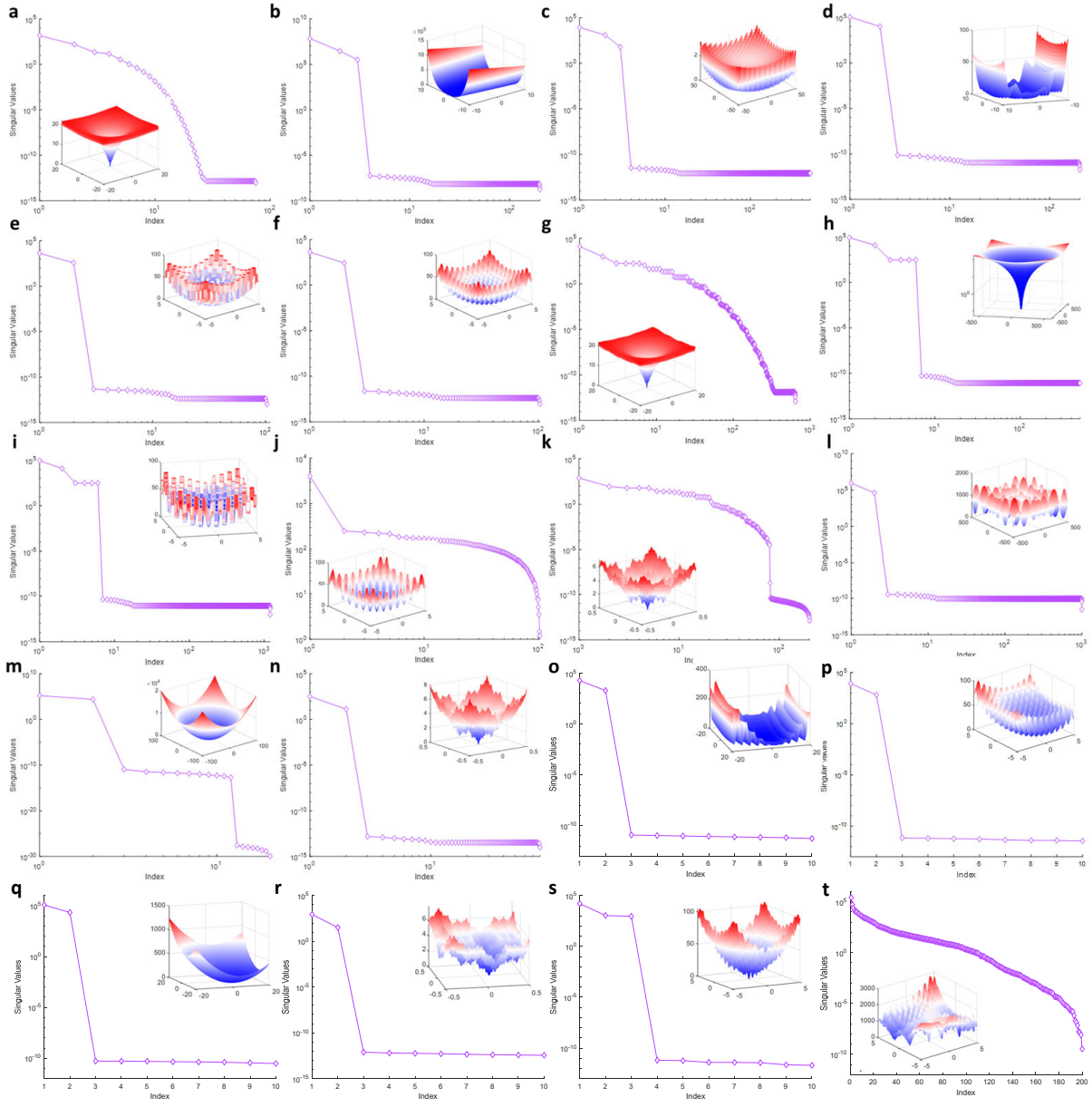
[27] X.-F. Xie, W.-J. Zhang, and Z.-L. Yang, "Dissipative particle swarm optimization," in *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*, vol. 2, pp. 1456–1461, IEEE, 2002.

[28] M. Clerc and J. Kennedy, "The particle swarm-explosion, stability, and convergence in a multidimensional complex space," *IEEE transactions on Evolutionary Computation*, vol. 6, no. 1, pp. 58–73, 2002.

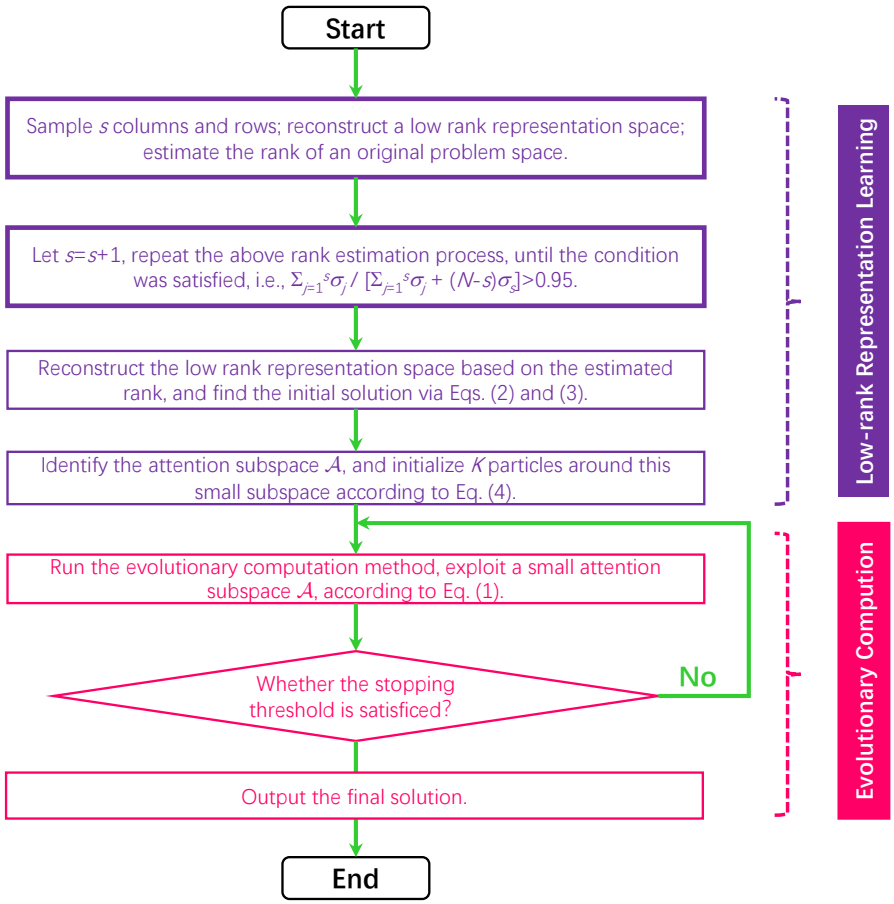
[29] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm intelligence*, vol. 1, no. 1, pp. 33–57, 2007.

[30] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *MHS'95. Proceedings of the sixth international symposium on micro machine and human science*, pp. 39–43, Ieee, 1995.

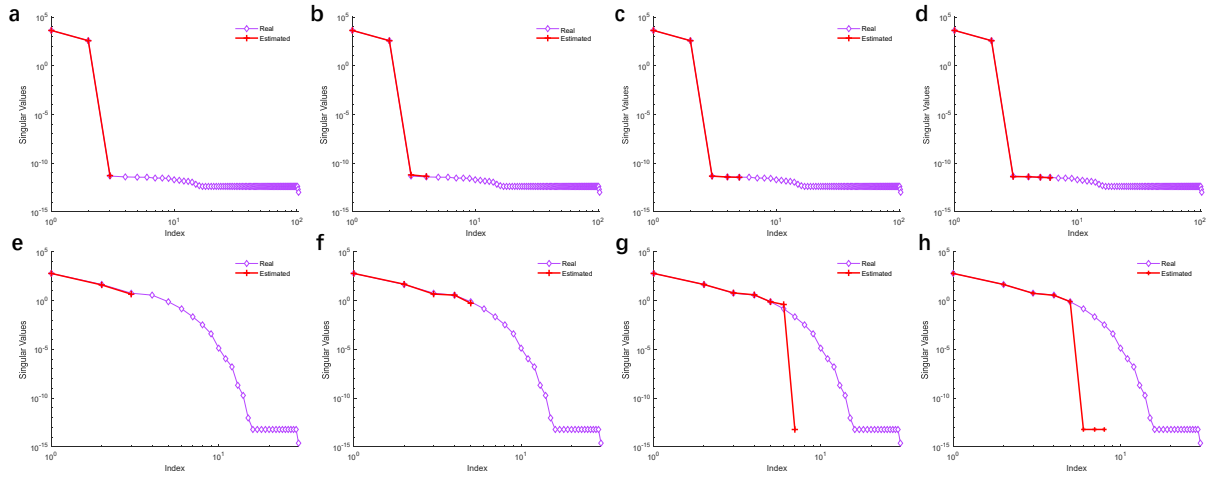
SUPPLEMENTARY FIGURE & TABLE



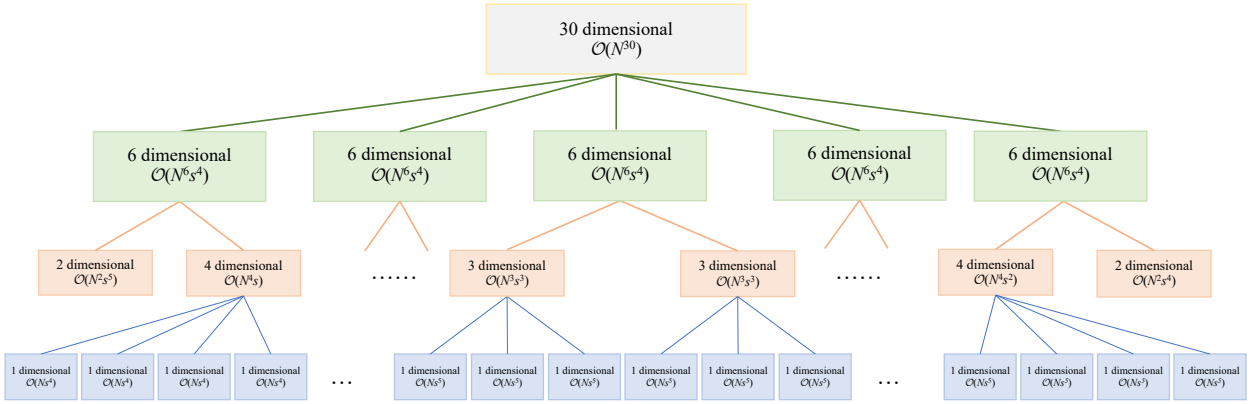
Supplementary Figure 1. **Ubiquitous low-rank structures in representative benchmark functions**. **(a)** Ackley function ($M = N = 75$); **(b)** Rosenbrock function ($M = N = 200$); **(c)** Griewank function ($M = N = 500$); **(d)** Levy function ($M = N = 200$); **(e)** Non-continuous Rastrigin function ($M = N = 102$); **(f)** Rastrigin function ($M = N = 102$); **(g)** Rotated Ackley function ($M = N = 640$); **(h)** Rotated Griewank function ($M = N = 500$); **(i)** Rotated Non-continuous Rastrigin function ($M = N = 1200$); **(j)** Rotated Rastrigin function ($M = N = 100$); **(k)** Rotated Weierstrass function ($M = N = 200$); **(l)** Schwefel function ($M = N = 1000$); **(m)** Sphere function ($M = N = 20$); **(n)** Shifted Levy function ($M = N = 10$). **(o)** Shifted function ($M = N = 80$). **(p)** Shifted Rastrigin function ($M = N = 10$). **(q)** Shifted Sphere function ($M = N = 10$). **(r)** Shifted Weierstrass function ($M = N = 10$). **(s)** Hybrid Composite function 1 ($M = N = 10$). **(t)** Hybrid Composite function 2 ($M = N = 200$). The geometric property of each benchmark function is given in sub-figures. The mathematical expressions as well as the variables range of these representative benchmarks are summarized in Table 1.



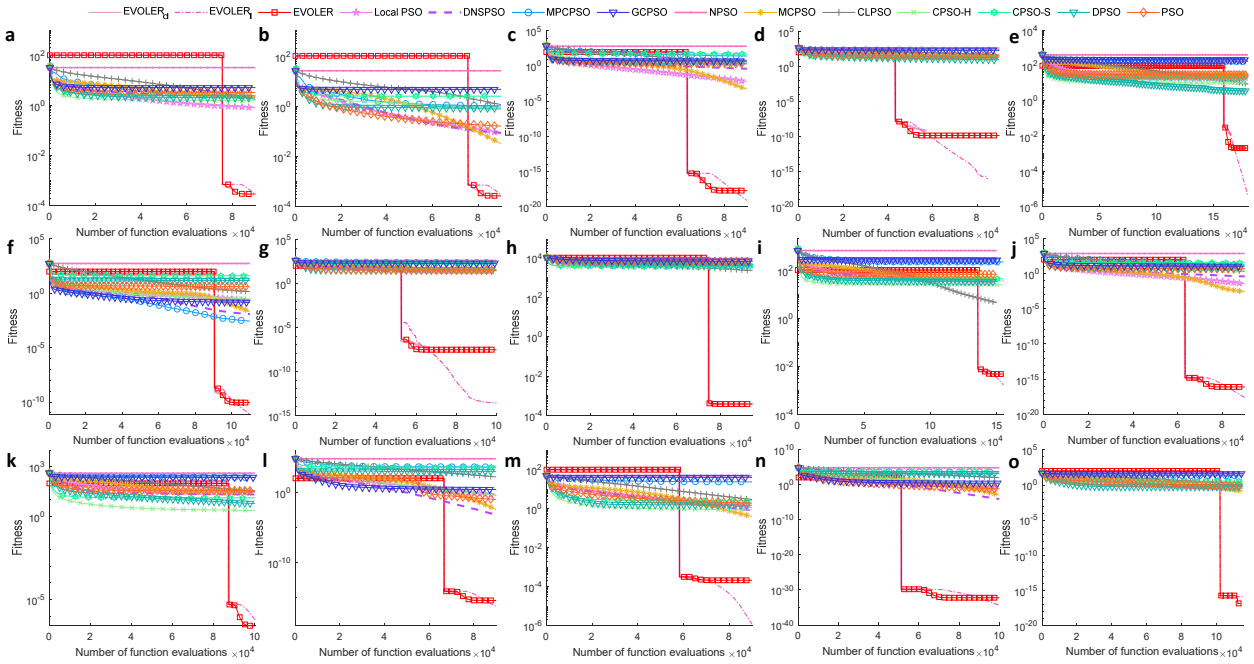
Supplementary Figure 2. **Algorithm flow of EVOLER.** Our method consists of two successive stages. The first stage of low-rank representation learning involves: (i) Rank estimation (see SI section I) and random sampling, (ii) reconstruction of low-rank representation space, and (iii) identification of one attention subspace. Then. in the next stage of evolutionary computation, K particles will exploit the small subspace after the proper initialization.



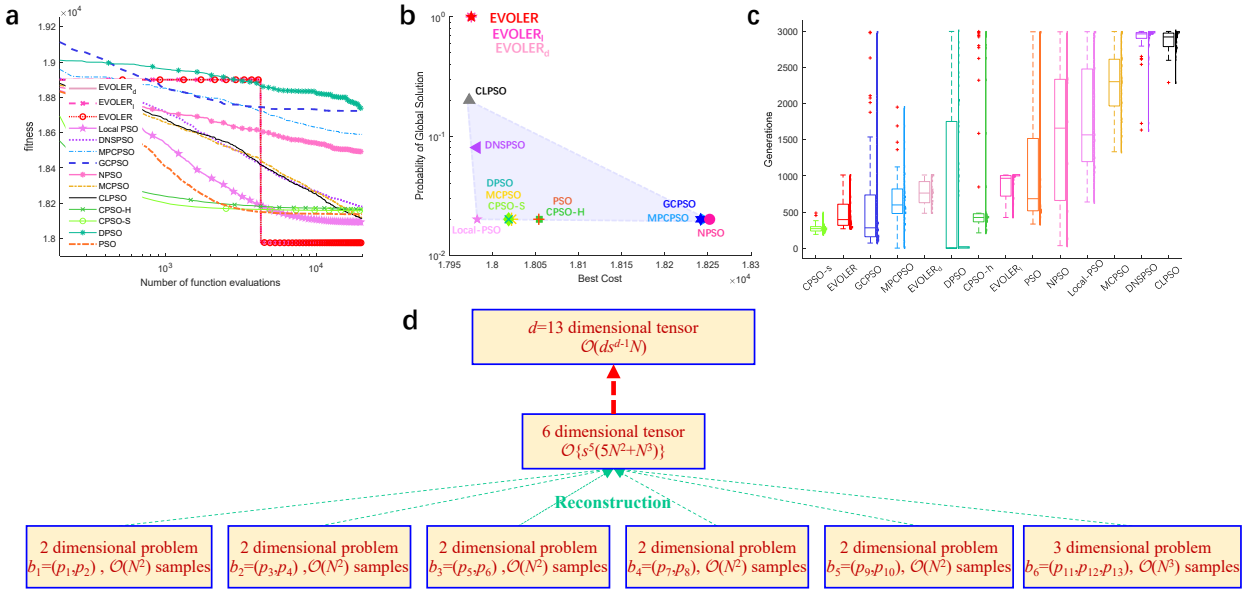
Supplementary Figure 3. **Online Rank Estimation of Unknown Problem Space**. For clarity, we again take the 2D problem for example. **(a)**-**(d)** Estimated and real singular values of non-continuous Rastrigin function, in the case of $s = 3, 4, 5, 6$; $M = N = 103$. As the sampling length s was increased from 3 to 6, we found the singular values would show the sharp decrease when $s = 3$, i.e., $\sigma_r^2 \gg \sigma_{r+1}^2$, thus the estimated rank is $\hat{r} = 2$ in this strict low-rank case. **(e)**-**(h)** Estimated and real singular values of Ackley function, in the case of $s = 3, 5, 7, 9$; $M = N = 30$. As the sampling length s was increased from 3 to 9, we found the singular values may show a sharp decrease when $s = 7$, and hence the estimated rank is $\hat{r} = 6$ in this approximate low-rank case. For example, the accumulated energy of the first $\hat{r}$ singular values accounts for $\sum_{j=1}^{r} \sigma_j^2 / \sum_{j=1}^{N} \sigma_j^2 > \sum_{j=1}^{r} \hat{\sigma}_j^2 / [\sum_{j=1}^{r} \hat{\sigma}_j^2 + (N - r)\hat{\sigma}_{r+1}^2] > 0.95$ of the total energy.



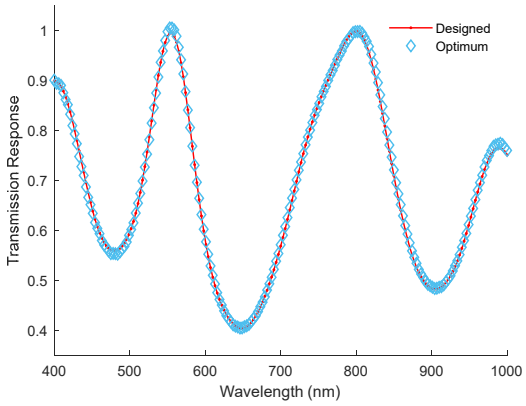
Supplementary Figure 4. **Generalized hierarchy reconstruction of a low-rank representation for high dimension problems**. Taking the $d = 30$ problem for example, a hierarchy concept can be applied to reconstruct a low-rank representation space. Firstly, we divide the original d =30 dimensional problem into $L = 5$ blocks; thus each block becomes one $N^{d/L} = N^6$ dimensional subspace. By applying the low-rank approximation, we are able to reconstruct the original 30 dimensional space, via $L = 5$ smaller tensors with the dimension of $N^{d/L}s^{L-1} = N^6s^4$. Following this line, then each 6 dimensional subspace can be divided into two 3 dimensional subspaces; and each subspace becomes the N^3s^2 dimensional one. Similarly, we are able to reconstruct the N^6s^4 dimensional problem, via 2 small tensors of $N^3s^2 \times s = N^3s^3$ dimensions. This hierarchy reconstruction procedure will be repeated, until a set of 1-dimensional spaces are obtained. In this manner, the number of samples, as well as the complexity, will be reduced to $\mathcal{O}\{ds^{(d/L-1)}N\}$; while a direct reconstruction of $d = 30$ dimensional representation space requires $\mathcal{O}\{ds^{d-1}N\}$ samples. Note that, the maximum number of blocks can be determined via the constraint $d/L > L - 1$, so that $L \leq \lfloor\sqrt{d}\rfloor$. In practice, the number of blocks can be configured to $2 \leq L \leq \lfloor\sqrt{d}\rfloor$.



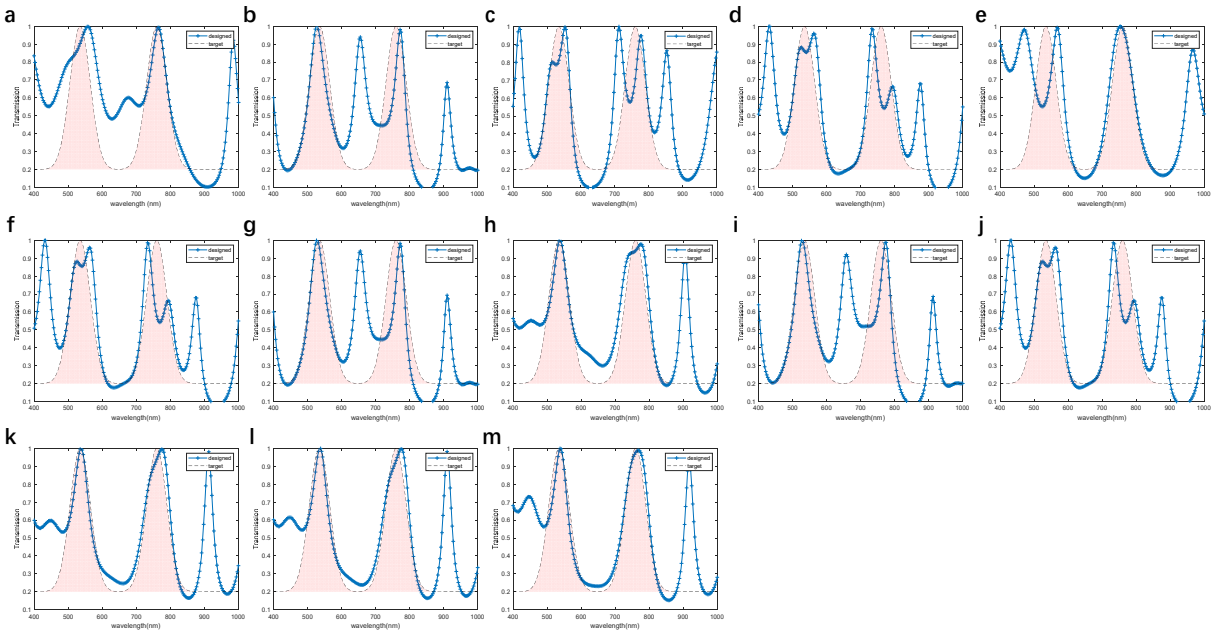
Supplementary Figure 5. **Finding the global optimum almost surely in more challenging 30 benchmark functions**. Averaged convergence curves are obtained, based on 500 independent trials with random initializations. The problem dimension is $d = 30$, and the number of blocks is $L = 5$. The number of particles is $K = 80$. The maximum generation is $T = 900$. In the experiments, the grid size N is configured firstly, and then the sampling length s is automatically configured, after the online rank estimation (see SI I-A). In order to minimize the number of samples, a hierarchy structured exploration was also applied; see more details in SI Section IV. **(a)** Hybrid composite function 1. **(b)** Hybrid composite function 3. **(c)** Levy function. **(d)** Rastrigin function. **(e)** Rastrigin function with noise. **(f)** Rotated Griewank function. **(g)** Rotated Rastrigin funcyion. **(h)** Schwefel function. **(i)** Shifted composite function. **(j)** Shifted Levy function. **(k)** Shifted Rastrigin function. **(l)** Shifted Sphere function. **(m)** Shifted Weierstrass function. **(n)** Sphere function. **(o)** Weierstrass function.



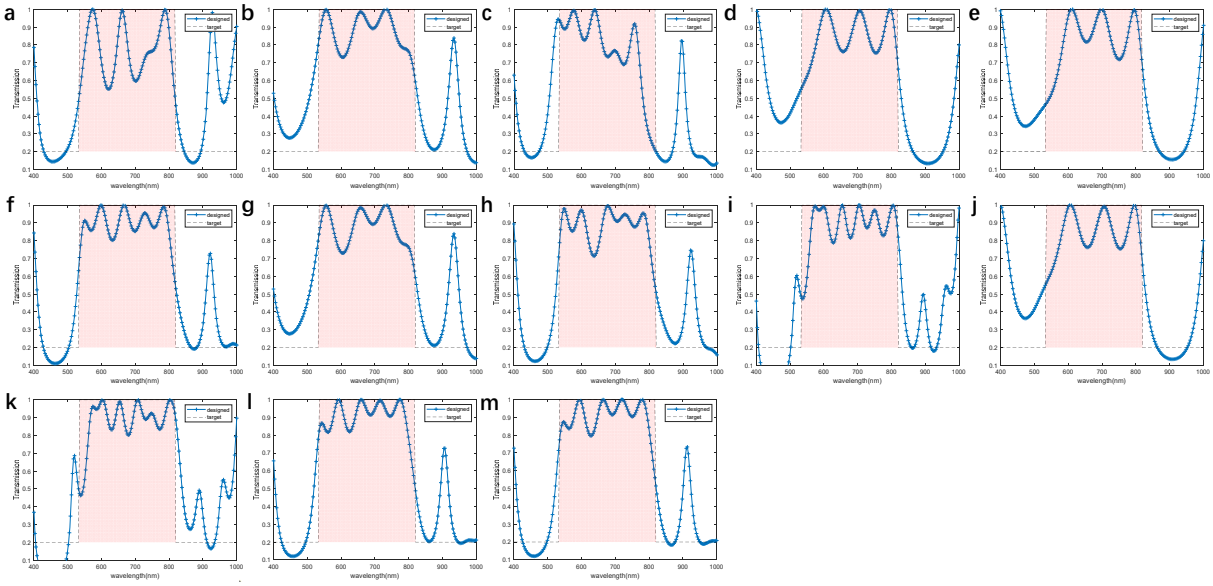
Supplementary Figure 6. **Computational power dispatch problem of** $d = 13$**. (a)** Averaged convergence results of various evolutionary computation methods; 50 independent trials, $K = 100$ particles. **(b)** Achievable minimal costs of different methods, as well as the probabilities of finding them. **(c)** Required generations for convergence. **(d)** The original problem of 13 dimension is transformed to another equivalent 6-dimensional problem. In each new variable b_j $(j = 1, 2, 3, 4, 5, 6)$, it merges 2 (or 3) dimension of the original problem. On this basis, the structured exploration is implemented on this new 6-dimensional problem $(b_1, b_2, b_3, b_4, b_5, b_6)$, with the time/sample complexity $\mathcal{O}\{s^5(5N^2 + N^3\}$, which is dramatically smaller than a direct structured exploration on 13 dimensional problem with the complexity of $\mathcal{O}\{13s^{12}N\}$.



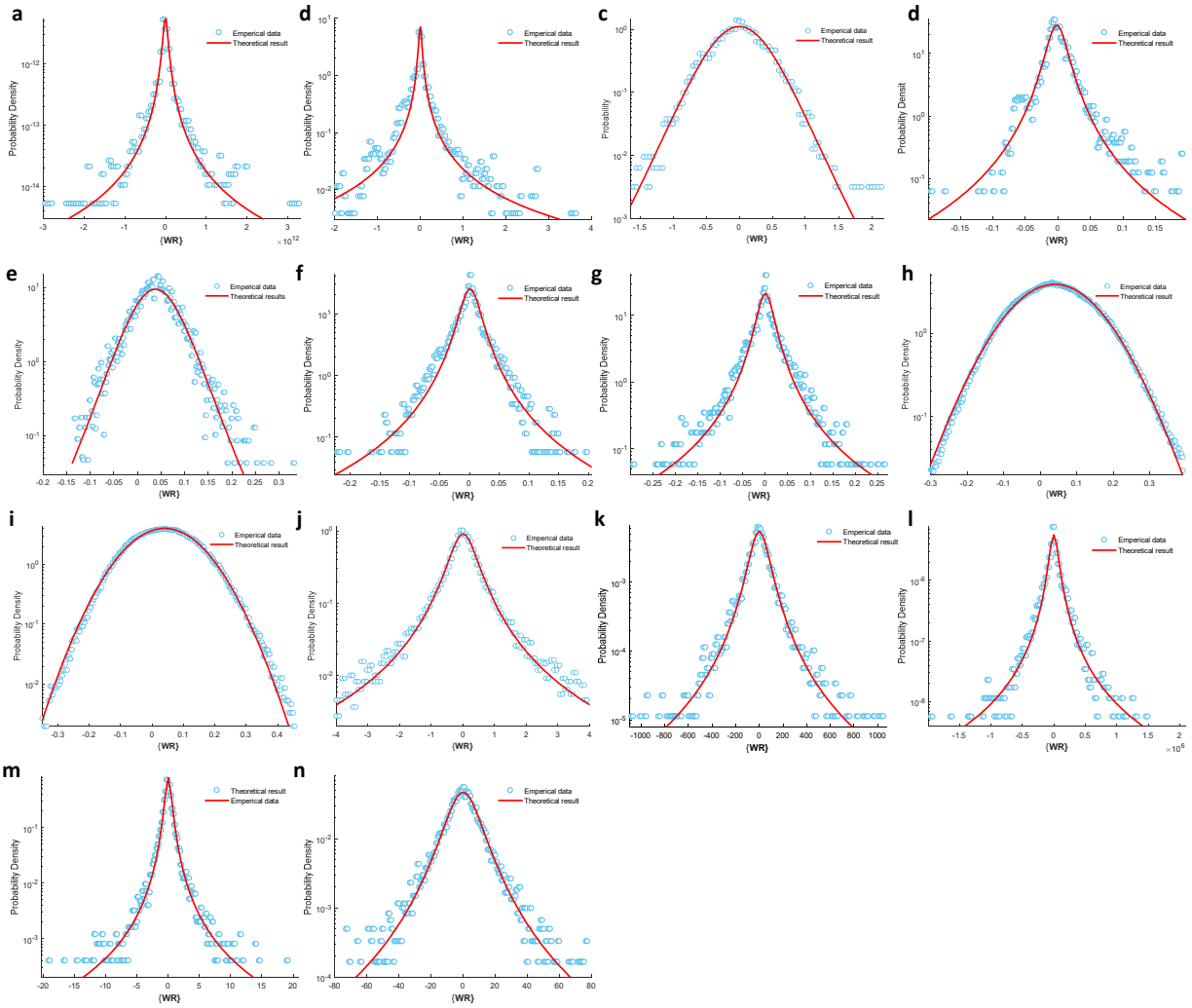
Supplementary Figure 7. **Finding the optimal nanophotonics structure with EVOLER.** To demonstrate the global optimality of our method, we consider one special case with the attainable optimal spectrum. I.e., this optimal transmittance spectrum corresponds to the global solution, i.e., $\mathbf{a}^* = [25\ 15\ 24\ 63\ 135]$ nm. In this case, the global solution $\mathbf{a}^*$ results in a fitness of 0. As seen, our method reliably gains the global optimum of this nano-photonics structure inverse design.



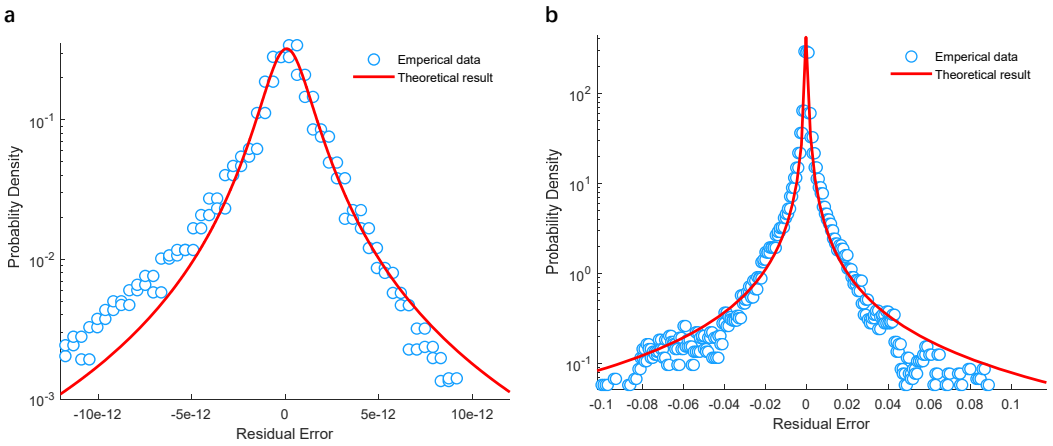
Supplementary Figure 8. **Inverse design of nano-photonics structures using classical evolutionary computation.** Designed transmittance response of 18-layers nano-photonics structure; the target spectrum involves two bell-shaped transmittance bands. Each classical evolutionary method was evaluated for 5 independent trials; see SI Table II for the parameter settings of various evolutionary computation methods. Here, we plot the designed transmittance response of each method, which corresponds to the median mean square error (MSE). **(a)** NPSO. **(b)** DNSPSO. **(c)** MPCPSO. **(d)** DPSO. **(e)** CPSO-S. **(f)** MCPSO. **(g)** CPSO-H. **(h)** CLPSO. **(i)** GCPSO. **(j)** Canonic PSO. **(k)** Local PSO. **(l)** EVOLER$_l$. **(m)** EVOLER$_d$.



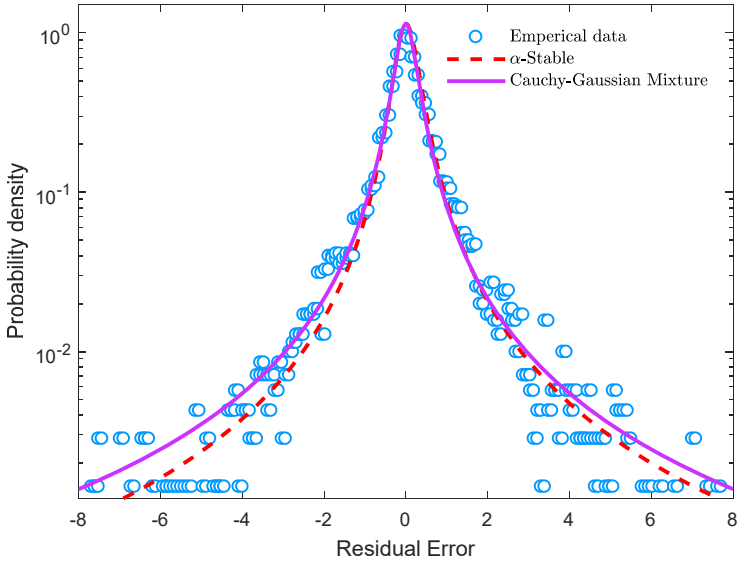
Supplementary Figure 9. **Inverse design of nano-photonics structures using classical evolutionary computation.** Designed transmittance response of 18-layers nano-photonics structure; the target spectrum involves one sharp transmittance bands. Each classical evolutionary method was evaluated for 5 independent trials; see SI Table II for the parameter settings of various evolutionary computation methods. Here, we plot the designed transmittance response of each method, which corresponds to the median mean square error (MSE). **(a)** NPSO. **(b)** DNSPSO. **(c)** MPCPSO. **(d)** DPSO. **(e)** CPSO-S. **(f)** MCPSO. **(g)** CPSO-H. **(h)** CLPSO. **(i)** GCPSO. **(j)** Canonic PSO. **(k)** Local PSO. **(l)** EVOLER$_l$. **(m)** EVOLER$_d$.



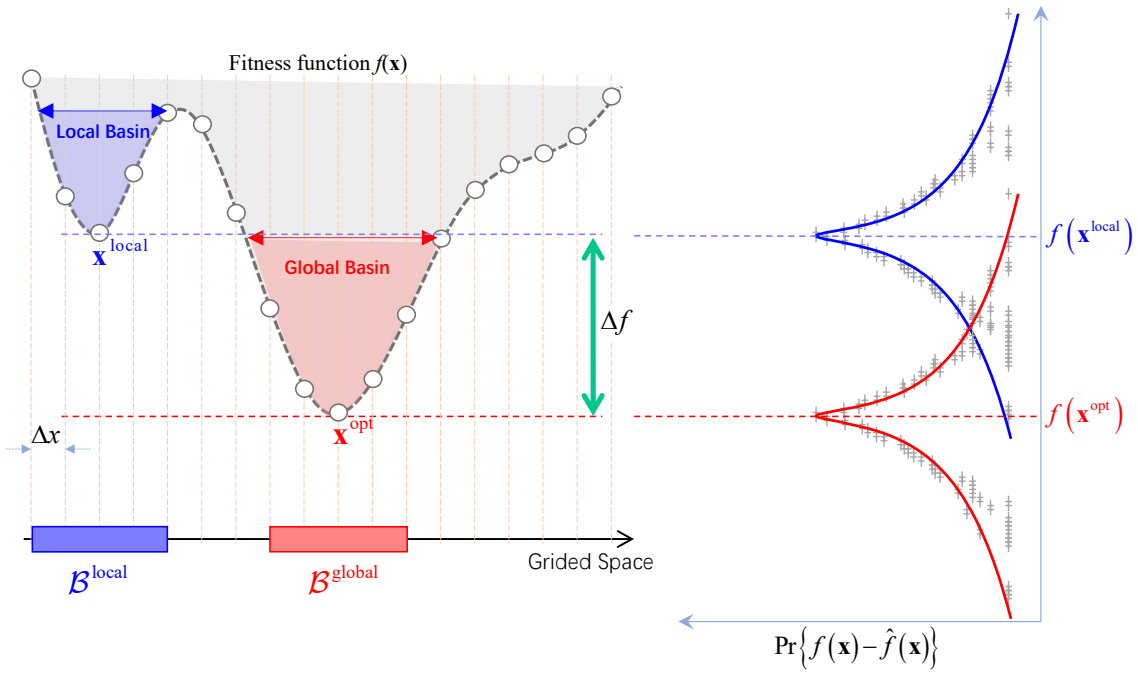
Supplementary Figure 10. **Theoretical probability density function (PDF) of the element in WR(:,j)** ($j = 1, 2, \cdots, N$)**.** Each benchmark function was independently evaluated for 10 trails. **(a)** Ackley function; $\mathbf{x} = (-5.32, 5.32)^d$, $M = N = 107$, $s = 25$. **(b)** Rosenbrock function; $\mathbf{x} = (-10, 10)^d$, $M = N = 201$, $s = 25$. **(c)** Griewank function; $\mathbf{x} = (-600, 600)^d$, $M = N = 1201$, $s = 25$. **(d)** Levy function; $\mathbf{x} = (-50, 50)^d$, $M = N = 201$, $s = 25$. **(e)** Non-continuous Rastrigin function; $\mathbf{x} = (-5.12, 5.12)^d$, $M = N = 103$, $s = 25$. **(f)** Rastrigin function; $\mathbf{x} = (-5.12, 5.12)^d$, $M = N = 103$, $s = 25$. **(g)** Schwefel function; $\mathbf{x} = (-500, 500)^d$, $M = N = 101$, $s = 25$. **(h)** Sphere function; $\mathbf{x} = (-100, 100)^d$, $M = N = 51$, $s = 25$. **(i)** Weierstrass function; $\mathbf{x} = (-0.5, 0.5)^d$, $M = N = 80$, $s = 25$. **(j)** Rotated Ackley function; $\mathbf{x} = (-32, 32)^d$, $M = N = 641$, $s = 25$. **(l)** Rotated Griewank function; $\mathbf{x} = (-600, 600)^d$, $M = N = 1201$, $s = 25$. **(l)** Rotated Non-continuous Rastrigin function; $\mathbf{x} = (-5.12, 5.12)^d$, $M = N = 205$, $s = 25$. **(m)** Rotated Rastrigin function; $\mathbf{x} = (-5.12, 5.12)^d$, $M = N = 103$, $s = 25$. **(n)** Rotated Weierstrass function; $\mathbf{x} = (-0.5, 0.5)^d$, $M = N = 201$, $s = 25$.



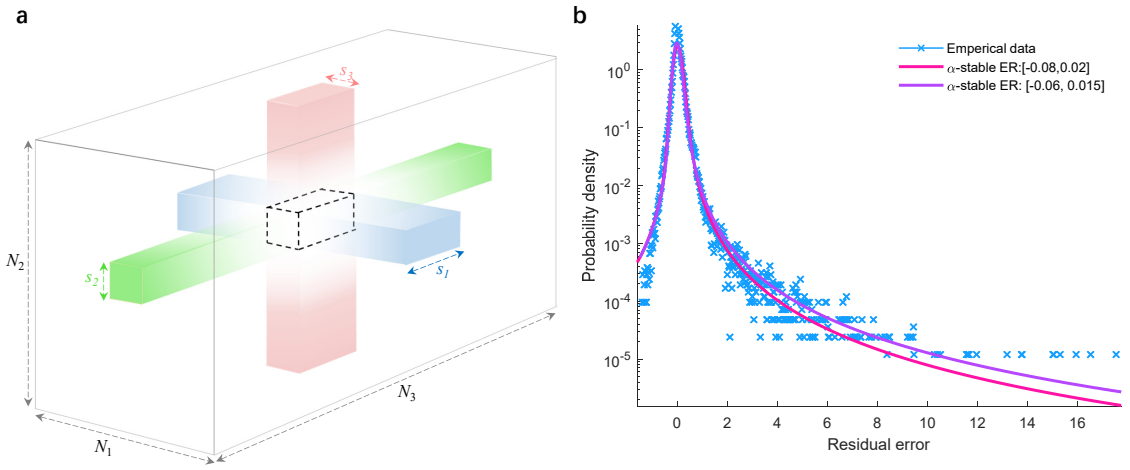
Supplementary Figure 11. **Theoretical Probability Distribution of Residual Error** $e_{m,n} = \{\hat{\mathbf{F}} - \mathbf{F}\}$**. (a)** Strict low rank case; Schwefel function, $d = 2$, $\mathbf{x} = (-500, 500)^d$; $M = N = 201$, $s = 25$; **(b)** Approximate rank-restricted case; Rotated Ackley function, $d = 2$, $\mathbf{x} = (-32, 32)^d$; $M = N = 321$, $s = 25$. In both cases, the theoretical α-stable distribution accurately models the empirical data of residual errors.



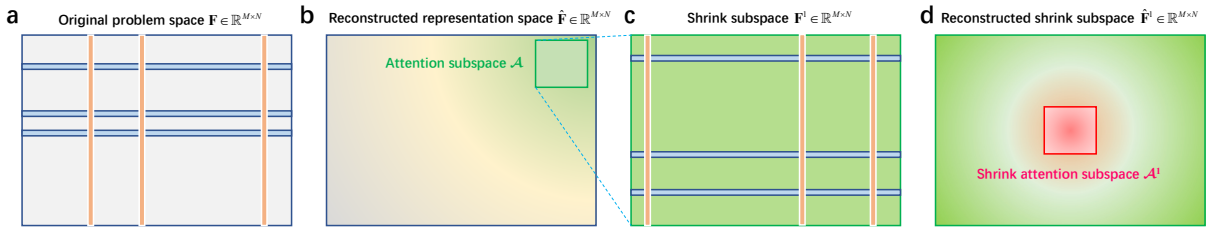
Supplementary Figure 12. **Approximating the α-stable density with the Cauchy-Gaussian Mixture.** Rotated Ackley function; $d = 2$, $\mathbf{x} = (-32, 32)^d$, $M = N = 321$, $s = 25$. The α-stable density with the estimated parameters: $\alpha = 1.1037$, $\beta = 0.0907$, $\gamma = 0.2688$, $\delta = 0.0168$. The Cauchy-Gaussian mixture density with the estimated parameters: $\lambda = 0.999$, $\sigma = 0.2936$.



Supplementary Figure 13. **Illustration of the global optimum and the next local optimum**. For clarity, we consider the 1 dimensional fitness function, with a grid resolution Δx. When the grid size N is sufficiently large, this grid resolution Δx would be very small; so that the global optimum in the discrete space, $\mathbf{x}^{\text{opt}} = \arg\min_{\mathbf{x}\in\mathbf{X}} f(\mathbf{x})$, would fall into the basin of the global optimum $\mathcal{B}^{\text{global}}$ in the continuous space, i.e., $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^2} f(\mathbf{x})$; $\mathbf{X} \in \mathbb{R}^{M\times N}$ denotes a discrete grid of continuous variables. The next local optimum $\mathbf{x}^{\text{local}} \in \mathbf{X}$ belongs to the basin of local optimum $\mathcal{B}^{\text{local}}$. In this case, the probability of mistaking the next local optimum to the global optimum can be derived, based on the theoretical PDF of residual errors (left subfigure). As seen, when the initial solution falls into the global basin, even if it is not equal to the global optimum, the global optimum can be still attainable. As such, the evolutionary computation method would converge to the optimum solution.



Supplementary Figure 14. **(a)** Illustration of the structured exploration of the high dimensional problems, taking the 3D tensor for example. Given the 3D tensor $\boldsymbol{\mathcal{F}} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$, three small sampled versions are firstly attained, which are denoted as $\boldsymbol{\mathcal{C}}_1 \in \mathbb{R}^{N_1 \times s_2 \times s_3}$ (green cuboid), $\boldsymbol{\mathcal{C}}_2 \in \mathbb{R}^{s_1 \times N_2 \times s_3}$ (red cuboid), $\boldsymbol{\mathcal{C}}_3 \in \mathbb{R}^{s_1 \times s_2 \times N_3}$ (blue cuboid). On this basis, a small intersection tensor $\boldsymbol{\mathcal{R}} \in \mathbb{R}^{s_1 \times s_2 \times s_3}$ ($s_1,\ s_2,\ s_3 \ll N$) is obtained (black cuboid at the center). For clarity, each tensor with random sampling positions is represented by one cuboid. **(b)** Theoretical probability distribution function of the residual error of reconstructed 5D tensor; Ackley function, $d = 5$, $N_1 = N_2 = N_3 = N = 14$, $s_1 = s_2 = s_3 = s = 4$. ER: exclusion region, which is used to exclude some singular data points.



Supplementary Figure 15. **Reducing samples via a hierarchy structured exploration**. **(a)** The first-level structured exploration on an original problem space $\mathbf{F} \in \mathbb{R}^{M_0 \times N_0}$, by randomly sampling s columns and rows (e.g., $s = 3$). **(b)** After the reconstruction of a low-rank representation space $\hat{\mathbf{F}} \in \mathbb{R}^{M_0 \times N_0}$, the first-level attention subspace $\mathcal{A}$ as well as the initial solution $\hat{\mathbf{x}}^{\text{opt}}$ are identified. **(c)** The second-level structured exploration on the shrink problem space $\mathbf{F}^1 \in \mathbb{R}^{M_1 \times N_1}$ which is located around an identified attention subspace $\mathcal{A}$, by randomly sampling s columns and rows (e.g., $s = 3$). **(d)** Identifying the second-level attention subspace $\mathcal{A}^1$, based on the reconstructed shrink subspace $\hat{\mathbf{F}}^1 \in \mathbb{R}^{M_1 \times N_1}$. This hierarchy structured explorations will be applied for Q times; and finally, we identify a very small attention subspace $\mathcal{A}^{Q-1}$, which then steers the next exploitation toward a smaller subspace. In this way, the required samples of EVOLER can be effectively reduced.

Supplementary Table I: Benchmark Functions.

Function Name	Function Expression	Variable Range	Optimum Solution	Successful threhold	Function Property
Ackley	$f(\mathbf{x}) = -a\exp\left(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d}x_i^2}\right) - \exp\left(-\frac{1}{d}\sum_{i=1}^{d}\cos\left(cx_i^2\right)\right) + a + \exp(1)$ $a = 20, b = 0.2$	$[-32, 32]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Continuous, Differentiable, Non-Separable, Non-Scalable, Unimodal
Rosenbrock	$f(\mathbf{x}) = \sum_{i=1}^{d-1}\left[100\left(x_{i+1} - x_i^2\right)^2 + (x_i - 1)^2\right]$	$[-10, 10]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Continuous, Differentiable, Non-Separable, Scalable, Unimodal
Griewank	$f(\mathbf{x}) = \sum_{i=1}^{d-1}\frac{x_i^2}{4000} - \prod_{i=1}^{d}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	$[-600, 600]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Continuous, Differentiable, Non-Separable, Scalable, Multimodal
Levy	$f(\mathbf{x}) = \sin^2(\pi w_i) + \sum_{i=1}^{d-1}(w_i - 1)^2\left[1 + 10\sin^2(\pi w_i + 1)\right] + (w_d - 1)^2\left[1 + \sin^2(2\pi w_d)\right]$ $w_i = 1 + \frac{x_i - 1}{4}$	$[-50, 50]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	
Non-continuous Rastrigin	$f(\mathbf{x}) = \sum_{i=1}^{d}\left(y_i^2 - 10\cos(2\pi y_i) + 10\right), y_i = \begin{cases} x_j & \lvert x_j\rvert < 1/2 \\ round(2x_j)/2 & \lvert x_j\rvert \geq 1/2 \end{cases}$	$[-5.12, 5.12]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Multimodal, Non-Separable, Asymmetric
Rastrigin	$f(\mathbf{x}) = 10d + \sum_{i=1}^{d}\left[x_i^2 - 10\cos(2\pi x_i)\right]$	$[-5.12, 5.12]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Multimodal, Non-Separable
Schwefel	$f(\mathbf{x}) = 418.9829d - \sum_{i=1}^{d}x_i\sin\left(\sqrt{\lvert x_i\rvert}\right)$	$[-500, 500]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Multimodal, Non-Separable, Asymmetric, Deceptive
Sphere	$f(\mathbf{x}) = \sum_{i=1}^{d}x_i^2$	$[-100, 100]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Unimodal, Separable
Weierstrass	$f(\mathbf{x}) = \sum_{i=1}^{d}\left(\sum_{k=0}^{k_{\max}}a^k\cos\left(2\pi b^k(x_i + 0.5)\right)\right) - d\sum_{k=0}^{k_{\max}}a^k\cos\left(2\pi b^k 0.5\right)$ $a = 0.5, b = 3, k_{max} = 20$	$[-0.5, 0.5]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Continuous, Undifferentiable, Separable, Scalable, Multimodal
Rotated Ackley	$f(\mathbf{x}) = -a\exp\left(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d}x_i^2}\right) - \exp\left(-\frac{1}{d}\sum_{i=1}^{d}\cos\left(cx_i^2\right)\right) + a + \exp(1), a = 20, b = 0.2$ $a = 20, b = 0.2, \mathbf{x} = \mathbf{x}\cdot\mathbf{M}$	$[-32, 32]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Multimodal, Non-Separable, Asymmetric
Rotated Griewank	$f(\mathbf{x}) = \sum_{i=1}^{d-1}\frac{x_i^2}{4000} - \prod_{i=1}^{d}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$ $\mathbf{x} = \mathbf{x}\cdot\mathbf{M}$	$[-600, 600]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Multimodal, Non-Separable
Rotated Non-continuous Rastrigin	$f(\mathbf{x}) = \sum_{i=1}^{d}\left(y_i^2 - 10\cos(2\pi y_i) + 10\right), y_i = \begin{cases} x_j & \lvert x_j\rvert < 1/2 \\ round(2x_j)/2 & \lvert x_j\rvert \geq 1/2 \end{cases}$ $\mathbf{x} = \mathbf{x}\cdot\mathbf{M}$	$[-5.12, 5.12]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Multimodal, Non-Separable, Asymmetric
Rotated Rastrigin	$f(\mathbf{x}) = 10d + \sum_{i=1}^{d}\left[x_i^2 - 10\cos(2\pi x_i)\right]$ $\mathbf{x} = \mathbf{x}\cdot\mathbf{M}$	$[-5.12, 5.12]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Multimodal, Non-Separable
Rotated Weierstrass	$f(\mathbf{x}) = \sum_{i=1}^{d}\left(\sum_{k=0}^{k_{\max}}a^k\cos\left(2\pi b^k(x_i + 0.5)\right)\right) - d\sum_{k=0}^{k_{\max}}a^k\cos\left(2\pi b^k 0.5\right)$ $a = 0.5, b = 3, k_{max} = 20, \mathbf{x} = \mathbf{x}\cdot\mathbf{M}$	$[-0.5, 0.5]^d$	$\{0\}^d$	$f(x) = 0$ $\zeta = 10^{-10}$	Multimodal, Non-Separable, Asymmetric
Shifted Levy	$f(x) = \sin^2(\pi x) + \sum_{i=1}^{d-1}(w_i - 1)^2\left[1 + 10\sin^2(\pi w_i + 1)\right] + (w_d - 1)^2\left[1 + \sin^2(2\pi w_d)\right]$ $w_i = 1 + \frac{x_i - 1}{4}, \mathbf{x} = \mathbf{x} - \mathbf{o}$	$[-20, 20]^d$	$\{\mathbf{o}\}^d \in [-2, 2]$	$f(x) = 0$ $\zeta = 10^{-10}$	Shifted

Shifted Rastrigin	$f(\mathbf{x}) = 10d + \sum_{i=1}^{d}\left[x_i^2 - 10\cos(2\pi x_i)\right]$ $\mathbf{x} = \mathbf{x} - \mathbf{o}$	$[-5,5]^d$	$\{\mathbf{o}\}^d \in [-1,1]$	$f(x) = 0$ $\zeta = 10^{-10}$	Shifted
Shifted Sphere	$f(\mathbf{x}) = \sum_{i=1}^{d} x_i^2$ $\mathbf{x} = \mathbf{x} - \mathbf{o}$	$[-100,100]^d$	$\{\mathbf{o}\}^d \in [-5,5]$	$f(x) = 0$ $\zeta = 10^{-10}$	Unimodal, Separable, Shifted
Shifted weierstrass	$f(\mathbf{x}) = \sum_{i=1}^{d}\left(\sum_{k=0}^{k_{\max}} a^k \cos\left(2\pi b^k (x_i + 0.5)\right)\right) - d\sum_{k=0}^{k_{\max}} a^k \cos\left(2\pi b^k 0.5\right)$ $a = 0.5, b = 3, k_{max} = 20, \mathbf{x} = \mathbf{x} - \mathbf{o}$	$[-0.5,0.5]^d$	$\{\mathbf{o}\}^d \in [-0.1,0.1]$	$f(x) = 0$ $\zeta = 10^{-10}$	Multimodal, Non-Separable, Asymmetric, Shifted
Hybrid Composition Function 1	$F(\mathbf{x}) = \sum_{i=1}^{n} w_i f_i\left(\mathbf{x}/\lambda(i)\right)$ $f_{1-2}(\mathbf{x}) = 10d + \sum_{i=1}^{d}\left[x_i^2 - 10\cos(2\pi x_i)\right]$ $f_{3-4}(\mathbf{x}) = \sum_{i=1}^{d}\left(\sum_{k=0}^{k_{\max}} a^k \cos\left(2\pi b^k (x_i + 0.5)\right)\right) - d\sum_{k=0}^{k_{\max}} a^k \cos\left(2\pi b^k 0.5\right)$ $a = 0.5, b = 3, k_{max} = 20$ $f_{5-6}(\mathbf{x}) = \sum_{i=1}^{d-1}\frac{x_i^2}{4000} - \prod_{i=1}^{d}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$ $f_{7-8}(\mathbf{x}) = \sum_{i=1}^{d}\left(10^6\right)^{\frac{i-1}{d-1}} x_i^2$ $f_{9-10}(\mathbf{x}) = \sum_{i=1}^{d} x_i^2$ $\sigma_i = 1, i = 1, ..., d, \lambda = [1, 1, 10, 10, 5/60, 5/60, 5/32, 5/32, , 5/100, 5/100]$ $w_i = \exp(-\sum_{k=1}^{d} x_k^2/(2d\sigma_i^2))$, $w_i = \begin{cases} w_i, & w_i == \max(w_i) \\ w_i\left(1 - \max(w_i)^{10}\right) & w_i \neq \max(w_i) \end{cases}$	$[-5,5]^d$	0^d	$f(x) = 0$ $\zeta = 10^{-10}$	Multimodal, Non-Separable, Scalable, Asymmetric
Hybrid Composition Function 2	$F(\mathbf{x}) = \sum_{i=1}^{n} w_i f_i\left(\mathbf{x}/\lambda(i)\right)$ $f_{1-2}(\mathbf{x}) = \sum_{i=1}^{d} x_i^2$, $f_{3-4}(\mathbf{x}) = \sum_{i=1}^{d}\left(\sum_{k=0}^{k_{\max}} a^k \cos\left(2\pi b^k (x_i + 0.5)\right)\right) - d\sum_{k=0}^{k_{\max}} a^k \cos\left(2\pi b^k 0.5\right)$ $a = 0.5, b = 3, k_{max} = 20$ $f_{5-6}(\mathbf{x}) == 10d + \sum_{i=1}^{d}\left[x_i^2 - 10\cos(2\pi x_i)\right]$ $f_{7-8}(\mathbf{x}) = \sum_{i=1}^{d-1}\frac{x_i^2}{4000} - \prod_{i=1}^{d}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$ $f_{9-10}(\mathbf{x}) = \sin^2(\pi x) + \sum_{i=1}^{d-1}(w_i - 1)^2\left[1 + 10\sin^2(\pi w_i + 1)\right] + (w_d - 1)^2\left[1 + \sin^2(2\pi w_d)\right]$ $w_i = 1 + \frac{x_i - 1}{4}, \mathbf{x} = \mathbf{x} + 1^d$ $\sigma_i = 1, i = 1, ..., d, \lambda = [1, 1, 10, 10, 5/60, 5/60, 5/32, 5/32, , 5/100, 5/100]$ $w_i = \exp(-\sum_{k=1}^{d} x_k^2/(2d\sigma_i^2))$, $w_i = \begin{cases} w_i, & w_i == \max(w_i) \\ w_i\left(1 - \max(w_i)^{10}\right) & w_i \neq \max(w_i) \end{cases}$	$[-5,5]^d$	0^d	$f(x) = 0$ $\zeta = 10^{-10}$	Multimodal, Non-Separable, Scalable, Asymmetric

Shifted Hybrid Composition Function 1	$F(\mathbf{x}) = \sum_{i=1}^{n} w_i f_i\left(\mathbf{x}/\lambda(i)\right)$ $f_{1-2}(\mathbf{x}) = 10d + \sum_{i=1}^{d}\left[x_i^2 - 10\cos\left(2\pi x_i\right)\right]$ $f_{3-4}(\mathbf{x}) = \sum_{i=1}^{d}\left(\sum_{k=0}^{k_{\max}} a^k \cos\left(2\pi b^k\left(x_i + 0.5\right)\right)\right) - d\sum_{k=0}^{k_{\max}} a^k \cos\left(2\pi b^k 0.5\right)$ $a = 0.5, b = 3, k_{max} = 20$ $f_{5-6}(\mathbf{x}) = \sum_{i=1}^{d-1}\frac{x_i^2}{4000} - \prod_{i=1}^{d}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$ $f_{7-8}(\mathbf{x}) = \sum_{i=1}^{d}\left(10^6\right)^{\frac{i-1}{d-1}} x_i^2$ $f_{9-10}(\mathbf{x}) = \sum_{i=1}^{d} x_i^2$ $\sigma_i = 1, i = 1, \ldots, d, \lambda = \left[1, 1, 10, 10, 5/60, 5/60, 5/32, 5/32, , 5/100, 5/100\right]$ $w_i = \exp(-\sum_{k=1}^{d} x_k^2/(2d\sigma_i^2)), \mathbf{x} = \mathbf{x} - \mathbf{o}$ $w_i = \begin{cases} w_i, & w_i == \max(w_i) \\ w_i\left(1 - \max\left(w_i\right)^{10}\right) & w_i \neq \max(w_i) \end{cases}$	$[-5, 5]^d$	$\{\mathbf{o}\}^d \in [-0.5, 0.5]$	$f(x) = 0$ $\zeta = 10^{-2}$	Multimodal, Non-Separable, Scalable, Asymmetric,Shifted

Note that, for the Schwefel function, the global optimum $f(\mathbf{x}^*)$ will be larger than 0, due to a truncated constant (i.e., 418.9829).

$\mathbf{M} \in \mathbb{R}^{d\times d}$ denotes an orthogonal random rotation matrix.

ζ denotes the acceptance threshold of the optimal solutions.

o denotes the shifted value of the optimal solution.

For the 30 dimensional benchmark functions, the successful threshold is set to 10^{-2}.

Supplementary Table II: Configuration of various PSO methods.

Mechanismse	Method	Main Concept	Parameter Configuration	Note
multiple population	CPSO-S [20]	Instead of employing one single particle to find the optimal d-D vector, the vector is split into d components, so that each particle needs only to optimize a 1-D vector.	$c_1 = 1.49$ $c_2 = 1.49$ $w = 0.9 \rightarrow 0.4$	
	CPSO-H [20]	CPSO-H combines CPSO-S with the standard PSO. CPSO-H executes for one iteration, followed by one iteration of the standard PSO. To implement the information exchange, some of the particles in one half of the algorithm are replaced with the best solution discovered by the other half algorithm.	$c_1 = 1.49$ $c_2 = 1.49$ $w = 0.9 \rightarrow 0.4$	
	MCPSO [21]	MPSO is based on a master-slave model. The population in MPSO consists of one master swarm and several slave swarms. The slave swarms execute PSO independently while the master evolves based on its own knowledge and slave swarms.	$c_1 = 2.05$ $c_2 = 2.05$ $c_3 = 2.0$ $w = 0.9 \rightarrow 0.4$ $\phi = 0, 0.5, 1$	
	MPCPSO [22]	In each iteration, the whole population is split into two subpopulations which is eltist population and general populationbased on the performance of particle. The elitist population uses the MDCLS for population evolution while the general population adopts DSMLS for population evolution. When the algorithm falls into a local optimum in some iterations, a differential mutation operatoris introduced to change the particle.	$w = 0.729$ $c = 1.49445$ $F = 0.5$ $t = 0.5$ $\mathrm{maxT} = 5$	
Improved learning objects	CLPSO [23]	For each dimension of the particle, it would learn from the other particles instead of its own p_{best}, with a certain probability.	$m = 7$ $c = 1.49445$ $Pc_i = 0.05 \rightarrow 0.5$ $w = 0.9 \rightarrow 0.4$	
	NPSO [24]	In contrast to canonical PSO whereby the particles move forward the best positions, NPSO drives the particles far from the worst positions.	$c_1 = 2$ $c_2 = 2$	
Parameter settings	DNSPSO [25]	In DNSPSO, there are four topology states. Each particlesstate can be determined according the particle's neighbor distribution and a Markov transmitting matrix. Thus, each particles can adaptively change their parameters and learn the target according its own state.	State 1: $c_1 = c_2 = 2$ State 2: $c_1 = 2.1$, $c_2 = 1.9$ State 3: $c_1 = 2.2$, $c_2 = 1.8$ State 4: $c_1 = 1.8$, $c_2 = 2.2$	
Population topology	GCPSO [26]	The velocity update equation is replaced from the standard PSO.The new equation only involves its own P_{best} and a scaling factor.	$w = 0.72$ $f_c = 5$ $s_c = 15$	
Refined position update	DPSO [27]	Each particle has a probability to generate one new position, in order to enhance the exploration capacity.	$c_v = 0$ $c_1 = 2$ $c_2 = 2$ $w = 0.4$ $c_l = 0.001$	
Canonical	PSO [28], [29]	Standard PSO	$w = 0.7968$ $c_1 = 1.4962$ $c_2 = 1.4962$	

Local version PSO	Local PSO [28], [30]	Rather than track the global best of the whole population, the particles only track the best position of its local neighborhood.	$w = 0.7968$ $c_1 = 1.4962$ $c_2 = 1.4962$ $n = 2$	
Machine Learning Assisted	EVOLER	EVOLER consists of two successive stages. First, it perform the structured random exploration by using the sample-efficient sampling. Second, it reconstructs the representative space and identifies the attention subspace, whereby the particles are properly initialized.	$w = 0.3$ $c_1 = 2 \rightarrow 1.5$ $c_2 = 1.5 \rightarrow 2$	See Method F in the main text for the settings of s and N.
Machine Learning Assisted	EVOLER$_l$	EVOLER$_l$ also contains two successive stages, i.e. low-rank representation learning and evolutionary computation. In the second stage, the local version PSO method is adopted.	$w = 0.3$ $c_1 = 2 \rightarrow 1.5$ $c_2 = 1.5 \rightarrow 2$	See Method F in the main text for the settings of s and N.
Machine Learning Assisted	EVOLER$_d$	EVOLER$_d$ still contains two successive stages, i.e. low-rank representation learning and evolutionary computation. In the second stage, the downhill simplex method is adopted..	$\alpha = 1$ $\beta = 0.5$ $\gamma = 2$	See Method F in the main text for the settings of s and N.