The structural landscape and diversity of *Pyricularia oryzae* MAX effectors revisited

Mounia LAHFA[1], Philippe BARTHE[1], Karine DE GUILLEN[1], Stella CESARI[2], Mouna RAJI[1], Thomas KROJ[2], Marie Le Naour--Vernet[2], François Hoh[1], Pierre GLADIEUX[2], Christian ROUMESTAND[1], Jérôme GRACY[1], Nathalie DECLERCK[1], André PADILLA[1]*

[1] Centre de Biologie Structurale, Univ Montpellier, CNRS UMR 5048, INSERM U 1054, Montpellier, France.

[2] PHIM Plant Health Institute, Univ Montpellier, INRAE, CIRAD, Institut Agro, IRD, Montpellier, France

* Corresponding author

E-mail: andre.padilla@cbs.cnrs.fr

Keywords: *Pyricularia oryzae*, MAX effectors, NMR, protein structure, AlphaFold

Short title: MAX structural family portrait

ABSTRACT

Plant pathogenic fungi secrete a wide variety of small proteins, named effectors. *Magnaporthe* AVRs and ToxB-like (MAX) effectors constitute a superfamily of secreted proteins widely distributed in *Pyricularia (syn. Magnaporthe) oryzae*, a devastating fungus responsible for blast disease in cereals such as rice. In spite of high evolutionary sequence divergence, MAX effectors share a common fold characterized by a ß-sandwich core often stabilized by a conserved disulfide bond.

In this study, we investigated the structural landscape and diversity within this effector family based on a previous phylogenetic analysis of *P. oryzae* protein sequences that identified 94 ortholog groups (OG) of putative MAX effectors. Combining protein structure modeling approaches and experimental structure determination, we validated the prediction of the conserved MAX core domain for 77 OG clusters. Four novel MAX effector structures determined by NMR were in remarkably good agreement with AlphaFold2 (AF) predictions. Based on the comparison of the AF-generated 3D models we propose an updated classification of the MAX effectors superfamily in 20 structural groups that highlight variation observed in the canonical MAX fold, disulfide bond patterns and decorating secondary structures in N- and C-terminal extensions. About one-third of the MAX family members remain single, showing no obvious structural relationship with other MAX effectors. Analysis of the surface properties of the AF MAX models also highlights the very high variability remaining within the MAX family when examined at the structural level, probably reflecting the wide diversity of their virulence functions and host targets.

Author summary

MAX effectors are a family of virulence proteins from the plant pathogenic fungus *Pyricularia (syn. Magnaporthe) oryzae* that share a similar 3D structure despite very low amino-acid sequence identity. Characterizing the function and evolution of these proteins requires a detailed understanding of their structural diversity. We used a combination of experimental structure determination and structural modeling to characterize in detail the MAX effector repertoire of *P. oryzae*. A prediction pipeline based on similarity searches and structural modeling using the AlphaFold2 (AF) software were used to predict MAX effectors in a collection of 120 *P. oryzae* genomes. We then compared AF models with experimentally validated NMR structures. The resulting models and experimental structures revealed that the preserved MAX core coexists with extensive structural variability in terms of structured N- or C-terminal extensions. For each of the AF models, we also analyzed the surfaces of the canonical fold that may be involved in protein-protein interactions. This work constitutes a major step in mapping the functional network of MAX effectors through their structure by identifying possible recognition sites that may help focusing studies of their putative targets in infected plant hosts.

Introduction

Fungal plant pathogens secrete small proteins, called effectors, which promote disease by targeting cellular processes in the host plant. These fungal effectors are usually identified based on their secretion signal and other features such as cysteine enrichment [1,2]. There are hundreds of predicted effectors in the genomes of plant pathogenic fungi and very few of them show sequence similarities and/or homology to known proteins or protein domains. Some effectors are of particular interest since they constitute avirulence (AVR) factors that can be detected by plant immune systems and render crops resistant to severe diseases. Because fungal effectors usually show no amino-acid sequence homology to known proteins or protein domains, their biological function cannot be inferred from systematic *in silico* analysis (such as domain searches) but has so far been elucidated on a case-by-case basis. As a huge and growing number of putative effectors is being discovered, deciphering the functions and adaptive evolution of fungal effectors thus requires robust predictive tools for analyzing protein sequences and identifying candidates on which prioritize functional and structural studies.

Recently, a combination of primary sequence pattern searches and structural modeling resulted in a major breakthrough in the study of effector biology by revealing that fungal effector repertoires are actually dominated by a limited number of families sharing common structures despite extensive sequence variability [3–5].

One such family is the superfamily of MAX (*Magnaporthe* AVRs and ToxB-like) effectors we identified in *Pyricularia oryzae* (synonym: *Magnaporthe oryzae*), the fungus causing blast disease in rice, wheat, and other cereals or grasses. This pathogenic fungus is both a major threat to global food security and a prime

experimental model in plant pathology because of its ability to rapidly evolve to escape recognition of the plant immune system. By solving the solution structure of two *P. oryzae* effectors, AVR1-CO39 and AVR-Pia, we discovered strong structural similarities between these sequence-unrelated effectors as well as with the ToxB effector from the wheat infecting fungus *Pyrenophora tritici-repentis* [6]. This latter showing sequence homology with a putative effector from *M. oryzae*, we solved the crystal structure of this MoToxB protein (supplementary materials and methods) and confirmed its structural similarity with ToxB as well as with the other structurally determined *P. oryzae* MAX effectors that together with MoToxB share less than 13% sequence identity (S1 Fig). Like other MAX effectors, MoToxB is structured as a 6-stranded ß-sandwich (ß1 to ß6) of two triple-stranded ß-sheets with a ß6ß1ß2-ß3ß4ß5 topology. The highly conserved disulfide bond forms a bridge between ß1 and the loop connecting ß4 and ß5 while a second disulfide bond connects ß2 and ß6 (Fig 1).

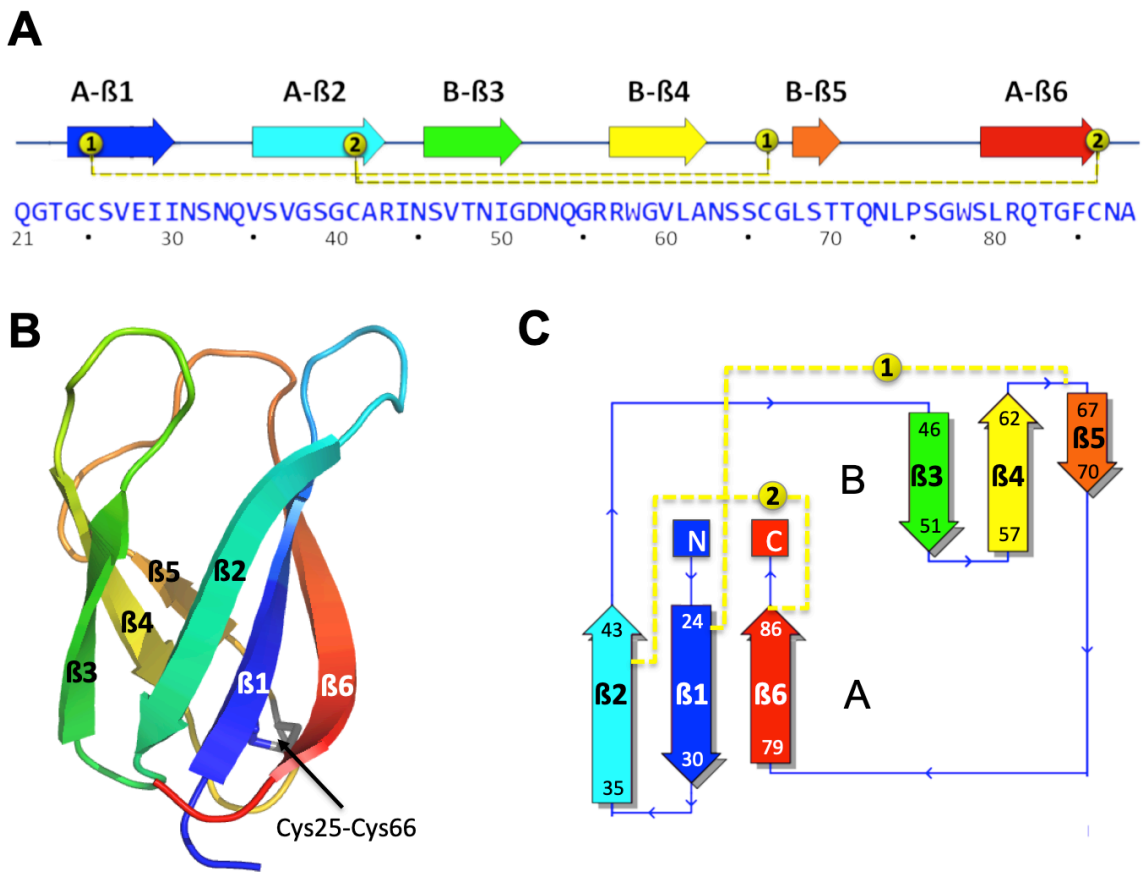


Fig 1. Structure of the M. oryzae ToxB (MoToxB) MAX effector

Primary and secondary structure of MoToxB showing the triple-stranded beta-sandwich forming the conserved MAX core with the two beta-sheets labeled by A and B, strands indicated by arrows and two disulfide bonds in yellow dotted lines. Disulfide bond "1" is almost strictly conserved in MAX effectors. (B) Cartoon representation of MoToxB crystal structure (PDB 6R5J) shown in rainbow color and the conserved disulfide bond "1" shown in sticks. (C) MoToxB topology diagram drawn by PDBsum and colored using the same color scheme as in A and B.

MAX effectors are specific to plant pathogenic ascomycetes fungi and they have undergone a major expansion in *P. oryzae*, suggesting an important role in the infection process. Our recent analysis of 120 isolates of *P. oryzae* identified 58 to 78 MAX effector genes per genome, corresponding to 5 to 10% of the full repertoire of predicted effectors. In addition to this genetic expansion, the importance of MAX

115 effectors in *P. oryzae* virulence is supported by our finding that many MAX effectors
116 are massively expressed during the early stages of plant infection in a host-
117 dependent manner [5–7]. They are also specifically targeted by the plant immune
118 system: nearly half of the cloned avirulence genes of *P. oryzae,* which encode
119 avirulence (AVR) factors recognized by host resistance proteins in blast disease
120 resistant cultivars, correspond to MAX effectors [6,8–11]. Analysis of the recognition
121 of the MAX effectors AVR-Pia, AVR1-CO39, and AVR-Pik by rice immune receptors
122 indicates that they may target a specific class of proteins called small HMAs (sHMAs)
123 [12], which show similarity to copper chaperones containing a heavy metal-
124 associated domain (HMA). Experimental evidence supports this hypothesis in AVR-
125 Pik, but the sHMA proteins' biological function and role in the infection process are
126 unknown. Another MAX effector, AvrPiz-t, targets four different host proteins involved
127 in different cellular processes [13,14].

128 In our previous study, we have explored the genetic diversity of the MAX
129 effector family in different *P. oryzae* host-specific lineages [7] in order to get insights
130 into the adaptive evolution and the selective forces that shape the molecular
131 variability of structurally analogous effectors. Focusing on the variations observed
132 within the predicted conserved MAX core, we identified a total of ~7800 orthologous
133 and paralogous sequences of putative MAX effectors clustered in 94 orthogroups
134 (OG). Pan-genome analyses showed that the MAX effector repertoire is highly plastic
135 compared to other secreted proteins, both in terms of the presence/absence of
136 orthogroups and the sequence variability within OG clusters. Interestingly, mapping
137 of polymorphic residues for the three effectors whose binding interfaces have been
138 experimentally characterized (AVR1-CO39, AVR-Pia and AVR-Pik) [7] suggests that

139 amino acid changes often co-localized with residues interacting with immune
140 receptors and, presumably, also with their host target proteins.

141 In order to further exploit the structure-based repertoire of *P. oryzae* MAX
142 effectors and get a complete view of their molecular diversity, a systematic and
143 robust analysis of their three-dimensional structure, especially outside the MAX core
144 domain, is still needed. Such analysis could allow a more comprehensive
145 classification of effector proteins within the MAX family and provide information on
146 how MAX effectors evolve and adapt to specific hosts. Indeed, many MAX effectors
147 possess N- and C-terminal extensions that could establish specific protein-protein
148 interactions playing a key role in the function of MAX effectors and their interference
149 with different host molecular processes. Examining in more details these extensions,
150 as well as other non-conserved structural features, may thus provide insights into the
151 mechanism by which MAX effectors acquire new virulence capabilities.

152 In the present study, we combined experimental and computational
153 approaches to finely characterize the structural diversity of MAX effectors. In order to
154 evaluate the reliability of the homology modeling approach we previously used for
155 predicting MAX effectors, we undertook structural studies of several MAX candidates
156 and succeeded in solving four new structures by Nuclear Magnetic Resonance
157 (NMR). Comparison of the new experimental MAX structures with their
158 corresponding 3D models generated by template-based or *ab initio* modeling
159 programs revealed the high accuracy with which AlphaFold2 (AF) can predict the
160 structure of MAX effectors, including for non-conserved side-chains in terminal
161 extensions that were not previously observed. We therefore revisited with the use of
162 AF the structural landscape of *P. oryzae* MAX effectors and validated the presence of
163 a MAX core in 77 of the 94 MAX OG clusters defined in Le Naour--Vernet et al. 2023

[7]. Structural alignment of the AF models allow us to describe the structural consensus and variability within the MAX family, from the canonical fold, but also including disulfide bond pattern variations, decorating secondary structures within the N- and C-terminal extensions that may be involved in protein-protein interactions as well as surface properties, including stickiness and electrostatics of the core domain of MAX effectors.

This work represents the most extensive structural analysis of a fungal effector family of a plant pathogen to date. It also provides valuable knowledge for analyses aimed at elucidating the function of MAX effectors, notably through the prediction of interaction sites within the MAX fold that could contribute to targeting host proteins during infection.

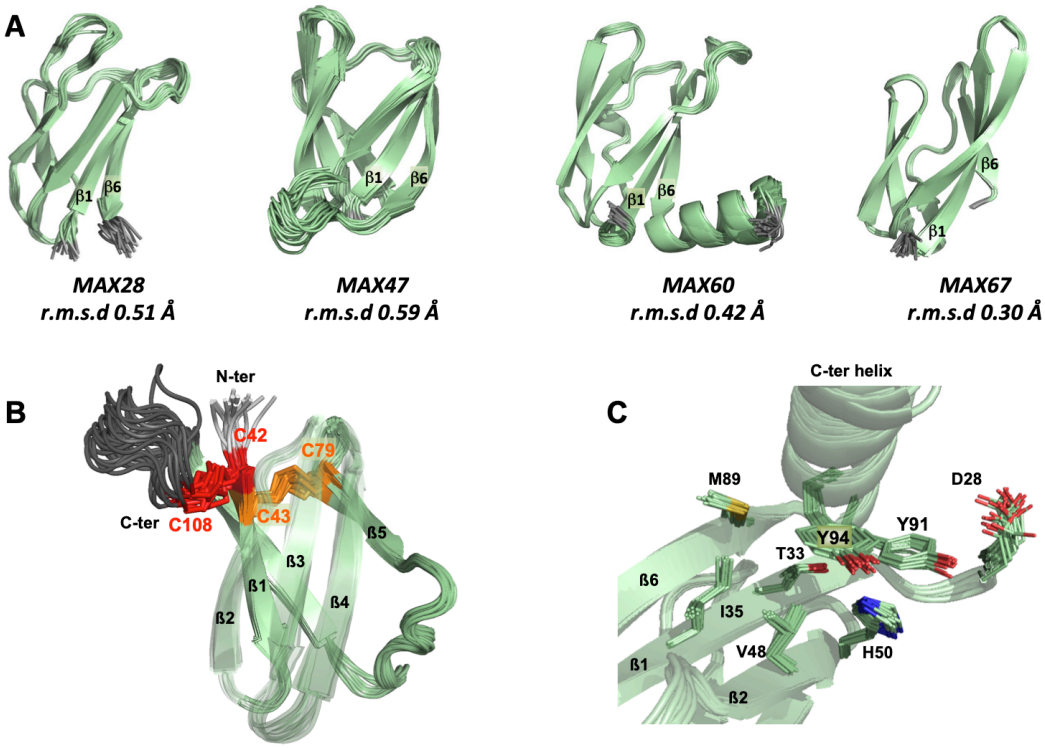
176 # Results

177 ## *NMR structures validate template-based modeling of MAX effectors*

178 In our previous phylogenic analysis of the MAX effector repertoire in *P. oryzae*,
179 we used a hybrid multiple Template Modeling (TM) strategy for predicting the 3D
180 structure of the conserved MAX core of representative sequences of 94 orthogroups
181 (OG) and we evaluated the reliability of these 3D models (referred as TM-pred
182 models) by calculating the TM-pred score [7]. The TM-pred score is an estimate of the
183 TM-score given by TM-align superimposition [15] that would be observed when
184 structurally aligning the TM-pred model with the corresponding experimentally
185 resolved structure. At the time this analysis was performed with only 8 experimental
186 structures of MAX effectors used as templates for homology modeling in a training
187 data set used to set-up of the TM-pred scoring function [7]. From the online table
188 https://pat.cbs.cnrs.fr/magmax/model/ (see also Materials and Methods) can be
189 downloaded the TM-pred models and their TM-pred scores calculated for the
190 predicted MAX core of 94 OG proteins (i.e. selected representative sequences of
191 MAX OG clusters, as listed in the S2 Table). Comparison of the TM-pred scores
192 shown in the S1 Fig indicates that ~90% of the OG proteins could be modeled with a
193 MAX structure having a predicted accuracy exceeding a TM-pred score > 0.6,
194 whereas only three TM-pred models of OG proteins (OG22, OG77 and OG85) had a
195 TM-pred score below 0.5 and were rejected as MAX effectors.

196 In order to assess the validity of these predictions, we aimed to resolve new
197 experimental structures not included in the set of templates used in training the TM-
198 pred score. We attempted to express genes corresponding to mature proteins from
199 10 OG clusters was attempted (S3 Table) for MAX effector candidates showing high
200 expression level during the biotrophic stage of infection [7]. Using NMR spectroscopy

201 we successfully determined the structure of four OG cluster representatives (OG28,
202 OG47, OG60 and OG67) and we confirmed that all four proteins are MAX effectors
203 (referred to as "MAX" instead of "OG clusters" from hereon). The 20 best-refined
204 conformers obtained for each of these effectors were superimposed (Fig 2A) and the
205 low root mean square deviations (r.m.s.d) support the high quality of the NMR
206 structures. The complete structural statistics are given in S4 Table.
207



208

209 ***Fig 2. NMR structures of four MAX effectors.*** *(A) NMR structures of MAX28, MAX47,*
210 *MAX60 and MAX67 showing the superimposition and the r.m.s.d. of their 20 best*
211 *conformers. The N- and C-terminal unstructured extensions before ß1 and after ß6,*
212 *respectively, are not shown. (B) View of the two disulfides bonds, C42-C108 (red) and C43-*
213 *C79 (orange) for NMR structure of MAX47. The loop between the end of ß6 and the C-*
214 *terminus is colored in dark grey. The ß2, ß3 and ß4 strands are transparent. (C) Local*
215 *environment of the two residues Y91 and Y94 in the C-terminal helix of MAX60.*
216

217 Among the four MAX effectors, three proteins (MAX28, MAX47 and MAX67)
218 displayed the characteristic MAX ß-sandwich fold consisting of two triple-stranded ß-
219 sheets stapled against each other through a conserved disulfide bridge linking ß1
220 and the ß4-ß5 loop (Fig. 1). The two cysteine residues forming this salt bridge are the
221 only residues that are highly conserved in MAX effectors. In MAX67, the ß1 and ß2
222 strands are longer than (10 a.a.) than their counterparts previously observed in other
223 MAX structures. A slight divergence from the canonical MAX fold was observed for
224 MAX60: the ß5 strand is replaced by a helical turn that prevents the formation of a
225 regular ß-sheet with ß4. In addition to the central MAX core, MAX28, MAX47 and
226 MAX60 presented remarkable N- and C-terminal extensions. For MAX47, the 23
227 residue-long N-terminal sequence extending before the ß1 strand was enriched in
228 serine residues and was not resolved in the NMR structure. The ß1 strand starts with
229 two consecutive cysteine residues. The first cysteine makes a disulfide bond with the
230 last C-terminal cysteine residue (C42-C108). The second cysteine in ß1 forms a
231 disulfide bond with the cysteine in the ß4-ß5 loop (C43-C79) as expected in
232 canonical MAX effectors. Both disulfide bonds are well defined in the NMR structure
233 (Fig 2B). MAX60 has a C-terminal extension, which forms a well-defined α-helix. This
234 C-terminal helix is attached to the structural core by hydrophobic contacts with the
235 rings of two tyrosine residues (Y91 and Y94). In the NMR experiments, close
236 contacts deduced from Nuclear Overhauser Effects (NOEs) data were found
237 between residues D28 to T33 and H50 for tyrosine Y91, and residues T33, I35, V48
238 and M89 for tyrosine Y94 (Fig 2C). In case of MAX28, cleavage of the His-tag after
239 affinity chromatography resulted in the precipitation of the protein that could not be
240 purified with sufficient quality for structural studies. Nonetheless, the N-terminal His-
241 tagged version was amenable to NMR analysis (Material and Methods). The 42

residue long C-terminal extension of MAX28 contains lysine-repeated motifs (KxxxK) and was predicted to be unstructured. The corresponding resonances were not assigned in the NMR spectra.

The four new NMR structures of MAX effectors were superimposed using TM-align with the corresponding TM-pred models that we previously generated by template homology modeling (Fig 3A). The quality of the models can be evaluated by the root-mean-square deviation (r.m.s.d) calculated between the observed and predicted structures and the TM-scores given by TM-align (a value of 1 meaning a perfect match). Comparison of the superimposed backbones shows that the overall MAX fold as well as the relative orientation of the two ß-sheets forming the central ß-sandwich were all well predicted, especially in the case of MAX28 for which the TM-pred model of the MAX domain matches very well the experimental structure (r.m.s.d.=2.11 Å), even for the loops joining the ß-strands. MAX28 was also the effector with the highest estimated TM-pred score (0.75), in remarkably good agreement with the true TM-score (0.74) of the TM-pred model aligned to the NMR structure. Structural predictions of the MAX core were also very good for MAX67, except for the long strands ß1 and ß2 observed in the NMR structure that could not be accurately modeled, explaining the rather low TM-pred value proposed for TM-pred_MAX67. Some ß-strands were also poorly defined in case of MAX47 and MAX60 whose models, compared to the NMR structures, exhibit also strong divergence for the connecting loop regions. For MAX47, the low reliability of the TM-pred model was anticipated by the TM-pred score (0.63).

267 ***AlphaFold2 reliably predicts MAX effectors core and extensions.***

268 The recent release of Artificial Intelligence (AI)-based methods have
269 demonstrated the high accuracy with which deep-learning software such as
270 AlphaFold2 (AF2) can predict the 3D structure of most *structured* proteins [16]. We
271 therefore decided to test the accuracy of AF2 to predict MAX effector structures by
272 submitting the amino acid sequences corresponding to MAX28, MAX47, MAX60 and
273 MAX67 for which the coordinates of the NMR structures we determined had not yet
274 been released in the PDB. We used three different implementations of AF2 to
275 generate the multiple sequence alignment (MSA) from which protein 3D models are
276 predicted. The MMseqs2 and Jackhmmer implementations use MSAs generated
277 automatically online or locally and that can include PDB templates, whereas the
278 *Custom* MSA implementation uses MSAs provided by the user and was
279 parameterized to not use PDB templates (see Materials and Methods). For each
280 sequence query, the three MSAs were independently submitted to AF2 and the
281 predictive quality of the top-ranked models was assessed thanks to the predicted
282 local distance difference test (pLDDT) [16,17]. The pLDDT score scaling from 0 to
283 100 is a residue-level accuracy score computed by AF2 that provides an estimate of
284 the confidence of each residue's predicted position in the protein structure. We
285 considered the average pLDDT score for the overall protein, or only for residues
286 within the predicted MAX core domain, herein called the MAX pLDDT score. The
287 MAX pLDDT score allows better filtering of the AF-generated models, since it
288 provides a quantitative measure of prediction quality focusing on the MAX core
289 domain. For each MAX effector, the AF model having the highest MAX pLDDT score
290 was selected and referred as its AF_MAX model (Fig 3).

The MAX core domain was very well predicted in AF_MAX47, AF_MAX60 and AF_MAX67 according to their high MAX pLDDT scores, close to or exceeding 90 (Fig 3B and S5 Table). All the three selected models were obtained from the Jackhmmer AF2 implementation. Lower confidence score (74.1) was retrieved for the best AF2 model of MAX28, generated using the *Custom* MSA implementation. Nevertheless, AF_MAX28 was very close to MAX28 experimental structure according to the average r.m.s.d. value (1.42 Å) calculated when superimposing the MAX core backbone atoms of the NMR conformers. Indeed, for all four MAX effectors, the MAX domain of the selected AF model displayed side-chain rotamers very similar to those observed in the experimental structures within the uncertainty of the NMR approach, i.e. density of NMR-derived constraints (Fig 3B).

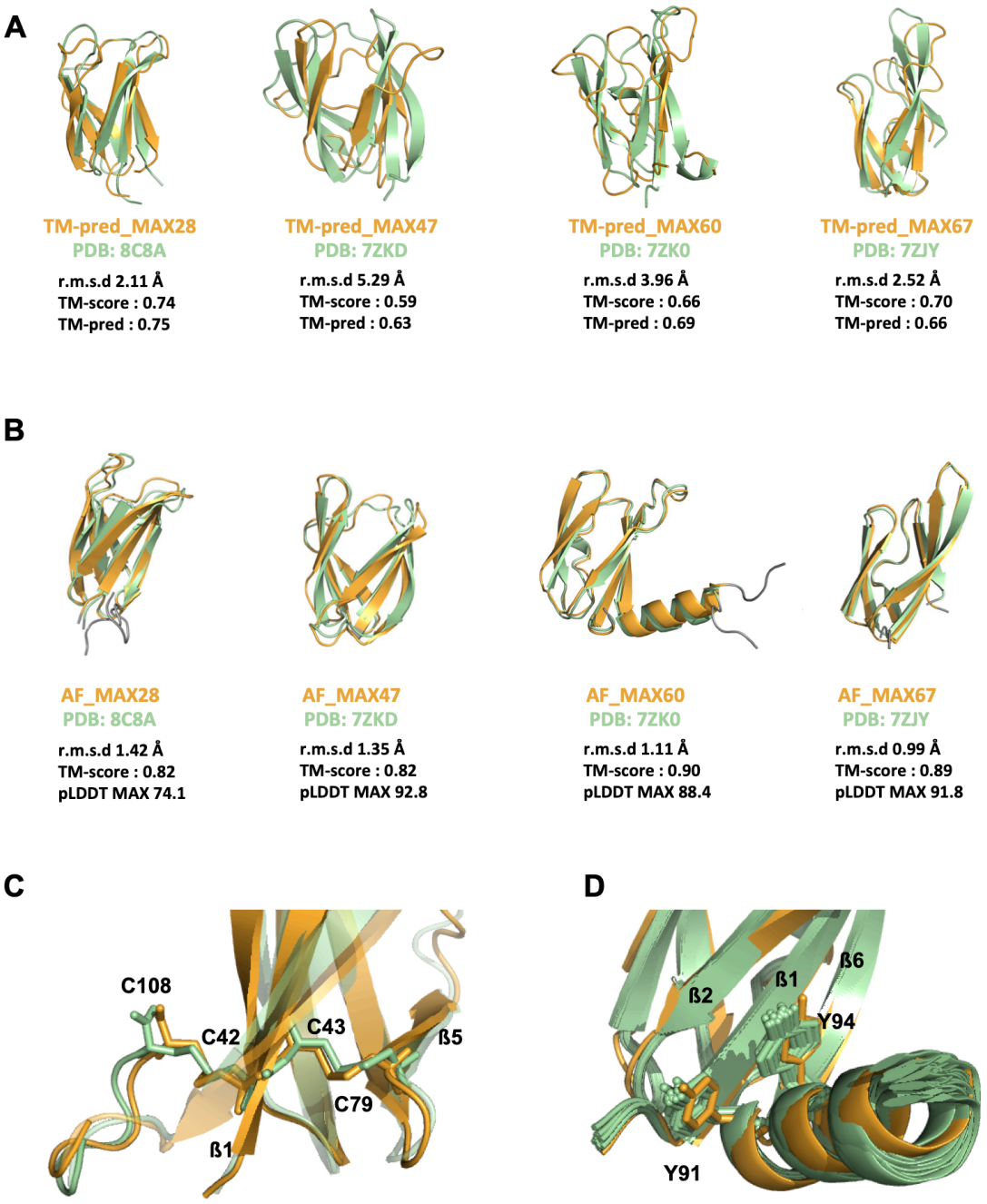


Fig 3. Comparison of newly determined NMR structures of MAX effectors with their TM-pred or AF predicted models. *Superimposition of the four MAX effectors determined in this study by NMR shown by the best model (in green) and of their corresponding 3D models (in orange) predicted by hybrid multiple template modeling (TM-pred models shown in A) or AlphaFold2 (AF models shown in B). Metrics used for the quantitative assessment of the similarities between the predicted 3D models and their respective experimental structures are indicated: the root mean square deviation (r.m.s.d; the lowest, the best), the template modeling score (TM-score from TM-align, a value of 1*

312 *corresponding to a perfect match), TM-pred score (a predictive estimate of the TM-score).*
313 *The N- and C-terminal boundaries were set according to the TM-pred models and did not*
314 *include extensions determined in the NMR structures. (B) The r.m.s.d. between backbone*
315 *heavy atoms of the superimposed NMR structure and AF models is given for the MAX*
316 *domain only (Table S5), as well as the predicted local distance difference test (pLDDT) that*
317 *was used to estimate the reliability of the AF predictions in the MAX core (MAX pLDDT*
318 *score, Table S5). (C) View of the two disulfides bonds of MAX47, C42-C108 and C43-C79,*
319 *as observed in the best NMR conformer and in the predicted AF_MAX47 model. (D) Local*
320 *environment of the two tyrosyl side-chains of Y91 and Y94 in the C-terminal helix of MAX60*
321 *in the 20 NMR conformers and in AF_MAX60.*
322
323 AlphaFold2 modeling also succeeded in predicting details within the core
324 domains. The cysteine residue side-chains forming the conserved disulfide bond "1"
325 are well defined in all four MAX effector models, as well as those forming an
326 additional SS bond bridging the N- and C-terminal extensions in the experimental
327 structure of MAX47 (Fig 3C). Another example of consistency between experimental
328 structures and AF models was the remarkably well-defined position and orientation of
329 the C-terminal helix of MAX60, in particular for two tyrosine residues whose aromatic
330 side chains stack over the ß1-ß2-ß6 ß-sheet of the MAX core (Fig 3D). For MAX28,
331 both AF model and NMR structure were consistent in predicting unstructured N- and
332 C-terminal extensions. Moderate deviations from the experimental structure were
333 observed for residues in the C-terminus of MAX67 (S2 Fig).
334

335 ## *Family portrait of AlphaFold2 MAX models*

336 Given the high predictive quality of the AF-generated models of MAX effectors,
337 we applied the same AF modeling strategy to all other *P. oryzae* OG representative
338 members. The models were visualized to check the presence of the characteristic
339 MAX core by inspecting the ß-strand topology and formation of the conserved SS "1"
340 disulfide bond (Pymol. v.1.6; Delano 2002). OG proteins showing significant

topological deviations from this canonical MAX fold were discarded. The strict presence of the short ß5 strand was not used as a filtering criterion. Among the 80 AF models that matched the canonical MAX structure (Fig 4), 57 had MAX pLDDT scores greater than 80 and 20 had MAX pLDDT scores ranging from 60 to 80. Only three OG proteins, OG26, OG73 and OG94, exhibiting a central core compatible with a MAX fold had a MAX pLDDT score below 60 and were not kept in our final selection of 77 validated MAX structures.

Among the 14 OG proteins that are not predicted to fold with the MAX topology, three (OG22, OG77 and OG85) were previously rejected based on their low TM-pred score and five (OG04, OG59, OG65, OG68 and OG81) had a TM-pred score under 0.6 (S1 Fig). Other OG proteins such as OG51 and OG54 that exhibited a TM-pred score above 0.6 compatible with a MAX structure displayed significant distortions from the canonical MAX fold when modeled by AlphaFold2.

For OG51, two models computed with MMseqs2 and the AF Jackhmmer implementation gave very similar models (backbone r.m.s.d. of 1.77Å) with pLDDT overall scores of 52.6 and 61.9, respectively. However, the C-terminal ß-strand of OG51 had a parallel orientation relative to the first ß1 strand that was not compatible with the MAX topology. The best OG54 model had a pLDDT score of 59.4 but deviated from the MAX topology by the absence of the C-terminal ß6 strand, which was not accurately modeled.

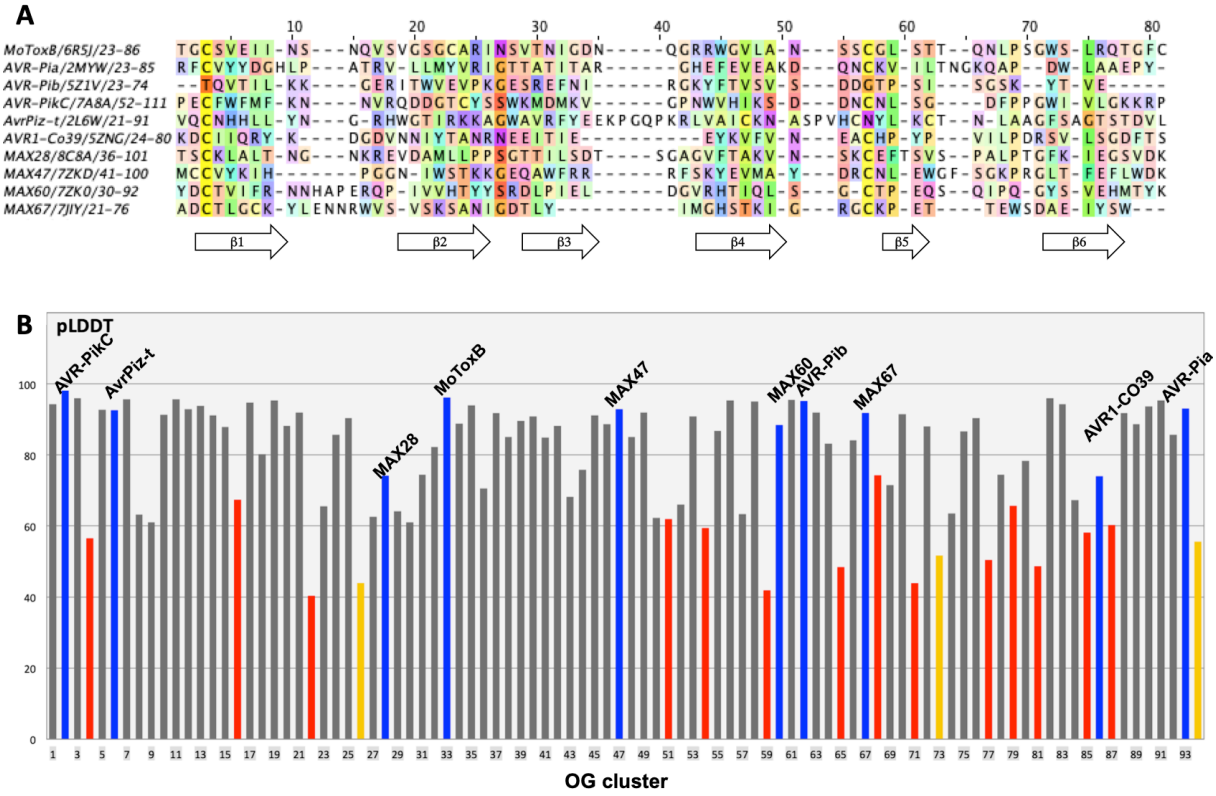


Fig 4. pLDDT scores of AF models for known and predicted MAX effectors.

(A) Structural alignments of experimentally determined MAX effector structures using MoToxB structure for reference. Residues are colored according to the Taylor scheme [18](Taylor, 1997) and conservation (above 15% threshold) is used as a shading factor.

(B) pLDDT scores of the best AF models selected for the 94 OG clusters of predicted MAX effectors with experimentally determined structures in blue, and OG clusters that displayed canonical MAX folds in grey. For these OG clusters we report the best MAX pLDDT score. OG clusters represented in red are the ones for which none of the models matched the canonical MAX fold, and in such cases we report the best overall pLDDT score of the model. MAX pLDDT scores of models that have a canonical MAX fold but a score below 60 are shown in orange. Full data is available in S5 Table.

The 14 OG clusters that gave inconclusive AF models were submitted to two other protein structure prediction web-servers, RaptorX [19] and RosettaFold [20]. None of the computed models displayed the canonical MAX fold, confirming that these OG cluster sequences were difficult to model and quite probably do not exhibit the expected MAX topology.

380 ***Variations around the canonical MAX fold***

381 A statistical overview of the general structural features characterizing the 77
382 AF models of validated MAX effectors is given in Tables S7a and S7b. A unique
383 member of the MAX family is MAX52, whose AF model consists of two MAX core
384 modules in tandem. The MAX52 model was cut into two chains (MAX52A and
385 MAX52B), each corresponding to a MAX core domain. In Fig 5A is indicated the
386 average size of the ß strands and connecting loops forming the conserved 6
387 stranded ß sandwich that defines the canonical MAX fold, well represented by the
388 MoToxB (MAX33) structure. ß1 and ß2 are usually of similar size and associate
389 together with ß6 to form the longest anti-parallel ß sheet, while ß3 and ß4 strands are
390 usually shorter. Substantial variations around this common fold can be observed,
391 notably in the C-terminal part of the MAX core where the short ß5 strand can be
392 totally absent (MAX78 and MAX83) or replaced by a helix (MAX20), whereas the ß5
393 strand can range from 2 to 8 residues following a loop that can have up to 25
394 residues (MAX29) or present a short helix (MAX15, MAX60). The number of disulfide
395 bonds stabilizing the protein can also greatly varied (Table S7a), from none in
396 MAX61 and MAX62 (AVR-Pib) to three in MAX46 or four in MAX52. Asides of the
397 conserved disulfide bond (SS “1”), which is a hallmark of the MAX domain, three
398 types of additional disulfide bonds (SS “2”, “3” and “4”) are found in the observed or
399 AF predicted MAX structures (Figs 5A and B). SS “4”, joining the N-terminus of ß1 to
400 a C-terminal cysteine, is well defined in both MAX47 and MAX92, although it was not
401 present in any of the previously determined MAX 3D structures that could serve as
402 template. This disulfide bond was experimentally validated by our NMR structure of
403 MAX47.

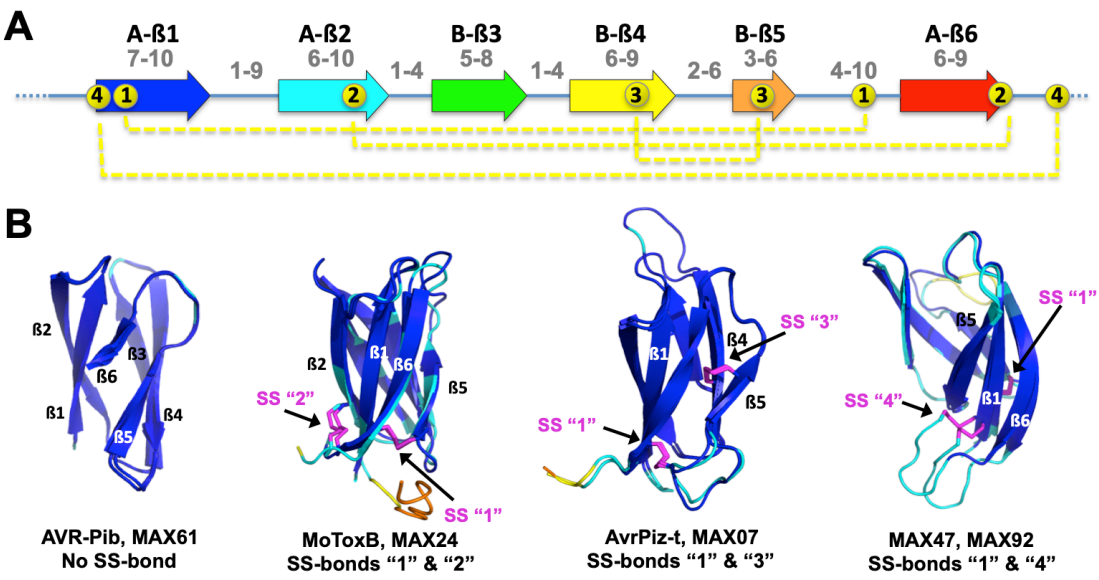


Fig 5. MAX domain structural features. *(A) Average size range (indicated in grey) of ß strands (arrows) and connecting loops forming the central ß sandwich of the canonical MAX core, and of the N-ter and C-ter extensions. Average ranges are calculated in Table S7b. The 4 different types of disulfide bonds observed in MAX structures and models are indicated by dotted yellow lines. (B) Four different sets of structural models illustrating the variability of disulfide bond patterns. The AF models are colored according to their pLDDT score, by blue for high accuracy (>90), cyan for backbone at good accuracy (> 70), yellow for low confidence (> 50 and < 70) and orange for disordered (< 50). The disulfide bonds are shown in magenta, except for the AVR-Pib structure, which does not have a disulfide bond.*

N- and C-terminal extensions

Over two-third of the 77 AF_MAX models possess peptide segments with 15 or more residues extending at one or both ends of the central MAX domain (Table S7a and S7b). Diverse length of these extensions, especially in C-terminal, can be observed among OG sequences belonging to the same OG cluster (e.g. OG01, OG02 or OG15 clusters in Table S1a). C-terminal extensions are also more numerous and usually longer than N-terminal extensions. They are often modeled by AF2 with well-defined secondary structures, such as additional ß strands extending the ß2ß1ß6 sheet by one or two strands (e.g. MAX08, MAX12, MAX25), or a terminal

424 helix as observed in the model and solution structure of MAX60. There are also many
425 examples where terminal extensions appear as unstructured regions that could not
426 be modeled with high confidence by AlphaFold2. Long intrinsic disordered regions
427 (IDRs) of more than 30 a.a. [21–23] may have diverse function in bacterial [24,25]
428 and fungal effectors [26,27] and we had therefore searched for IDR signatures in
429 MAX effector sequences using ESpritz prediction software [28]. Long IDRs were
430 predicted for only a few MAX effectors and their presence was cross-validated for six
431 AF_MAX models: in MAX15 (118 a.a.), MAX27 (36 a.a.) and MAX43 (43 a.a) as N-
432 ter extensions, and in MAX28 (42 a.a.), MAX53 (49 a.a.) and MAX78 (43 a.a.) as C-
433 ter extensions. It thus appears that long IDRs is not a frequent feature among *P.*
434 *oryzae* MAX effectors, present in less than 8% of all modeled structures. The NMR
435 solution structure of MAX28 validated the unstructured nature of its C-terminal
436 extension.
437

438 ***Structural classification of MAX effectors***

439 Structural similarities between proteins can help to elucidate the function of an
440 uncharacterized protein and infer its molecular evolution. Here we performed
441 hierarchical clustering of the selected 77 MAX effector models with two protein
442 structure alignment software, Dali [29] and TM-align [15,30] which use different
443 criteria for similarity scoring of superimposed structures. The Dali Z-score relies on
444 secondary structure pairing and is a good estimate of topological conservation while
445 the TM-score is computed for the whole alignment and weights paired residues with
446 low r.m.s.d. more strongly than those that are more distant. When analyzed
447 independently, the structural alignment trees retrieved from these two clustering
448 approaches did not reveal clear sub-families of MAX structures, as shown by the lack

of long internal branches in Fig 6. To facilitate the comparison of the trees, we differently colored the lines connecting each MAX model in both trees (see Materials and Methods).

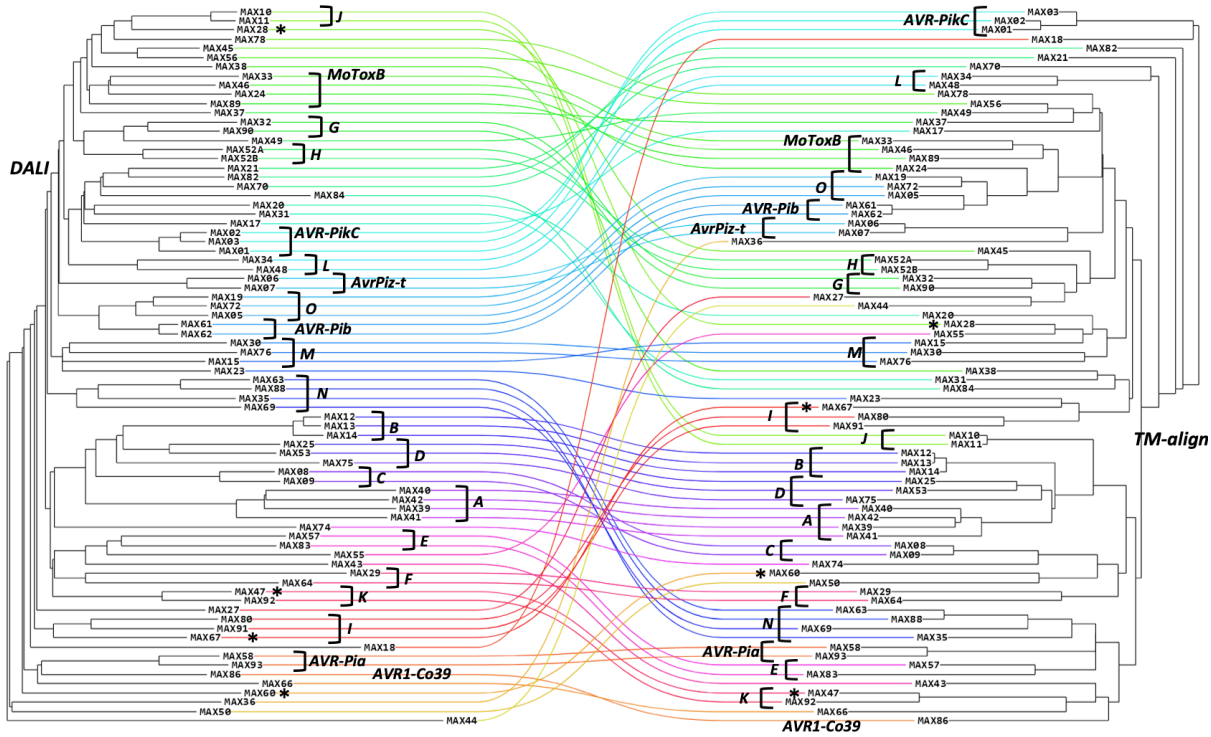
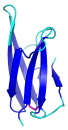
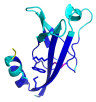
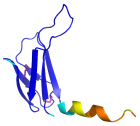


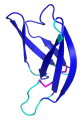
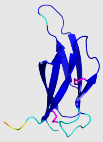
Fig 6. Comparison of the structural similarity trees of MAX effectors based on the Dali Z-score (left) and TM-alignTM-score (right) of their superimposed AF models. *Unstructured N- and C-terminal regions were removed from the AF_MAX models prior to the analysis. A line of a specific color connects each AF_MAX model in the Dali and TM-align trees. The MAX effectors with an experimental 3D structure are indicated by their name, or a star for the four novel MAX structures. The structural groups to which the MAX effectors were assigned are indicated by brackets and illustrated in Table 1 and S3 Fig.*

In this representation, any bundle of lines of similar colors highlights a possible structural similarity between the models that was common to both clustering methods and that could define a group of MAX models. Each group of at least two models was visually inspected for additional secondary structures that could add to the MAX

467 core, the disulfide bond pattern as well as for specific structural features that fall
468 outside the statistical average values reported in Fig 5A. Using this dual clustering
469 method, we defined 15 groups of MAX models sharing common structural features,
470 in addition to the 5 groups of well-established MAX effectors (AVR-Pia, AVR-Pib,
471 AVR-PikC, AvrPiz-t and MoToxB). Groups A to G gather MAX models exhibiting
472 major structured elements adding to the canonical MAX core: groups A and B contain
473 models with 2 extra strands, groups C and D contain models with 1 extra strand, and
474 groups E to G contain models with C-terminal helical extensions after ß6. Group H
475 consists of the MAX52 tandem domains connected by a structured linker. In contrast,
476 groups I to O correspond to plain MAX structures, without ostentatious decoration but
477 presenting variations of the MAX fold specific to each group. The characteristics of
478 this total of 20 MAX structural groups are summarized in Table 1 and illustrated with
479 more details in S3 Fig. Only 6 groups exhibits a disulfide bond pattern with a specific
480 SS bond adding to the conserved SS"1", and only the AVR-Pib group has no SS
481 bond. The MAX models listed in Table 1 represent about two-third of the 77 MAX
482 effectors that we validated in the present analysis. Several of them had been
483 identified in a previous computational structural genomics study of *M. oryzae* strain
484 70-15 (see below) and are also listed in Table 1.
485

486 ***Table 1. Structural groups identified in* P. oryzae *MAX effector family***

Group	representative structure	MAX Id	SS(*)	Structural Specificities and Extensions
A		*MAX39, MAX40,* ***MAX41,*** *MAX42, MGG_16939T0 MGG_18062T0*	*1*	additional ß7ß8 strands flanking one side of the ß2ß1ß6 sheet and C-terminal extensions after ß8 that are in helical conformations
B		***MAX12,*** *MAX13, MAX14 MGG_16619T0*	*1*	additional ß7ß8 strands flanking one side of the ß2ß1ß6 sheet
C		***MAX08,*** *MAX09*	*1*	presence of the ß7 additional strand running antiparallel to the ß6
D		***MAX25,*** *MAX53, MAX75 MGG_02635T0 MGG_15207T0 MGG_11304T0*	*1*	presence of the ß7 additional strand running antiparallel to the ß6
E		*MAX57,* ***MAX83***	1	long ß1 and ß6 strands,C-terminal helical extensions that are stabilized by hydrophobic interactions
F		*MAX29,* ***MAX64*** *MGG_17255T0*	*1*	C-terminal helical extensions that are stabilized by local residue-specific interactions (disulfide bond or hydrophobic residues)
G		*MAX32,* ***MAX90*** *MGG_08469T0*	*1, 3*	long loop joining the ß1 and ß2 strands and helical C-terminal extensions that are not strictly conserved in their lengths and orientations
H		***MAX52***	*1, 3*	two MAX domains
AVR-Pia		*MAX58,* ***MAX93***	*1*	ß2 strand interaction with HMA (**)
AVR-PikC		*MAX01,* ***MAX02****, MAX03 MGG_15972T0*	*1*	ß3 strand and N-terminal extension interacting with HMA (**)

Group	representative structure	MAX Id	SS(*)	Structural Specificities and Extensions
I		***MAX67,*** *MAX80, MAX91 MGG_16175T0*	*1*	long ß1 strand paired with long ß2 or ß6 strands
J		*MAX10,* ***MAX11***	*1*	presence of a short helix in the connecting loop between ß5 and ß6
K		***MAX47,*** *MAX92*	*1, 4*	disulfide bond "4" joining the N-terminus of ß1 and the C-terminal extension
L		***MAX34,*** *MAX48 MGG_15625T0 MGG_08992T0*	*1, 3*	malformed ß2 strand and anti-parallel ß4ß5 strands stabilized by a disulfide bond
M		*MAX15,* ***MAX30****, MAX76 MGG_10282T0*	*1*	longer ß1 and ß2 strands and their connecting loops
N		***MAX63,*** *MAX88, MAX35, MAX69 MGG_02207T0*	*1*	unusually long loop making the connection between strands ß2 and ß3
O		*MAX05,* ***MAX19****, MAX72 MGG_12426T0 MGG_14793T0 MGG_18060T0*	*1*	Structuraly similar to AVR-Pib.
AVR-Pib		*MAX61,* ***MAX62***		Structuraly similar to group O models.
AvrPiz-t		***MAX06,*** *MAX07 MGG_18041T0*	*1, 3*	very similar structures with the main difference for the ß3ß4 connecting loop
MoToxB		*MAX24,* ***MAX33****, MAX46, MAX89 MGG_17132T0*	*1, 2*	Structuraly similar to ToxB with disulfide bond "2"

487 *The AF models are colored by their pLDDT scores (see legend of Fig 5). The groups with*
488 *major structural variations (addition of secondary structures) are listed in the left-hand*
489 *panels, including the MAX domain duplication of MAX52. The remaining groups, from I to O*
490 *that do not have additional secondary structures but displayed specific structural variations of*
491 *the MAX core domain itself are shown in the right-hand panels. To complete the overview,*
492 *the five groups of well-established MAX effectors (AVR-Pia, AVR-Pib, AVR-PikC, AvrPiz-t*

493 *and MoToxB) are shown at the bottom and highlighted in grey color. The MAX Id column*
494 *gives the AF model identifiers of the MAX effector list reported in the present study*
495 *(representative MAX model indicated in bold) and the corresponding MAX effectors reported*
496 *by Seong & Krasileva, 2021 [3]; Seong and Krasileva, 2023 [4]; Yan and Talbot, 2023 [5] in*
497 *the M. oryzae strain 70-15 are referred to by their MGG identifier (S8 Table).*
498 *(*) Type of disulfide bond as defined in Fig 5*
499 *(**) from crystallographic structures of complexes*
500

501 ## *Many MAX effectors are singletons*

502 Singletons of MAX effectors are defined as those with no obvious structural
503 relationship to other MAX effectors. They constitute a third (26/77) of the MAX
504 effectors we modeled. In the majority of cases (17), these MAX effectors consist of a
505 simple MAX core with unstructured N- and/or C-terminal extension(s). Among them is
506 AVR1-CO39 (MAX86), which is specifically absent from the *Oryza*-infecting lineage,
507 and whose structure has been experimentally determined. The remaining 9
508 singletons MAX effectors possess very diverse structured extensions asides the
509 folded MAX core domain as illustrated in S4 Fig. Among these MAX effectors,
510 MAX49 was already reported in *M. oryzae* strain 70-15 corresponding to
511 *MGG_08482T0* [5]. In MAX49, N-terminal and the C-terminal helices form a helix
512 bundle structure, not observed in other MAX effector models. MAX74 has an
513 additional ß7 strand running anti-parallel to ß6, as observed for group D models.
514 Three other models have a ß7 strand, i.e. MAX36, MAX43 and MAX66, the latter
515 having long ß1 and ß6 strands. Additional C-terminal helices were modeled with
516 good accuracy in MAX55 and MAX60, but not in MAX44 and MAX84.
517

518 ## *MAX domains exhibit very variegated surface properties*

519 Comparison of the molecular surfaces of homologous proteins can highlight

520 common or specific features related to their function. This can be challenging in case
521 of proteins that differ in size or have structural elements adding to their common fold.
522 Here we have performed a detailed comparative analysis of the surface properties of
523 the standalone MAX domains extracted from 49 AF_MAX models in which no
524 structured regions are predicted to interact with the central core (S7a Table). For this
525 subset of MAX domains, we computed SURFMAP [31] 2D projections of their
526 molecular envelop and compare the distribution of different surface features i.e.
527 exposed secondary structures, electrostatic potential, stickiness (Fig 7) and amino-
528 acid polymorphism within the OG cluster to which belongs each representative MAX
529 model.

530 The ß1-exposed surface is mostly discontinuous around the ß2 surface that is
531 located at the *northern* pole of the projection (Fig 7C). The continuous ß3 surface lies
532 below the ß2 surface and the discontinuous ß4 surface is found at the bottom of the
533 projection. The ß6 surfaces are found to the *west* and *east* of the ß1-ß2 surface
534 areas whereas ß5 surfaces are found on the most *eastern* areas. The electrostatic
535 potential maps (Fig 7D and S2 files) reveal that MAX domain surfaces are more often
536 positively charged than negatively charged or neutral, and that the molecular
537 surfaces can appear entirely positive (e.g. MAX06 (AvrPiz-t), MAX23, MAX34,
538 MAX47) or negative (e.g. MAX10, MAX38, MAX78, MAX86 (AVR1-Co39)), or
539 present intense electrostatic patches (e.g. MAX02 (AVR-PikC), MAX58, MAX62
540 (AVR-Pib), MAX80). It is well known that positively charged regions in proteins are
541 important for interaction with negatively charged macromolecules, such as nucleic
542 acids and lipopolysaccharides [32], whereas negatively charged protein surfaces can
543 be involved in membrane attachment or DNA mimicking functions [33–35]. In
544 MAX47, we noticed that its unstructured N-terminal extension is rich in Asp residues,

545 suggesting that it could make transient interactions with the positively charged MAX
546 core in the absence of its cellular target. In MAX62 (AVR-Pib), a surface loop region
547 forms a strong positive patch (Fig 7C), which has been shown to be essential for the
548 avirulence function of AVR-Pib and its nuclear localization in host cells [35].
549 Interestingly, a similar positive patch is visible on the surface of its structural
550 homologs MAX61 belonging to the same structural group, as well as in MAX05 and
551 MAX72 belonging to group O (Table1), suggesting that these effectors may also rely
552 on this positive surface loop for their function. Similarly, AvrPiz-t displays a positively
553 charged surface partially involving lysine residues that are required for AvrPiz-t
554 avirulence and virulence functions in rice [14]. Inversely, while the surface of MAX86
555 (AVR1-CO39) is strongly negative, that of MAX93 (AVR-Pia) is neutral, yet both
556 AVR1-CO39 and AVR-Pia interact similarly through their ß2 strand with the HMA
557 domain from the rice immune receptors RGA5 [36] and Pikp-1 [37], respectively.

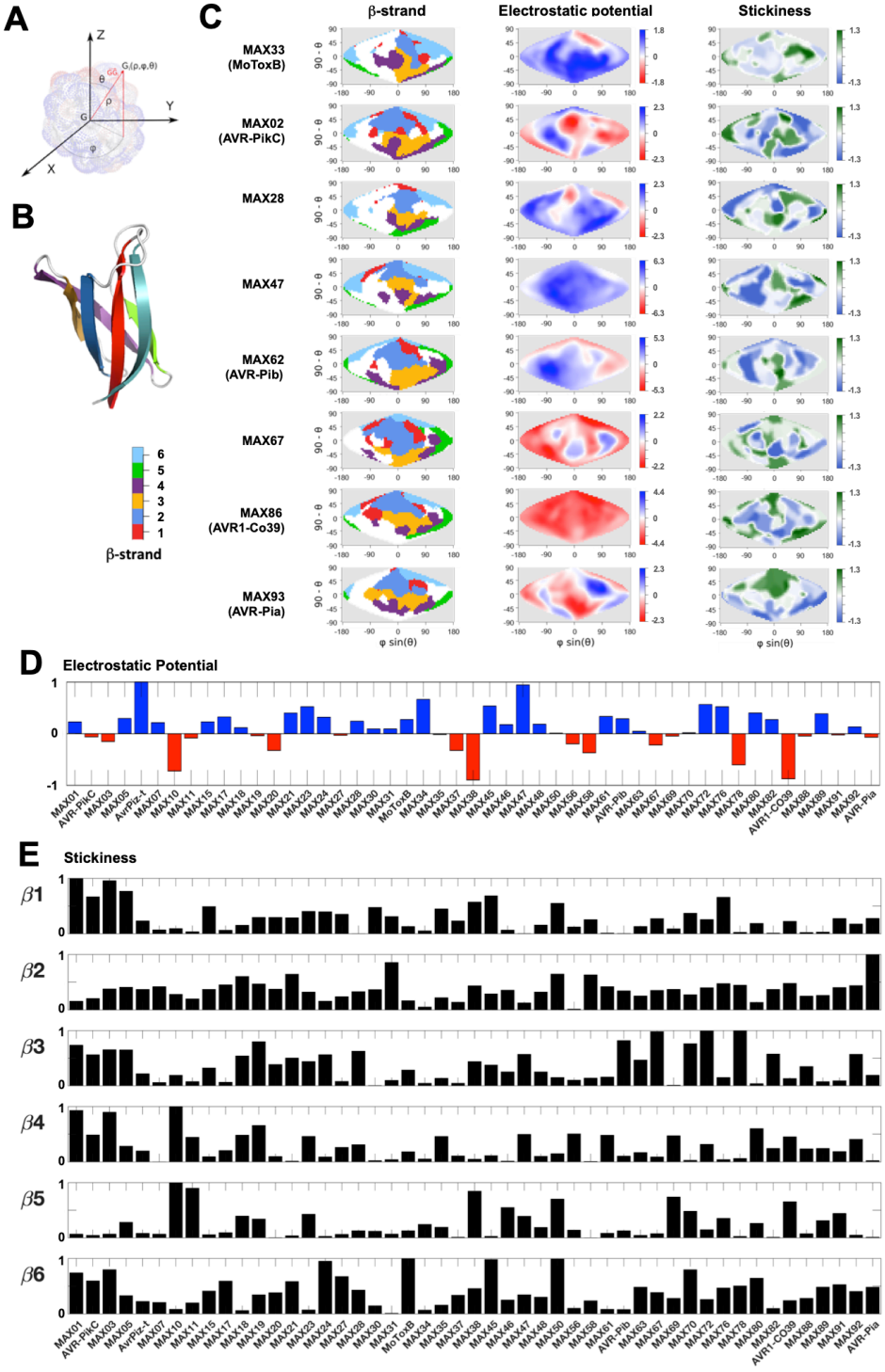


Fig 7. Surface properties of the MAX core domains.

*Surface properties of MAX core domains computed and represented using SURFMAP. (**A**) Schematic of calculation of the spherical coordinates (from Schweke et al., 2022). The coordinates of each surface particle G_i is expressed in spherical coordinates (ρ, φ, θ), where ρ represents the distance of the particle G_i to the center of mass G of the protein, φ is the angle between the X axis and the projected vector $\overrightarrow{GG_i}$ in the plan ($\overrightarrow{GX}$, $\overrightarrow{GY}$), while θ is the angle between the vector $\overrightarrow{GG_i}$ and the Z axis.*

565 ***B**) MoToxB structure showing the 6 ß-strands of the MAX core with the color code used for the*
566 *surface representation in panel C.* ***C**) 2D maps of exposed ß-strands, electrostatic potential and*
567 *stickiness of the molecular surfaces calculated by SURFMAP for MAX domains extracted from AF2*
568 *models of MAX effectors with known structure. N- and C-terminal extensions were discarded and the*
569 *MAX core 3D models were all superimposed to AF_MAX33 (MoToxB) giving a reference frame for the*
570 *Sanson-Flamsteed 2D projection computed surfaces. The ß-strand maps use the color code given in*
571 *panel B, the electrostatic potential maps are scaled in the indicated kT/e units and the stickiness*
572 *(related to hydrophobicity) scale is that defined by Levy E. D., 2010* [38]. ***D**) Comparison of overall*
573 *surface electrostatic potentials of MAX core domains, summed over the entire molecular surface and*
574 *normalized by the highest absolute value calculated for the subset of 49 MAX effector models (S2*
575 *files).* ***E**) Relative surface stickiness of MAX core ß-strands. Stickiness values were summed for*
576 *residues forming each of the 6 ß-strands of the MAX core domains and normalized by the highest*
577 *value calculated for each strand in the subset of 49 MAX core models.*
578

579 Wide divergence is also observed in the surface stickiness of the MAX
580 domains. Surface stickiness is mostly related to surface hydrophobicity and reflects
581 the propensity of each amino acid to be involved in molecular interfaces [38]. For
582 AVR-Pia, the HMA binding site correlates well with the presence of a large
583 hydrophobic patch on the surface (Fig 7C and panel C in S5 Fig) and a very sticky ß2
584 strand, also present in MAX31 (Fig 7E). This is however not the case for AVR1-CO39
585 whose surface hydrophobicity is limited except in strand ß5. This strand is particularly
586 variable in length and surface properties, exhibiting high stickiness in only a few MAX
587 effectors (e.g. MAX10, MAX11 and MAX38) where it could contribute to the binding
588 site of host target proteins. For the AVR-Pik group (MAX01, MAX02 and MAX03)
589 surface stickiness is high in ß1 and sticky patches are also observed in strands ß3,
590 ß4 and ß6. In all these effectors, the ß3 stickiness could serve in an interaction with
591 strand ß4 of HMA domains, as observed in complexes of different MAX02 effectors
592 with the HMA domain from Pikp-1 or from the rice protein OsHIPP19 targeted by
593 AVR-PikF [8–10]. Other MAX effectors (e.g. MAX67, MAX72 and MAX78) possess a
594 highly sticky ß3 strand that could also associate with the ß-strand of an HMA domain

595 or other type of protein domain. In the crystal structures of the AVR-Pik effectors, the
596 anti-parallel ß1-ß6 strands of the MAX core are making hydrophobic contacts with
597 residues in their N-terminal extension which adopts a conserved extended
598 conformation and considerably expands the binding interface with the HMA (S5 Fig).
599 In MAX structural groups A to D of Table 1 (not included in the present subset), a
600 sticky ß6 strand often associates with an anti-parallel ß7 extending the MAX core ß-
601 sheet. Similarly, the highly hydrophobic ß6 strand present in MAX33 (MoToxB),
602 MAX45 and MAX50 could interact with target proteins through an antiparallel ß-
603 strand arrangement.

604 Altogether, this analysis highlights the very variegated surface properties
605 exhibited by the *P. oryzae* MAX effectors. In spite of sharing a common fold, these
606 sequence diverse proteins retain very high diversity at the structural level. On the
607 other hand, sequence conservation is high inside each OG cluster with average
608 conservation scores in strands and loops close to the maximum conservation score
609 of 9 (S7c Table). Three clusters display low conservation scores in ß2, ß3 and ß6
610 strands represented by MAX47, in ß2 and the loop joining ß4 to ß5 for MAX63, and
611 in strands ß4 and ß5 for MAX70, and accordingly, higher polymorphism on their
612 surfaces (S6 Fig and S2 Files).

613

614 # Discussion

615 We initially used hybrid multiple template modeling that combines information
616 from multiple template structures to predict the core of MAX effector structures. The
617 best models were selected according to their predicted TM-score (TM-pred), a
618 commonly used measure to assess the structural similarity between predicted
619 models and the true native structure. By comparing new experimental MAX effector

620 NMR structures with their corresponding TM-pred models indicated the low accuracy
621 of the models, pointing to the limits of template-based homology modeling [39] and
622 the need of alternative modeling strategies to model MAX effectors.
623
624 *Ab initio modeling*
625 Recently, *ab initio* modeling, specifically using Rosetta and the two web
626 servers Robetta and QUARK, was applied for predicting the structures of already
627 known MAX effector proteins of *Magnaporthe oryzae* [40]. The Rosetta *ab initio*
628 modeling approach generated models with the best overall TM-scores, with AVR-Pib
629 achieving the highest TM-score. The exception to this trend was AVR-Pik, for which
630 the QUARK method produced the best model with a TM-score of 0.89, compared to a
631 TM-score of 0.51 in Rosetta. Overall, the study highlights the strengths and limitations
632 of *ab initio* modeling for predicting the structures of fungal effector proteins. While *ab*
633 *initio* methods can provide valuable insights into the overall fold and topology of
634 proteins, accurately predicting detailed secondary structures remains challenging,
635 especially in the cases where the TM-scores are relatively low. The performance of
636 different *ab initio* methods can also vary depending on the specific protein being
637 modeled. To address the issue of missing secondary structures or other structural
638 defects, further refinement techniques or hybrid methods can be employed. These
639 approaches aim to improve the accuracy of the predicted models by incorporating
640 additional information, such as experimental data, molecular dynamics simulations
641 [41], fragment assembly [42], or consensus modeling that can be used to refine and
642 improve the quality of predicted models. Although this can improve the robustness
643 and accuracy of the predicted models, it involves additional computational steps and
644 requires the availability of multiple predictions.

645 *Modeling of MAX effectors from M. oryzae strain 70-15 using deep learning*

646 TrRosetta and AlphaFold are novel computational methods based on deep
647 learning techniques to predict protein structures from amino acid sequences,
648 bypassing some of the computationally-intensive steps required by traditional
649 methods. The MAX family repertoire of *M. oryzae* strain 70-15 was investigated and
650 models were generated by computational methods mostly using TrRosetta [3] and
651 AlphaFold [4,5]. The first study (TrRosetta models) identified 11 MAX candidates and
652 similar population sizes of 26 and 32 MAX candidates were reported for the two other
653 AlphaFold studies, respectively. While 10 MAX of these candidate models were
654 commonly identified and systematically assigned to MAX effectors by all three
655 studies, another set of 22 MAX effectors were identified in less than three studies (S8
656 Table), indicating the difficulties in model selection based on pTM-score and in
657 detecting structural similarity using methods like CATH [43] and SCOPe [44].

658 Most of the MAX effectors that were previously reported in the TrRosetta and
659 AlphaFold studies were also identified in our analysis (Table 1 and S8 Table) with the
660 exception of OG54 that we did not select as a MAX effector. MAX members of the
661 groups C, E and H were not found in these previous reports while the equivalent
662 sequences in the *M. oryzae* strain 70-15 for MAX members of groups J and K could
663 not be found.

664 In our study we used the pLDDT score (predicted *Local Distance Difference*
665 *Test*), which provides a residue-level assessment of the confidence or accuracy of
666 each residue's predicted position in the protein structure. By focusing on the pLDDT
667 scores within the MAX folded region, we specifically evaluated the local accuracy of
668 the predicted structure within the MAX core domain. This localized scoring approach
669 enables a more focused evaluation and filtering of the protein structure predictions,

670 and allows identification and selection of the best models with higher accuracy and
671 reliability within the specific region of interest. By comparing the experimental results
672 from NMR with the corresponding AF models, we were able to assess the accuracy
673 and reliability of the AF predictions.
674
675 *Venturia inaequalis MAX-like effectors*
676 *V. inaequalis* is an ascomycete fungus, in the *Venturiaceae* family, responsible
677 for apple scab disease. Assessing models of *V. inaequalis* MAX-like effectors [45]
678 against the criteria we defined here to assign MAX effectors to subfamilies, revealed
679 that none of the models fitted into any of the MAX subfamilies. Indeed, *V.*
680 *inaequalis* MAX effectors present three remarkable disulfide bonds. One is the
681 canonical "1" bond found in other MAX effectors, while the remaining two were not
682 found in any of the MAX effectors from *P. oryzae*. Moreover, MAX-like effectors of *V.*
683 *inaequalis* usually possess a C-terminal helical extension connected to the MAX core
684 domain *via* a specific disulfide bond that was not observed in any MAX from *P.*
685 *oryzae*. This makes these newly discovered MAX-like effectors from *V. inaequalis*
686 another subfamily with unique sequence/structure features [45]. As *V.*
687 *inaequalis* exclusively colonizes and releases effectors in the subcuticular host
688 environment without penetrating the underlying epidermal cells, their function, and
689 thus targets, are probably fundamentally different from the other studied MAX
690 effectors from *P. oryzae*.
691
692 *MAX effector AlphaFold models predicted in 120 P. oryzae genomes*
693 Clustering of the 77 MAX effector models successfully predicted by AlphaFold
694 allowed us to distinguish 15 subfamilies. Our study showed that, beyond the well-

695 known conserved MAX core, the MAX superfamily is characterized by an important
696 structural diversity shown by additional structured regions in the C-terminal portions
697 of several of its members. Specifically, these include additional strands observed in
698 groups A to D and helical extensions observed in groups E to G. These regions may
699 play a role in protein-protein interactions or contribute to the effector's overall
700 functionality. On the other hand, long intrinsically disordered regions (IDRs) were
701 rarely observed in only six MAX effectors. IDR regions lack a stable 3D structure and
702 exhibit conformational flexibility, allowing them to interact with multiple binding
703 partners and fulfill various functions [46] but are difficult to predict from the sequence
704 [47].

705 Furthermore, the study also identified more modest structural variations within
706 the MAX core for seven groups, labeled from I to O. These variations indicate that
707 even within the core structural framework of the MAX effectors, there are subtle
708 differences that distinguish these groups from each other. The classification given in
709 Table 1 can be improved by incorporating surface properties. Thus, the AVR-Pib and
710 O groups could be agglomerated, since they are structurally similar and members of
711 the O group have electrostatic surfaces related to AVR-Pib.

712 Notably, our study uncovered a domain duplication event within one of the
713 MAX effector clusters (group H). Other dual-domain effectors have been described in
714 recent studies, i.e. the Fol dual-domain effectors (FOLD) [48] and effectors predicted
715 from *Puccinia graminis* [4]. The discovery of dual-domain effectors, including the
716 domain duplication found in one of the MAX effector cluster, adds to our
717 understanding of the diversity and complexity of effector proteins. These dual-domain
718 effectors likely have evolved to possess multiple functional domains that contribute to
719 their virulence or interaction with host plants.

720 *ß1 and ß6 surface stickiness of AVR-Pik effectors*

721 The AVR-Pik AF models display high surface stickiness on ß1 and ß6 strands
722 that could serve to anchor residues from the N-terminal extensions of these effectors
723 to their MAX core domains. While there are multiple experimental structures of AVR-
724 Pik in complex with HMA domains, whether the structure of the free AVR-Pik effector
725 differs when not in complex with an HMA domain remains an open question. In the
726 AF models of MAX01, MAX02 and MAX03, the first 10-residues at the N-terminus
727 are unstructured. These are followed by a stretch of 20 residues, which are well
728 defined according to their pLDDT scores but without regular secondary structure, and
729 which adopt the conformation observed when the AVR-Pik effectors are in complex
730 with an HMA domain. Alternative AF models computed with the *Custom* MSA
731 implementation that does not use PDB templates showed different conformations for
732 this stretch of 20 residues, from totally unstructured (MAX01) to helical tendency
733 (MAX02). Based on these AF calculations we could expect that when free, AVR-Pik
734 effectors have a N-terminal extension alternating between different conformations but
735 we cannot exclude that this extension, while not having canonical secondary
736 structures, could participate to the overall stability of the AVR-Pik effectors.
737 Truncation mutants may serve to validate this hypothesis, however mutations within
738 or truncations of *a priori* unfolded protein regions may have deleterious effects not
739 only on the structure stability, as shown for N-terminal truncations of AVR-PikD
740 (pages 87-88 in ref. [49]) but also and more specifically on the overall folding process
741 of the protein. Recently, we used HP-NMR (high Hydrostatic Pressure NMR) [50,51]
742 to study the folding/unfolding of AVR-Pia, AVR-Pib [52] and MAX60 [53] . In the case
743 of MAX60, we observed an early folding intermediate involving ß1, ß6 and the C-
744 terminal helix that is a specific extension of this MAX effector, and mutants lacking

745 this helix were not sufficiently stable to be purified. This demonstrates that the
746 presence or absence of additional sequences outside the core can have profound
747 impacts on the folding of these effectors.
748
749 *Concluding Remarks*
750 Beyond the scope of information derived from the primary sequence, structural
751 information expands our understanding of biological processes at the atomic level.
752 Moreover, the three-dimensional structures of the different partners can be indicative
753 of the evolutionary pressure exerted on a biological system. This is far to be trivial
754 especially in the case of sequence divergent effectors that are also found to form a
755 significant structural family. The presented prediction pipeline, associated with
756 modeling using AlphaFold already showed a successful prediction rate of MAX
757 effectors of nearly 80%. The quality of these results has been proven by the
758 experimental resolution of four new MAX effectors MAX28, MAX47, MAX60 and
759 MAX67, which revealed the outstanding accuracy of the predictions when
760 superimposed with their respective model, especially in the position and orientation
761 of the side chains and secondary structures. A subset of 20% of the OG cluster
762 sequences resulted in ambiguous models (unreliable or with major structural
763 deviation of the expected MAX fold) pointing out the limits of *de novo* prediction on
764 such divergent sequences. This emphasizes the significance of the experimental
765 structural elucidation, which remains the key evidence even if known to be
766 challenging and resource consuming. Nonetheless, structure modeling will provide
767 solid foundations for the structure characterization or the study of interacting
768 biological entities, when limited or no information at all is available. The surface
769 stickiness mapping while not unambiguously informing of what could be the protein

interface on the MAX domain surface, which is in contact with the host target, may help to build hypothesis and to design plant biology experiments for validation. Linked to these predictive studies of interfaces between MAX effectors and their targets are new developments in Alphafold2, such as improved prediction of protein-protein interactions [54] or screening of interactions between proteins [55].

The MAX family effector proteins provide an example where the core structural platform has the plasticity to adapt extensions constituting new potential interacting modules with target proteins. This structural variability within effector families is important for their functional diversity and adaptability. By adding extensions or new modules, effectors can acquire novel functionalities, interact with different host targets, or modulate different host immune responses.

783 # Materials and Methods

784 ## *Experimental Structures*

785 *MAX28 Protein expression and purification.*

786 Protein expression and purification experimental details for MAX47, MAX60
787 and MAX67 are available in [56] Lahfa et al. 2022. For MAX28 we followed
788 essentially the same protocol than for the other MAX effectors for producing the ^{15}N-
789 labelled NMR sample but the His_6-tag was not cleaved to keep the protein soluble.
790 Uniformly labeled ^{15}N MAX28 was expressed in E. coli BL21 (DE3) cells (Invitrogen,
791 Thermo Fisher Scientific, Waltham, USA) from a homemade plasmid pDB-his-CCDB-
792 3C (courtesy of Frederic Allemand, CBS Montpellier, France). Protein expression
793 was carried out in $^{15}NH_4Cl$ (1 g/l) enriched M9 medium. Cells were grown at 37 °C
794 until reaching an OD600 = 0.8 and then, expression proceeded overnight at 30 °C
795 after induction by addition of 0.3 mM IPTG. Cells were harvested by centrifugation,
796 re-suspended in denaturing buffer (50 mM Tris, 300 mM NaCl, 1 mM DTT
797 (dithiothr.itol), 8 M urea, pH 8) and lysed by ultra-sonication. The supernatant
798 containing the unfolded protein was applied to a HisTrap HP 5 ml affinity column
799 (Cytiva, Freiburg im Breisgau, Germany). The His_6-tagged protein was eluted in 50
800 mM Tris, 300 mM NaCl, 1 mM DTT, 8 M urea, pH 8 with an imidazole gradient up to
801 500 mM. At this step, MAX28 was directly dialyzed against 10 mM Na Phosphate, 2
802 mM DTT, 150 mM NaCl, pH 6.8 buffer in order to remove imidazole and urea,
803 allowing the refolding of the protein. The MAX28 samples were then concentrated
804 using Amicon Ultra Centrifugal Filter Devices (MW cutoff 3000 Da), (Merck Millipore,
805 Burlington, USA) prior to size exclusion chromatography (SEC) using HiLoad 16/600
806 Superdex 75 pg column (Cytiva). Fractions containing protein were pooled,
807 concentrated to 0.4 mM and stored at −20°C. All NMR experiments were carried out

808 at 27°C on a Bruker Avance III 800 MHz or Bruker Avance III 700 MHz spectrometer,
809 both equipped with 5 mm z-gradient TCI cryoprobe. NMR samples consisted on
810 approximately 0.4 mM ^{15}N-labeled protein dissolved in 10 mM Na-Phosphate buffer
811 (pH 6.8) and 150 mM NaCl with 5% D_2O for the lock.
812
813 *NMR Structure determination of MAX28, MAX47, MAX60 and MAX67.*
814 ^{1}H chemical shifts were directly referenced to the methyl resonance of DSS, while
815 ^{15}N chemical shifts were referenced indirectly to the absolute ^{15}N/^{1}H frequency ratio.
816 All NMR spectra were processed with Topspin 3.6 (Bruker) and analyzed with Cindy
817 2.1 (Padilla, www.cbs.cnrs.fr). Assignments for MAX28, MAX47, MAX60 and MAX67
818 have been deposited to and are available from the BMRB data bank under the
819 accession entry 34782, 34731, 34730 and 34729, respectively.
820 The NMR structures were determined from the NMR constraints listed in S4 Table
821 that were obtained as follow. NOE cross-peaks identified on 3D [^{1}H, ^{15}N] NOESY-
822 HSQC (mixing time 150 ms) were assigned through automated NMR structure
823 calculations with CYANA*3* [57,58]. Hydrogen bond restraints were derived using
824 standard criteria on the basis of the amide ^{1}H / ^{2}H exchange experiments and NOE
825 data. When identified, the hydrogen bond was enforced using the following restraints:
826 ranges of 1.8–2.0 Å for d(N-H,O), and 2.7–3.0 Å for d(N,O). Dihedral restraints were
827 obtained from TALOS-N [59] analysis of backbone atom chemical shifts for MAX47,
828 MAX60 and MAX67. For the final list of restraints, distance values redundant with
829 covalent geometry were eliminated and disulfide bonds that were consistent with
830 short distances between cysteine residues were added.
831 A total of 200 three-dimensional structures were generated using the torsion angle
832 dynamics protocol of CYANA*3* from NOEs, hydrogen bonds and disulfide bond

833 restraints (S4 Table). The 20 best structures (based on the final target penalty
834 function values) were minimized with CNS 1.2 according to the RECOORD procedure
835 [60] and analyzed with PROCHECK [61]. The rmsds were calculated with MOLMOL
836 [62]. All statistics are given in S4 Table.
837 The structure coordinates have been deposited at the Protein Data Bank under the
838 following accession codes: MAX28 (PDB_8C8A), MAX47 (PDB_7ZKD), MAX60
839 (PDB_7ZK0), MAX67 (PDB_7ZJY),).
840

841 ***TM-pred MAX models Web Table***

842 Homology models of each OG representative sequence relative to each 3D
843 template were built using MODELLER v9.1 [63] with several alternative query-template
844 threading alignments as described in [7]. The top-5 models were selected according
845 to the TM-pred scoring function and are given in
846 https://pat.cbs.cnrs.fr/magmax/model/ Web page.
847 The Web table has the following columns: Group (OG cluster), Score (composite
848 evaluation score of the best model -best scores are in red, worst are in blue-), Dfire,
849 Goap, Qmean, E1D, E2D, E3D (individual evaluation scores of the best model),
850 Alignment (aligned identifier sequences), Identifier (protein identifier of the orthologs
851 prioritizing *Oryzae* infecting strains). Structural models and alignments are available
852 by clicking on each OG cluster in the first column. A multiple sequence alignment of
853 non-redundant orthologous proteins is displayed at the top of the page, starting with
854 the modeled representative sequence. For each OG cluster given in [7] we further
855 filtered out redundant sequences using CD-HIT [64] (S1 Table). The representative
856 sequence of each OG cluster was determined to be the sequence sharing the
857 highest sequence identity with a consensus sequence derived from the OG cluster

858 sequence alignment by MAFFT [65] (S2 Table). Signal peptide prediction before the
859 cleaved residue (cs.SIGNALP41), mean hydrophobic index (ih.MEAN), sequence
860 conservation (1D.CSRV), alternative alignments and model secondary structures
861 (2d.STRIDE) are displayed below the sequence alignment. Sequence conservation
862 scoring implemented in PAT [66] was calculated according to Sonnhammer and
863 Hollich, 2005 [67] with a Gonnet matrix [68]. At the bottom of the page, the 5 best
864 models can be displayed using different representations and color schemes. Models
865 can also be downloaded in PDB format.
866

867 ***Modeling by AlphaFold.***

868 For each OG representative sequence we computed three AF models differing
869 by the way of building multiple sequence alignment (MSA). The MMseqs2 MSA was
870 obtained from the MMseqs2 [69,70] web server as implemented in the ColabFold
871 version of AlphaFold 2.0 [71]. We also used the version of AlphaFold 2.2.0 that builds
872 MSAs by Jackhmmer on uniclust, mgnify and uniref90 databases. These two
873 implementations use PDB templates. Finally, a *Custom* MSA was build from
874 Muscle_v3.8.31 [72] by inserting (-profile option) the query sequence on top of a
875 previously computed MSA, termed ß1ß4_MSA. The ß1ß4_MSA was build from a
876 Muscle alignment of the OG sequences (S1 Table) by filtering out those having the
877 two flanking cysteine residues in the ß1 and in the loop between ß4 and ß5 strands
878 not correctly aligned to the 8 3D template sequences. The ß1ß4_MSA was further
879 processed by truncating the sequences by eliminating residues (-2 included) before
880 and (+2 included) after the first and last aligned cysteine residues, respectively, and
881 filtering out for redundant sequences by CD-HIT, giving a total size of 247 aligned
882 sequences (S6 Table). For each query, the consistency of the *Custom* MSA was

883 determined by checking the correct alignment of the cysteine residues in the query
884 and in the appended ß1ß4_MSA. When consistent the *Custom* MSA was converted
885 to a3m format by the reformat.pl script [73] and directly used as input in the
886 ColabFold implementation of AlphaFold 2.0 calculations that was setup without the
887 use of PDB templates. *Custom* MSAs could not be built for OG61 and OG62 from
888 absence of cysteine residues in their primary sequences and were not consistent for
889 OG15, OG27, OG71, OG81, OG85 and OG92.

890 The quality of each model was assessed by the pLDDT overall score [74]. The
891 correct MAX topology was verified by visual inspection (Pymol v.1.6 Delano 2002).
892 For models having the MAX topology, a MAX pLDDT score that was an average
893 score of residues in the MAX core domain (including residues from ß1 to ß6) was
894 calculated. For each query, the best AF MAX model was selected when the MAX
895 pLDDT score was above 60.

896

897 *Dali implementation and search*

898 A standalone implementation of DaliLite.v5 [75] was used for this work. For the
899 all-to-all clustering by Dali we first discarded *unstructured* stretches in each model.
900 For this, the *structured* domain of each model was defined by taking the STRIDE [76]
901 output, and filtering for the first residue in the first and last residue in the last
902 secondary structure (helix or strand), respectively (S7 Table). The model of MAX52
903 was split in two chains A and B each containing a MAX domain. All these *structured*
904 domains were used for clustering with Dali Z-scores excluding *de-facto* unresolved
905 protein regions without loosing important structural information.

906 *TM-align scoring and side-by-side plot with Dali Z-score tree.*

907 The distance between each pair of AF models that were used for Dali
908 clustering was estimated by the TM-score obtained from TM-align after pairwise
909 model superposition. A classification tree was then inferred from these pairwise
910 distances using FastME [77]. Finally, Fig 6 was obtained by joining identical models
911 in the FastME tree and in the Dali tree, respectively, by a line of the same color.
912

913 ***Surface properties of MAX core domains.***

914 A subset of 49 MAX effector AF models, each consisting of a MAX core
915 domain and optional N- and/or C- *unstructured* extensions (S7 Table) was defined by
916 discarding AF models having N- and/or C-terminal *structured* extensions (listed in the
917 groups A to H in Table 1 and in S4 Fig). All MAX core domains of the 49 AF models
918 were superimposed to the reference MoToxB structure with their ß1 strand vertically
919 aligned to the Z Cartesian axe giving a reference frame for the Sanson-Flamsteed 2D
920 projection computed using SURFMAP [31]. Their surface properties including
921 stickiness and electrostatics (APBS) [78] were computed by SURFMAP and are given
922 in S2 File. The temperature factor column of the PDB files was used to encode the
923 color of the exposed surface of the six ß-strands, from 1 to 6, respectively. The sum
924 of the surface stickiness positive values of each individual ß-strand was computed by
925 filtering SURFMAP surface stickiness output and are reported in Fig 7D. The amino-
926 acid conservation scores given for each OG cluster to which belongs each
927 representative MAX model were used to color encode the surface from white for high
928 conservation score of 9, light blue colors for intermediate conservation scores (from 8
929 to 6), sky-blue for low conservation score of 5 and darker blue colors indicating highly
930 polymorphic positions with conservation scores of 4 and below.

933

934 **Author Contributions:** M.L., and K.d.G. prepared the ^{15}N-^{13}C-labeled protein
935 samples (sub-cloning, protein expression and purification). P.B. made the NMR
936 resonance assignment of the four proteins in the study conditions. J.G. wrote scripts
937 for protein sequence analysis and homology modeling. C.R. and A.P. supervised the
938 project, conceived experiments and modeling, participated in the interpretation, and
939 M.L., A.P., N.D., C.R., S.C., P.G. wrote the article. A.P. also contributed to the
940 funding acquisition with P.G. and T.K. All authors have read and agreed to the
941 published version of the manuscript.

942

943 **Funding:** This research project was funded by the ANR project MagMAX (ANR-18-
944 CE20-0016-02), the European Research Council (ERC-2019-STG-852482-ii-MAX),
945 and supported by French Infrastructure for Integrated Structural Biology (FRISBI)
946 grant No. ANR-10-INSB-05.

947

948 # Acknowledgements

949 We are grateful to Liisa Holm for feedback in implementing Dali. The authors
950 are grateful to J. Maidment (CBS, Montpellier) for careful proof reading and language
951 correction of our manuscript. We are grateful to Matthew Bowler and Didier Nurizzo
952 at the European Synchrotron Radiation Facility (ESRF), Grenoble, France for
953 providing assistance in using beamline ID30A-1. The authors gratefully acknowledge
954 the ESRF for provision of synchrotron radiation facilities via Block Allocation Group
955 beamtime.

956

957

958 # Supporting information

959 S1 Fig. Template-based modeling of MAX effector sequences

960 S2 Fig. NMR structures and AF models

961 S3 Fig. Groups of AlphaFold models

962 S4 Fig. Singletons MAX effectors with structured extensions

963 S5 Fig. Hydrophobic residues on the surface of AVR-Pik and AVR-Pia

964 S6 Fig. Amino-acid polymorphism mapped on the surface of AF MAX models

965 S1 Table. List of Pyricularia (syn. Magnaporthe) oryzae entries in the 94 MAX
966 effector orthogroup (OG) clusters

967 S2 Table. Representative sequences of the 94 MAX effector orthogroup clusters

968 S3 Table. Report on structural studies of 10 putative MAX effectors (OG proteins)

969 S4 Table. Refinement statistics of MAX effector NMR structures and AF model
970 superimposition

971 S5 Table. Summary of AlphaFold2 modeling statistics and list of validated MAX
972 effectors

973 S6 Table. ß1ß4 MSA

974 S7 Table. Structural features of 77 AF-modeled structures of validated MAX effector,
975 statistics and conservation scores.

976 S8 Table. Comparizon of our selection of MAX effectors with previous modelisation
977 studies

978 Supplementary Materials_and_Methods. Crystallographic structure determination of
979 MoToxB

980 S1 File. Zipped folder of AF MAX models

981 S2 File. Zipped folder of AF MAX models surfaces

982

References

1. Lo Presti L, Lanver D, Schweizer G, Tanaka S, Liang L, Tollot M, et al. Fungal Effectors and Plant Susceptibility. Annu Rev Plant Biol. 2015;66: 513–545. doi:10.1146/annurev-arplant-043014-114623
2. Sperschneider J, Williams AH, Hane JK, Singh KB, Taylor JM. Evaluation of Secretion Prediction Highlights Differing Approaches Needed for Oomycete and Fungal Effectors. Front Plant Sci. 2015;6. doi:10.3389/fpls.2015.01168
3. Seong K, Krasileva KV. Computational Structural Genomics Unravels Common Folds and Novel Families in the Secretome of Fungal Phytopathogen *Magnaporthe oryzae*. Mol Plant-Microbe Interactions®. 2021;34: 1267–1280. doi:10.1094/MPMI-03-21-0071-R
4. Seong K, Krasileva KV. Prediction of effector protein structures from fungal phytopathogens enables evolutionary analyses. Nat Microbiol. 2023;8: 174–187. doi:10.1038/s41564-022-01287-6
5. Yan X, Tang B, Ryder LS, MacLean D, Were VM, Eseola Bisola A, et al. The transcriptional landscape of plant infection by the rice blast fungus Magnaporthe oryzae reveals distinct families of temporally co-regulated and structurally conserved effectors. Plant Cell. 2023;35: 1360–1385. doi:https://doi.org/10.1093/plcell/koad036
6. de Guillen K, Ortiz-Vallejo D, Gracy J, Fournier E, Kroj T, Padilla A. Structure Analysis Uncovers a Highly Diverse but Structurally Conserved Effector Family in Phytopathogenic Fungi. PLOS Pathog. 2015;11: e1005228.
7. Le Naour—Vernet M, Charriat F, Gracy J, Cros-Arteil S, Ravel S, Veillet F, et al. Adaptive evolution in virulence effectors of the rice blast fungus Pyricularia oryzae. Seidl MF, editor. PLOS Pathog. 2023;19: e1011294. doi:10.1371/journal.ppat.1011294
8. Maqbool A, Saitoh H, Franceschetti M, Stevenson CEM, Uemura A, Kanzaki H, et al. Structural basis of pathogen recognition by an integrated HMA domain in a plant NLR immune receptor. eLife. 2015;4: e08709.
9. De la Concepcion JC, Franceschetti M, Maqbool A, Saitoh H, Terauchi R, Kamoun S, et al. Polymorphic residues in rice NLRs expand binding and response to effectors of the blast pathogen. Nat Plants. 2018 [cited 10 Jul 2018]. doi:10.1038/s41477-018-0194-x
10. Maidment JHR, Franceschetti M, Maqbool A, Saitoh H, Jantasuriyarat C, Kamoun S, et al. Multiple variants of the fungal effector AVR-Pik bind the HMA domain of the rice protein OsHIPP19, providing a foundation to engineer plant defense. J Biol Chem. 2021;296: 100371. doi:10.1016/j.jbc.2021.100371
11. Zhang S, Wang L, Wu W, He L, Yang X, Pan Q. Function and evolution of Magnaporthe oryzae avirulence gene AvrPib responding to the rice blast resistance gene Pib. Sci Rep. 2015;5: 11642. doi:10.1038/srep11642
12. Oikawa K, Fujisaki K, Shimizu M, Takeda T, Saitoh H, Hirabuchi A, et al. The blast pathogen effector AVR-Pik binds and stabilizes rice heavy metal-associated (HMA) proteins to co-opt their function in immunity. Plant Biology; 2020 Dec. doi:10.1101/2020.12.01.406389
13. Park CH, Shirsekar G, Bellizzi M, Chen S, Songkumarn P, Xie X, et al. The E3 Ligase APIP10 Connects the Effector AvrPiz-t to the NLR Receptor Piz-t in Rice. Dinesh-Kumar SP, editor. PLOS Pathog. 2016;12: e1005529. doi:10.1371/journal.ppat.1005529
14. Bai P, Park C, Shirsekar G, Songkumarn P, Bellizzi M, Wang G. Role of lysine residues of the *Magnaporthe oryzae* effector AvrPiz-t in effector- and PAMP-triggered immunity. Mol Plant Pathol. 2019;20: 599–608. doi:10.1111/mpp.12779
15. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score =

1032 0.5? Bioinformatics. 2010;26: 889–895. doi:10.1093/bioinformatics/btq066
1033 16. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly
1034 accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589.
1035 doi:10.1038/s41586-021-03819-2
1036 17. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved
1037 protein structure prediction using potentials from deep learning. Nature. 2020;577:
1038 706–710. doi:10.1038/s41586-019-1923-7
1039 18. Taylor WR. Residual colours: a proposal for aminochromography. Protein Eng
1040 Des Sel. 1997;10: 743–746. doi:10.1093/protein/10.7.743
1041 19. Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, et al. Template-based protein
1042 structure modeling using the RaptorX web server. Nat Protoc. 2012;7: 1511–1522.
1043 doi:10.1038/nprot.2012.085
1044 20. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al.
1045 Accurate prediction of protein structures and interactions using a three-track neural
1046 network. Science. 2021;373: 871–876. doi:10.1126/science.abj8754
1047 21. Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al.
1048 Classification of Intrinsically Disordered Regions and Proteins. Chem Rev. 2014;114:
1049 6589–6631. doi:10.1021/cr400525m
1050 22. Liu Y, Chen S, Wang X, Liu B. Identification of Intrinsically Disordered Proteins
1051 and Regions by Length-Dependent Predictors Based on Conditional Random Fields. Mol
1052 Ther - Nucleic Acids. 2019;17: 396–404. doi:10.1016/j.omtn.2019.06.004
1053 23. Monzon AM, Necci M, Quaglia F, Walsh I, Zanotti G, Piovesan D, et al.
1054 Experimentally Determined Long Intrinsically Disordered Protein Regions Are Now
1055 Abundant in the Protein Data Bank. Int J Mol Sci. 2020;21: 4496.
1056 doi:10.3390/ijms21124496
1057 24. Marín M, Uversky VN, Ott T. Intrinsic Disorder in Pathogen Effectors: Protein
1058 Flexibility as an Evolutionary Hallmark in a Molecular Arms Race. Plant Cell. 2013;25:
1059 3153–3157. doi:10.1105/tpc.113.116319
1060 25. Sun H, Zhu X, Li C, Ma Z, Han X, Luo Y, et al. Xanthomonas effector XopR hijacks
1061 host actin cytoskeleton via complex coacervation. Nat Commun. 2021;12: 4064.
1062 doi:10.1038/s41467-021-24375-3
1063 26. Aparicio Chacón MV, Van Dingenen J, Goormachtig S. Characterization of
1064 Arbuscular Mycorrhizal Effector Proteins. Int J Mol Sci. 2023;24: 9125.
1065 doi:10.3390/ijms24119125
1066 27. Jaswal R, Rajarammohan S, Dubey H, Kiran K, Rawal H, Sonah H, et al. Intrinsically
1067 Disordered Kiwellin Protein-Like Effectors Target Plant Chloroplasts and are
1068 Extensively Present in Rust Fungi. Mol Biotechnol. 2023.
1069 doi:https://doi.org/10.1007/s12033-023-00717-y
1070 28. Walsh I, Martin AJM, Di Domenico T, Tosatto SCE. ESpritz: accurate and fast
1071 prediction of protein disorder. Bioinformatics. 2012;28: 503–509.
1072 doi:10.1093/bioinformatics/btr682
1073 29. Holm L. DALI and the persistence of protein shape. Protein Sci. 2020;29: 128–
1074 140. doi:10.1002/pro.3749
1075 30. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on
1076 the TM-score. Nucleic Acids Res. 2005;33: 2302–2309. doi:10.1093/nar/gki524
1077 31. Schweke H, Mucchielli M-H, Chevrollier N, Gosset S, Lopes A. SURFMAP: A
1078 Software for Mapping in Two Dimensions Protein Surface Features. J Chem Inf Model.
1079 2022;62: 1595–1601. doi:10.1021/acs.jcim.1c01269
1080 32. Giraldo MC, Valent B. Filamentous plant pathogen effectors in action. Nat Rev

1081 Microbiol. 2013;11: 800–814. doi:10.1038/nrmicro3119
1082 33. Wang H, Chou C, Hsu K, Lee C, Wang AH -J. New paradigm of functional regulation
1083 by DNA mimic proteins: Recent updates. IUBMB Life. 2019;71: 539–548.
1084 doi:10.1002/iub.1992
1085 34. Dennehy R, Duggan N, Dignam S, McCormack S, Dillon E, Molony J, et al. Protein
1086 with negative surface charge distribution, Bnr1, shows characteristics of a DNA-mimic
1087 protein and may be involved in the adaptation of *Burkholderia cenocepacia*.
1088 MicrobiologyOpen. 2022;11. doi:10.1002/mbo3.1264
1089 35. Goldenberg NM, Steinberg BE. Surface Charge: A Key Determinant of Protein
1090 Localization and Function. Cancer Res. 2010;70: 1277–1280. doi:10.1158/0008-
1091 5472.CAN-09-2905
1092 36. Guo L, Zhang Y, Ma M, Liu Q, Zhang Y, Peng Y, et al. Crystallization of the rice
1093 immune receptor RGA5A_S with the rice blast fungus effector AVR1-CO39 prepared *via*
1094 mixture and tandem strategies. Acta Crystallogr Sect F Struct Biol Commun. 2018;74:
1095 262–267. doi:10.1107/S2053230X18003618
1096 37. Varden FA, Saitoh H, Yoshino K, Franceschetti M, Kamoun S, Terauchi R, et al.
1097 Cross-reactivity of a rice NLR immune receptor to distinct effectors from the rice blast
1098 pathogen *Magnaporthe oryzae* provides partial disease resistance. J Biol Chem.
1099 2019;294: 13006–13016. doi:10.1074/jbc.RA119.007730
1100 38. Levy ED. A Simple Definition of Structural Regions in Proteins and Its Use in
1101 Analyzing Interface Evolution. J Mol Biol. 2010;403: 660–670.
1102 doi:10.1016/j.jmb.2010.09.028
1103 39. Fiser A. Template-Based Protein Structure Modeling. In: Fenyö D, editor.
1104 Computational Biology. Totowa, NJ: Humana Press; 2010. pp. 73–94. doi:10.1007/978-
1105 1-60761-842-3_6
1106 40. Rozano L, Mukuka YM, Hane JK, Mancera RL. Ab Initio Modelling of the Structure
1107 of ToxA-like and MAX Fungal Effector Proteins. Int J Mol Sci. 2023;24: 6262.
1108 doi:10.3390/ijms24076262
1109 41. Radom F, Plückthun A, Paci E. Assessment of ab initio models of protein
1110 complexes by molecular dynamics. MacKerell A, editor. PLOS Comput Biol. 2018;14:
1111 e1006182. doi:10.1371/journal.pcbi.1006182
1112 42. Mortuza SM, Zheng W, Zhang C, Li Y, Pearce R, Zhang Y. Improving fragment-
1113 based ab initio protein structure assembly using low-accuracy contact-map predictions.
1114 Nat Commun. 2021;12: 5011. doi:10.1038/s41467-021-25316-w
1115 43. Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P, et al. CATH: expanding
1116 the horizons of structure-based functional annotations for genome sequences. Nucleic
1117 Acids Res. 2019;47: D280–D284. doi:10.1093/nar/gky1097
1118 44. Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins—
1119 extended, integrating SCOP and ASTRAL data and classification of new structures.
1120 Nucleic Acids Res. 2014;42: D304–D309. doi:10.1093/nar/gkt1240
1121 45. Rocafort M, Bowen JK, Hassing B, Cox MP, McGreal B, de la Rosa S, et al. The
1122 Venturia inaequalis effector repertoire is dominated by expanded families with
1123 predicted structural similarity, but unrelated sequence, to avirulence proteins from
1124 other plant-pathogenic fungi. BMC Biol. 2022;20: 246. doi:10.1186/s12915-022-01442-
1125 9
1126 46. Bondos SE, Dunker AK, Uversky VN. On the roles of intrinsically disordered
1127 proteins and regions in cell communication and signaling. Cell Commun Signal. 2021;19:
1128 88, s12964-021-00774–3. doi:10.1186/s12964-021-00774-3
1129 47. Zhao B, Kurgan L. Deep learning in prediction of intrinsic disorder in proteins.

1130 Comput Struct Biotechnol J. 2022;20: 1286–1294. doi:10.1016/j.csbj.2022.03.003
1131 48. Yu DS, Outram MA, Smith A, McCombe CL, Khambalkar PB, Rima SA, et al. The
1132 structural repertoire of *Fusarium oxysporum* f. sp. *lycopersici* effectors revealed by
1133 experimental and computational studies. biorxiv; 2021 Dec.
1134 doi:10.1101/2021.12.14.472499
1135 49. Varden FA. Structure/function studies of effectors from the potato late blight and
1136 rice blast pathogens. University of East Anglia. 2019. Available:
1137 https://ueaeprints.uea.ac.uk/id/eprint/70234/1/FAVarden_100085474_PhD_Feb2019.
1138 pdf
1139 50. Dubois C, Herrada I, Barthe P, Roumestand C. Combining High-Pressure
1140 Perturbation with NMR Spectroscopy for a Structural and Dynamical Characterization of
1141 Protein Folding Pathways. Molecules. 2020;25: 5551. doi:10.3390/molecules25235551
1142 51. Roche J, Dellarole M, Royer CA, Roumestand C. Exploring the Protein Folding
1143 Pathway with High-Pressure NMR: Steady-State and Kinetics Studies. In: Akasaka K,
1144 Matsuki H, editors. High Pressure Bioscience. Dordrecht: Springer Netherlands; 2015.
1145 pp. 261–278. Available: http://link.springer.com/10.1007/978-94-017-9918-8_13
1146 52. Dubois C, Lahfa M, Pissarra J, de Guillen K, Barthe P, Kroj T, et al. Combining High-
1147 Pressure NMR and Geometrical Sampling to Obtain a Full Topological Description of
1148 Protein Folding Landscapes: Application to the Folding of Two MAX Effectors from
1149 Magnaporthe oryzae. Int J Mol Sci. 2022;23: 5461. doi:10.3390/ijms23105461
1150 53. Lahfa M, Mouhand A, de Guillen K, Barthe P, Kroj T, Roumestand C. Does a Similar
1151 3D Structure Mean a Similar Folding Pathway? The Presence of a C-Terminal α-Helical
1152 Extension in the 3D Structure of MAX60 Drastically Changes the Folding Pathway
1153 Described for Other MAX-Effectors from Magnaporthe oryzae. 2023.
1154 54. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein
1155 interactions using AlphaFold2. Nat Commun. 2022;13: 1265. doi:10.1038/s41467-022-
1156 28865-w
1157 55. Yu D, Chojnowski G, Rosenthal M, Kosinski J. AlphaPulldown—a python package
1158 for protein–protein interaction screens using AlphaFold-Multimer. Cowen L, editor.
1159 Bioinformatics. 2023;39: btac749. doi:10.1093/bioinformatics/btac749
1160 56. Lahfa M, Padilla A, De Guillen K, Pissarra J, Raji M, Cesari S, et al. 1H, 13C, 15 N
1161 backbone and side-chain NMR assignments for three MAX effectors from Magnaporthe
1162 oryzae. Biomol NMR Assign. 2022;16: 305–309. doi:10.1007/s12104-022-10095-2
1163 57. Güntert P. Automated NMR structure calculation with CYANA. Methods Mol Biol
1164 Clifton NJ. 2004;278: 353–378. doi:10.1385/1-59259-809-9:353
1165 58. Güntert P, Buchner L. Combined automated NOE assignment and structure
1166 calculation with CYANA. J Biomol NMR. 2015 [cited 18 Apr 2015]. doi:10.1007/s10858-
1167 015-9924-9
1168 59. Shen Y, Bax A. Protein Structural Information Derived from NMR Chemical Shift
1169 with the Neural Network Program TALOS-N. In: Cartwright H, editor. Artificial Neural
1170 Networks. New York, NY: Springer New York; 2015. pp. 17–32. doi:10.1007/978-1-
1171 4939-2239-0_2
1172 60. Nederveen AJ, Doreleijers JF, Vranken W, Miller Z, Spronk CAEM, Nabuurs SB, et
1173 al. RECOORD: A recalculated coordinate database of 500+ proteins from the PDB using
1174 restraints from the BioMagResBank. Proteins Struct Funct Bioinforma. 2005;59: 662–
1175 672. doi:10.1002/prot.20408
1176 61. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK—a program to
1177 check the stereochemical quality of protein structures. : 283–291.
1178 62. Koradi R, Billeter M, Wüthrich K. MOLMOL: A program for display and analysis of

macromolecular structures. J Mol Graph. 1996;14: 51–55.
63. Webb B, Sali A. Protein Structure Modeling with MODELLER. Methods Mol Biol. 2020;2199: 239–255. doi:10.1007/978-1-0716-0892-0_14
64. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152. doi:10.1093/bioinformatics/bts565
65. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010
66. Gracy J, Chiche L. PAT: a protein analysis toolkit for integrated biocomputing on the web. Nucleic Acids Res. 2005;33: W65–W71. doi:10.1093/nar/gki455
67. Sonnhammer EL, Hollich V. Scoredist: A simple and robust protein sequence distance estimator. BMC Bioinformatics. 2005;6: 108. doi:10.1186/1471-2105-6-108
68. Gonnet GH, Cohen MA, Brenner SA. Exhaustive Matching of the Entire Protein Sequence Database. Science. 1992;256: 1443–1445. doi:10.1126/science.1604319
69. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35: 1026–1028. doi:10.1038/nbt.3988
70. Mirdita M, Steinegger M, Söding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Hancock J, editor. Bioinformatics. 2019;35: 2856–2858. doi:10.1093/bioinformatics/bty1057
71. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19: 679–682. doi:10.1038/s41592-022-01488-1
72. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797. doi:10.1093/nar/gkh340
73. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011;9: 173–175. doi:10.1038/nmeth.1818
74. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29: 2722–2728. doi:10.1093/bioinformatics/btt473
75. Holm L. Using Dali for Protein Structure Comparison. In: Gáspári Z, editor. Structural Bioinformatics. New York, NY: Springer US; 2020. pp. 29–42. doi:10.1007/978-1-0716-0270-6_3
76. Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res. 2004;32: W500–W502. doi:10.1093/nar/gkh429
77. Lefort V, Desper R, Gascuel O. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program: Table 1. Mol Biol Evol. 2015;32: 2798–2800. doi:10.1093/molbev/msv150
78. Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE, et al. Improvements to the APBS biomolecular solvation software suite. Protein Sci. 2018;27: 112–128. doi:10.1002/pro.3280