Img2Mol - Accurate SMILES Recognition from Molecular Graphical Depictions

Djork-Arné Clevert,_[a] Tuan Le,[a] Robin Winter,[a‡] and Floriane Montanari [a‡]

Supplementary information

A The *Img2Mol* encoder network architecture and training procedure

The network architecture is depicted and described in Figure A.1 and consists in total of 8 convolution layers arranged in 5 stacks, followed by three fully-connected layers. We used AdamW optimizer with initial learning rate of 10^{-4}. We trained the *Img2Mol* network for 300 epochs with a batch-size of 256. The learning rate was updated and multiplied by 0:7 according to a plateau scheduler with a patience of 10 epochs with respect to the validation metric. Additionally, we applied early stopping with a patience of 50 epochs with respect to the validation metric. Throughout all training experiments, the validation metric was the loss on the validation set.

B Error analysis

To better quantify the influence of the regression accuracy of the *Img2Mol* encoder and the molecule size on the *CDDD* reconstruction accuracy, we performed the following simulation. Molecules of different sizes were randomly selected from the test data set and embedded in the *CDDD* space. Then, their *CDDD*-embedding was perturbed by increasing additive Gaussian noise (drawn from $\mathcal{N}(0, \sigma^2)$) and evaluated how often the perturbed *CDDD*-embeddings could correctly be decoded to the initial molecule. The left plot in Figure B.1 shows that the MSE of the *Img2Mol* encoder increases with the molecule size, which is obvious since the complexity of the task increases with the number of features (bonds or atoms) to be detected. The right plot in Figure B.1 clearly shows that the reconstruction accuracy of the *CDDD*-decoder decreases with the molecule size and the noise-level. However, for noise-levels of $\sigma \leq 0.15$, even large molecules can be reconstructed with an accuracy of about 90%. We postulate that this decrease in the reconstruction accuracy of the *CDDD*-decoder is a peculiarity of the *CDDD*-space. It appears that for larger molecules, the volume of decodable latent space is smaller and that the decoder is therefore more sensitive to noise perturbation for large molecules.

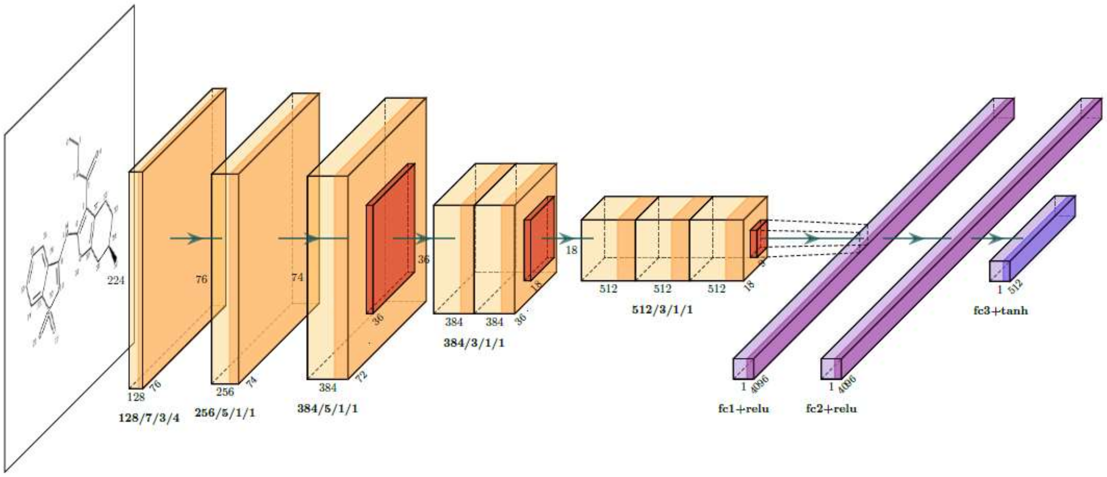

Fig. A.1 *Img2Mol* network consists in total of 8 convolution layers arranged in 5 stacks, followed by three fully-connected layers. It starts with 3 convolution layers followed by a max-pooling, then it has 1 contiguous block of 2 convolution layers followed by max-pooling, then it has 1 contiguous block of 3 convolution layers followed by max-pooling, and at last, three fully-connected layers. Each convolution layer is fully specified, the format is explained using the first layer. The expression “128/7/3/4“, is to be understood as follows: “128“ is the number of units, “7“ is the size of the convolutional filter in pixels, “3“ is the stride and stands for the number of pixels with which the filter is moved over the image and “4“ is the number of pixels with which the input is padded.

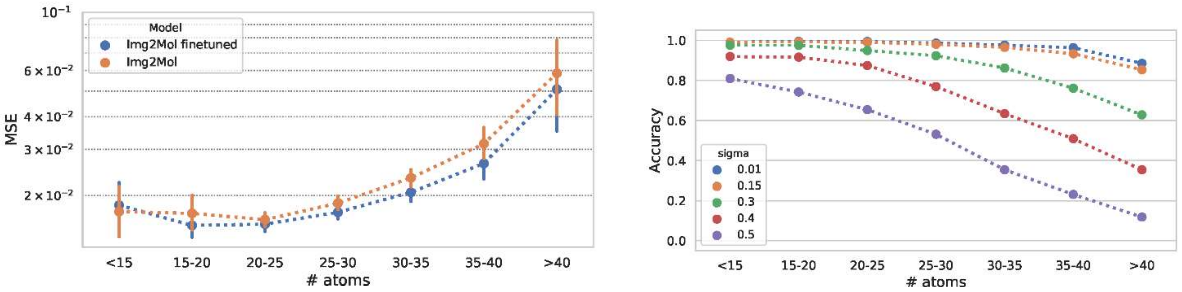

Fig. B.1 The left plot shows the mean squared error (MSE) of the *Img2Mol* network on the test benchmark data as a function of number of atoms in the molecule. The plot clearly shows that the MSE increases with the molecule size. The plot on the right shows the reconstruction accuracy of the pretrained CDDD decoder for perturbations with additive noise-levels drawn from $\mathcal{N}(0, \sigma^2)$ as function of number of atoms in the molecule.