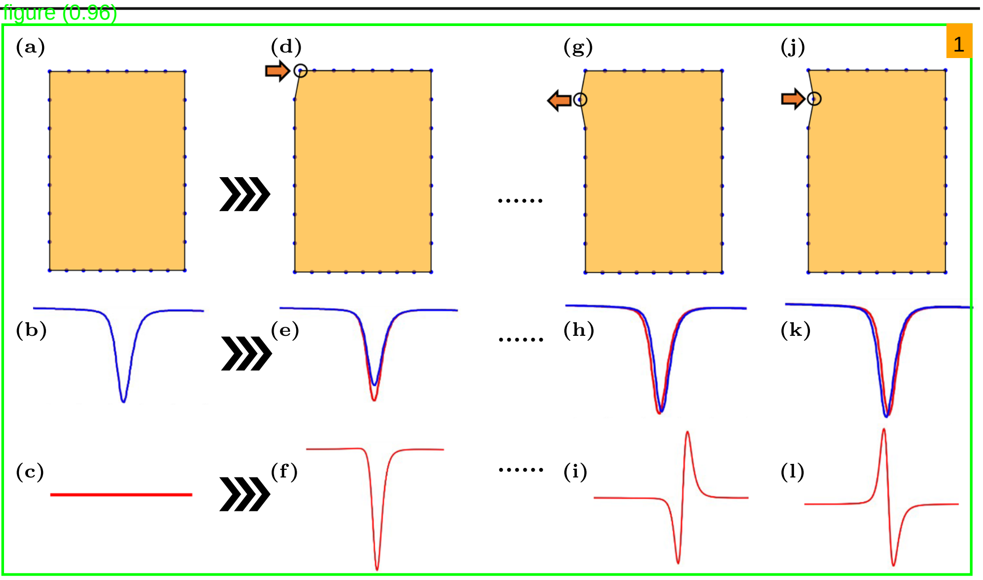
Fig. 4 S-parameter curve clustering instruction. **a-c.** The circuit grid parameter matrix in the original state is shown in a, the S_{11} curve in the original state is shown in b, and the differential S_{11} curve is shown in c. **d,g and j.** Circuit grid parameter matrices in the new state after a single grid point shift. **e,h and k.** The red curve is S_{11} in the new state, and the blue curve is S_{11} in the original state. **f,i and l.** The differential S_{11} curve



$\{r_i, r_{i+1} \ldots, r_n\}$ negative, which indicates that it is difficult for the agent to explore a better solution. An "end" action was added to a typical action cluster to avoid the disadvantages mentioned in our work. The agent can choose to end the circuit design in any state. If the agent selects the "end" action in state s_i, its reward $r_i = 0$ in state s_i and the end of an iteration.

For example, the upward shift clustering results of the S_{11} curvehas been excluded, and an "end" action has been added. The computational space requirements will be reduced from 116^n to 4^n. A clustering algorithm is added to reduce the action space of the filter to achieve fast convergence in adequate time. The space is greatly reduced in this method and is the highlight of filter design automation.

(d) Circuit Characteristic Extraction

Circuit characteristics are clustered into c different clusters of $G_i\,(i = 1, 2, ..., c)$ by using a partition-based k-means algorithm with n vectors $x_i\,(i = 1, 2, ..., n)$ as input (n is related to S_{11}). The Euclidean distance between the vector x_i of the selection group G_i and the cluster centre $C_i = a_1, a_2, a_3, ..., a_c$ is calculated as follows [43]:

$$J\left(G_i, C_i\right) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \left\|x_i - a_k\right\|^2 \tag{9}$$

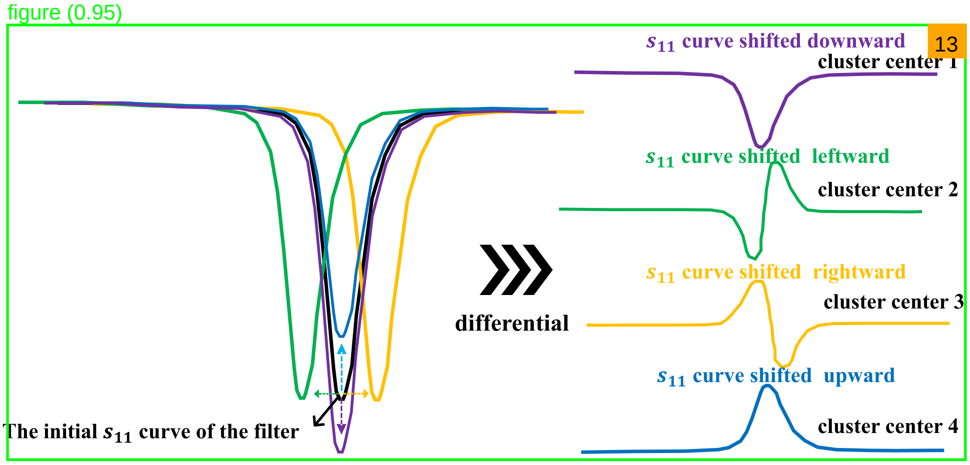
The distance from each object to each cluster centre is compared, and c objects are assigned to the nearest cluster centre C_i.

3.3.2 Reinforcement learning

The reinforcement learning part of the PAAC-K model is composed of actors and critics [42]. When the critic judges an action as beneficial, the agent to increase the probability of the action occurring. Conversely, the probability decreases.

The algorithm interacts with the environment several times. The value function $V\,(s_t; \theta_v)$ is estimated based on the environment's reward for updating the policy model $\pi\,(a_t \mid s_t; \theta)$. Each work shares one policy $\pi\,(a_t \mid s_t; \theta)$. The advantage function $A\left(s_{n,t}, a_{n,t}; \theta, \theta_v\right)$ and value Q-function

Fig. 5 Four clusters based the differential S_{11} curves


s_{11} curve shifted downward
cluster center 1
s_{11} curve shifted leftward
cluster center 2
s_{11} curve shifted rightward
cluster center 3
s_{11} curve shifted upward
cluster center 4
differential
The initial s_{11} curve of the filter