

Computer generation of fruit shapes from DNA sequence

M. Pérez-Enciso^{1,2}, C. Pons³, A. Granell³, S. Soler³, B. Picó³, A.J. Monforte⁴, L.M. Zingaretti¹

¹ Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, 08193 Bellaterra, Spain.

⁸ ² Catalan Institute for Research and Advanced Studies (ICREA), Passeig de Lluís Companys,
⁹ 23, 08010 Barcelona, Spain.

³ Instituto de Conservación y Mejora de la Agrodiversidad Valenciana (COMAV-UPV),
Universitat Politècnica de València, 46011 València, Spain.

12 ⁴ Instituto de Biología Molecular y Celular de Plantas (IBMCP), Consejo Superior de
13 Investigaciones Científicas (CSIC), Universitat Politècnica de València, 46022 València
14 Spain.

15

16 Correspondence:

17 M. Pérez-Enciso

18 CRAG, 08193 Bellaterra, Spain

19 mperezenciso@gmail.com

20

21 **Running title:** Computer generation of fruit shapes from DNA

22

23 **Keywords:** Cucurbits, Deep learning, Fruit shape, Morphometrics, Genomic prediction,
24 Tomato

25

26 **Abstract**

27 The generation of realistic plant and animal images from marker information could be a main
28 contribution of artificial intelligence to genetics and breeding. Since morphological traits are
29 highly variable and highly heritable, this must be possible. However, a suitable algorithm has
30 not been proposed yet. This paper is a proof of concept demonstrating the feasibility of this
31 proposal using ‘decoders’, a class of deep learning architecture. We apply it to Cucurbitaceae,
32 perhaps the family harboring the largest variability in fruit shape in the plant kingdom, and to
33 tomato, a species with high morphological diversity also. We generate Cucurbitaceae shapes
34 assuming a hypothetical, but plausible, evolutive path along observed fruit shapes of *C. melo*.
35 In tomato, we used 353 images from 129 crosses between 25 maternal and 7 paternal lines for
36 which genotype data were available. In both instances, a simple decoder was able to recover
37 expected shapes with large accuracy. For the tomato pedigree, we also show that the
38 algorithm can be trained to generate offspring images from their parents’ shapes, bypassing
39 genotype information. Data and code are available at
40 <https://github.com/miguelperezenciso/dna2image>.

41

42 **Introduction**

43 Shape and color patterns of animals and plant fruits are not only aesthetic features, but they
44 also convey essential information on animal welfare or fruit quality and can be critical for
45 consumer appreciation. Besides, plant and animal appearance have played a major role ever
46 since domestication and many breeds and plant varieties were created based on morphology.
47 Even today, breeders’ associations can spend much time in defining the ‘racial standard’.
48 Often, domestication and breeding have untapped a range of shapes that was not present in
49 the wild. The variability in morphology and colors in the dog is amazing compared to that of
50 its wild ancestor the wolf. In plants, domestic squashes and gourds exhibit an enormous
51 diversity in shapes whereas its wild counterparts produce small, rounded fruits only
52 (Xanthopoulou *et al.* 2019). Today, dairy bull catalogs, a business worth millions of euros
53 worldwide, usually present a picture of the bull in addition to its genetic evaluation. Bull
54 catalogs usually include information on a ‘global’ conformation score that is an important
55 part of the genetic value of the bull, and an indication of longevity. In many vegetables
56 breeding programs, experienced breeders rely on their ‘eye’ to quickly discard unpromising
57 experimental crosses.

58

59 Shape is easily modified by artificial selection and, unsurprisingly, has received much
60 attention from the genetics community and the breeding industry (Tanksley 2004; Monforte
61 *et al.* 2014). Tomato is perhaps the best studied species from a morphological point of view;
62 numerous quantitative trait loci (QTL) and some causative genes affecting shape have been
63 identified (Monforte 2014; Snouffer *et al.* 2020). Cucurbitaceae in turn have been less well
64 studied, yet they allegedly display the largest morphological variability in the plant kingdom
65 (Paris 2001). For instance, a whole sequencing effort of the different *C. pepo* morphotypes
66 did reveal numerous SNP differences but no clear clue on causative loci for shape
67 (Xanthopoulou *et al.* 2019).

68

69 The statistical analysis of shape has a long history in Evolution, which has fostered most of
70 the analysis tools available today (Zelditch *et al.* 2004; Claude 2008; Klingenberg 2010).
71 Traditional morphometrics is based on the analysis of summary statistics such as length,
72 width, ratios, and areas (Brewer *et al.* 2006). Modern morphometrics, in turn, is based on the
73 concept of 'landmarks' (Zelditch *et al.* 2004). A landmark is an anatomical position that can
74 be identified in all samples, e.g., the tip of the nose in cattle. In landmark-based geometric
75 morphometrics, the spatial information is contained in landmark coordinates. Shapes can then
76 be compared once a common reference scale is found. This can be done via Generalized
77 'Procrustes' Analysis (GPA, Gower 1975), which consists of finding an optimal
78 superimposition of several shapes such that distances between them are minimized.

79

80 In breeding, morphology research has focused so far on detection of quantitative trait loci
81 (QTL) of shape-derived statistics (e.g., Monforte *et al.* 2014). These QTL often explain only
82 a percentage of observed variability. This is not unexpected; a large body of literature shows
83 that significant loci identified from genomewide association studies (GWAS) explain but a
84 small percentage of genetic variability in complex traits (Wood *et al.* 2014; Robinson *et al.*
85 2017; Visscher *et al.* 2017). Therefore, GWAS is not optimum for prediction. An alternative
86 is to use all markers for prediction of some of the shape metrics (Tong *et al.* 2022).
87 Nevertheless, shape is highly dimensional, and the QTL or genomic prediction approaches
88 restrict the list of potential candidate genes by focusing on single univariate statistics. In
89 addition, these summary statistics do not allow reconstructing the original shape and hampers
90 the prediction of global appearance changes induced by selection.

91

92 Here we approach this issue from a holistic, opposite angle. We propose to reproduce
93 expected shapes and textures that would result from a given individual's DNA sequence. To
94 that end, we explore algorithms based on deep learning tools. Note that, in contrast to
95 standard descriptors of shape, the goal here is prediction given new DNA information rather
96 than QTL search. Breeding is mainly concerned with prediction of future offspring
97 performance and this proposal aligns with this target. This novel theoretical framework can
98 have an important impact in breeding.

99

100 This paper is a proof of concept that the proposed approach is feasible, at least in simplified
101 scenarios. We use a class of deep architectures, called 'decoders', to reproduce the expected
102 shapes given a linear vector of causative polymorphisms and random SNPs. First, we show
103 how a trained decoder is able to generate simple geometric forms (2D and 3D ellipses)
104 followed by more realistic applications in cucurbitas and tomato fruits. We end by showing
105 that, provided shapes are inherited through an 'additive' mechanism, the algorithm can
106 predict offspring shapes based on parents' images, bypassing genotype information. More
107 sophisticated algorithms would be needed if shapes are not inherited 'additively'.

108

109 **Material and methods**

110 *Generation of simple 2D and 3D images*

111 We first performed a simple experiment using 2D ellipse and 3D ellipsoid shapes to verify
112 that the proposed architecture is useful. An ellipse can be defined by the lengths of its
113 horizontal (x) and vertical (y) axes, plus a third axis z for 3D shapes. We drew 2D ellipses
114 with cv2.ellipse() function from OpenCV python package (Bradski 2000) randomly varying x
115 and y axis lengths, that is, ellipses differed in shape, size, and orientation. Images were black
116 and white of size 64 x 64 pixels. The decoder network (described below) was trained using an
117 input vector containing x/y ratio and ellipse size, i.e., the two 'causative loci', and 100
118 random uniformly distributed variables. The 100 random numbers were aimed at representing
119 noise from DNA information that is unrelated to the 'phenotype', i.e., the image containing
120 the ellipse.

121

122 We generated 3D ellipsoids as three-dimensional binary arrays using pymrt package (Metere
123 and Möller 2017), array size was 32 x 32 x 32. As in the previous example, images were
124 predicted from x, y, and z axes lengths plus 100 random uncorrelated variables. For
125 representation of the 3D shapes, ellipsoid projections were drawn using the marching cubes

126 algorithm as implemented in skimage (van der Walt *et al.* 2014) and the plot_trisurf package.
127 However, since these 3D plots were not too accurate, we also plotted the ellipsoid sections
128 across the x, y, and z axes. Both observed and predicted shapes were plotted.

129

130 *Cucurbit shapes*

131 *C. pepo* fruits can adopt an enormous diversity of shapes (Figure 1A). This variability
132 appeared only after domestication, since all wild fruits are small and round (Paris 1986).
133 According to (Paris 1989, 2001), *C. melo* shapes may have followed several evolutive
134 pathways. One pathway would be wild gourd (akin to pumpkin shape) → scallop → acorn; a
135 second pathway would be wild gourd → marrow → straightneck → zucchini → cocozelle
136 (Figure 1B). See also Figure 17 in (Paris 1989). We extracted contours from the
137 ‘contours.png’ file, based in (Paris 1989) and available in GitHub
138 (<https://github.com/miguelperezenciso/dna2image/blob/main/images/contours.png>), using
139 OpenCV library (Bradski 2000). Contours were centered and 500 pseudo-landmarks were
140 obtained with the algorithm in Zingaretti *et al.* (2021). Next, contours were aligned with a
141 generalized procrustes algorithm implemented in python package ‘procrustes’ (Meng *et al.*
142 2022) and images were resized to 64 x 64 pixels.

143

144 To generate *C. pepo* shapes along the putative evolutive gradient, we first sampled a random
145 number from a uniform distribution $s \sim U(-1, 1)$, where $s = -1$ defines an ‘acorn’ form; 0, a
146 ‘pumpkin’, and 1 corresponds to ‘cocozelle’ (Figure 1B). Using the sampled s value, the two
147 closest basic shapes were identified, and we defined a function that drew an intermediate
148 shape between the nearest basic shapes, weighted by the proximity to each of the bounding
149 contour (Figure 1C, see code in GitHub
150 <https://github.com/miguelperezenciso/dna2image/blob/main/dna2img.cucurbita.ipynb>). The
151 fruit corresponding to shape s was drawn in a 64 x 64 pixel image and noise was added to
152 mimic rugosity of naturally observed fruits. This was done by adding an autoregressive noise
153 to the contour (see code in GitHub). The decoder was trained using the ‘true’ s value and 100
154 random uncorrelated variables as input and the cucurbit shape images as output; 1,000 images
155 were used for training and 100 for testing.

156

157 *Tomato shapes from experimental crosses*

158 We used 353 tomato images from 129 crosses between 25 traditional varieties and 7 modern
159 inbreds (Table S1). Traditional varieties were a subset of the TRADITOM project, which

160 collected a wide sample of traditional tomato varieties from Southern Europe (Pons *et al.*
161 2022; Blanca *et al.* 2022). Longitudinal cuts from about three fruits per parental or crossed
162 line were photographed. Fruit images were segmented using a cluster algorithm ($k=3$) and
163 contours were identified using a thresholding algorithm, as implemented in openCV.
164 Contours were centered, cropped, and resized to 128 x 128 pixel binary images.

165

166 Modern inbred and traditional varieties were genotyped by sequence (GBS) previously
167 (Blanca *et al.* 2022). Sixty eight segregating SNPs located within fruit shape candidate genes
168 (Pons *et al.* 2022) were extracted. Hybrid offspring GBS genotypes were inferred from their
169 parental genotypes. In addition, 48 biochemical, color and morphological metrics obtained
170 with tomato analyzer had been obtained from each of the hybrid tomato fruits (Pons *et al.*
171 2022) were also used for prediction. These metrics were not available for the 32 founder lines
172 and were inferred with linear regression assuming additivity. This was done separately for
173 each metric. The final network was trained using the 116 (68 + 48) 'DNA' measures as input
174 for each of the accessions and the 353 tomato images as output. Input values were the same
175 for images pertaining to the same accession.

176

177 *Shape prediction*

178 We used a simple decoder architecture made-up of a first fully connected layer, followed by a
179 reshaping layer and by three transposed convolutional layers (Figure 2). Code was
180 implemented in keras and tensorflow (<https://keras.io/>, Abadi *et al.* 2015; Chollet 2015) and
181 is inspired in autoencoder architectures (Brownlee 2019; Chollet 2021). The same decoder
182 architecture was used for ellipse, cucurbita or tomato shape prediction, except that layer
183 dimensions were adjusted according to image size (Figure 2). For ellipsoid 3D predictions,
184 3D transposed convolution layers were used instead of 2D transposed convolutions, but
185 architecture was otherwise identical (see code in
186 <https://github.com/miguelperezenciso/dna2image>).

187

188 *From phenotype to phenotype*

189 Modern phenomics has sparked interest in 'phenomic selection', which consists in replacing
190 genotyping by high throughput phenotyping to predict future offspring performance (Rincent
191 *et al.* 2018; Cuevas *et al.* 2019; Robert *et al.* 2022). Here we considered two scenarios. In the
192 first scenario we predicted 2D ellipses given two 'parents' ellipses. To do that, we first need

193 to specify inheritance rules for images. Four arbitrary ‘image inheritance’ actions were
194 defined:

195

196 - Additivity: the ‘offspring’ ellipse x and y coordinates are obtained by averaging
197 coordinates of ‘parent’ ellipses.

198 - Dominance: for any pair of parent coordinates, the maximum of the two coordinates is
199 selected as offspring coordinate.

200 - Imprinting: the offspring ellipse is identical to the first parental ellipse.

201 - Epistasis: the offspring ellipse is drawn by swapping the x and y coordinates of an
202 ellipse intermediate between parents’ coordinates. That is, the epistatic offspring
203 ellipse is a transposed additive ellipse.

204

205 We generated ~ 1,000 ellipse trios for each inheritance pattern to train the network. We
206 trained the network for each inheritance pattern separately.

207

208 In the second, more realistic scenario, we used all combinations of male, female and
209 offspring tomato images in a given cross from the previously described dataset. This resulted
210 in a dataset of 2,325 tomato image trios. We utilized the same autoencoder architecture in
211 both ellipse and tomato scenarios. Input consisted of two images that fed two separate CNN
212 layers, one for each parental image, that were next concatenated (Figure 3).

213

214 **Results**

215 *Shape prediction*

216 We first show, as proof of concept in a toy example, that the simple decoder architecture in
217 Figure 2 is able to learn and generate 2D and 3D simple forms from ‘genotype’ data. To train
218 the decoder, we generated ~ 1000 2D ellipses and 3D ellipsoids with varying axis ratios and
219 sizes (volumes) and the network was validated in 100 additional test images. Figure 4 show a
220 sample of observed and predicted 2D ellipses, while results for 3D shapes are in Figure 5. In
221 this latter case, sections across the three axes are shown for clarity since the 3D figure drawn
222 with python package trisurf was not too accurate. Prediction is remarkably accurate also in
223 the case of 3D shapes, especially when one considers the high dimensionality of the output
224 image: $32 \times 32 \times 32 = 32,768$ float numbers. Albeit in a simplistic scenario, we can see a
225 naïve decoder is quite effective in predicting shapes conditional on text (DNA) information.

226

227 To investigate whether the decoder network can be applied to more complex and realistic
228 scenarios, we simulated cucurbit images from *C. pepo* as described in methods. We trained
229 the same decoder as in the previous toy example using the shape causative locus *s* plus 100
230 random SNPs as input and the simulated cucurbit images as output. An example of five
231 randomly predicted images is in Figure 6. Overall, prediction was quite reasonable, and
232 predicted shapes can be easily recognized. Note the ‘rugosity’ induced by the autoregressive
233 model, which is also reproduced in the prediction. We found the maximization algorithm can
234 have a large influence on results. RMSprop performed best, whereas Adam failed often and
235 Adagrad did not seem to work.

236

237 Prediction of a random set of tomato shapes based on the 116 metrics is shown in Figure 7.
238 Predictions were very good overall, except of hybrids involving TR_MO_004 (Figure 7,
239 sample 1). This traditional variety belongs to the horticultural group “Coeur de Boeuf”,
240 which fruits are big with irregular shapes.

241

242 *From phenotype to phenotype*

243 Can we bypass genotype information altogether? If shapes are highly heritable, the network
244 could learn inheritance patterns and predict offspring shape directly from parents’ forms,
245 without resorting to genotypes. Figure 8 shows examples of the four image ‘inheritance’
246 behaviors defined: ‘additivity’, ‘dominance’, ‘imprinting’ and ‘epistasis’. We observe that
247 predictions were reasonably accurate for additivity and epistasis but were worse for
248 dominance and, especially, for imprinting. It seems the network can accurately find additive
249 and non-linear patterns but is less adapted to predictions where the order of inputs is relevant.
250 We conjecture then that recurrent neural networks (RNNs, e.g., Hill et al. 2018) could be
251 better suited to this problem.

252

253 In the second example, we used the images from crosses between traditional and modern
254 inbred tomato lines described. Predictions (Figure 9) were remarkably accurate overall,
255 proving fruit appearances can be predicted from ancestor images. It also suggests that the
256 predominant action seems to be additive.

257

258 **Discussion**

259 Being able to predict highly dimensional objects such as appearance can revolutionize
260 breeding by merging genome and phenome information in a coherent framework. Here we

261 present a proof of concept that this is possible, even using very simple network architectures.
262 We show that 2D, but also 3D, shapes can be accurately predicted and generated.

263

264 The problem posed here is similar to the ‘text-to-image’ challenge, where algorithms are
265 trained to generate images from figure captions. Some works have recently reported highly
266 accurate results (Ramesh *et al.*; Radford *et al.* 2021) and we foresee that ‘dna-to-image’
267 should follow. There are some differences between text and DNA that require specific
268 developments though. First, text is divided in a finite, relatively small number of items
269 (words) which relationships can be inferred by automatically parsing large available
270 databases. DNA sequence can be split into coding / noncoding, introns / exons but cannot be
271 assimilated to ‘words’ with specific meanings. DNA or marker data are not segmented; their
272 relationships are much more intricate than those in words from human languages and are
273 unknown to a large extent. For instance, most discovered causative mutations that affect
274 shape are located outside coding regions (Wu *et al.* 2018; Martínez-Martínez *et al.* 2022).
275 Second, large corpuses of images and figure captions are available for training text-to-image
276 problems; these datasets are not readily available for fruits or other agricultural scenarios.
277 Finally, texts used to generate images are short and simple; algorithms usually fail and
278 generate unpredictable results if input text is slightly changed. In the case of DNA, the
279 number of differences between strain or individual genotypes is very large; we still do not
280 know how dna-to-image algorithms will cope with this issue.

281

282 Text-to-image methods rely on text encoding, also called ‘embedding’, i.e., in finding an
283 optimum numeric representation of text elements in a reduced n-dimensional space. DNA
284 encoding is to be critical in dna-to-image problems as well. Previous research on DNA
285 encoding has utilized small DNA sequences, e.g., taking exons as ‘words’ (Zou *et al.* 2019; Ji
286 *et al.* 2021). However, this cannot be applied to generic marker data or complete sequence.
287 We hypothesize that standard dimension reduction techniques, such as classical principal
288 component analysis (PCA), can be a useful alternative especially when shape is controlled by
289 numerous loci of small effect.

290

291 For simulation purposes of cucurbit shapes, we assumed an underlying continuous gradient
292 that results in a continuous morphological variation (Figure 1C). We assumed this for
293 computational and illustrative purposes, although we reckon there is no clear biological
294 evidence on this hypothesis. Modern cultivars adopt discrete shapes and intermediate shapes

295 are rarely observed. However, traditional unimproved varieties and their crosses do show a
296 number of intermediate features (Montero-Pau *et al.* 2017).

297

298 Numerous genes that influence shape have been discovered (Monforte *et al.* 2014; Grumet
299 and Colle 2016; Snouffer *et al.* 2020). These genes act in concerted action during
300 development (Wu *et al.* 2018). Note the method proposed here does not require causative loci
301 to be identified, as prediction methods rely on linkage disequilibrium between causative and
302 genotyped markers. Nevertheless, known causative polymorphisms could be given larger
303 weights than the rest of SNPs. There are several approaches that can be used to achieve this.
304 One option is the ‘attention’ mechanism, which is used to underline words of particular
305 relevance in text analysis (Vaswani *et al.*). Another possibility is to define a specific input
306 layer for causative mutations and merging with the rest of SNPs in a separate layer. This is
307 straightforward with standard software such as Keras (Chollet 2015).

308

309 Further work is warranted to overcome limitations of this work and continue this area of
310 research. First of all, appropriate datasets of large size in 2D and 3D must be generated. In
311 fact, one of the limiting steps for this methodology to be applied is the lack of datasets of
312 enough size containing high density genotypes and good quality images. The simplest
313 scenario should be fruits, as is the TRADITOM initiative in tomato (Pons *et al.* 2022; Blanca
314 *et al.* 2022), but many other applications can be envisaged: animal conformation (e.g., dairy
315 bull catalogs, dog breeds), whole plant appearance, leaf and root morphology, color patterns,
316 ... Second, more complex network architectures inspired in current text-to-image algorithms
317 must be adapted to the dna-to-image scenario. Finally, generative models, such as conditional
318 generative adversarial networks (CGANs; Goodfellow *et al.* 2014; Mirza and Osindero
319 2014), conditional on DNA information, could be used to produce images of high quality. On
320 top of that, new tools for dealing with 3D objects are needed.

321

322 In summary, we have shown that very simple networks can be successfully trained in small
323 datasets to accurately predict fruit images. Although much work remains to be done, this
324 research opens new possibilities in the area of prediction of complex traits.

325

326 **Data availability statement**

327 All data and code are available at <https://github.com/miguelperezenciso/dna2image>.

328

329 **Acknowledgments**

330 Authors acknowledge the technical support of Gorka Perpiña and Eva María Pérez. Modern
331 inbreds and their hybrids with traditional tomato varieties were provided by Meridiem Seeds
332 (<https://meridiemseeds.com/>, Murcia, Spain).

333

334 **Funding**

335 Work funded by Ministry of Science and Innovation-State Research Agency (AEI, Spain,
336 10.13039/501100011033) grant numbers PID2019-108829RB-I00 and CEX2019-000902-S,
337 European Commission H2020 research and innovation program through TRADITOM grant
338 agreement No.634561, and HARNESSSTOM grant agreement No.101000716, and
339 PROMETEO projects 2017/078 and 2021/072 to promote excellence groups by the
340 Conselleria d'Educació, Investigació, Cultura i Esports (Generalitat Valenciana, Spain) and
341 by the CERCA Programme / Generalitat de Catalunya (Spain).

342

343 **Author contributions**

344 MPE and LMZ conceived computational research and developed code; AG, BP and AJM
345 conceived and discussed empirical research; MPE, CP and SS performed research. MPE and
346 AJM wrote the draft with help from rest of authors.

347

348 **References**

349 Abadi M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, *et al.*, 2015 TensorFlow: Large-scale
350 machine learning on heterogeneous systems. tensorflow.org 1.
351 [https://doi.org/10.1016/0076-6879\(83\)01039-3](https://doi.org/10.1016/0076-6879(83)01039-3)

352 Blanca J., J. Montero-Pau, C. Sauvage, G. Bauchet, E. Illa, *et al.*, 2015 Genomic variation in
353 tomato, from wild ancestors to contemporary breeding accessions. *BMC Genomics* 16:
354 257. <https://doi.org/10.1186/s12864-015-1444-1>

355 Blanca J., C. Pons, J. Montero-Pau, D. Sanchez-Matarredona, P. Ziarsolo, *et al.*, 2022
356 European traditional tomatoes galore: a result of farmers' selection of a few diversity-
357 rich loci. *J Exp Bot* 73: 3431–3445. <https://doi.org/10.1093/jxb/erac072>

358 Bradski G., 2000 The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

359 Brewer M. T., L. Lang, K. Fujimura, N. Dujmovic, S. Gray, *et al.*, 2006 Development of a
360 controlled vocabulary and software application to analyze fruit shape variation in tomato
361 and other plant species. *Plant Physiol* 141: 15–25.
362 <https://doi.org/10.1104/pp.106.077867>

363 Brownlee J., 2019 *Deep Learning for Computer Vision*.

364 Chollet F., 2015 Keras: Deep learning library for theano and tensorflow. <https://keras.io/> 7:
365 T1.

366 Chollet F., 2021 *Deep Learning with Python*. Manning Publications.

367 Claude J., 2008 *Morphometrics with R*. Springer.

368 Cuevas J., O. Montesinos-López, P. Juliana, C. Guzmán, P. Pérez-Rodríguez, *et al.*, 2019
369 Deep Kernel for Genomic and Near Infrared Predictions in Multi-environment Breeding
370 Trials. *G3 Genes|Genomes|Genetics* 9: 2913–2924.
371 <https://doi.org/10.1534/G3.119.400493>

372 Goodfellow I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, *et al.*, 2014 Generative
373 adversarial nets. *ArXiv arXiv:1011*. <https://doi.org/10.1016/B978-0-408-00109-0.50001-8>

375 Gower J. C., 1975 Generalized procrustes analysis. *Psychometrika* 40: 33–51.
376 <https://doi.org/10.1007/BF02291478>

377 Grumet R., and M. Colle, 2016 Genomic Analysis of Cucurbit Fruit Growth, pp. 321–344 in
378 Springer, Cham.

379 Hill S. T., R. Kuintzle, A. Teegarden, E. Merrill, P. Danaee, *et al.*, 2018 A deep recurrent
380 neural network discovers complex biological rules to decipher RNA protein-coding
381 potential. *Nucleic Acids Res* 46: 8105–8113. <https://doi.org/10.1093/nar/gky567>

382 Ji Y., Z. Zhou, H. Liu, and R. v Davuluri, 2021 DNABERT: pre-trained Bidirectional
383 Encoder Representations from Transformers model for DNA-language in genome.
384 *Bioinformatics* 37: 2112–2120. <https://doi.org/10.1093/bioinformatics/btab083>

385 Klingenberg C. P., 2010 Evolution and development of shape: Integrating quantitative
386 approaches. *Nat Rev Genet* 11: 623–635. <https://doi.org/10.1038/nrg2829>

387 Martínez-Martínez C., M. J. Gonzalo, P. Sipowicz, M. Campos, I. Martínez-Fernández, *et al.*,
388 2022 A cryptic variation in a member of the Ovate Family Proteins is underlying the
389 melon fruit shape QTL fsqs8.1. *Theoretical and Applied Genetics* 135: 785–801.
390 <https://doi.org/10.1007/s00122-021-03998-6>

391 Meng F., M. Richer, A. Tehrani, J. La, T. D. Kim, *et al.*, 2022 Procrustes: A python library to
392 find transformations that maximize the similarity between matrices. *Comput Phys
393 Commun* 276: 108334. <https://doi.org/10.1016/j.cpc.2022.108334>

394 Metere R., and H. E. Möller, 2017 PyMRT and DCMPI: Two New Python Packages for MRI
395 Data Analysis, in *Proceedings of the 25th Annual Meeting & Exhibition of the
396 International Society for Magnetic Resonance in Medicine (ISMRM)*, Honolulu.

397 Mirza M., and S. Osindero, 2014 Conditional Generative Adversarial Nets. *ArXiv
398 1411.1784v*.

399 Monforte A. J., A. Diaz, A. Caño-Delgado, and E. van der Knaap, 2014 The genetic basis of
400 fruit morphology in horticultural crops: lessons from tomato and melon. *J Exp Bot* 65:
401 4625–4637. <https://doi.org/10.1093/jxb/eru017>

402 Montero-Pau J., J. Blanca, C. Esteras, E. Ma. Martínez-Pérez, P. Gómez, *et al.*, 2017 An
403 SNP-based saturated genetic map and QTL analysis of fruit-related traits in Zucchini
404 using Genotyping-by-sequencing. *BMC Genomics* 18: 94.
405 <https://doi.org/10.1186/s12864-016-3439-y>

406 Paris H., 1986 A proposed subspecific classification for *Cucurbita pepo*. *Phitologia* 61: 133–
407 138.

408 Paris H. S., 1989 Historical records, origins, and development of the edible cultivar groups
409 of *Cucurbita pepo* (Cucurbitaceae). *Econ Bot* 43: 423–443.
410 <https://doi.org/10.1007/BF02935916>

411 Paris H. S., 2001 History of the Cultivar-Groups of *Cucurbita pepo*. *Horticultura reviews* 25:
412 71–170.

413 Pons C., J. Casals, S. Palombieri, L. Fontanet, A. Riccini, *et al.*, 2022 Atlas of phenotypic,
414 genotypic and geographical diversity present in the European traditional tomato. *Hortic
415 Res.* <https://doi.org/10.1093/hr/uhac112>

416 Radford A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, *et al.*, 2021 Learning Transferable
417 Visual Models From Natural Language Supervision. *ArXiv*.

418 Ramesh A., A. Nichol, and M. Chen, Hierarchical Text-Conditional Image Generation with
419 CLIP Latents. ArRxiv.

420 Rincent R., J. P. Charpentier, P. Faivre-Rampant, E. Paux, J. le Gouis, *et al.*, 2018 Phenomic
421 Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions:
422 Proof of Concept on Wheat and Poplar. *G3 Genes|Genomes|Genetics* 8: 3961–3972.
423 <https://doi.org/10.1534/G3.118.200760>

424 Robert P., J. Auzanneau, E. Goudemand, F. X. Oury, B. Rolland, *et al.*, 2022 Phenomic
425 selection in wheat breeding: identification and optimisation of factors influencing
426 prediction accuracy and comparison to genomic selection. *Theoretical and Applied
427 Genetics* 135: 895–914. <https://doi.org/10.1007/S00122-021-04005-8/FIGURES/7>

428 Robinson M. R., G. English, G. Moser, L. R. Lloyd-Jones, M. A. Triplett, *et al.*, 2017
429 Genotype–covariate interaction effects and the heritability of adult body mass index. *Nat
430 Genet* 49: 1174–1181. <https://doi.org/10.1038/ng.3912>

431 Snouffer A., C. Kraus, and E. van der Knaap, 2020 The shape of things to come: ovate family
432 proteins regulate plant organ shape. *Curr Opin Plant Biol* 53: 98–105.

433 Tanksley S. D., 2004 The Genetic, Developmental, and Molecular Bases of Fruit Size and
434 Shape Variation in Tomato. *Plant Cell* 16: S181–S189.
435 <https://doi.org/10.1105/tpc.018119>

436 Tong H., A. N. Nankar, J. Liu, V. Todorova, D. Ganeva, *et al.*, 2022 Genomic prediction of
437 morphometric and colorimetric traits in Solanaceous fruits. *Hortic Res* 9.
438 <https://doi.org/10.1093/hr/uhac072>

439 Vaswani A., G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, *et al.*, Attention Is All You Need
440 Visscher P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, *et al.*, 2017 10 Years of
441 GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 101: 5–22.

442 Walt S. van der, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, *et al.*, 2014
443 Scikit-image: Image processing in python. *PeerJ* 2014: e453.
444 <https://doi.org/10.7717/PEERJ.453/FIG-5>

445 Wood A. R., T. Esko, J. Yang, S. Vedantam, T. H. Pers, *et al.*, 2014 Defining the role of
446 common variation in the genomic and biological architecture of adult human height. *Nat
447 Genet* 46. <https://doi.org/10.1038/ng.3097>

448 Wu S., B. Zhang, N. Keyhaninejad, G. R. Rodríguez, H. J. Kim, *et al.*, 2018 A common
449 genetic mechanism underlies morphological diversity in fruits and other plant organs.
450 *Nat Commun* 9: 4734. <https://doi.org/10.1038/s41467-018-07216-8>

451 Xanthopoulou A., J. Montero-Pau, I. Mellidou, C. Kissoudis, J. Blanca, *et al.*, 2019 Whole-
452 genome resequencing of *Cucurbita pepo* morphotypes to discover genomic variants
453 associated with morphology and horticulturally valuable traits. *Hortic Res* 6: 1–17.
454 <https://doi.org/10.1038/s41438-019-0176-9>

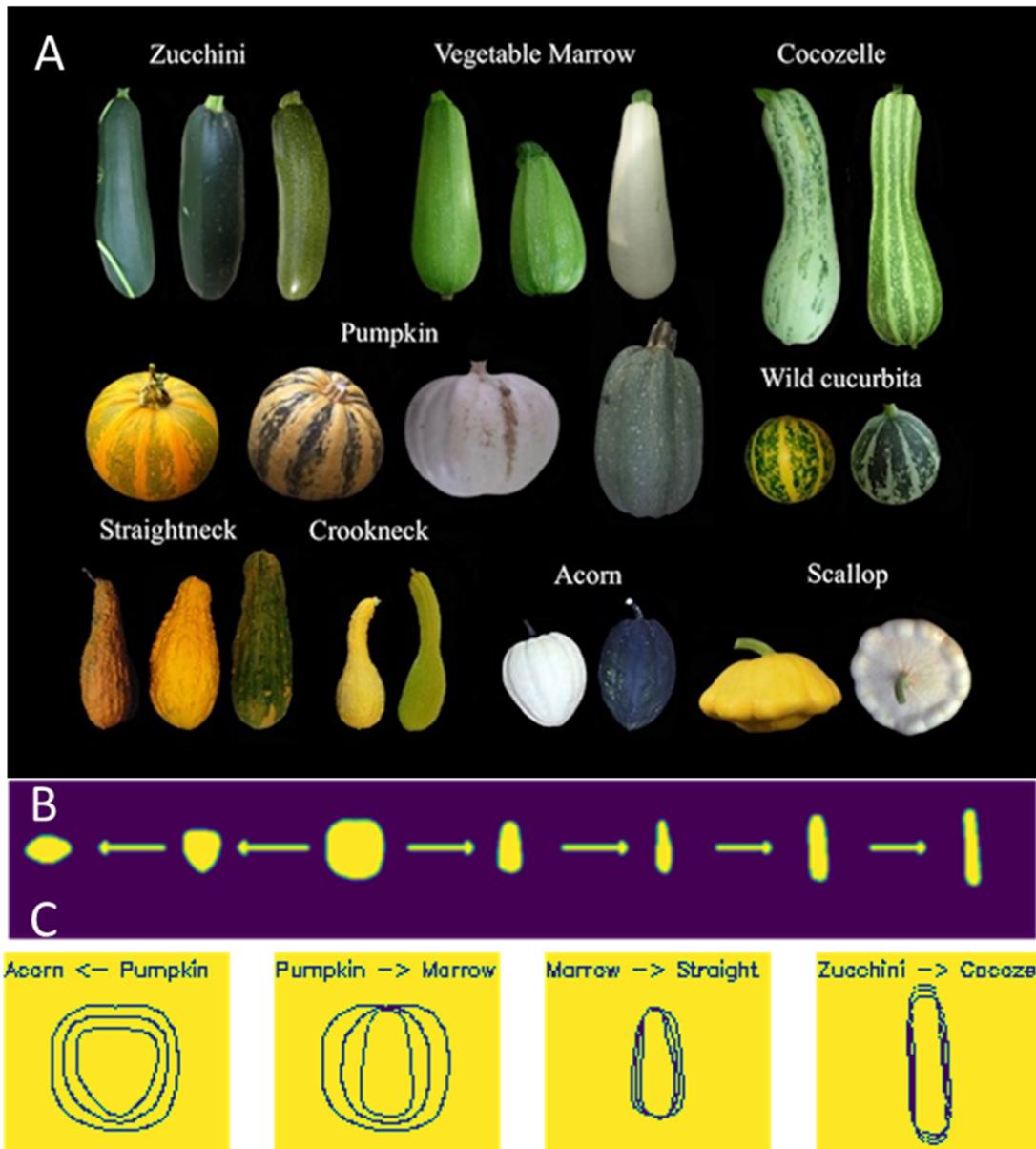
455 Zelditch M. L., D. L. Swiderski, and H. D. Sheets, 2004 *Geometric Morphometrics for
456 Biologists: A primer*. Academic Press.

457 Zingaretti L. M., A. Monfort, and M. Pérez-Enciso, 2021 Automatic Fruit Morphology
458 Phenome and Genetic Analysis: An Application in the Octoploid Strawberry. *Plant
459 Phenomics* 2021. <https://doi.org/10.34133/2021/9812910>

460 Zou Q., P. Xing, L. Wei, and B. Liu, 2019 Gene2vec: gene subsequence embedding for
461 prediction of mammalian N^6 -methyladenosine sites from mRNA. *RNA* 25: 205–218.
462 <https://doi.org/10.1261/rna.069112.118>

463

464



465

466

467 **Figure 1: A)** Variability found in *C. pepo* fruit shapes. **B)** Assumed evolutive pathways for
468 shape simulation: scallop \leftarrow acorn \leftarrow pumpkin / wild gourd \rightarrow marrow \rightarrow straightneck \rightarrow
469 zucchini \rightarrow cocozelle. **C)** Each panel shows contours of two observed shapes and an
470 intermediate shape, illustrating how a continuous evolutive gradient corresponds to a given
471 shape.

472

```
# decoder network dna --> image: generates images out of snp data
def dna2image(n_snp,image_size):
    input = tf.keras.layers.Input(shape=(n_snp))

    x = tf.keras.layers.Dense(np.prod(image_size))(input)
    x = tf.keras.layers.Reshape(image_size+(1,))(x)
    x = tf.keras.layers.Conv2DTranspose(16, (3,3), activation='relu', padding='same')(x)
    x = tf.keras.layers.Conv2DTranspose(8, (3,3), activation='relu', padding='same')(x)

    output = tf.keras.layers.Conv2DTranspose(1, (5,5), activation='relu', padding='same')(x)

    return tf.keras.Model(input, output)
```

473

474 *Figure 2: Keras code with the decoder used for image prediction. Function requires number*

475 *of SNPs and output image size as input parameters.*

476

```
# decoder network dna --> image: generates images out of image pairs
def img22img(image_size):
    input1 = keras.Input(shape=image_size+(1,))
    input2 = keras.Input(shape=image_size+(1,))
    x1 = layers.experimental.preprocessing.Rescaling(1.0 / 255)(input1)
    x2 = layers.experimental.preprocessing.Rescaling(1.0 / 255)(input2)

    x = layers.concatenate([x1, x2])
    x = keras.layers.Conv2D(16, (5,5), activation='relu', padding='same')(x)
    x = keras.layers.Conv2D(8, (3,3), activation='relu', padding='same')(x)
    x = keras.layers.Flatten()(x)
    x = keras.layers.Dense(16)(x)

    embed = keras.layers.Dense(2)(x)

    x = keras.layers.Dense(np.prod(image_size))(embed)
    x = keras.layers.Reshape(image_size+(1,))(x)
    x = keras.layers.Conv2DTranspose(16, (3,3), activation='relu', padding='same')(x)
    x = keras.layers.Conv2DTranspose(8, (3,3), activation='relu', padding='same')(x)

    output = keras.layers.Conv2DTranspose(1, (5,5), activation='relu', padding='same')(x)

    return keras.Model(inputs=[input1, input2], outputs=output)
```

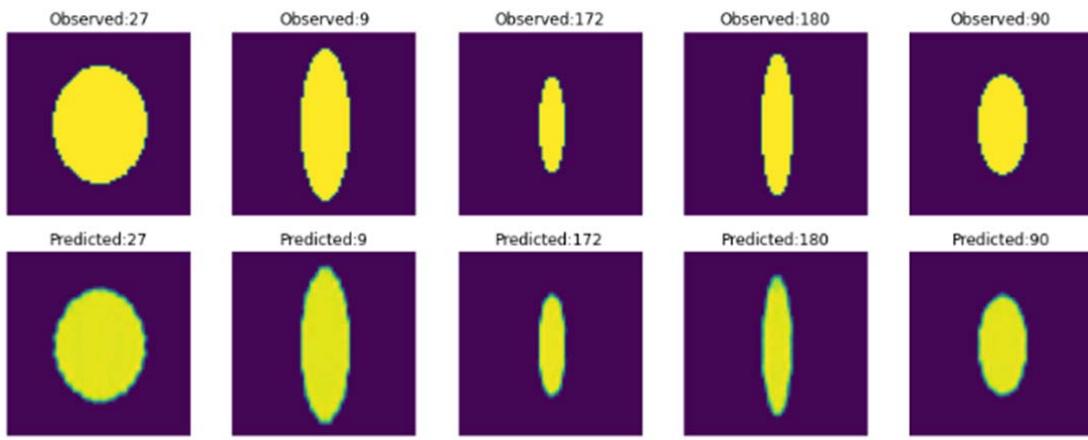
477

478 *Figure 3: Keras code used for offspring image prediction based on parents' images. It*

479 *requires image size as input, which should be the same in input and output images. Size of*

480 *embed vector can be fine-tuned for better performance.*

481



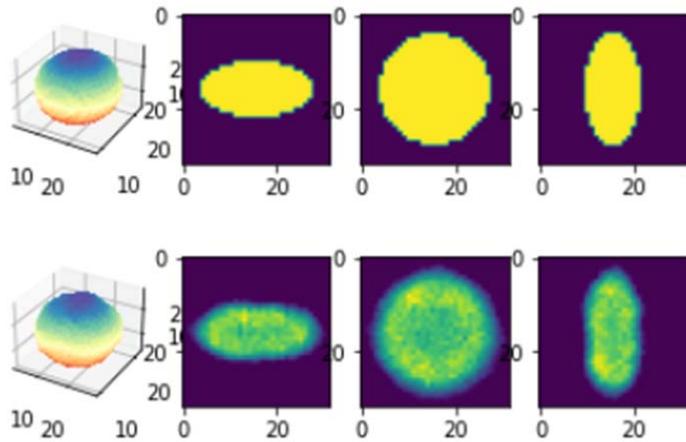
482

483 **Figure 4:** Top row: random sample of simulated ellipses; bottom row: predicted images
484 using decoder in Figure 2.

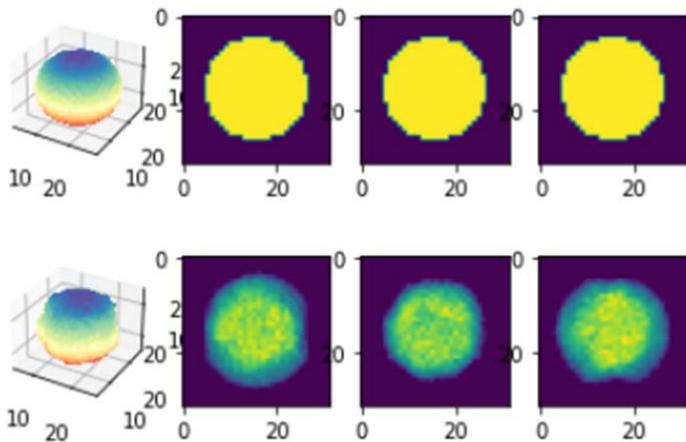
485

486

ellipsoid 0.1123046875 0.10406453907489777



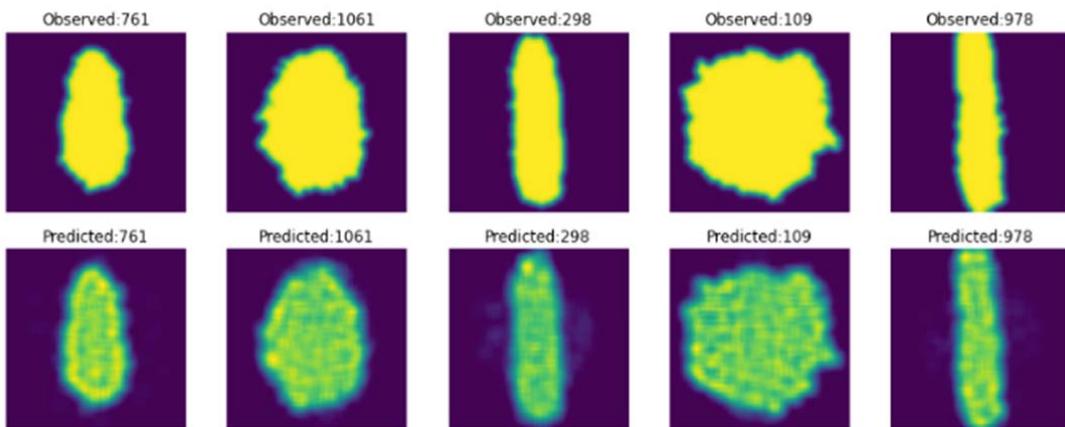
sphere 0.17138671875 0.10233276337385178



487

488 **Figure 5:** Generated (top rows) and predicted (bottom rows) of two 3D ellipsoids. The left
489 column represents observed and predicted 3D representation, and the following columns are
490 transversal cuts along the three axes.

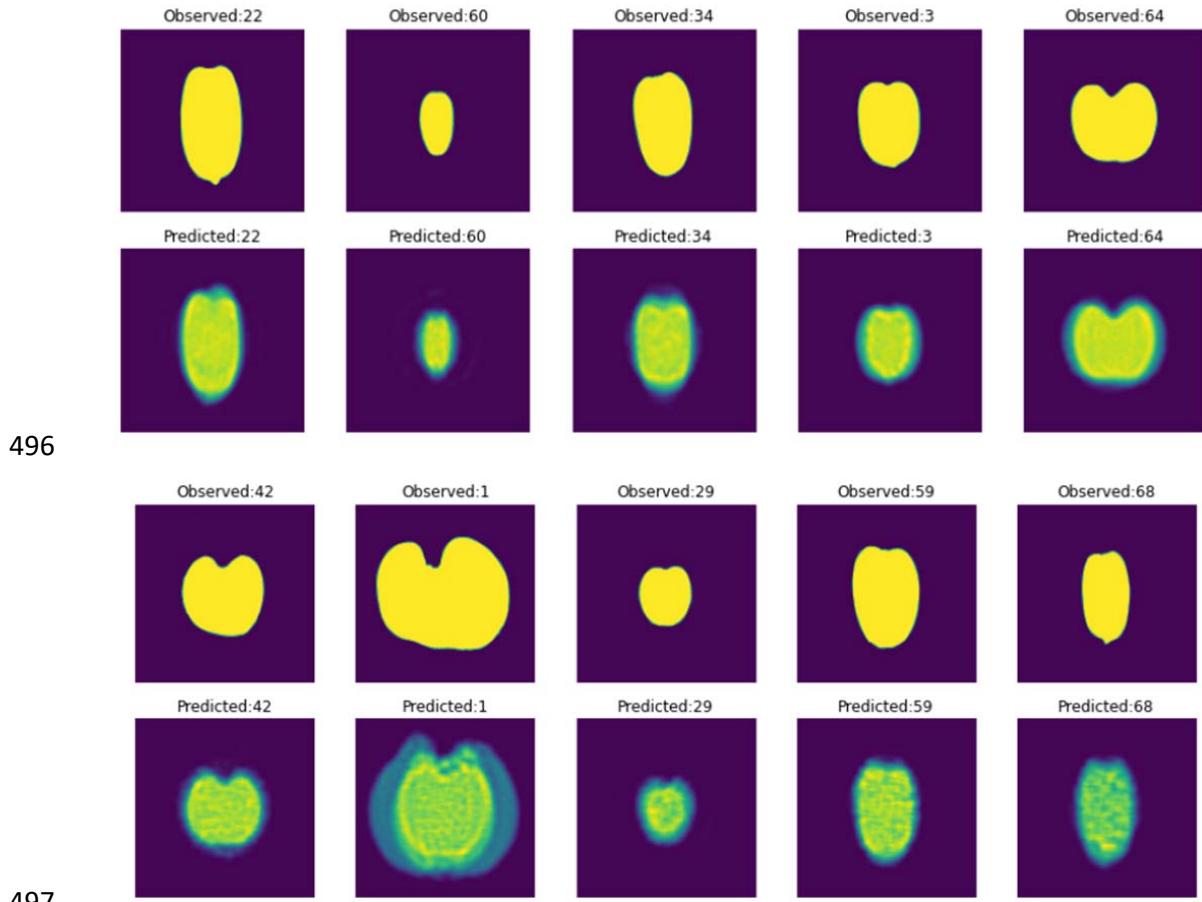
491



492

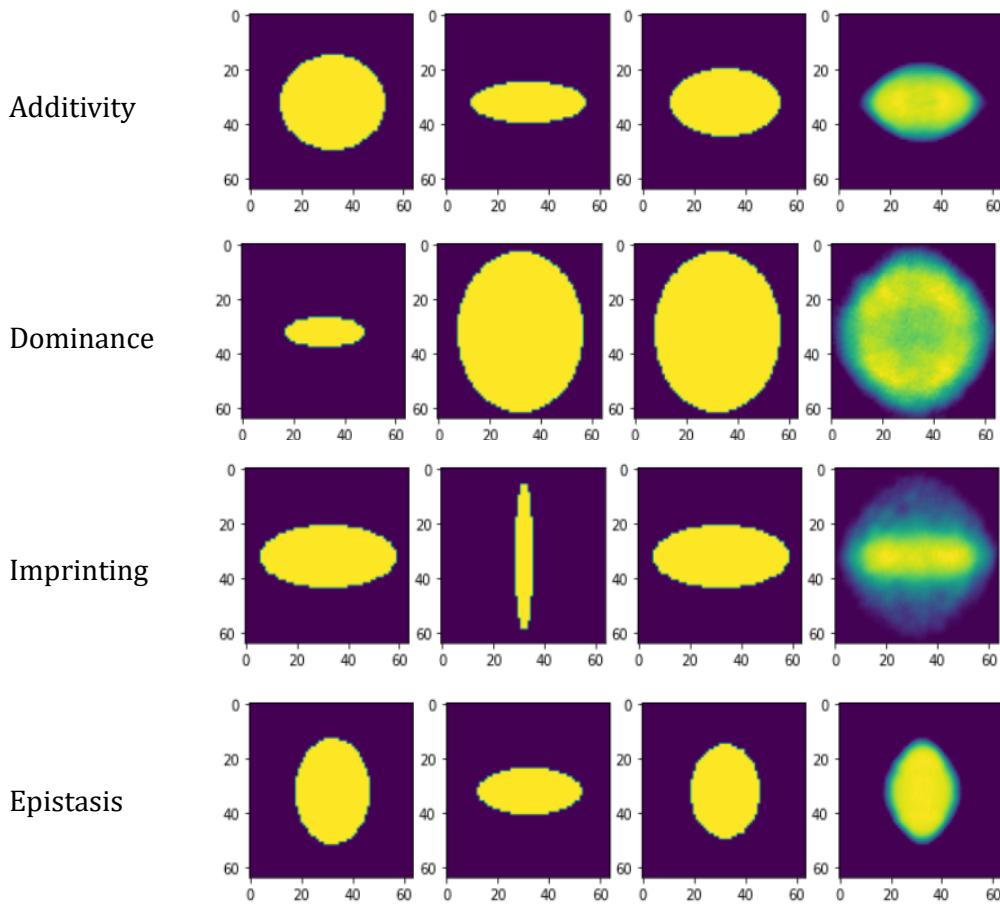
493 **Figure 6:** Top row: sample of simulated cucurbit images including autoregressive noise;
494 bottom row: predicted images using decoder in Figure 2.

495



498 **Figure 7:** Sample of observed tomato images (first and third rows) and the corresponding
499 predicted images using decoder in Figure 2.

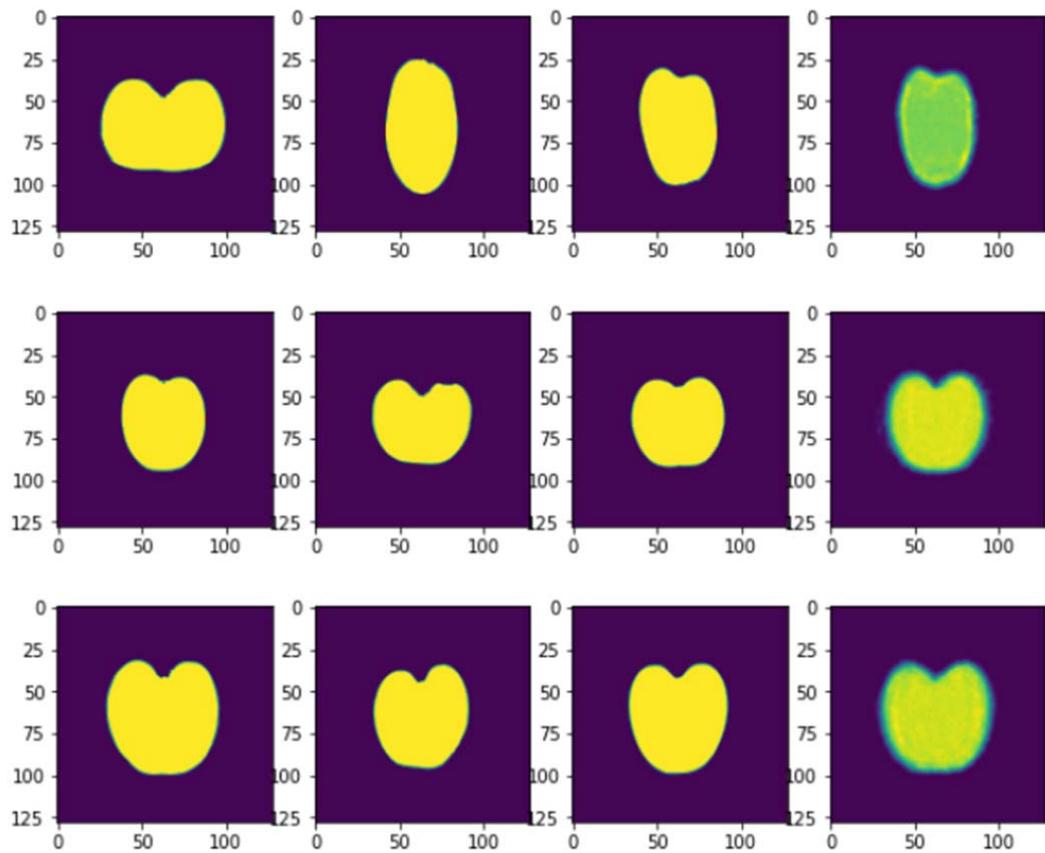
500



501

502 **Figure 8:** Examples of the four arbitrary image inheritance patterns defined: ‘additivity’,
503 ‘dominance’, ‘imprinting’ and ‘epistasis’. Columns show ‘paternal’, ‘maternal’, ‘offspring’
504 and predicted images.

505



506

507 **Figure 9:** Observed tomato trios in three random crosses and predicted offspring based on
508 network in Figure 4. Columns are paternal, maternal, offspring and predicted offspring
509 images. Images' size is 124 x 124 pixels.

510

511 **Table S1:** Parental tomato lines.

Code	Type	Fruit type
MS_1	Modern Inbred	Salad tomato
MS_2	Modern Inbred	Salad tomato
MS_3	Modern Inbred	Salad tomato
MS_4	Modern Inbred	Long processing
MS_5	Modern Inbred	Cocktail round
MS_6	Modern Inbred	Cherry round
MS_7	Modern Inbred	Cherry round
TR_TH_001	Traditional	round
TR_TH_002	Traditional	round
TR_TH_003	Traditional	flattened
TR_CA_001	Traditional	ovoid
TR_CA_002	Traditional	flat
TR_VA_001	Traditional	flat
TR_VA_002	Traditional	oxheart
TR_VA_003	Traditional	round
TR_MO_001	Traditional	flat
TR_MO_002	Traditional	round
TR_MO_003	Traditional	round
TR_MO_004	Traditional	oxheart
TR_VI_001	Traditional	Long
TR_VI_002	Traditional	round
TR_VI_003	Traditional	round
TR_VI_004	Traditional	ellipsoid
TR_VI_005	Traditional	ovoid
TR_VI_006	Traditional	rectangular
TR_PO_001	Traditional	ellipsoid
TR_PO_002	Traditional	long
TR_PO_003	Traditional	ellipsoid
TR_PO_004	Traditional	ellipsoid
TR_IS_001	Traditional	long
TR_IS_002	Traditional	ovoid
TR_IS_003	Traditional	round

512 MS_1 to MS_7 correspond to modern inbred lines provided by Meridiem Seeds. Codes for
513 the traditional varieties are according TRADITOM project (Blanca et al. 2022).