Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information: Optimization of Chemical Synthesis with Heuristic Algorithms

Jialu Chen,[1] Wenjun Xu[1], Ruiqin Zhang*[1,2]

[1]Department of Physics, City University of Hong Kong, Hong Kong SAR, People's Republic of China

[2]Beijing Computational Science Research Center, Beijing 100193, People's Republic of China

*Corresponding author.

Email address: aprqz@cityu.edu.hk

S1. Supplementary Method

S1.1 Genetic algorithm

In the genetic algorithm, each value in search space is denoted by a chromosome to mimic the natural selection process of living beings. The algorithm starts with a set of chromosomes called population. New chromosomes are generated in each iteration, called a generation. The optimizing object is called fitness, and a chromosome with higher fitness is likely to be chosen for the next generation. In each iteration, two chromosomes are picked, and they switch their genetic information to generate two new chromosomes. Then, new chromosomes are modified by the mutation operation. These steps repeat until the same number as the population size is picked. Detailed steps used in this work are shown as follows.

Encoding: The search space x_i was transferred to bits of chromosomes by $\lceil \log_2 [(x_i^u - x_i^l)/p] \rceil$. The superscripts u and l denote upper and lower limits in each dimension, while p is the precision and "⌈ ⌉" indicates the ceiling function. For integers and real numbers, p values were chosen as 1 and 10^{-7}. The Gray code was used to convert numbers into bits because of its Hamming distance properties to keep bits of closer numbers similar.

Selection: Tournament selection with a size of 3 was used for choosing parents' chromosomes to generate the next generation. 3 chromosomes were randomly chosen, and the one with maximum fitness was picked as a parent's chromosome.

Crossover: Two-point crossover was used. Two points were randomly chosen, and bits between the two points were swapped between two parents' chromosomes.

Mutation: Each bit (0 or 1) was randomly changed to the counterpart with a probability of 0.005.

S1.2 Particle swarm optimization

The particle swarm optimization is based on the social behavior of the movement of organisms. The algorithm searches the optimized solution by a population of particles, and each particle denotes a point in the search space. The motion of each particle is determined by its local best-known position and the global best-known position. The position of particles moves by the following equations:

$$v_{p,d}^{k+1} = \omega v_{p,d}^{k} + c_1 r_1 \left(x_{p,d}^{loc} - x_{p,d}^{k}\right) + c_2 r_2 \left(x_d^{glo} - x_{p,d}^{k}\right) \#(S1)$$

$$x_{p,d}^{k+1} = x_{p,d}^{k} + v_{p,d}^{k+1} \#(S2)$$

In Equations S1 and S2, v, p, d, and k denote the velocity, particle identification, dimension, and iteration number. x is the position, while $x_{p,d}^{loc}$ and x_d^{glo} denote the optimal local position of the particle p and optimal global position. $r_1$and r_2 are two uniform random numbers in the range of 0 and 1 for each dimension and particle. Inertial weight ω, cognitive coefficient c_1, and social coefficient c_2 are adjustable model parameters. Here, these parameters were fixed to 0.8, 2, and 2 during the iteration process. The initial velocities were randomly generated between $x_i^l - x_i^u$ and $x_i^u - x_i^l$.

S1.3 Simulated annealing

The simulated annealing algorithm originates from annealing in metallurgy, as the particles of materials become orderly by slow cooling. Very fast simulated annealing method[1] was used with an initial temperature T_0 as 1. In each step, the current position x_d^k transfers to the next position x_d^{k+1} by the following equation:

$$x_d^{k+1} = \begin{cases} x_d^k, \ y^k \geq y^{k+1} \ and \ s > e^{(y^{k+1} - y^k)/T_k} \\ x_d^k + sgn(r - 0.5) * \left[T_k \left(1 + \frac{1}{T_k}\right)^{|2r-1|} - 1\right], else \end{cases} \#(S3)$$

y^k and y^{k+1} are the current and next value respecting to positions. The "sgn" means the sign function, and s or r is from the uniform distribution U[0, 1]. The temperature decreases with the iteration number k as:

$$T_k = T_0 e^{-c\left\lfloor \frac{k}{l} \right\rfloor^d} \#(S4)$$

The annealing process was performed every l step, and "⌊ ⌋" indicates the floor function. c and d are two parameters set as 1 to control the cooling speed.

S1.4 Descriptor of ligands

$\%V_{bur}$ (min) - 3 · HOMO-LUMO gap (eV)[2] was used as the descriptor for ligands in the used dataset. $\%V_{bur}$ (min) is the minimum value of the percent buried volume ($\%V_{bur}$) of a ligand's conformers. $\%V_{bur}$ denotes the percent of the overlap volume between a metal sphere and ligand relative to the volume of the sphere. The steric descriptor $\%V_{bur}$ (min) effectively classifies active and inactive ligands[3], and is a parameter of the empirical formula for predicting reaction outcomes[4]. All ligands exist in the *kraken* library[5], so conformers of these ligands can be directly obtained from the library, where conformers were optimized at the PBE(D3BJ)/6-31+G(d,p)[6-8] level with the SMD[9] model to present the $CHCl_3$ environment implicitly. In this library, putative metal atoms were introduced to construct ligand-metal structures, as shown in Figure S2. In detail, the metal sphere with a radius of *R* is centered on the metal atom, bonded with the ligand with a bond length *d*. In addition, the ligand's volume is the sum of non-hydrogen atoms' volumes. The Bondi radii scaled by 1.17, 3.5 Å, and 2.28 Å were chosen as the atom radii, *R*-value, and *d*-value.

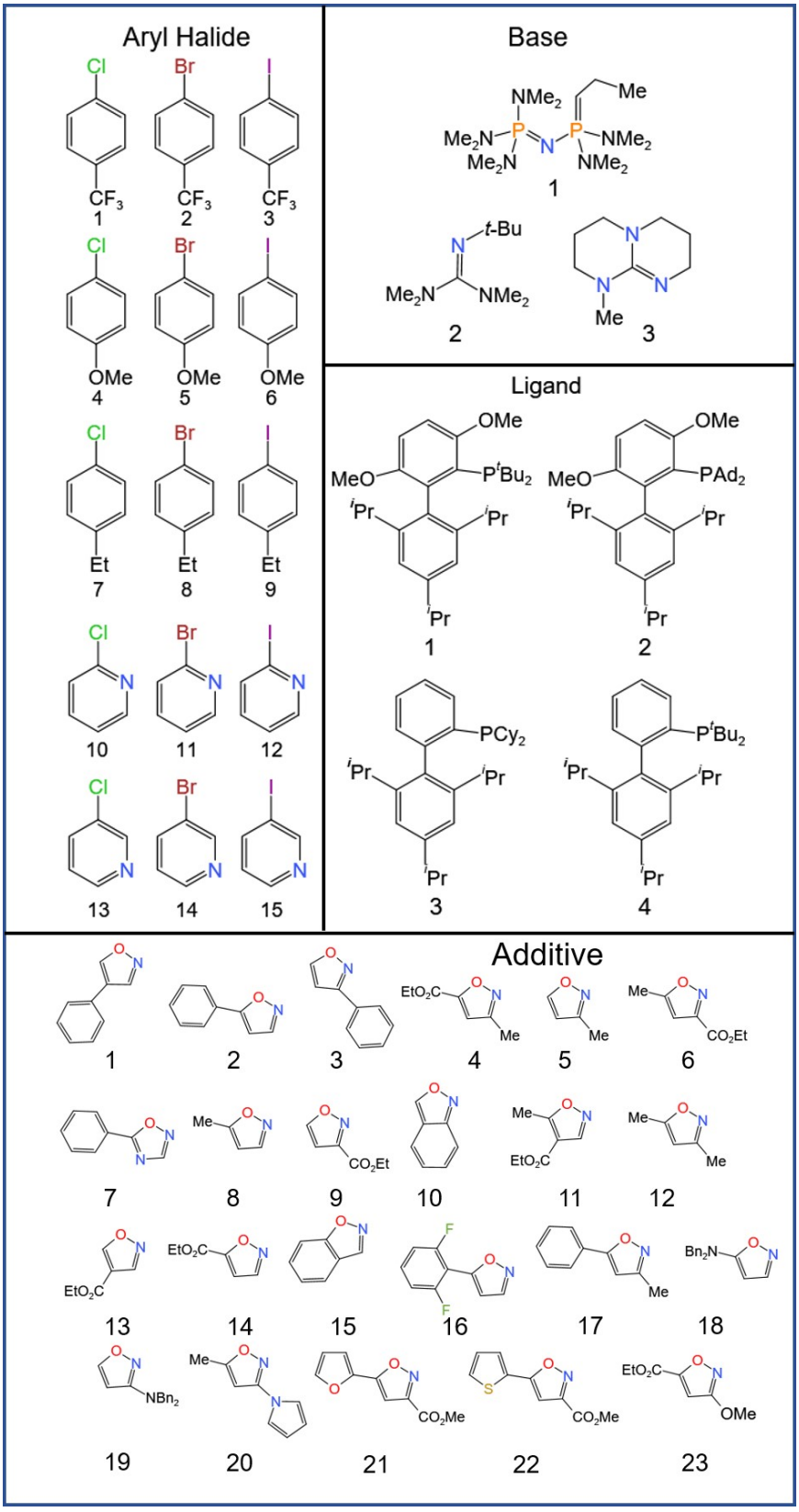

Figure S1. Chemical structures of reagents in Buchwald-Hartwig reactions.

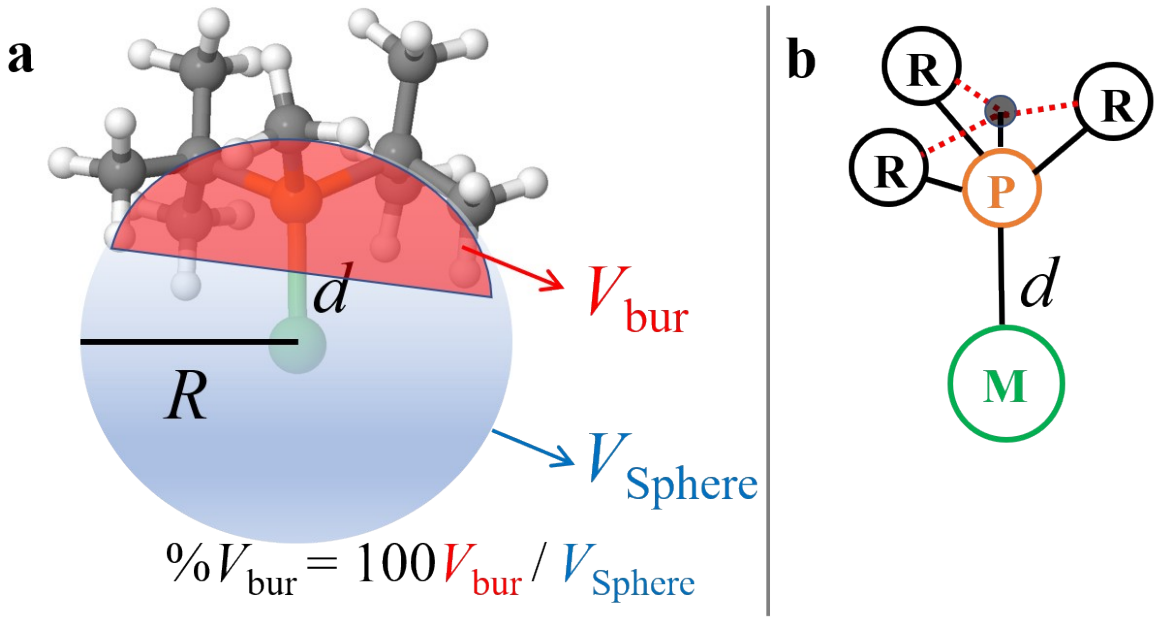

Figure S2. The definition of $\%V_{bur}$ (a). The grey, white, orange, and green balls denote C, H, P, and metal atoms, respectively. The method of locating the putative metal atom (b). The black, orange, and green circles denote atoms bonded with P, P, and metal atoms, respectively. The grey-filled circle is the center of three atoms bonded with P.

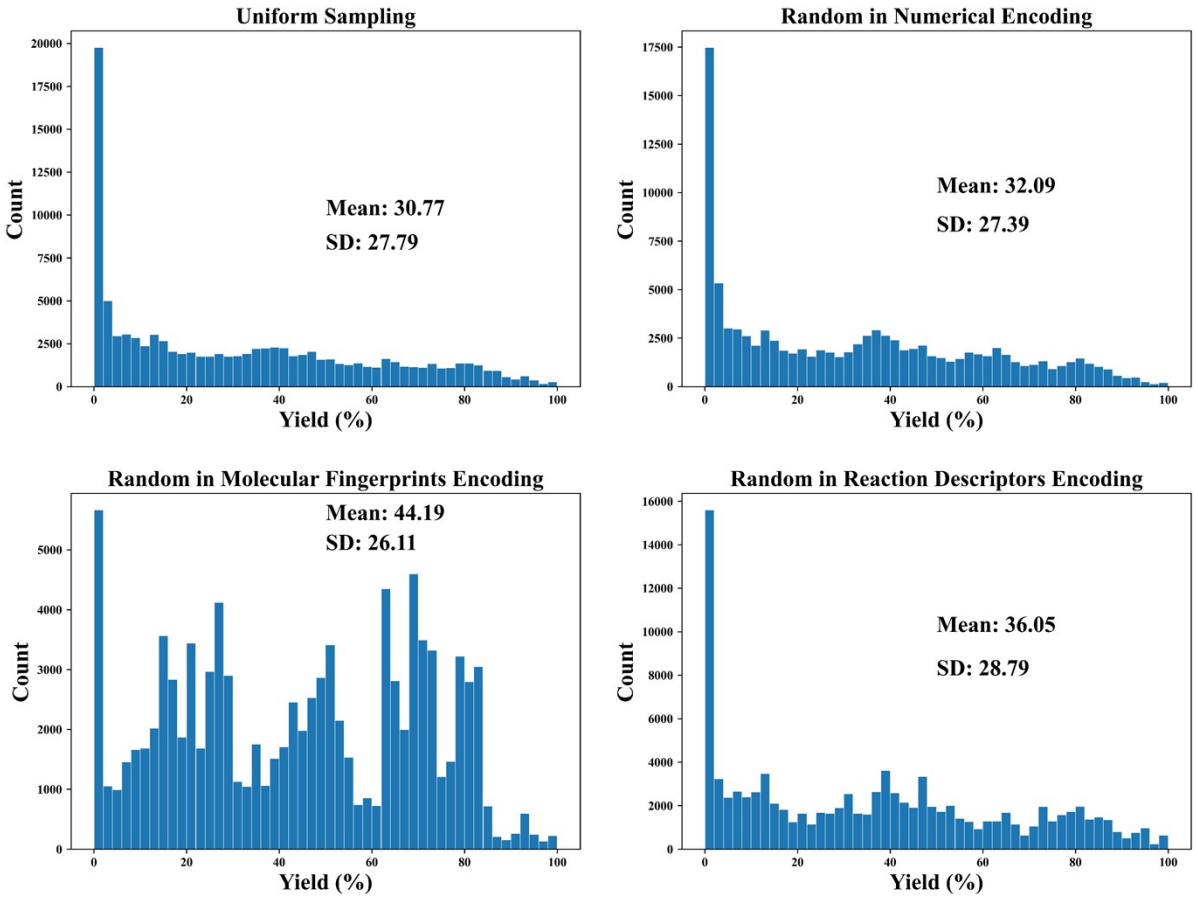

Figure S3. The distributions of yields of 10^5 samplings by four different methods. The mean and standard deviation (SD) of yields (%) are shown.

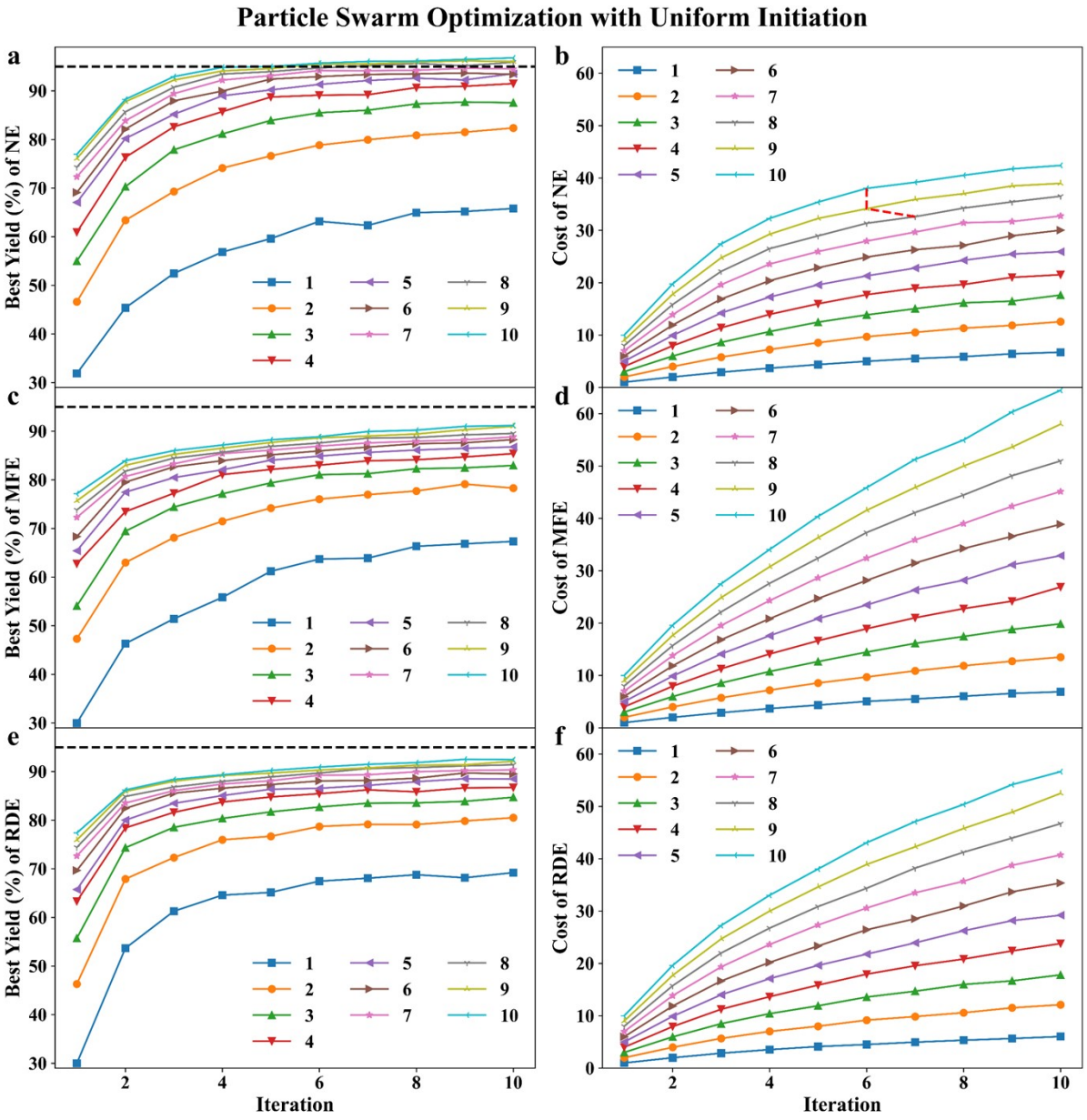

Figure S4. Best yield (%) and cost change of PSO versus the iteration with uniform initiation averaged by repeating 1000 times. The legends with different colors denote the population sizes. NE (a, b), MFE (c, d), and RDE (e, f) denote numerical, molecular fingerprint, and reaction descriptor encoding, correspondingly. The black dashed line represents the yield of 95%, and the red dashed line (b) denotes the lowest cost to reach the 95% yield in each set of parameters.

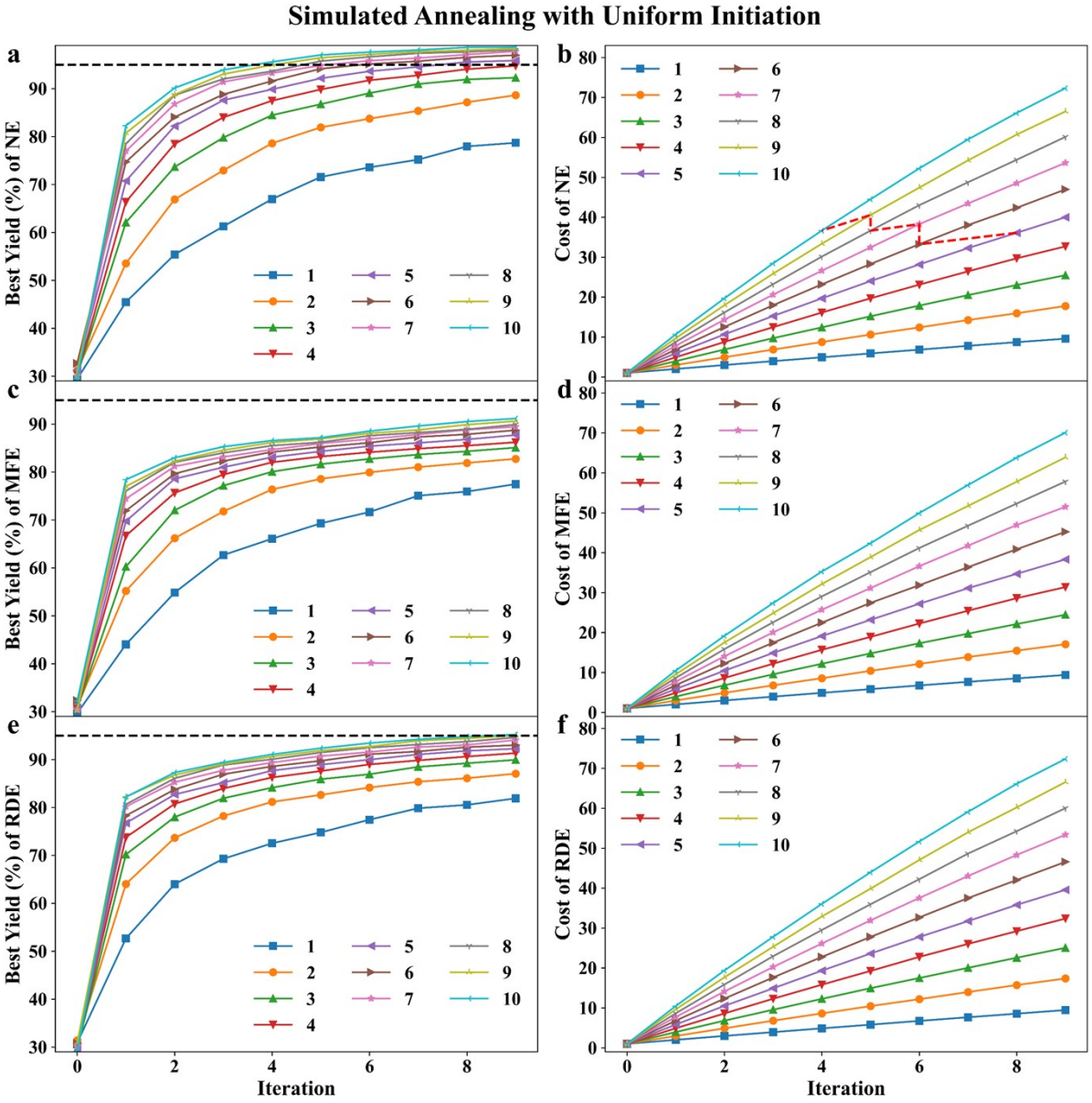

Figure S5. Best yield (%) and cost change of SA with the iteration averaged by repeating 1000 times. The legends with different colors denote the required step for decreasing temperature one time, and zero iteration means the initiation. NE (a, b), MFE (c, d), and RDE (e, f) denote numerical, molecular fingerprint, and reaction descriptor encoding, correspondingly. The black dashed line represents the yield of 95%, and the red dashed lines (b) and (f) denote the lowest cost to reach the 95% yield in each set of parameters.

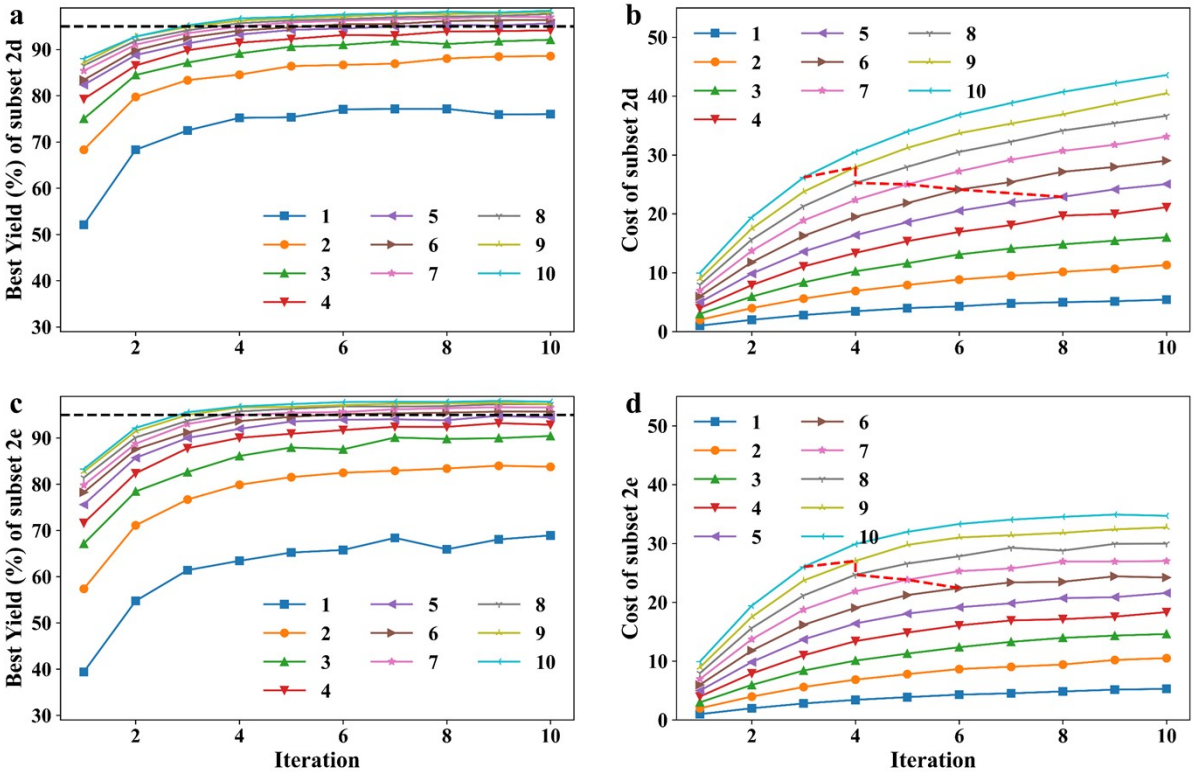

Figure S6. Best yield (%) and cost change of PSO with NE encoding versus the iteration in the subsets 2d (a, b) and 2e (c, d) by repeating 1000 times. The legends with different colors denote the population sizes. The black dashed line represents the yield of 95%, and the red dashed lines (b) and (d) denote the lowest cost to reach the 95% yield in each set of parameters.

Table S1. The reaction conditions reached a 95% yield for the Buchwald–Hartwig reactions. These numbers denote molecular structures shown in Figure S1, while 0 means no additive is added.

Base	Ligand	Aryl Halide	Additive	Yield (%)
3	2	11	4	97.29
3	2	12	0	95.42
3	2	12	1	100.00
3	2	12	2	97.57
3	2	12	3	95.39
3	2	12	4	99.62
3	2	14	0	98.04
3	2	14	2	98.73
3	2	15	0	99.69
3	2	15	2	98.29
3	3	11	4	96.59
3	3	12	0	98.03
3	3	12	1	100.00
3	3	12	4	97.95
3	3	12	5	98.85
3	3	12	6	98.18
3	3	15	0	96.15
3	4	11	4	95.56
3	4	12	2	95.07
3	4	12	4	99.03
3	4	12	5	95.75
3	4	12	6	95.68
3	4	12	9	96.13
3	4	14	3	95.13
3	4	15	0	96.92

References

1. Ingber, L., Very fast simulated re-annealing. *Math. Comput. Model.* **1989,** *12* (8), 967-973.
2. Chen, J.; Zhang, R., Volcano Plots of Reaction Yields in Cross-Coupling Catalysis. *J. Phys. Chem. Lett.* **2022,** *13*, 520-526.
3. Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Peters, E.; Gensch, T.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G., Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. *Science* **2021,** *374* (6565), 301-308.
4. Niemeyer, Z. L.; Milo, A.; Hickey, D. P.; Sigman, M. S., Parameterization of phosphine ligands reveals mechanistic pathways and predicts reaction outcomes. *Nat. Chem.* **2016,** *8* (6), 610-7.
5. Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S., A comprehensive discovery platform for organophosphorus ligands for catalysis. *J. Am. Chem. Soc.* **2022,** *144* (3), 1205-1217.
6. Grimme, S.; Ehrlich, S.; Goerigk, L., Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011,** *32* (7), 1456-1465.
7. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996,** *77* (18), 3865.
8. Hehre, W. J.; Ditchfield, R.; Pople, J. A., Self—consistent molecular orbital methods. XII. Further extensions of Gaussian—type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* **1972,** *56* (5), 2257-2261.
9. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009,** *113* (18), 6378-6396.