A EXPERIMENT SETTINGS DETAILS

In this section we provide additional information about the experiments settings.

Design in the Ras Protein Family Input sequences are restricted to a maximum length of 186 with 25 possibilities (20 natural amino acids and 5 additional). We use the *esm2_t30_150M_UR50D* architecture from the ESM2 repository [2], which is made of 30 attention layers and 150 millions of parameters in total. We use the $120,000$ elements of the PFAM database to initalize the repertoire in every experiment. We use a batch size of 126 at each iteration and perform 7937 iterations to evaluate $1e6$ elements in total. Every method is run for a total of 5 trials. ME-GIDE was run for three values of target entropy: 0.4, 0.6 and 0.8. MAP-ELITES was run with different values of number of mutations at each iteration but the best results were obtained with $1-$point mutation at each iteration. ME-GDP was run for values of $\sigma_g = 0.1, 1, 10, 100, 1000$ and the best results were obtained for $\sigma_g = 100$.

Binarized digits Input images are treated as vectors of length 784 with 2 possibilites (0 or 1). We use an RBM with 500 hidden units and we train it with the contrastive divergence algorithm. We initialize the repertoire uniformly at random. We use a batch size of 512 at each iteration and perform 2000 iterations to evaluate $1,024,000$ elements in total. Every method is run for a total of 5 trials. ME-GIDE was run for three values of target entropy: 0.4, 0.6 and 0.8. MAP-ELITES was run with different values of number of mutations at each iteration and with crossover but the best results were obtained with $1-$point mutation at each iteration and no crossover. ME-GDP was run for values of $\sigma_g = 1, 10, 100$ and the best results were obtained for $\sigma_g = 10$ in value.

Discrete LSI The latent space is made of $32 \times 32 = 1024$ codes with 512 possibilites. We use a VQ-VAE which architecture is detailed in Appendix C. We initialize the repertoire uniformly at random. We use a batch size of 560 at each iteration and perform 10000 iterations to evaluate $5,600,000$ elements in total. Every method is run for a total of 5 trials. ME-GIDE was run for three values of target entropy: 0.4, 0.6 and 0.8. MAP-ELITES was run with different values of number of mutations at each iteration and with crossover but the best results were obtained with $1-$point mutation at each iteration and no crossover. ME-GDP was run for values of $\sigma_g = 1, 10, 100$ and the best results were obtained for $\sigma_g = 100$. Concerning the descriptors and objective range, the CLIP-based descriptors are scalar values ranging from 0 to 10, lower descriptor indicating stronger similarity. To compute the fitness score, we transform the score associated with the fitness prompt by applying the function $x \mapsto (10 - x) \times 10$. Thus we obtain a score ranging from 0 to 100 as displayed in the Figure 5.

B DETAILED RESULTS FOR DIFFERENT TARGET ENTROPIES

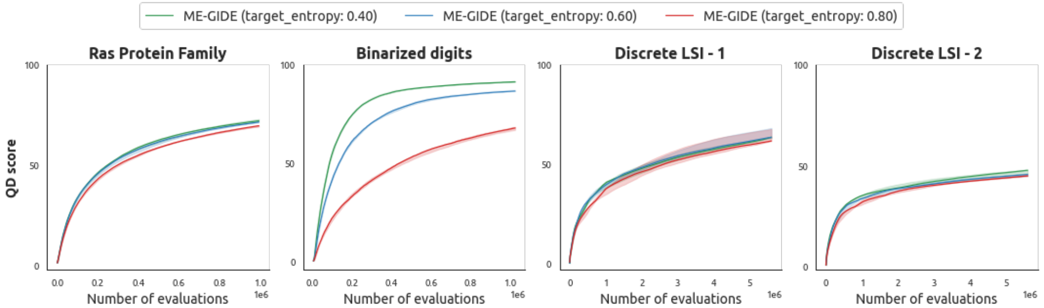


Figure 6: QD-score evolution on the different domains (median and interquartile range over 5 seeds) for different values of target entropy. On three out of the four domains, different values of target entropy yield similar results. It demonstrates that the use of a target entropy eases the hyperparameter tuning procedure.

---

[2] https://github.com/facebookresearch/esm

In all of our experiments we run our method for 3 values of entropy and only plot the best one on Figure 2. We show here that the conclusions are unchanged even with other values of target entropies, meaning that the range $[0.4, 0.8]$ seems a reasonable starting point for most problems.

C VQ-VAE ARCHITECTURE

To train our VQ-VAE, we follow guidelines of Van Den Oord et al. (2017). We train our VQ-VAE on ImageNet[3] where images are pre-processed to be of size 128×128. We use the same architecture as the authors of the aforementioned paper and use a latent vector of size $32 \times 32 = 1024$ with a codebook size of 512. We use 3 convolutional layers for the encoder and 3 layers for the decoder. We use *Adam* optimizer with a learning rate $\eta = 2\text{e} - 4$ we set the commitment loss coefficient to $\beta = 0.25$.

D VISUALIZATION OF THE MNIST DATA IN THE DESCRIPTOR SPACE

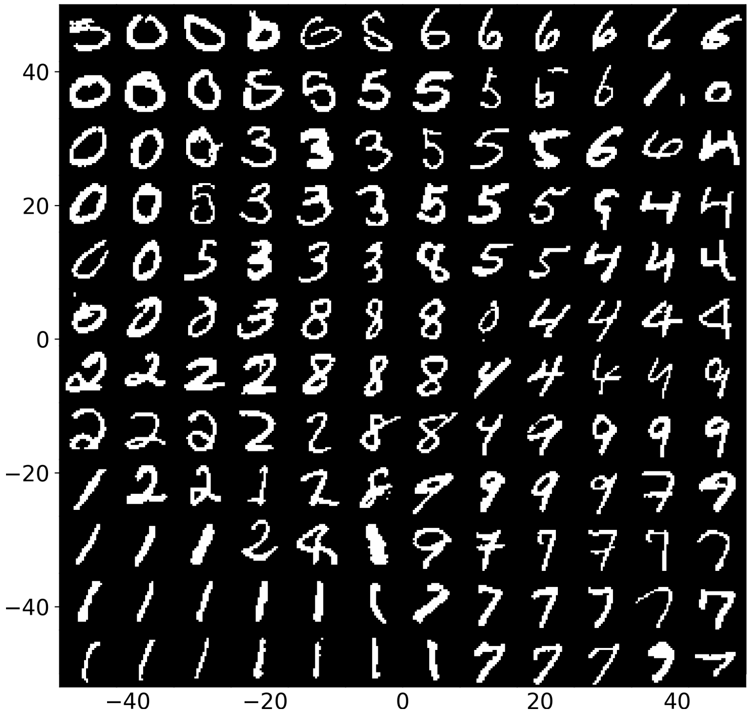


Figure 7: MNIST images projected in a T-SNE reduction of our descriptor space. The whole MNIST dataset is embedded into a the 20-dimensional space defined by the PCA over the hidden units of the RBM. Then we obtain a 2-dimensional representation with T-SNE and we sample images uniformly in this space.

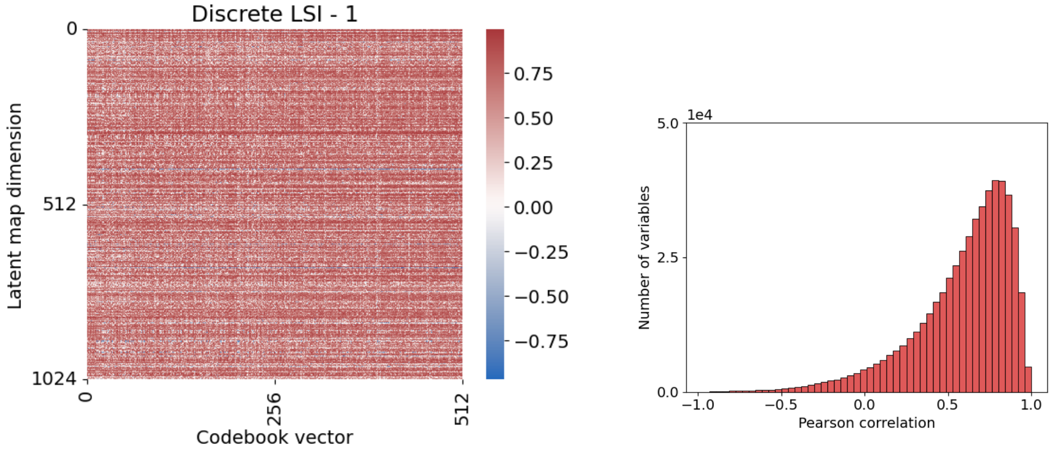
On complex data such as MNIST binary images, it is no easy task to define a relevant descriptor space. To characterize the diversity of different images of digits, one possible solution is to use the features extracted by a Deep Neural Network trained for the classification task. We instead chose to use the features implicitly embedded in the hidden layer of the RBM trained on the MNIST data. As the RBM model is trained for likelihood estimation, we expect it to learn a robust representation of the data, that efficiently separe the different classes of digits. To further validate this choice, we visualize a projection of the MNIST dataset in our embedding space. To do so, we first embed the whole dataset in the 20-dimensional descriptor space, then we project it in dimension 2 using the T-SNE algorithm. On Figure 7 we showcase MNIST images sampled uniformly in the projection space. It demonstrates the fact that our descriptor space properly spread the MNIST different classes and is able to characterize the diversity in the digits' space.

---

[3] https://www.image-net.org/

E CORRELATIONS BETWEEN GIDE ESTIMATE IMPROVEMENT AND TRUE IMPROVEMENT

To assess the quality of the approximation we make in Equation 2, which is a an estimation of the true improvement $g(x^{(i,k)}) - g(x)$ denoted $d_{i,k}$ here. For this analysis, we consider the Discrete LSI - 1 setting and we set g to the CLIP fitness function associated with our prompt "A labrador". We sample 10 random elements from the VQ-VAE latent space and we compute the true improvement $d_{i,k}$ for every possibility, leading to $5,242,880$ evaluations in total. Then we compute our estimated improvement $\tilde{d}_{i,k}$ on the same data. We display on the left of the Figure 8, the heatmap of the Pearson correlation for each dimension of the latent space $i < 1024$ and each possible position $k < 512$ of the variable $\tilde{d}_{i,k}$ and $d_{i,k}$. On the right of the same figure, we plot the histogram showing the distribution of the Pearson correlations between the $1024 \times 512 = 524,288$ variables $\tilde{d}_{i,k}$ and $d_{i,k}$, the median correlation is $m = 0.65$. Overall we measure high positive correlations over the majority of variables, which demonstrate that the approximation used by ME-GIDE contains relevant information for optimization. Interestingly, some directions are clearly better approximated than others, justifying that keeping randomness in the choice of the direction to update is sound.



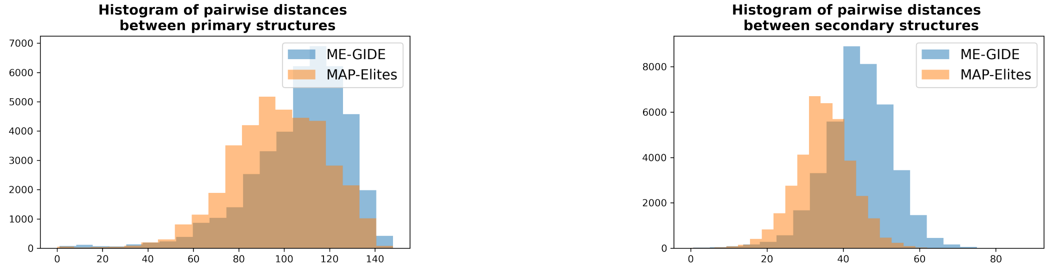
(a) Heatmap of the Pearson correlations between the 1024×512 estimates values $\tilde{d}_{i,k}$ and true improvement values $g(x^{(i,k)}) - g(x)$.

(b) Histogram of the Pearson correlations between our estimate values and the true improvement values over the $524,288$ approximated variables.

Figure 8: The gradient-informed estimate used by ME-GIDE shows a positive correlation with the true improvement $g(x^{(i,k)}) - g(x)$. It demonstrates the capability of our method to leverage gradients to determine best updates directions.

F VALIDATION ON PROTEIN DATA

We visualize the diversity obtained on proteins with two information: primary structure and secondary structure. Firstly, we sub-sample 300 the repertoire obtained with MAP-ELITES and ME-GIDE by performing a $K = 300$-means clustering on the centroids of the repertoire. Then every protein from the original repertoire is added to the new one, keeping only the most fit in every region. We evaluaete diversity in the primary structure (amino acid sequence) by computing the pairwise Levenshtein distances between each protein. We display the histograms of the distributions of pairwise distances on MAP-ELITES and ME-GIDE on Figure 9.



(a) Histogram of edit distances between sequence data over the ME-GIDE and MAP-ELITES repertoires.

(b) Histogram of edit distances between secondary structure over the ME-GIDE and MAP-ELITES repertoires.

Figure 9: ME-GIDE finds more diverse solutions in the sequence space.