Comparaison de différents codecs pour la compression d'image sur ADN synthétique

Claire COUVREUR[1] Michela TESTOLINA[1] Théo LADUNE[2] Pierrick PHILIPPE[2] Marc ANTONINI[1]

[1]Laboratoire I3S UMR 7271 Université Côte d'Azur, CNRS, I3S, France

[2]Orange Innovation, Cesson-Sévigné, France

Résumé – En réponse à la croissance exponentielle de la demande en stockage de données, la recherche s'est tournée vers des alternatives plus durables en remplacement des méthodes traditionnelles (e.g. HDD). Une approche prometteuse consiste à stocker des données froides, c'est-à-dire les données rarement utilisées, sur de l'ADN synthétique, qui offre une densité importante d'information, une faible consommation d'énergie ainsi qu'une longue durabilité. Ce papier explore et compare diverses techniques de compression d'image adaptées au stockage de données sur ADN, en évaluant à la fois les méthodes conventionnelles basées sur les transformations et les approches modernes basées sur l'apprentissage.

Abstract – In response to the ever-growing data storage demands, research has been increasingly focusing on more sustainable solutions as alternative to traditional storage methods (e.g. HDD). One promising approach is storing cold data —that is, data that is rarely accessed— on synthetic DNA, which offers exceptional density, low energy consumption, and long-term durability. This paper explores and compares various image compression techniques tailored for DNA-based storage, evaluating both conventional transform-based methods and state-of-the-art learning-based approaches.

1 Introduction

Au cours de ces dernières années, le besoin de stockage en données numériques a connu une augmentation significative. Les dispositifs de stockage conventionnels actuels ne parviennent pas à suivre cette demande en raison de leur durée de vie limitée ainsi que de leur coût énergétique et environnemental [8]. Une alternative, implémentée pour la première fois en 1988 [6], consiste à utiliser l'ADN synthétique comme support de stockage des données numériques. Cette solution prometteuse se démarque par sa haute densité d'information, sa faible consommation d'énergie et sa longue durabilité. Le processus de stockage de données sur ADN repose sur quatre étapes clés : l'encodage des données, la synthèse des brins d'ADN (ou écriture), le séquençage (ou lecture) de ces brins et le décodage. Contrairement aux encodages binaires conventionnels, l'encodage ADN repose sur un système quaternaire correspondant aux quatre nucléotides de l'ADN (A, C, G et T). L'état de l'art a généralement adopté deux approches pour compresser les données en séquences ADN : (1) encoder les données brutes dans un format binaire avant de les transcoder en système quaternaire [7], ou (2) encoder directement les données brutes dans un format quaternaire [3, 11].

Cet article analyse le problème de la compression d'image sur ADN synthétique, en adoptant une approche en deux étapes : l'image est d'abord compressée à l'aide d'un encodeur binaire performant, puis transcodée en un format de représentation moléculaire (fichier fasta). Le paradigme suivi est celui présenté en Fig. 1. Dans des applications réelles, le séquençage de l'ADN est sujet à du bruit moléculaire, principalement dû au processus de séquençage des brins d'ADN. Cependant, dans ce papier, nous supposons un canal sans bruit pour concentrer l'analyse uniquement sur la performance pure des codecs.

Le travail a été réalisé dans le cadre du projet PEPR MolecularXiv (ANR-22-PEXM-003) financé par France 2030.

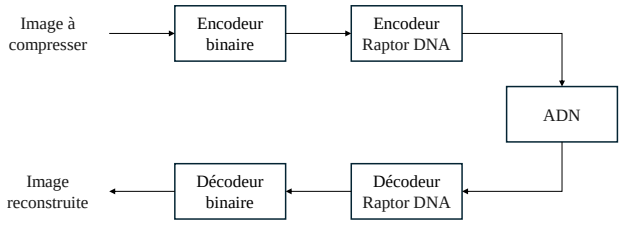

FIGURE 1 : Étape de codage/décodage. La présente architecture est une version simplifiée du modèle de vérification (VM) de JPEG DNA. Dans cet article, nous proposons de remplacer l'encodeur/décodeur binaire par soit le codeur JPEG XL, soit JPEG AI ou encore Cool-Chic.

Dans le cadre de cet article, nous étudions trois méthodes de compression d'image : JPEG XL [1], JPEG AI [2] et Cool-Chic [9].

L'encodeur JPEG XL [1], méthode conventionnelle étudiée dans cet article, a été normalisé en 2021 par le comité JPEG (Joint Photographic Experts Group) [1] et, en tant que norme JPEG, il adopte une approche de codage par transformée par blocs basée sur la transformée en cosinus discrète (DCT). L'activité JPEG DNA, quant à elle, envisage une architecture adoptant un codage binaire avec JPEG XL, suivi d'un transcodage en codes ADN quaternaires à l'aide de codes DNA Raptor [10], pour la création de la première norme de codage d'images compatible avec le stockage sur le support ADN synthétique. En particulier, l'un des flux de travail adopté par l'activité JPEG DNA, également pris en compte dans cet article, est illustré en Fig. 1.

Avec le développement du Deep Learning, le groupe JPEG s'est tourné vers de nouvelles méthodes de codage basées

[1] https://jpeg.org

sur l'apprentissage. JPEG AI [2] est la première norme de compression d'image à adopter des réseaux neuronaux pour l'encodage et le décodage bout en bout. Elle offre une efficacité de compression améliorée par rapport aux méthodes traditionnelles, avec une qualité visuelle supérieure à des débits binaires comparables. De plus, elle fournit une représentation compacte en flux unique, idéale pour les tâches de vision par ordinateur et de traitement d'images, permettant un traitement direct dans l'espace latent de JPEG AI.

Malgré les performances de débit-distorsion améliorées démontrées par les méthodes de codage d'images basées sur l'apprentissage, leur complexité de calcul reste grande. Une solution à ce problème est l'utilisation de Représentation Neuronale Implicite (INR). Le modèle est `<< overfitté >>` sur une image et évite ainsi l'utilisation de base d'apprentissage, permettant une diminution du coût de calcul et une plus grande facilité d'implémentation. Un exemple d'INR est le codec Cool-Chic (COOrdinate-based Low Complexity Hierarchical Image Codec) [9] inspiré du codeur COIN [4]. Cool-Chic représente une image à l'aide de grilles latentes multi-résolutions, converties en intensité RGB via une transformée de synthèse basée sur un réseau neuronal convolutif (CNN). Un modèle de probabilité autorégressif est utilisé pour véhiculer efficacement les grilles latentes.

Le flux binaire généré par les codecs d'image binaires peut être converti en ADN en utilisant différentes méthodes, notamment les codes DNA Raptor [10]. Cette approche permet de produire des séquences d'ADN synthétique conformes aux contraintes biochimiques, tout en offrant la flexibilité de minimiser le débit de nucléotides ou d'introduire de la redondance pour la correction d'erreurs à l'aide d'en-têtes fixes. DNA Raptor applique d'abord des codes Fountain [5] pour ajouter, si nécessaire, de la redondance à la séquence, suivis d'un simple mappage binaire-quaternaire, où les séquences résultantes sont filtrées selon un ensemble de contraintes biochimiques prédéfinies.

Cette étude vise donc à comparer les performances de trois méthodes de compression d'images binaires, à savoir JPEG XL, JPEG AI et Cool-Chic, dans le contexte de la compression en ADN synthétique en adoptant le même encodeur DNA Raptor. L'article est organisé comme suit : la section 2 présente la mise en œuvre suivie pour chaque codec et pour l'encodeur quaternaire. La section 3 présente ensuite le jeu de données utilisé ainsi que les notations et les résultats accompagnés de remarques. Enfin, la section 4 présente une revue des performances des différents codecs ainsi que des suggestions pour des travaux futurs.

2 Mise en œuvre du codeur ADN

2.1 Implémentation des codeurs binaires

La configuration adoptée pour l'étude des différents codecs est la suivante :

JPEG XL : Le modèle de vérification (VM) JPEG DNA [2] a été utilisé pour compresser les images en ADN. Plus précisément, le logiciel intègre la version 0.8.1 de la bibliothèque `libjxl` pour la compression des images, en conservant tous les paramètres par défaut et en ajustant uniquement le paramètre de qualité via l'option `-q`.

[2]URL accordée sur demande

JPEG AI : Le logiciel de référence JPEG AI [3] a été exploité pour compresser l'ensemble de données avec JPEG AI. Plus précisément, des modèles pré-entraînés optimisés pour l'erreur quadratique moyenne (MSE) et le Multi-Scale Structural Similarity Index Measure (MS-SSIM) à différentes étapes ont été utilisés dans les expériences. Le codec a été configuré en désactivant tous les outils optionnels, en sélectionnant *High Operation Point (hop)*, en exploitant l'architecture d'encodeur/décodeur la plus avancée qui inclue des blocs d'attention et des transformeurs, et en activant la fonctionnalité d'ajustement automatique du débit binaire pour optimiser la compression.

Cool-Chic : L'implémentation open-source de Cool-Chic 3.4[4] est utilisée pour en évaluer les performances de compression. La configuration d'encodeur *slow_100k* et la configuration de décodeur *hop* (1430 multiplications par pixel décodé) sont utilisées afin d'obtenir la meilleure efficacité de compression. La métrique de distorsion minimisée par ces encodages Cool-Chic est MSE dans le domaine couleur YUV444.

2.2 Réalisation du codage quaternaire

Les images ont d'abord été encodées en format binaire à l'aide des trois méthodes présentées en section 2.1. Les débits binaires ont été harmonisés entre ces méthodes afin d'assurer une évaluation équitable. Le choix de ces débits cibles a été réalisé en fonction des plages typiques d'intérêt pour les applications de stockage sur ADN, notamment pour des scénarios d'archivage à long terme sans contraintes de coût de stockage, où une qualité visuelle sans perte ou quasi sans perte est privilégiée. Par ailleurs, un ensemble d'images de qualité visuelle plus faible a été inclus pour représenter des scénarios d'application avec des limitations liées au coût de stockage. L'unité de débit nucléotide par pixel (ntpp) sera utilisée en complément de l'unité binaire. Elle est définie comme le ratio entre le nombre de nucléotides dans le fichier fasta et le nombre de pixels dans l'image d'entrée. Le ratio introduit par le DNA Raptor étant relativement constant, les débits nucléotides rapportés sont directement comparables aux débits binaires traditionnels, ce qui permet une évaluation cohérente des performances de compression dans les deux domaines.

L'intervalle de débits varie pour chaque image du jeu de données, mais tous les codecs ont été utilisés avec des débits comparables. Plus précisément, le débit binaire le plus élevé du jeu de données est de 4,17 bits par pixel (bpp), soit environ 2,28 nucléotides par pixel (ntpp), tandis que le débit médian le plus élevé est de 2,4 bpp (environ 1,31 ntpp). À l'autre extrémum, le débit binaire le plus faible du jeu de données est de 0,04 bpp (environ 0,02 ntpp), tandis que le débit médian le plus bas est de 0,1 bpp (environ 0,05 ntpp).

Pour la conversion en ADN quaternaire, nous avons utilisé l'implémentation de l'encodeur DNA Raptor, intégrée dans JPEG DNA VM. Pour notre étude, nous avons adopté le mode `-direct_transcode` permettant une transcription direct d'un fichier binaire déjà encodé en un fichier fasta. Nous avons aussi activé l'option `-probe_overhead` qui minimise l'entête, et donc le nombre de paquets utiles pour décoder le fichier. Cette option permet ainsi d'obtenir le débit minimum, mais en l'utilisant il n'y a plus de correction d'erreur possible. Ce qui s'avère pertinent dans notre contexte non bruité.

[3]`https://gitlab.com/wg1/jpeg-ai/jpeg-ai-reference-software`

[4]`https://orange-opensource.github.io/Cool-Chic`

TABLE 1 : BD-rate avec JPEG DNA VM comme codec de référence

	Codecs	MS-SSIM Y	PSNR Y	VIF Y	IW-SSIM Y	VMAF YUV
Image 0003	JPEG AI DNA	$\mathbf{-41,58\,\%}$	$-19,64\,\%$	$\mathbf{-27,55\,\%}$	$\mathbf{-43,10\,\%}$	$\mathbf{-46,12\,\%}$
	Cool-Chic DNA	$-19,06\,\%$	$\mathbf{-23,74}\,\%$	$-20,84\,\%$	$-23,35\,\%$	$-38,05\,\%$
Image 0006	JPEG AI DNA	$\mathbf{-42,37\,\%}$	$\mathbf{-22,56\,\%}$	$\mathbf{-28,54\,\%}$	$\mathbf{-40,77\,\%}$	$\mathbf{-40,07\,\%}$
	Cool-Chic DNA	$-22,71\,\%$	$-16,03\,\%$	$-21,02\,\%$	$-25,30\,\%$	$-22,76\,\%$
Moyenne	JPEG AI DNA	$\mathbf{-44,58\,\%}$	$-23,39\,\%$	$\mathbf{-30,23\,\%}$	$\mathbf{-43,65\,\%}$	$\mathbf{-45,13\,\%}$
	Cool-Chic DNA	$-25,66\,\%$	$\mathbf{-24,87\,\%}$	$-25,60\,\%$	$-27,86\,\%$	$-33,42\,\%$

3 Résultats et discussions

3.1 Jeu de données et notations

L'ensemble de données JPEG AIC-3 [12] a été utilisé pour les expériences. Il comprend 10 images aux contenus variés, incluant des objets, des visages, de la nourriture, des animaux, des scènes naturelles et des images synthétiques, avec des résolutions allant de 560×888 à 2592×1946 pixels.

Chaque codec d'image binaire, lorsqu'il est combiné avec l'encodeur Raptor en système quaternaire pour l'ADN, est désigné comme suit : la combinaison de JPEG XL avec l'encodeur Raptor, utilisée dans l'activité JPEG DNA, est appelée JPEG DNA VM. De même, la combinaison de JPEG AI avec l'encodeur Raptor est nommée JPEG AI DNA, tandis que la combinaison de Cool-Chic avec l'encodeur Raptor est désignée sous le nom de Cool-Chic DNA.

3.2 Analyse des résultats

Afin d'évaluer les performances de chaque codec, nous avons retenu les métriques de qualité objectives suivantes : MS-SSIM, PSNR, VIF et IW-SSIM toutes calculées sur le canal de luminance Y, et VMAF estimée dans l'espace colorimétrique YUV. L'évaluation de ces métriques a été réalisée à l'aide du logiciel de calcul de JPEG AI [5].

Comparaison de JPEG DNA VM avec les codecs basés sur l'apprentissage : La Table 1 recense les performances en termes de Bjøntegaard Delta (BD)-rate pour les images 0003 et 0006 du jeu de données JPEG AIC-3, ainsi que la moyenne des métriques sur l'ensemble des 10 images, en prenant le modèle de vérification JPEG DNA VM comme référence. L'image 0003 contient plusieurs zones homogènes, dont un large fond uni de couleur sombre, tandis que l'image 0006 présente un niveau élevé de texture, avec des éléments tels que des arbres et de l'eau. Les résultats montrent que les méthodes basées sur l'apprentissage, JPEG AI DNA et Cool-Chic DNA, surpassent tous deux le modèle de référence JPEG DNA VM pour l'ensemble des métriques objectives de qualité considérées.

Comparaison des codecs basés sur l'apprentissage : De manière générale, la Table 1 met en évidence que JPEG AI DNA offre des performances supérieures à Cool-Chic DNA sur la plupart des métriques, à l'exception du PSNR Y, où Cool-Chic DNA présente une amélioration de 1,5 % en moyenne sur l'ensemble du jeu de données. Cela peut s'expliquer par le fait que Cool-Chic est optimisé pour MSE, ce qui conduit à de meilleurs résultats pour cette métrique spécifique. Nous retrouvons des conclusions similaires en analysant les résultats des images 0003 et 0006 : les performances des deux méthodes basées sur l'apprentissage restent globalement constantes d'une image à l'autre, à l'exception de la métrique PSNR Y. Dans ce cas-là, Cool-Chic DNA affiche de meilleures performances que JPEG AI DNA pour l'image 0003 (qui contient principalement des basses fréquences), mais pas pour l'image 0006 (qui contient principalement des hautes fréquences).

Comparaison par gamme de débits des trois codecs : La Fig. 2 présente les courbes débit-distorsion pour les images 0003 et 0006 du jeu de données JPEG AIC-3. Afin d'améliorer la lisibilité, les résultats du MS-SSIM Y sont exprimés en dB selon l'équation suivante : $MS\text{-}SSIM(dB) = -10\log_{10}(1 - MS\text{-}SSIM)$.

Aux faibles débits, c'est-à-dire en dessous de 0,4 ntpp, les résultats des deux métriques de qualité objective indiquent que JPEG DNA VM présente généralement des performances inférieures à celles des deux codecs basés sur l'apprentissage. Il est important de noter que le codec JPEG XL a été conçu principalement pour des applications de haute qualité, ce qui explique ses performances plus faibles aux bas débits. Selon le MS-SSIM Y, JPEG AI DNA présente de meilleures performances que Cool-Chic DNA, mais cette différence n'est pas visible lorsque l'on considère le PSNR Y.

Dans la plage de qualité intermédiaire (entre 0,4 et 1 ntpp), l'écart de performance entre les codecs semble rester stable lorsqu'on considère le MS-SSIM Y. En revanche, selon la métrique PSNR Y, un changement de tendance est observé : Cool-Chic DNA semble surpasser ou obtenir des résultats comparables à JPEG AI DNA. Il convient de noter que pour ces débits de ntpp, les valeurs de PSNR Y dépassent 40 dB, ce qui suggère que les différences visuelles entre les images pourraient être à peine perceptibles par l'œil humain. Par conséquent, une analyse plus approfondie prenant en compte la qualité visuelle perçue des images est nécessaire pour une discussion plus précise.

Aux plus hauts débits analysés dans cette étude (au-delà de 1 ntpp), JPEG DNA VM obtient souvent de meilleures performances en termes de métriques objectives que les deux méthodes basées sur l'apprentissage. Dans cette plage de qualité, Cool-Chic DNA présente des performances comparables à celles de JPEG DNA VM. En revanche, les valeurs de PSNR Y pour JPEG AI DNA semblent atteindre un plateau aux plus hauts débits, une tendance particulièrement visible pour l'image 0003 du jeu de données.

[5] https://gitlab.com/wg1/jpeg-ai/jpeg-ai-qaf

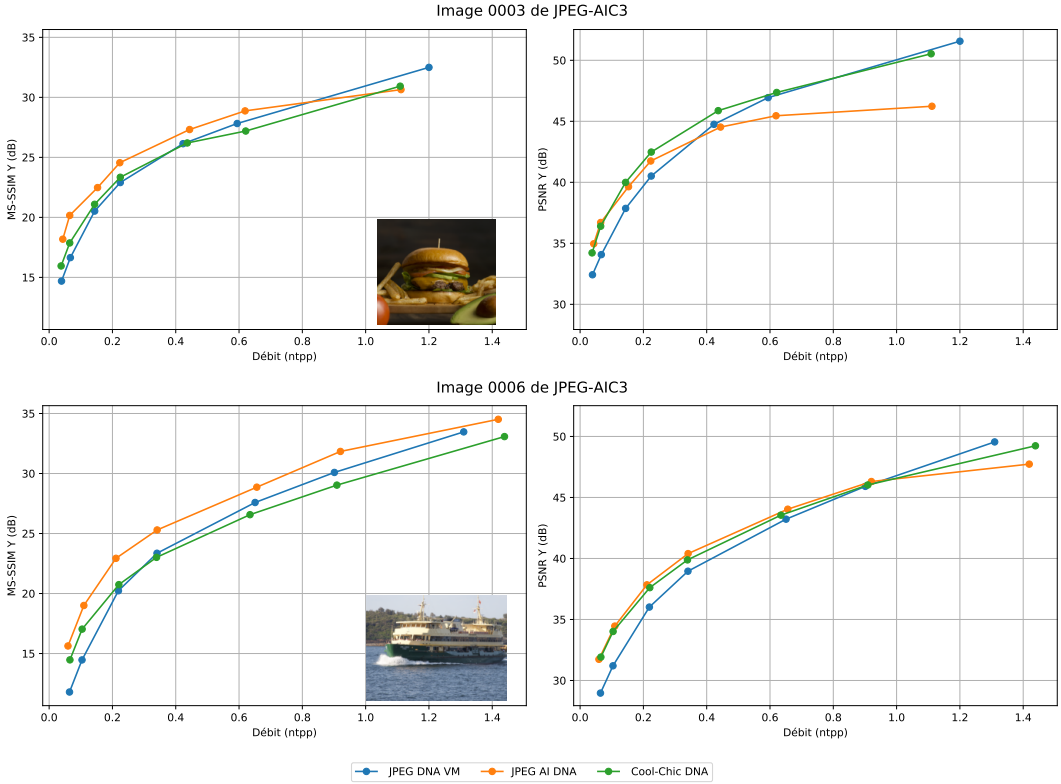

FIGURE 2 : Courbes de qualité en fonction du débit pour les images 0003 et 0006 du jeu de données JPEG AIC-3.

4 Conclusion

Dans cet article, nous avons comparé différents codecs de compression d'images dans le contexte du stockage de données basé sur l'ADN. Nos résultats suggèrent que les méthodes basées sur l'apprentissage surpassent les approches conventionnelles, représentées ici par JPEG XL, avec une différence particulièrement marquée pour les faibles débits. Parmi les méthodes testées, JPEG AI semble offrir de meilleures performances que Cool-Chic dans le régime des faibles débits. Toutefois, à des débits plus élevés, la comparaison reste incertaine, indiquant la nécessité d'études supplémentaires, en particulier sur l'intervalle de haute qualité. De plus, il convient de noter que, contrairement aux méthodes de compression basées sur l'apprentissage comme JPEG AI, qui nécessitent des procédures de réentraînement longues et coûteuses, Cool-Chic permet d'optimiser n'importe quelle métrique de qualité d'image objective directement lors de la phase d' `<< overfitting >>` pendant l'encodage.

Les travaux futurs pourraient étendre cette analyse en intégrant une comparaison visuelle des images reconstruites et en évaluant la complexité computationnelle des différentes méthodes. De plus, une version optimisée de Cool-Chic en MSE et MS-SSIM pourrait être envisagée afin d'examiner plus en détail son potentiel dans ce contexte.

Références

[1] Jyrki ALAKUIJALA, Ruud VAN ASSELDONK, Sami BOUKORTT *et al.* : JPEG XL next-generation image compression architecture and coding tools. *In Applications of digital image processing XLII*, volume 11137, pages 112–124. SPIE, 2019.

[2] Elena ALSHINA, João ASCENSO et Touradj EBRAHIMI : Jpeg ai : The first international standard for image coding based on an end-to-end learning-based approach. *IEEE MultiMedia*, 31(4):60–69, 2024.

[3] Melpomeni DIMOPOULOU, Marc ANTONINI, Pascal BARBRY et Raja APPUSWAMY : A biologically constrained encoding solution for long-term storage of images onto synthetic DNA. *In 2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

[4] Emilien DUPONT, Adam GOLIŃSKI, Milad ALIZADEH, Yee Whye TEH et Arnaud DOUCET : COIN : COmpression with Implicit Neural representations, avril 2021. arXiv :2103.03123 [eess].

[5] Yaniv ERLICH et Dina ZIELINSKI : DNA fountain enables a robust and efficient storage architecture. *Science*, 355(6328):950–954, 2017.

[6] Andy EXTANCE : How DNA could store all the world's data. *Nature*, 537(7618):22–24, septembre 2016. Publisher : Nature Publishing Group.

[7] Nick GOLDMAN, Paul BERTONE, Siyuan CHEN, Christophe DESSIMOZ, Emily M LEPROUST, Botond SIPOS et Ewan BIRNEY : Towards practical, high-capacity, low-maintenance information storage in synthesized dna. *nature*, 494(7435):77–80, 2013.

[8] Min GU, Xiangping LI et Yaoyu CAO : Optical storage arrays : a perspective for future big data storage. *Light : Science & Applications*, 3(5):e177–e177, 2014.

[9] Théo LADUNE, Pierrick PHILIPPE, Félix HENRY, Gordon CLARE et Thomas LEGUAY : COOL-CHIC : Coordinate-based low complexity hierarchical image codec. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13515–13522, 2023.

[10] Davi LAZZAROTTO, Jorge ENCINAS RAMOS, Michela TESTOLINA et Touradj EBRAHIMI : Storing images and point clouds on DNA support with fountain codes. *In* Andrew G. TESCHER et Touradj EBRAHIMI, éditeurs : *Applications of Digital Image Processing XLVII*, page 39, San Diego, United States, septembre 2024. SPIE.

[11] Trung Hieu LE, Xavier PIC, Jeremy MATEOS et Marc ANTONINI : Implicit neural multiple description for DNA-based data storage. *In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8000–8004. IEEE, 2024.

[12] Michela TESTOLINA, Vlad HOSU, Mohsen JENADELEH, Davi LAZZAROTTO, Dietmar SAUPE et Touradj EBRAHIMI : JPEG AIC-3 Dataset : towards defining the high quality to nearly visually lossless quality range. *In 2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 55–60. IEEE, 2023.