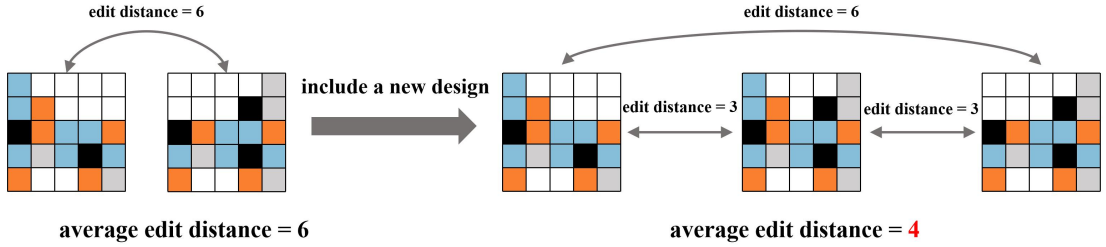
Supplementary material 9:

Further discussion on diversity measurement

Diversity is an important aspect for evaluating robotic systems and, in turn, the performance of robot design algorithms, as diversified design alternatives are crucial for handling dynamic environments and increasing robustness of robotic systems. To our knowledge, previous studies have predominantly employed two methods for quantifying diversity: (a) averaged measures of distinctiveness within a group of robots, such as per-voxel entropy [] and pair-wise edit distance []; (b) manual categorization of robot designs into distinct classes, followed by the calculation of the Simpson index, which is analogous to an entropy measure of class distribution []. The latter method becomes impractical when dealing with more abstract morphologies without clear subpopulations. The former, on the other hand, presents a paradox (**Supplementary figure 9**): including a new robot design into an existing collection can reduce diversity, even if the new design is distinct, provided that it falls within the existing distribution of this collection. Here, by "falling into the distribution" we mean that the distance between the new design and existing ones is on average smaller than that within the original collection.



Supplementary figure 9. A paradox with diversity measurement. The inclusion of a new, distinct robot design decreases, rather than increases, the diversity when solely measured as the edit distance. This is counter-intuitive as the addition of a distinct alternative should benefit diversity.

To address the above issue, we incorporate the number of distinct robot designs into our measurement as a ***correction***. Thus, our two measures -- **edit distance** (measuring the distinctiveness of evolved designs) and **the number of distinct designs** -- complement each other, providing a more comprehensive and reasonable characterization of diversity. However, we acknowledge that the weights assigned to these quantities (1.0 and 0.1) are somewhat expedient and primarily intended to bring them onto the same scale. This is based on our preliminary experiments where we found that the number of distinct high-performing designs obtained in a single run of experiments typically ranged from several dozens to around two hundred, while the edit distance is defined to range between 0 and 25. Given the lack of universally accepted metrics for measuring morphological diversity, we hope our approach could inspire future work to devise even more reasonable and comprehensive approaches.