



HAL
open science

Minimum information guidelines for experiments structurally characterizing intrinsically disordered protein regions

Bálint Mészáros, András Hatos, Nicolas Palopoli, Federica Quaglia, Edoardo Salladini, Kim van Roey, Haribabu Arthanari, Zsuzsanna Dosztányi, Isabella C Felli, Patrick D Fischer, et al.

► **To cite this version:**

Bálint Mészáros, András Hatos, Nicolas Palopoli, Federica Quaglia, Edoardo Salladini, et al.. Minimum information guidelines for experiments structurally characterizing intrinsically disordered protein regions. *Nature Methods*, In press, 10.1038/s41592-023-01915-x . hal-04154454

HAL Id: hal-04154454

<https://hal.science/hal-04154454v1>

Submitted on 7 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Minimum information guidelines for experiments structurally** 2 **characterising intrinsically disordered protein regions**

3
4 Bálint Mészáros^{1,S}
5 András Hatos^{2*#}
6 Nicolas Palopoli^{3*}
7 Federica Quaglia^{2,4*}
8 Edoardo Salladini^{2*}
9 Kim Van Roey^{5*}
10 Haribabu Arthanari^{6,7}
11 Zsuzsanna Dosztányi⁸
12 Isabella C. Felli⁹
13 Patrick D Fischer^{6,7}
14 Jeffrey C. Hoch¹⁰
15 Cy M. Jeffries¹¹
16 Sonia Longhi¹²
17 Emiliano Maiani^{13,14}
18 Sandra Orchard¹⁵
19 Rita Pancsa¹⁶
20 Elena Papaleo^{13,17}
21 Roberta Pierattelli⁹
22 Damiano Piovesan²
23 Iva Pritisana^{18,19}
24 Luigi Tenorio²
25 Thibault Viennet^{6,7}
26 Peter Tompa^{16,20,21}
27 Wim Vranken^{5,21}
28 Silvio CE Tosatto²
29 Norman E Davey^{22,†}

30
31

32 1. Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg
33 69117, Germany.

34 2. Department of Biomedical Sciences, University of Padova, Italy.

35 3. Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes - CONICET, Bernal,
36 Buenos Aires B1876BXD, Argentina

37 4. Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research
38 Council (CNR-IBIOM), Bari, Italy.

39 5. Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, 1050 Brussels, Belgium;
40 Structural Biology Brussels, Vrije Universiteit Brussel, 1050 Brussels, Belgium.

41 6. Harvard Medical School (HMS), Boston, MA, USA;

42 7. Department of Cancer Biology, Dana-Farber Cancer Institute (DFCI), Boston, MA, USA.

43 8. Department of Biochemistry, Eötvös Loránd University, Pázmány Péter stny 1/c, Budapest H-1117,
44 Hungary.

- 45 ⁹. Department of Chemistry "Ugo Schiff" and Magnetic Resonance Center, University of Florence,
46 Via Luigi Sacconi 6, 50019 Sesto Fiorentino (Florence), Italy.
- 47 ¹⁰. Department of Molecular Biology and Biophysics, UConn Health, Farmington, Connecticut.
- 48 ¹¹. European Molecular Biology Laboratory (EMBL), Hamburg Unit, c/o Deutsches Elektronen-
49 Synchrotron, Hamburg, Germany.
- 50 ¹². Laboratory Architecture et Fonction des Macromolécules Biologiques (AFMB), UMR 7257, Aix
51 Marseille University and Centre National de la Recherche Scientifique (CNRS), 163 Avenue de
52 Luminy, Case 932, 13288, Marseille, France.
- 53 ¹³. Cancer Structural Biology, Danish Cancer Society Research Center, 2100, Copenhagen, Denmark
- 54 ¹⁴. UniCamillus - Saint Camillus International University of Health and Medical Sciences, Rome, Italy
- 55 ¹⁵. European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Hinxton,
56 Cambridgeshire CB10 1SD, UK.
- 57 ¹⁶. Institute of Enzymology, Research Centre for Natural Sciences, 1117, Budapest, Hungary.
- 58 ¹⁷. Cancer Systems Biology, Section for Bioinformatics, Department of Health and Technology,
59 Technical University of Denmark, 2800, Lyngby, Denmark
- 60 ¹⁸. Hospital for Sick Children, Toronto, ON, Canada.
- 61 ¹⁹. Medical University of Graz, Auenbruggerplatz 2, 8036, Graz, Austria.
- 62 ²⁰. VIB-VUB Center for Structural Biology, Pleinlaan 2, 1050, Brussels, Belgium.
- 63 ²¹. Structural Biology Brussels, Department of Bioengineering Sciences, Vrije Universiteit Brussel,
64 Pleinlaan 2, 1050, Brussels, Belgium.
- 65 ²². Division Of Cancer Biology, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham
66 Rd, Chelsea, London, UK

67

68 * These authors contributed equally to the manuscript.

69 [§] Current address: *Department of Structural Biology and Center for Data Driven Discovery, St Jude*
70 *Children's Research Hospital, Memphis, TN, 38105, USA.*

71 [#] Current address: *Department of Oncology, Lausanne University Hospital, Lausanne 1011,*
72 *Switzerland; Department of Computational Biology, University of Lausanne, Lausanne 1015,*
73 *Switzerland; Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland; Swiss Cancer Center*
74 *Leman, Lausanne 1011, Switzerland.*

75 [†] Corresponding author: *norman.davey@icr.ac.uk*

76

77 **Abstract**

78 An unambiguous description of an experiment, and the subsequent biological observation, is vital for
79 accurate data interpretation. Minimum information guidelines define the fundamental complement of
80 data that can support an unambiguous conclusion based on experimental observations. We present the
81 Minimum Information About Disorder Experiments (MIADE) guidelines to define the parameters
82 required for the wider scientific community to understand the findings of an experiment studying the
83 structural properties of intrinsically disordered regions (IDRs). MIADE guidelines provide
84 recommendations for data producers to describe the results of their experiments at source, for curators
85 to annotate experimental data to community resources and for database developers maintaining
86 community resources to disseminate the data. The MIADE guidelines will improve the interpretability
87 of experimental results for data consumers, facilitate direct data submission, simplify data curation,
88 improve data exchange among repositories and standardise the dissemination of the key metadata on an
89 IDR experiment by IDR data sources.

90 **Introduction**

91 The intrinsically disordered protein (IDP) field is generating increasingly large amounts of biophysical
92 data on the structural properties of intrinsically disordered regions (IDRs)¹⁻³. The complexity of the
93 produced IDP-related data continues to increase, and in recent years there has been a noticeable growth
94 in the number of analyses describing complex structural properties, conditional disorder and disorder-
95 function relationships⁴⁻⁸. Whereas a decade ago most IDP papers characterised disorder as a binary
96 state, now many papers contain comprehensive analyses describing multiple conditional states using
97 several complementary experimental methods^{9,10}. Moreover, the improved experimental tools now
98 enable the investigation of increasingly complex IDRs, IDPs, and multi-domain proteins. A key
99 responsibility of the IDP community is the development of minimum information guidelines to improve
100 the description, interpretation, storage and dissemination of data generated in the rapidly evolving IDP
101 field¹¹. In this document, we introduce the Minimum Information About Disorder Experiments
102 (MIADE) guidelines for the definition and interpretation of experimental results from IDP experiments.

103
104 Minimum information guidelines define the fundamental unit of information for the unambiguous
105 definition of experimental metadata to the level required for the key results of an experiment to be
106 comprehended by the wider scientific community¹². The role of minimum information guidelines is to
107 minimise data loss by preserving essential data and removing ambiguity while avoiding redundancy.
108 There are several requirements for a functional minimum information guideline. Firstly, the core
109 information conveyed by the experiment should be unequivocally defined. This should include the
110 observation itself but also any information that would change our understanding or confidence in the
111 biological or physical relevance of the observation. Second, adhering to the guidelines should be as
112 effortless as possible to enable its widespread adoption, i.e. the guidelines should avoid any excessive
113 burden in the description of an experiment while capturing the most important information to fulfil the
114 first requirement. Thirdly, it should be equally applicable to all IDR analysis methods so that the
115 experimental metadata is comparable across all sources of primary data, regardless of the experimental
116 approach. To fulfil these criteria, the MIADE guidelines recommend an unambiguous description of
117 the protein and the construct of the region(s) being studied at amino acid resolution, other components
118 of the sample, the experimental approach applied and the interpretation of the results of the experiment.
119 Importantly, any information about the experimental protocols, sample components or sequence
120 properties that might affect the interpretation of the results are an essential part of the unambiguous
121 description of the experimental results.

122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168

Minimum information guidelines are a compromise between the necessary depth of information to unambiguously describe an IDP experiment, and the reporting burden placed on researchers producing the metadata. MIADE-compliant data records should allow users to quickly assess an IDP experiment and the associated data, and point to the source data for the complete experimental context, but do not require annotation to a level of detail that allows the experiment to be reproduced. Therefore, unless their definition is essential to unambiguously interpret the results of the experiment, several aspects of the experimental setup are not required by the MIADE guidelines; this includes a complete description of the experimental constructs, a complete description of the sample and a complete description of the experimental protocol. In addition, minimum guidelines focus on the description of a single experiment and therefore cannot define how multiple experiments should be integrated to describe more complex features of proteins, such as conformational ensembles. Furthermore, minimum information guidelines are abstract recommendations that do not specify the technical details of the structured data types that are guideline compliant. In this document, we provide examples of data adhering to MIADE recommendations for multiple use cases including providing details of the updates to allow the storage of MIADE-compliant data in the DisProt IDP database¹. However, the technical specification of data storage is defined by exchange formats used to standardise and store compliant data and therefore it is outside the scope of this document.

The MIADE guidelines provide a community consensus created by experimentalists, curators and data scientists on the minimum information required to appropriately describe metadata on experimentally and computationally derived structural state(s) of IDPs or IDRs. The aim is to increase the accuracy, accessibility and usability of published IDP data, to comply with FAIR (findability, accessibility, interoperability, and reusability) data principles¹³, to support rapid and systematic curation of such IDP data in public databases and to improve interchange of IDP data between these IDP resources. We believe that these guidelines will provide an important roadmap to the thousands of data producers, curators and database developers in the IDP field and increase the utility of published IDP data for the larger biological community.

Where should MIADE be applied?

The vast majority of IDP experiments yield information about the structure or the function of IDPs. Functional IDP studies most commonly analyse their interactions with other molecules. Since the Minimum Information about a Molecular Interaction experiment (MIMIX) guidelines¹⁴, on which the MIADE guidelines have been modelled, already cover the molecular interaction aspects of these experiments, MIADE only focuses on the description of the structural aspects of the studied IDPs.

Experimental data can follow many paths to the final data consumer (Figure 1A). At each point in the flow of data valuable information can be lost, misinterpreted or misrepresented. After data production, the primary data are analysed by field-specific specialists (typically the research group that conducted the experiment) who interpret these complex experimental results to provide a biological observation. These specialists will author a publication that describes the novel observations and, ideally, they will directly submit the findings to a core IDP data resource. Currently, much of the data in the IDP field passes into a branch where biocurators interpret the description of the experiments and observations in the publication and then annotate the information into manually curated resources. The role of MIADE is to provide general recommendations that can be applied at each potential point of data loss to maximise the precision with which information is transferred.

169 The MIADE guidelines should be applied to free text descriptions when reporting on the experiment,
170 to data extraction from the primary literature and to structured metadata for dissemination. Therefore,
171 the MIADE guidelines provide a recommendation to unambiguously describe structural information on
172 IDRs inferred from experimental or computational analysis, intended for: (i) researchers authoring an
173 article on the structural state(s) of an IDR; (ii) researchers who want to directly submit such data to an
174 IDP resource, e.g. prior to peer reviewed publication of the data; (iii) biocurators who want to
175 define/curate data on structural state(s) of an IDR within an IDP resource; (iv) database developers who
176 want to disseminate IDR structural state data; and (v) data users who need to achieve full comprehension
177 by clarity of the meaning and origin of each piece of data (Table 1).

178 **What information is required by MIADE guidelines**

179 Both the biological and the methodological contexts are required to understand and compare
180 experimental data. Consequently, MIADE guidelines recommend the clear definition of four
181 components for IDP structural experiment reporting: the protein region that was studied, the structural
182 state of that region as inferred from the experiment, the experimental or computational approach applied
183 and the data source. Each region of a protein for which a structural state was inferred from an experiment
184 should be described separately. The exact application of the guidelines is use case specific, however,
185 when possible stable identifiers of external resources should be referenced, for example, UniProt for
186 protein definitions¹⁵, ECO (Evidence and Conclusion Ontology) for experimental definitions¹⁶ and
187 IDPO (Intrinsically Disordered Proteins Ontology) for structural state definitions
188 (<https://disprot.org/ontology>).

189 MIADE Checklist - minimising ambiguity in the definition of an experiment

190 The following information is required to create MIADE compliant description of an experiment
191 characterising the structural properties of an IDR:
192

193
194 **Protein Region:** definition of the region for which a structural state was experimentally determined or
195 computationally predicted. If several regions of a protein were inferred to be disordered, each region
196 should be defined separately. The definition should be unambiguous and concise, and should leave no
197 doubt about the identity of the protein that contains the region. The source organism and isoform should
198 always be specified. If the sequence is synthetic and not mappable to an existing protein this should be
199 stated explicitly. The experimental sequence of the protein region being studied should always be
200 defined. Similarly, any tags, labels, post-translational modifications or mutations present in the sample
201 under study should be described. Each region should be characterised by:
202

- 203 ● Definition of the source protein from which the region was derived:
 - 204 ● The common name for the source molecule. Both the protein name and gene name
205 should be added whenever possible. Ideally, this should be the official name provided
206 by a nomenclature committee such as the HGNC symbol from the HUGO Gene
207 Nomenclature Committee for human genes¹⁷. In cases where the field-specific name is
208 used, and it differs from the official name, the official name should be mentioned in
209 the first definition of the molecule.
210 *Example: Mitotic checkpoint serine/threonine-protein kinase BUB1 beta (BUBR1, also*
211 *known as BUB1B)*
 - 212 ● Scientific name, common name or NCBI taxonomy ID of the species of origin for the
213 source protein (or free text for chemical synthesis, unknown, and *in silico* origins).

214 *Example: Budding Yeast (Saccharomyces cerevisiae strain ATCC 204508 / S288c,*
215 *NCBI Taxon ID: 559292)*

216 • Accession or identifier for the source protein in a reference database. If an isoform of
217 a protein was used in the experiment, the accession or identifier specifically identifying
218 that isoform should be used whenever possible. The version number of the protein
219 sequence in the database can be added to further reduce ambiguity.

220 *Example: UniProt:P13569 (P13569-2 in case isoform 2 was used)*

221

222 • Definition of the protein region(s) for which a structural state was determined:

223 • Start and stop positions of the region: the position of the first and last residue of the
224 region, based on i) the sequence as described in the database annotating the source
225 protein from which the region was derived (i.e. positions should refer to the natural
226 sequence and should not consider added purification and solubility tags), or ii) in the
227 case of a sequence that is not mappable to a natural sequence, the sequence provided
228 by the data producer.

229 *Example: residues 708-831 of BUBR1*

230 • The amino acid sequence of the experimental construct encoding the region(s) in
231 IUPAC one-letter codes¹⁸.

232

233 • Definition of the experimental molecule (i.e., any tags in the construct that have been removed
234 before the sample has been studied can be ignored) including any alterations and additions to
235 the defined protein region:

236 • Tags and labels that are present in the experimental construct.

237 *Example: C-terminal 6xHis tag*

238 • Experimental proteoform including mutations, insertions, deletions or post-
239 translational modifications.

240 *Example: phosphorylation of BUBR1 on serine 21*

241

242 Structural state: structural state of the construct or a region(s) within the construct, as defined by the
243 experimental data or as inferred by the experimentalist.

244 • Classically, structural states in IDP experiments would be defined based on a binary “order”
245 and “disorder” description, however, as more complex structural properties are now being
246 experimentally defined, the structural properties of the region and subregions should be defined
247 to the highest resolution possible. The position of a structurally distinct subregion of a construct,
248 such as the observation of partially populated secondary structural elements, should be defined
249 explicitly as described for the protein region definition. If the boundaries of the structure state
250 elements within a construct are not clear this should be stated. When possible the corresponding
251 term and term ID for that structural state in the IDPO controlled vocabulary should be given. If
252 the observed structural property is not widely known by a general readership, for example,
253 describing more complex attributes than a binary order/disorder definition such as dynamics,
254 secondary structure propensity or compaction, the property should be clearly defined.

255 *Example: disorder (IDPO:00076)*

256

257 Experimental and computational approaches: definition of the experiment or computational approach
258 used to determine the structural state of the region. Each experimental setup should be described
259 separately. For studies that derive structural information from the integration of data from several
260 experiments, each individual experimental observation should be expressed in a MIAD-compliant
261 manner. The following parameters should be included in the experiment description:

262 ● The experimental or computational methods used to determine the structural state of the region.
263 If possible, this should be annotated with the corresponding term and term ID for that
264 experimental method in the ECO controlled vocabulary. The name of the computational or
265 experimental method(s) used to define the structural state of the protein region(s) should be
266 defined to the most detailed level possible. If relevant, any software used in the post-processing
267 of experimental data, or to define the structural state directly, should be defined including the
268 software version.

269 *Example: far-UV circular dichroism (ECO:0006179)*

270 ● The scientific name, common name or NCBI taxonomy ID of the host organism in which the
271 experiment was performed (or free text for *in vitro*, unknown, *in vivo*, and *in silico* experimental
272 environments); further specification of cell line or tissue is recommended. Special care should
273 be taken in defining experimental details for in-cell or cell extract studies.

274 *Example: in vitro*

275 ● Any experimental deviation that could alter the interpretation of the results and any condition
276 that could impact on the results should be clearly described. These deviations are generally
277 method specific, for example, *in vitro* experimental parameters (e.g. pH, pressure, protein
278 concentrations, temperature, buffer, salt, additional components including other proteins),
279 computational parameters (e.g. non-default options), Molecular Dynamics (MD) simulation
280 parameters (e.g. force field used) and integrative structural study parameters (e.g. experimental
281 sources, integration approach). See next section and Table 2 for details.

282 *Example: experiment was performed at 4°C.*

283 ● Any additional components in the sample that could alter the interpretation of the results. This
284 attribute is important to clearly capture structural changes induced by binding partners.
285 However, it also includes other components such as reducing agents, cofactors and crowding
286 agents which may trigger a structural change on the protein of interest. Each component should
287 be defined unambiguously, and if possible, include the concentration of the sample components
288 and refer to external databases including a definition of the molecule (e.g. Uniprot, ChEMBL).
289 Additional protein components should be defined to the same level of detail as the experimental
290 region being studied. See next section and Table 2 for details.

291 *Example: experiment was performed in the presence of 10 g/L polyethylene glycol 400*
292 *(PEG400) (ChEMBL:1201478).*

293

294 In the case of data being stored in a database, transferred between resources or defined in the absence
295 of a paper, it is important to also include the source of the data.

296

297 Data source: a reference to where the data were originally described.

298 ● In cases where data were published in a paper, the following information should be provided:
299 ○ publication database and identifier

300 *Example: PMID: 35055108*

301

302 ● In cases where data were directly submitted to a data resource, the following information should
303 be provided:

- 304 ○ the name of the data resource
305 ○ the accession number of the record holding the data in that resource
306 ○ the data creator who submitted the data
307 ○ contact details of the data creator

308 **Key factors that can influence the interpretation of structural IDR data**

309 Numerous factors connected to the protein region, protein construct or the experimental setup can
310 influence the structural state of the protein region being studied and, consequently, our confidence in
311 the biological relevance of the observed structure (see Table 2)^{19,20}. These factors can be technical
312 perturbations, to allow experimental measurements to be collected (e.g., changes in temperature or pH),
313 or perturbations related to the biological question under investigation (e.g. proteoforms with a PTM or
314 disease relevant mutation, or the presence of an interacting partner). In these cases, any description of
315 the structural state is only meaningful when the relevant factors that influence the observed state are
316 specified. While the minimum information requires the protein region and the experimental method to
317 be defined, it is up to the discretion of the authors to report deviations from the established protocol,
318 sample or sequence that could alter the interpretation of the results. Consequently, an explicit statement
319 by an author will simplify the task of the curator or reader to make a judgement of the importance of a
320 given deviation. In complex cases the meaningful description of the inferred structural states can include
321 several pieces of information that go beyond the specification of the protein region and the experimental
322 method applied. In Table 2, we provide pointers on which factors might be considered important
323 deviations based on known biological cases of conditional protein disorder and common experimental
324 perturbations.

326 **Example use cases**

327 There are several use cases for MIADE (Table 1), however, in practice there are two major distinct
328 applications: (i) creating an unambiguous description of an experiment in free text and (ii) encoding the
329 fundamental unit of metadata for an experiment in a standardised format. In this section, we will give
330 examples of how MIADE can be applied in each of these cases.

332 MIADE for authors

333 A key step in data capture is the unambiguous description of the specialist interpretation of the primary
334 data. Consequently, an accurate and unequivocal definition of the experimental observation in the text
335 of an article that adheres to the MIADE guidelines will simplify all downstream data interpretation.
336 Defining an experiment in free text requires detail that allows the experiment to be fully reproduced.
337 Consequently, most articles describe the experimental detail at a level of granularity that far exceeds
338 the requirements of a MIADE compliant entry. However, a comprehensive description of an
339 experiment's design and results does not mean that the data is accessible to the wider biological
340 community. A common issue amongst non-specialist readers and curators is that the data is described
341 in a manner that is highly technical, requires extensive knowledge of the experimental method or uses
342 field-specific jargon. Furthermore, important details are often not apparent as they are in materials and
343 methods sections, supplementary materials or even a previously published paper. Consequently,
344 MIADE guidelines recommend an explicit and unambiguous description of the experimental design,
345 the proteins under analysis and the interpretation of the results.

346
347 Consideration should be given to the fact that the description should be understandable to the wider
348 biological community and the key data should be explicitly stated. This will improve the clarity of the
349 document and allow rapid annotation by curators for community resources. In many cases, writing
350 engaging and readable scientific prose, and writing unequivocal descriptions of complex experiments
351 are conflicting goals. However, in any case where such conflicts occur, substance should take
352 precedence over style. For example, the definition of a protein as "*Budding Yeast (Saccharomyces*
353 *cerevisiae strain ATCC 204508 / S288c (TaxID:559292)) Spindle assembly checkpoint component*
354 *MAD3 (MAD3) (UniProt:P47074)*" may be rather awkward when compared to "yeast MAD3".

355 However, it removes ambiguity from the protein definition. By following the examples in the checklist
356 and understanding that a reader may not be familiar with terminology on IDRs and IDR experiments,
357 data can be presented in a manner that is both accurate and globally accessible.

358

359 MIADE implementation in DisProt

360 An important aspect to represent experimentally determined structural states of IDPs and IDRs in a
361 standard format is the use of stable external identifiers and controlled vocabularies (CV) to
362 unambiguously describe the captured data. In the future, IDP-specific exchange formats should be
363 developed to define these attributes for experimental metadata, however, for the moment it is useful to
364 consider how DisProt stores MIADE compliant data.

365

366 DisProt is a manually curated resource of intrinsically disordered regions (IDRs) and proteins (IDPs)
367 from literature, that relies on both professional and community curation. All DisProt entries correspond
368 to a specific UniProt entry (or one of its isoforms) and describe the structural state(s) of the region(s)
369 of the protein. When available, information on the presence of transitions between states, interactions
370 and functions, is also curated. The annotation of structural states and transitions makes use of specific
371 IDPO terms (<https://disprot.org/ontology>). As part of the development of the MIADE guidelines we
372 have updated the DisProt database and curation framework to allow the annotation of MIADE-
373 compliant entries¹. An improved construct definition was required to encode tags, labels, mutations or
374 modifications and the experimental setup definition was updated to allow complex experimental
375 samples to be described. Importantly, these additions will allow DisProt to annotate the observations of
376 complex experiments defining conditional multistate IDRs that are becoming increasingly common in
377 the literature.

378

379 *Proteoform definition:* The DisProt resource already included an unambiguous definition of the protein
380 or protein isoforms (using UniProt accession numbers) and its regions by mapping to the UniProt
381 sequence. The updated implementation can now define non-canonical and modified proteoforms. The
382 MIADE integration allows the possibility to encode deviations from the wildtype UniProt defined
383 protein sequence. Furthermore, the complete sequence of the experimental construct can now be
384 annotated if available. Annotatable construct alterations include *tags* and *labels* (using the PSI-MI
385 ontology (<https://www.ebi.ac.uk/ols/ontologies/mod>)²¹), mutations (using the HGVS nomenclature
386 (<https://varnomen.hgvs.org/>)) and *PTMs* and non-standard amino acids (using the PSI-MOD ontology
387 (<https://www.ebi.ac.uk/ols/ontologies/mod>)²²).

388

389 *Experimental conditions definition:* DisProt uses the Evidence and Conclusion Ontology (ECO,
390 <https://www.evidenceontology.org/>)²³ to annotate experimental methods. In addition, the DisProt
391 database can now store a range of experimental parameters that can influence our understanding of the
392 biological relevance of an experimental observation, i.e. *pH*, *temperature*, *pressure*, *ionic strength*, and
393 *oxidation-reduction potential*. The parameter can be quantified in cases where this information is
394 available. All parameters are defined in the NCI Thesaurus OBO Edition controlled vocabulary
395 (<https://ncit.nci.nih.gov/ncitbrowser/>) and their units in the Units of Measurement Ontology
396 (<https://biportal.bioontology.org/ontologies/UO>). Deviations from the expected value in the
397 experiment parameter (e.g. within normal range, increased, decreased, not specified or not relevant) can
398 also be added. All information is annotated with the text description taken directly from the scientific
399 article and curators' statements can be added to further clarify annotation.

400

401 *Experimental components definition:* The DisProt database can now describe experimental sample
402 components such as lipids, nucleic acids, small molecules, metal ions or proteins present during the

403 characterisation of the structural state of an IDR. The concentration of the components and cross
404 reference to the specific database, i.e. CheBI²⁴, ENA²⁵, RNACentral²⁶ and UniProt¹⁵ can also be added.
405 Similar to the other MIADE fields, a text description can be added into the corresponding *Statement*
406 field.

407

408 A representative list of DisProt use cases highlighting novel information covered by the addition of
409 fields from the MIADE update is described in Table 3.

410 **Case studies**

411 While MIADE only captures the core structural inferences derived from structural experiments on
412 IDRs, it can be applied to the description of experimental data with a very wide range of complexity in
413 terms of experimental design and studied system. In the following section we demonstrate how MIADE-
414 compliant information can be created using extracts from three papers that serve as examples of good
415 practice. These experiments are accompanied by a MIADE compliant entry in the DisProt resource
416 (Table 3). We chose these papers to provide a set of examples of increasing complexity that represented
417 several of the key issues tackled by the MIADE guidelines. A wide range of techniques are used to
418 characterise the structural properties of IDRs, however, for simplicity, both due to the available
419 literature and the wider understanding of the experimental approach, all examples describe NMR
420 experiments. We highlight the three key areas covered by MIADE from each paper: the definition of
421 the protein construct used; the deviation from the wildtype proteoform (including mutations, post-
422 translational modifications, tags, labels, and dyes); and the definition of the experimental setup,
423 including the environmental conditions and sample compositions that might have relevance for the
424 structural state.

425 The first paper describes the disordered structural state of human calpastatin (CAST), an inhibitor of
426 calpain, the Ca²⁺ activated cysteine protease²⁷. The authors unambiguously define two protein
427 constructs they used by referencing the common name of the protein and source organism, together
428 with a UniProt accession (*'¹⁵N-labeled and ¹³C-labeled full-length hCSD1 [corresponding to A137-
429 K277 of human calpastatin, SwissProt entry P20810]'* and *'C-terminal half of calpastatin (position in
430 whole calpastatin P204-K277)'*). The constructs are defined by providing residue numbers in
431 reference to the UniProt entry. However, the wording 'C-terminal half of calpastatin' could be
432 misleading as the construct under investigation is the C-terminal region of the first domain of
433 calpastatin. In addition, when providing UniProt residue start and stop numbering, the authors
434 erroneously state that the construct is P204-K277 rather than P203-K277. This example highlights
435 a common problem that stems from the custom of providing relative residue position within a
436 region of interest or domain when defining constructs, instead of absolute residue position in
437 reference to the full sequence. The authors clearly define the experimental method with different types
438 of NMR experiments, including heteronuclear single quantum coherence (HSQC), calculation of the
439 secondary chemical shift and ³J_{HNH α} scalar coupling constants determined with 3D HNCA-E.COSY.
440 For these experiments, the relevant environmental conditions are temperature and pH, which the authors
441 define in the materials and methods sections (*'HSQC spectra collected at 298 K and at pH 4.3, 5.23,
442 and 6.17 for hCSD1(67-141) as well as pH 3.85, 5.53, 6.07, and 7.25 for hCSD1. The temperature
443 dependence of the same type of resonances was measured at 280, 300, and 320 K in aqueous solution
444 for hCSD1(67-141)'*) - note that the authors use the relative numbering inside the domain being studied
445 as opposed to the absolute numbering in the full-length UniProt sequence, which would be 203-277).
446 Using these setups, the authors then determine that both constructs are essentially disordered and that

447 this observation is largely independent of temperature and pH in the ranges explored. The manuscript
448 also includes more refined observations about the structural properties of the protein, such as:
449 ‘subdomains A and B, two characteristic binding and functional sites of the inhibitor, have some helical
450 character’ or ‘restricted motions on a subnanosecond time scale indicated by larger than average $J(0)$
451 values are observed for G13-M17, K68-L72, S101-C105, and S128-V132. These residues of restricted
452 mobility also present some residual local structural features highlighted both by secondary chemical
453 shifts, SCS, and by their hydrophobicity pattern’.

454 The second paper details experiments performed on Eukaryotic translation initiation factor 4E-binding
455 protein 2 (EIF4EBP2), an interacting partner of Eukaryotic translation initiation factor 4E (eIF4E)¹⁰.
456 The authors define the protein construct as the full-length human protein by referencing its common
457 name (4E-BP2). The HUGO Gene Nomenclature Committee (HGNC) gene name is EIF4EBP2, and no
458 unambiguous identifier is provided, however, the naming is specific enough to unambiguously identify
459 the protein being studied, given that the protein has no known alternative isoforms. In addition,
460 throughout the paper the authors reference several key residues in the protein (such as T37, T46, S65,
461 T70 and S83) based on which readers and curators can confirm whether they map to the correct UniProt
462 sequence. As opposed to the previous example where conditions were changed, in this case,
463 measurements were performed on distinct proteoforms of the protein. The main structural conclusion
464 of the paper is that the structural state of EIF4EBP2 is dependent on its phosphorylation state. HSQC
465 NMR spectrum shows that ‘non-phosphorylated 4E-BP2 has intense peaks with narrow ¹HN chemical
466 shift dispersion characteristic of IDPs [...] However, wild-type 4E-BP2 uniformly phosphorylated at
467 T37, T46, S65, T70 and S83 shows widespread downfield and upfield chemical shifts for residues
468 spanning T19–R62, suggesting folding upon phosphorylation’. Using partial phosphorylation, the
469 authors then disentangle the individual contribution of each phosphorylation to the induced folding,
470 stating: ‘No significant change in global dispersion was observed for 4E-BP2 phosphorylated only at
471 S65/T70/S83, demonstrating that it remains disordered, while phosphorylating T37 and T46
472 (pT37pT46) induces a 4E-BP2 fold identical to phosphorylated wild type. Interestingly, when
473 phosphorylated individually, pT37 or pT46 result in a partly folded state, with some chemical shift
474 changes indicative of ordered structure (pT37). [...] Thus, phosphorylation of both T37 and T46 is
475 necessary and sufficient for phosphorylation-induced folding of 4E-BP2’. The authors also measure the
476 structural effect of binding to eIF4E and find that the interaction induces partial folding of the
477 phosphorylated 4E-BP2: ‘The spectrum of pT37pT46 in isolated and eIF4E-bound states demonstrate
478 an order-to-disorder transition upon eIF4E binding. [...] pT37pT46 undergoes an order-to-disorder
479 transition upon binding to eIF4E’. Therefore, both phosphorylation and the presence of a binding
480 partner can induce a structural transition of EIF4EBP2 through different mechanisms, and therefore the
481 inference that EIF4EBP2 is disordered is dependent on the exact proteoform as well as the presence of
482 other proteins. In addition to the structural state, the authors also directly address the connection
483 between phosphorylation and the interaction capacity: ‘non-phosphorylated or minimally
484 phosphorylated 4E-BPs interact tightly with eIF4E, while the binding of highly phosphorylated 4E-BPs
485 is much weaker and can be outcompeted by eIF4G’. While this piece of information is key to
486 understanding the biological regulatory role of EIF4EBP2, it cannot be captured in the structural state-
487 focused framework of MIADE and should be encoded as additional information in interactomics
488 databases.

489 In the third example, the authors study the human Cellular tumour antigen p53 (TP53) focusing on the
490 structural features of the disordered N-terminal region²⁸. The authors clearly define the protein being
491 studied by stating it is human TP53. In addition, they also provide an overview figure that contains the
492 UniProt region boundaries of various p53 regions and domains that are used in the constructs. In contrast

493 to the previous examples, the main construct used in this study is not a full-length protein or an isolated
494 protein region, but a chimeric protein consisting of an isotopically labelled N-terminal and a non-
495 labelled C-terminal region. The authors use a split intein splicing to produce the isotopically labelled
496 disordered N-terminal region and fused to the unlabelled central C-terminal regions (*'we utilized intein
497 splicing to segmentally label the NTAD within tetrameric p53 [...] NTAD (residues 1–61) labeled with
498 an NMR-active isotope (¹⁵N), while residues 62–393 remained unlabeled and NMR invisible'*). As a
499 result of this technique, the final construct has a short insertion where the intein was located, the position
500 of which was carefully chosen: *'The intein splice site was selected as D61/E62, a site that is distant in
501 the amino acid sequence from interaction sites or well-folded domains. Careful selection of the splice
502 site is important, since the Npu DnaE intein system inserts nonnative residues (GSCFNGT in the p53
503 constructs used here) at the splice site.'* This construct enables the assessment of the structural state of
504 the disordered NTAD in the context of the full length tetrameric TP53 by NMR HSQC spectra. For
505 technical reasons, the authors further introduced mutations to the sequence outside the disordered
506 regions being studied: *'To improve expression levels, stabilizing mutations
507 (M133L/V203A/N239Y/N268D) were introduced into the DNA-binding domain'*. The definition of the
508 environmental conditions covers the temperature and salt concentrations, with all other parameters
509 supposedly being in the normal range of similar NMR measurements: *'unless otherwise stated, all
510 spectra were recorded at 25 °C for samples in NMR buffer'* and *'salt titrations for p53(1–312) and
511 p53(1–61) were carried out with protein concentrations of 150 μM. The initial titration point had a
512 NaCl concentration of 150 mM, and NaCl from a 5-M concentrated stock was added to this sample at
513 50-mM increments up to 500 mM NaCl'*. Apart from unambiguously defining the protein construct, the
514 proteoform, the techniques and the environmental conditions, the main conclusion about the structural
515 state is also clearly stated as: *'the HSQC spectrum of the NTAD-p53 tetramer shows that the NTAD
516 remains dynamically disordered in the full-length protein'*.

517

518 **MIADE-compliant metadata capture at source**

519

520 To date, direct submission of data to community resources is underutilised by the IDP community. IDP
521 resources should improve their capacity to receive data pre-publication including the possibility to
522 embargo data until the time of final publication (similar to the PDB model) and develop tools and
523 resources that simplify MIADE-compliant reporting. Furthermore, the IDP community should enforce
524 the deposition of experimental data and metadata as a required component of the publication process.
525 The ideal situation would include the pre-publication submission of primary source data directly to the
526 corresponding field specific resource (Table 4). Subsequently, a reference to primary source data and
527 MIADE-compliant experimental metadata should then be submitted to a community resource such as
528 DisProt or IDEAL^{1,2}. This benefits the databases, as the efficiency of data collection and verification is
529 increased. This in turn benefits the IDP community and wider biological community, as more and more
530 precise data, linked to related primary data in field-specific databases, are readily available. Currently,
531 several databases allow pre- or post-publication submission of data related to IDR experiments, each
532 with their own submission process and data formats (Table 4). However, the proportion of data created
533 that is captured by these resources varies widely and no resource is successful in capturing all data
534 produced that fall within their scope. To facilitate data capture, as part of this work, the DisProt resource
535 has added a MIADE-compliant form for the submission of metadata from experiments structurally
536 characterising IDRs (<https://disprot.org/biocuration>).

537

538 **Discussion**

539

540 Over the past 10 years, the development of new and improved methods and technologies to study IDPs
541 has increased the complexity of the experiments characterising the structural properties of IDRs (Figure
542 1C). However, this revolution has not been reflected by advances in the data standardisation of the field.
543 Consequently, at all levels there is a requirement to improve the description, curation, storage and
544 dissemination of the fundamental data from these analyses. Guidelines to unambiguously define the key
545 information from an experiment simplifies data capture, minimise key data loss, standardise data
546 transfer and maximise data utilisation. The argument against standardised reporting guidelines has
547 always been the unbalanced burden placed on the reporter. However, the advantages far outweigh the
548 effort, allowing relevant data to be easily identified, recovered and reused, leading to improved data
549 management, minimised data loss and simplification of data sharing within and between groups.
550 Method independent metadata also allows data to be aggregated and to be analysed in subsets based on
551 data quality (Figure 1D). Furthermore, data aggregation across complementary methods simplifies
552 cross-validation of data permitting quality to be defined by consensus. Finally, improved data
553 management and upgrades to data deposition processes, will improve data transfer to community
554 resources accelerating the open science efforts of the IDP field.

555
556 Data capture should have the flexibility to cover old, new and future experimental approaches. The
557 MIADE guidelines store observations together with details on the experiment to allow data to be
558 reinterpreted in the future. While adding experimental parameters and sample components can add
559 considerably to the curation burden, they also allow for more nuanced observations to be captured. As
560 IDP experiments become increasingly complex by studying the modulatory effects of proteoforms,
561 concentrations, conditions and binding partners, it is imperative that these rich data on the context of
562 the studied protein region are captured wherever they are needed to faithfully interpret the reported
563 observations. These details can describe observations beyond binary order/disorder structure definition,
564 to quantitative measures that include dynamics, secondary structure propensity and compaction. To
565 capture every relevant detail, MIADE guidelines will need to evolve over time based on community
566 requirements. Controlled vocabularies and ontologies are a key component of this evolution. These
567 definitions standardise the meaning of the terms used to describe IDP data allowing the complete
568 unambiguous annotation of an IDP experiment and results. Ontologies such as the Intrinsically
569 Disordered Proteins Ontology (IDPO) and the Evidence and Conclusion Ontology (ECO) will need to
570 continually add terms as required to include novel experimental approaches, computational methods,
571 non-binary structural classifications (i.e. more detailed than order/disorder including dynamics,
572 secondary structure propensity and compaction), structural transition definitions and conditionality.

573
574 The MIADE guidelines are only an initial step towards standardised and lossless IDP data
575 representation within the biological community. Three key developments are still required: standardised
576 exchange formats for reporting IDP metadata and raw data, simplified pre- and post-publication data
577 deposition mechanisms for the IDP data repositories, and a community wide agreement to deposit data.
578 The diversity of the methodologies and data in the IDP community has proved to be a barrier to data
579 collection, and MIADE will allow the key data to be collected and aggregated across the field. In
580 parallel, each experimental approach in the field can develop method-specific storage and exchange
581 formats and standards for raw data. However, given the parallel requirements across many of these
582 approaches, efforts should be made to collaborate and reuse structured data formats when possible.
583 These exchange formats should hold experimental data at a range of detail from a MIADE-compliant
584 definition to a description of the experiment and results that would allow the experiment to be
585 reproduced (Figure 1B). Ultimately, the interpretation of raw experimental data will evolve as analysis
586 methods improve. Consequently, the best long-term strategy to safeguard the knowledge accumulated
587 by the IDP community is the standardised deposition of raw and processed experimental measurements

588 in addition to interpreted structural observations derived from the data. Enforcing data deposition is a
589 complex process, however, pressure at the point of publication by journals and reviewers can drive
590 compliance.

591

592 We see this document as the initial steps to open the discussion to standardise the controls, experimental
593 parameters and vocabulary used for each method applied by the IDP community. We advocate for the
594 importance of a clear and unambiguous description of an IDR experiment, and we hope this document
595 will encourage each experimental community to extend the guidelines to specify and enforce the
596 reporting of the important information for their experimental methods. It is important that data
597 producers, curators and database developers in the IDP field are conscious of the expanding interest in
598 IDRs by the wider biological community. The growing understanding of the functional significance of
599 IDPs by researchers outside the IDP field has increased the importance of making high quality and
600 understandable IDP data accessible to the wider community such as cell biologists studying the function
601 of IDRs, computational biologists developing tools to analyse IDRs and curators transferring IDR data
602 into community resources.

603

604 **Acknowledgements**

605 This work was funded by: ELIXIR, the research infrastructure for life-science data; a Cancer Research
606 UK Senior Cancer Research Fellowship (C68484/A28159 to N.E.D); Carlsberg Foundation
607 Distinguished Fellowship (CF18-0314); Danmarks Grundforskningsfond (DNRF125); National
608 Research, Development and Innovation (NRDI) Fund Young researchers' excellence programme
609 research grant (project FK128133 and FK142285 to R.P.); European Union's H-2020 MSCA-RISE
610 programme (grant agreement No. 778247 'IDPfun'); Fondazione CR Firenze; Fondazione Umberto
611 Veronesi; the Italian Ministry of University and Research (to EM and RP), PRIN 2017 under grant
612 agreement No 2017483NH8; EC H2020-WIDESPREAD-2020-5 Twinning grant (no. 952334
613 PhasAGE, to PT and SCET), VUB Strategic Research Program on Microfluidics (SRP51) at Vrije
614 Universiteit Brussel (VUB, Brussels, Belgium, to PT). The authors would like to thank Julie Forman-
615 Kay for her feedback on the MIADE guidelines and the manuscript.

616

617 **Author Contributions**

618 All authors contributed to the development of the MIADE guidelines. N.E.D, B.M, A.H, N.P, N.P, F.Q,
619 E.S and K.VR were the MIADE workgroup and defined the details of the MIADE guidelines. N.E.D
620 and B.M wrote the document with the help of the MIADE workgroup. All authors edited and refined
621 the final manuscript. N.E.D coordinated the project.

622

623 **Ethics Declaration**

624 **Competing interests**

625 The authors declare no competing interests.

626

627

628 **Figures**

629 **Figure 1.** (A) Scheme of data flow from primary data capture by the experimentalist to data dissemination to the
630 end consumer. (B) Definition of the scope of the MIADE guidelines and the requirements of a comprehensive
631 standard for IDP data. (C) Representation of the evolution of complexity of cutting edge experimental IDP papers.
632 (D) Representation of the requirement for data aggregation across analyses to build high confidence consensus
633 data on a region.

634

635

636 **Tables**

637

638 **Table 1. Cases where the MIADE guidelines should be applied to improve data interpretability**

639 **and minimise the loss of key data.**

640

Storing experimental metadata	<ul style="list-style-type: none">- Allows storage of high-level metadata.- Allows the integration and comparison of data from distinct experiments and experimental approaches.
Direct submission of IDR data pre-publication	<ul style="list-style-type: none">- Promotes early data capture by providing a standard with a low barrier for data entry to directly submit experimental results prior to publication to an IDP database.- Facilitates collection of IDR data in light of increasing data management and open science efforts.- Increases data available for community blind testing of computational IDP tools (CAID).
Defining key findings about IDRs in a publication	<ul style="list-style-type: none">- Defines the requirements for unambiguous description of the specialist interpretation of primary source data.- Increases the clarity of the paper and simplifies all downstream data interpretation.- Allows the reported structural IDR data to be rapidly captured in an IDP resource, where it is readily available to the community.
Curation of IDRs from a publication	<ul style="list-style-type: none">- Provides clear guidelines to unambiguously and efficiently define the fundamental details and results of the experiment(s) to facilitate data curation by non-specialist.
Transfer of IDR data	<ul style="list-style-type: none">- Standardised transfer of the key metadata on an IDR experiment to, between, and from IDR data repositories.- Promote/facilitate the implementation of FAIR principles with the IDR community.

641

642
643

Table 2. Key factors that can influence the interpretation of structural IDR data.

Deviation Type	Deviation Description	Example	
<u>Deviations from the canonical protein sequence</u>	<i>Definition of the construct used in the experiment</i>		
Post-translational modifications <i>Covalent modification of a residue side chain</i>	PTMs can change the physicochemical properties of a sequence and thereby alter the structural state, compaction or dynamics of an IDR. The structures of several IDRs have been shown to be modulated by the addition or removal of a PTM. Studies aimed at investigating these mechanisms will characterise modified proteoforms to understand the structural changes.	4E-BP2 folds into a 4 stranded beta structure upon phosphorylation of residues T37 and T46.	¹⁰
Substitutions, insertions and deletions <i>Replacement, addition or removal of residues of the canonical protein sequence</i>	Substitutions, insertions and deletions can affect local and global physicochemical properties of a region (for example, the charge, hydrophobicity, interaction capacity and size) potentially affecting the structural properties of a protein. Studies altering the protein sequence can enable the testing of the effect of indels, polymorphisms or disease variants, certain PTMs (such as phosphomimetics) or isoforms (by addition/removal of an exon).	A F82K mutation in Ferricytochrome c induces localised unfolding of a distal site in the ferric state.	²⁹
Tags and labels <i>Covalent attachment of an entity that enables analysis, identification, purification or solubility of the protein</i>	Tags and labels can have a measurable influence on the dynamics and stability of the protein they are attached to. The addition of tags is almost always a technical necessity, and its aim is not to measure a biological phenomenon. Tags play three major roles in IDR experiments: (i) for purification (e.g. FLAG tag), (ii) for solubility (e.g. maltose-binding protein (MBP)), (iii) for experimental readout (e.g. fluorescent tags for fluorescent microscopy or paramagnetic tags for NMR).	The addition of a His tag influences myoglobin short time scale (picoseconds) dynamics.	³⁰

<p>Proteolytic cleavage <i>Cleavage of the protein chain induced by a protease</i></p>	<p>Cleavage can disrupt both local structural elements and long range contacts by increasing the distance between residue pairs. Cleavage also introduces new N- and C-termini in the protein chain changing the polarity, solubility and interaction capacity of regions. Many proteins, especially extracellular proteins, are known to undergo cleavage, often in many subsequent steps. Cleavage products can be created in response to signalling events and often have very different biological activity, interaction capacity and structural states.</p>	<p>Cleavage of the disordered osteopontin removes long range intramolecular interactions, changing the structural state and the accessibility of the integrin binding site.</p>	<p>31</p>
<p><u>Experimental parameters</u></p>	<p><i>Parameters of the experimental setup for a sample</i></p>		
<p>pH <i>pH of the sample</i></p>	<p>The pH can affect the strength of ionic and hydrogen bonds and can so modulate the structural state of a protein³². Experimental parameters are often tweaked to find the optimal experimental parameters for the study of a specific protein, sometimes resulting in the use of non-physiological pH. Furthermore, comparison of a physiological state with a non-physiological pH state can be used to probe the structural properties of the region of interest. For example, forcing the complete unfolding of a construct with harsh experimental conditions to allow comparison to a "ground state"</p>	<p>NhaA, a sodium proton antiporter of the inner membrane of <i>Escherichia coli</i>, is activated at pH values between 6-7, with a maximal activity at pH 8.5, and inactivated by acidic pH.</p>	<p>33,34</p>
<p>Temperature <i>Temperature of the sample</i></p>	<p>The temperature has an explicit role in determining the strength of entropic terms in the Gibbs free energy that controls the stability of protein structures and complexes. Thus, changing the temperature can drastically change the stability of folded proteins and dynamics of IDRs. Changing the temperature of a protein sample in an experiment can serve to explore its folding/unfolding kinetics, its stability and oligomerization. For calorimetric techniques, such as Differential Scanning Calorimetry (DSC), regulating temperature is what provides the measurable signal. For certain experiments, such as NMR, changing the temperature is performed for technical reasons to improve the signal to</p>	<p>Hp26 becomes active with increased temperature in a two-step mechanism that firstly activates the protein and then unfolds it.</p>	<p>35,36</p>

	noise ratio.		
Pressure <i>Hydrostatic pressure of the protein sample</i>	High hydrostatic pressure (HHP) can induce unfolding by breakage of intramolecular interactions and exposure of cavities allowing binding of water. High hydrostatic pressure (HHP) is used to study the structure of partially structured intermediate transition states and the monomeric forms of oligomeric and aggregated proteins.	The 1D ¹ H NMR spectra support the proposed molten-globule state of Arc repressor under high pressure, moreover the ¹ H NMR spectra at a pressure range of 3.5-5 kbar are substantially different from those of the native state (1 bar, 20°C) and the fully denatured state (1 bar, 70°C).	^{37,38}
Force <i>Mechanical force applied to the protein</i>	Opposing forces applied to different parts of the protein can mechanically unfold the structure (either partially or completely), converting mechanical signals into biochemical ones. The most typical information provided are the number of steps in which a protein unfolds (reflecting the number of domains or intermediate structural states) and the force required for unfolding. For proteins undergoing force-induced unfolding in biological settings, these measurements explore their biological function. Atomic force microscopy and high-speed force spectroscopy are used to assess the stability and the folding/unfolding kinetics of proteins.	Mechanical unfolding of TTN-1 and twitchin of <i>Caenorhabditis elegans</i> affects the auto-inhibitory region and the catalytic core of the protein.	³⁹

<p>Redox potential <i>Redox potential of the sample</i></p>	<p>The redox potential affects the behaviour of residues, especially that of cysteine. Under oxidising conditions, cysteines can form disulphide bridges, while under reducing conditions, they can coordinate cations. Redox potential parameters are often tweaked to find the optimal experimental parameters for the study of a specific protein. Various cellular compartments have drastically different redox potentials (e.g. the extracellular space is oxidising while the cytoplasm is reducing), thus changing the redox potential in a sample can model various compartments or the transport between them.</p>	<p>The nuclear export signal (NES) of Yap1 is masked by a structured domain held together by disulphide bridges in the oxidised state. In reducing conditions, the domain unfolds, the NES becomes exposed and functional.</p>	<p>40</p>
<p>Light <i>Irradiating the protein with visible, UV or infrared light</i></p>	<p>Many light-sensitive proteins contain additional chromophores that can undergo structural changes (most often trans/cis isomerization) that consequently alter the structure and/or dynamics of the protein they are embedded in. Light-induced folding/unfolding of photosensitive proteins as a response to light is studied by altering these conditions.</p>	<p>Light-induced unfolding of the water-soluble photoactive yellow protein (PYP) allows it to become functionally active and bind partners.</p>	<p>41</p>
<p>Protein concentration <i>Concentration of the protein being tested in the sample</i></p>	<p>Increased protein concentration can promote aggregation, liquid-to-liquid phase separation and liquid-to-solid phase transition. Consequently, the structural state of an IDR can be concentration dependent. The solubility limit defines the concentration in which molecules are miscible in solution. If the protein concentration increases trespassing that limit, the macromolecule:macromolecule interactions are energetically more favourable than the macromolecule/solute interactions.</p>	<p>Several phase separation drivers (i.e. FUS and hnRNPA1) can undergo percolation or liquid-liquid phase separation in a concentration dependent manner.</p>	<p>42</p>
<p>Protein source <i>Details of the protein purification</i></p>	<p>An important element of the experimental setup is how the protein was generated, as prior history may have a significant effect on its structural state by determining the exact proteoform, including post-translational modifications, partial proteolysis, etc. Best practice is to check the final proteoform used in the structural studies, either by mass spectrometry or, if possible, by the structural experimental method itself (such as NMR structure determination).</p>	<p>Important information includes the cell type in which the protein was expressed (e.g., E. coli, yeast, insect cells (e.g., SF9) or human cells (e.g. HEK-293) not the source genome where the protein is encoded), the method of extraction (e.g. by sonication) and subsequent purification, especially if it included an intermittent heat-treatment and the application of agents</p>	

		for solubilization and/or denaturation (e.g. tween-20, urea), protease inhibitors and/or reducing agents.	
Computational parameters <i>Details of the parameters used in computational processing of the</i>	Complex processing of experimental data is commonly required for data interpretation in the IDP field. Any software used and computational parameters that can influence the results should be described.	The interpretation of the results of raw data post-processing, residue-specific intrinsic disorder prediction, molecular dynamics and integrative structural studies all rely heavily on software and parameterisation used.	
<u>Experimental sample components</u>	<i>Components added to the sample that are required for technical aspects of the experiment</i>		
Crowding agents <i>Addition of crowding agents to a sample to mimic the molecular concentrations found in cells</i>	Quinary interactions can have a strong effect on both the structural properties and interactions of a protein. Consequently, proteins behave differently in different contexts: for example, in the cell, in high concentrations of crowding agents and in a buffer. Few experiments have been performed to probe the effect of crowding on structure and interactions, however, the limited data available have suggested that the contribution can be significant and that it is largely protein specific. Biophysical measurements taken <i>in vitro</i> may not reflect the actual dynamics in the cellular milieu; consequently, the crowding agents are added to partially mitigate biases introduced by the non-physiological conditions.	Experiments studying the effects of a range of crowding agents at different concentrations on IDRs from PUMA, Ash1, E1A and p53 reveal that the induced structural changes depend on both protein sequence and crowding agent used.	¹⁹
Solubility agents <i>High ionic strength, amino acids, organic solvent</i>	Solubility agents (or hydrotropic agents) are typically small molecules that have both a hydrophobic and a hydrophilic region, and can increase the solubility of proteins by shielding their local hydrophobic regions from the solvent. Molecules added to a sample in a structural analysis to improve the solubility of the protein to be studied may alter its structural state.	Ionic strength and glycerol are used to mirror protein charges or increased repulsions, respectively. These two experimental components were both used to keep proteins stable in solution.	⁴³

<p>Folding/unfolding agents <i>Small molecules, organic solvents, high salt or non-ionic detergents</i></p>	<p>Folding and unfolding agents constitute a diverse set of molecules used in the structural characterisation of an IDR. They are used to modulate the structural state of a protein by shifting it towards either a folded or unfolded state. This is then used as a reference state with known properties that can be compared to other states, helping understand a structural property of the region under investigation.</p>	<p>Several cosolvents were used to perturb protein's stability: guanidine hydrochloride (GdnHCl) and urea are used to denature/partially unfold proteins, whereas hexafluoroisopropanol (HFIP) and trifluoroethanol (TFE) induce secondary structure formation.</p>	<p>44,45</p>
<p>Preservatives <i>Protease inhibitors, chelating agents and sodium azide</i></p>	<p>Protease inhibitors, chelating agents and sodium azide are often used to improve the overall stability of samples (e.g. against proteolysis) and might have an impact on protein's behaviour</p>		
<p>Biological background <i>Cell lysate, cell extract or in-cell sample</i></p>	<p>IDPs are increasingly investigated in biological backgrounds rather than <i>in vitro</i>. For example, isotopically labelled samples can be specifically studied by NMR in cell lysates, cell extract (nuclear/cytoplasmic extract) or even in cells or organelles. Fluorescently labelled proteins can also be studied in cells.</p>	<p>A range of cell lines, cell extracts and organelles have been used to characterise IDPs in their microenvironments. However, specific information on the amount of sample inside cells, the potential manipulation of cells with genetic engineering or drugs, etc. should be defined. Proper controls for intracellular pH and crowding should be provided for these data to be comparable.</p>	<p>46,47</p>
<p><u>Biological sample components</u></p>	<p><i>Components added to the sample that are directly related to the biological hypothesis being tested</i></p>		
<p>Binding partners <i>Known or predicted binding partners or ligands</i></p>	<p>Binding an interaction partner including ions, small molecules, proteins, nucleic acids or lipids/membranes can modulate the dynamics, compaction or secondary and tertiary structure of an IDR. Many disordered regions will form distinct conformations in the presence of a specific binding partner. These conformational changes can be drastic, shifting the protein from disordered to highly ordered, or to partially ordered with residual large amounts of disorder. In all cases they result in a shift in the sampled conformations.</p>		

- proteins		In isolation, p27 is disordered with nascent secondary structure. Upon binding to Cdk2-cyclin A complex p27 becomes ordered.	48
- nucleic acids		In isolation, HMG-1 is intrinsically disordered, however, upon binding to DNA the protein becomes ordered and adopts a well defined conformation in the minor groove.	49
- lipids/membranes		The intrinsically disordered N-terminal region of Hsp12 adopts a folded conformation comprising four α -helices upon micelle binding.	50
- small molecules		The dynamic KIX domain of the coactivator CBP/p300 can be stabilised by the addition of a small molecule.	51
Co-factors <i>Metal ions, iron-sulphur (Fe-S) clusters or organic cofactors (vitamins and their derivatives or fatty acids)</i>	Co-factors acting as cell state signals can heavily modulate the behaviour of an IDP. These observations can include folding and unfolding in the presence or absence of specific metal ions, protein aggregation by negatively charged cofactors compensating positively charged repeat regions or induction of liquid-liquid phase separation.	The calcium-binding Repeat-in-ToXin (RTX) of <i>Bordetella pertussis</i> adenylate cyclase toxin (CyaA) is disordered in the absence of calcium but folds upon calcium binding. This region acts as a switch integrating the differing calcium concentrations between the extracellular and intracellular environment.	52

644

645

646 **Table 3. Extra data curated by DisProt to allow a MIADE compliant annotation for the case study examples.**

Protein definition	Observed region (for which structural observations are made)	Construct region (present in the experiment)	DisProt Identifier	MIADE field	Relevant parameters controlled in the experiment	Experimental Method	Details	Structural state	
--------------------	--	--	--------------------	-------------	--	---------------------	---------	------------------	--

Calpastatin, UniProt: P20810	137-277	Same as observed region	DP00196	r011 - Experimental condition	pH, temperature	NMR ¹	pH = 3.85 - 7.25 T = 298K	Disordered	²⁷
	204-277	Same as observed region	DP00196	r011 - Experimental condition	pH, temperature	NMR ¹	pH = 4.3 - 6.17 T = 280 - 320K	Disordered	²⁷
eIF4E-binding protein 2, UniProt: Q13542	1-120	Same as observed region	DP01293	-	-	-	-	Disordered	¹⁰
	1-120	Same as observed region	DP01293	r007 - Construct alteration	Protein modification	NMR ¹	phosphoSer 65; phosphoThr 70; phosphoSer 83	Disordered	
	19-62	1-120	DP01293	r007 - Construct alteration	Protein modification	NMR ¹	phosphoThr 37	Molten globule	
	19-62	1-120	DP01293	r007 - Construct alteration	Protein modification	NMR ¹	phosphoThr 46	Molten globule	
	19-62	1-120	DP01293	r007 - Construct alteration	Protein modification	NMR ¹	phosphoThr 37; phosphoThr 46	Ordered	
	19-62	1-120	DP01293	r007 - Construct alteration	Protein modification	NMR ¹	phosphoThr 37; phosphoThr 46; phosphoSer 65; phosphoThr 70; phosphoSer 83	Ordered	
Cellular tumor antigen p53, UniProt: P04637	1-61	1-393	DP00086	r077; r078 - Construct alteration	Labels and dyes	NMR ¹	¹⁵ N label position: 1-61	Disordered	²⁸

				r077 - Experimental components	Salt concentration		[NaCl] = 150- 500mM		
				r081; r082 - Construct alteration	Protein mutation	NMR ¹	insertion p.Asp61_Glu62ins GlySerCysPheAsn GlyThr; substitutions p.Met133Leu; p.Val203Ala; p.Asn239Tyr; p.Asn268Asp		

647 ¹Nuclear magnetic resonance spectroscopy

648

649

650

Table 4. Representative set of databases for the submission of IDR experimental metadata and data

Data type	Database	Website	Submission Process	
IDR experiment metadata	DisProt <i>database of disordered regions manually curated from literature</i>	https://www.disprot.org/	Deposition, Curation Post Publication	¹
Nuclear magnetic resonance spectroscopy (NMR)	BMRB (Biological Magnetic Resonance Bank) <i>database of data on NMR spectroscopic investigations of biological macromolecules and metabolites</i>	https://bmr.io/	Deposition	⁵³
Circular dichroism	PCDDDB (Protein Circular Dichroism Data Bank) <i>database of circular dichroism (CD) and synchrotron radiation CD (SRCD) spectral data and their associated experimental metadata</i>	https://pcddb.cryst.bbk.ac.uk/	Deposition	⁵⁴

Small-angle X-ray and neutron scattering (SAXS and SANS)	SASBDB (Small-Angle Scattering Biological Data Bank) <i>database of small-angle scattering (SAS) experimental data and models</i>	https://www.sasbdb.org/	Deposition	⁵⁵
Protein Ensembles	PED (The Protein Ensemble Database) <i>database for the deposition of structural ensembles</i>	https://proteinensemble.org/	Curation Post Publication	⁵⁶

652

653

654 **Bibliography**

655 1. Quaglia, F. *et al.* DisProt in 2022: improved quality and accessibility of protein intrinsic disorder
656 annotation. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab1082.

657 2. Fukuchi, S. *et al.* IDEAL in 2014 illustrates interaction networks composed of intrinsically
658 disordered proteins and their binding partners. *Nucleic Acids Res.* **42**, D320-5 (2014).

659 3. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev.*
660 *Mol. Cell Biol.* **6**, 197–208 (2005).

661 **Highly influential review describing the structure-function relationship in the context of IDRs.**

662 4. Davey, N. E. The functional importance of structure in unstructured protein regions. *Curr. Opin.*
663 *Struct. Biol.* **56**, 155–163 (2019).

664 5. Schad, E. *et al.* DIBS: a repository of disordered binding sites mediating interactions with
665 ordered proteins. *Bioinformatics* **34**, 535–537 (2018).

666 6. Tompa, P. The interplay between structure and function in intrinsically unstructured proteins.
667 *FEBS Lett.* **579**, 3346–3354 (2005).

668 7. Jakob, U., Kriwacki, R. & Uversky, V. N. Conditionally and transiently disordered proteins:
669 awakening cryptic disorder to regulate protein function. *Chem. Rev.* **114**, 6779–6805 (2014).

670 8. Mitrea, D. M. & Kriwacki, R. W. Regulated unfolding of proteins in signaling. *FEBS Lett.* **587**,
671 1081–1088 (2013).

672 **Excellent review introducing concepts related to conditional IDRs and their functional role.**

673 9. Borgia, A. *et al.* Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61–66
674 (2018).

675 10. Bah, A. *et al.* Folding of an intrinsically disordered protein by phosphorylation as a regulatory
676 switch. *Nature* **519**, 106–109 (2015).

677 **Landmark study on the phospho-dependent folding of IDP EIF4EBP2 providing a exemplar for**
678 **future studies on conditional IDRs.**

679 11. Davey, N. E. *et al.* An intrinsically disordered proteins community for ELIXIR. [version 1; peer

- 680 review: 2 approved]. *F1000Res.* **8**, (2019).
- 681 12. Taylor, C. F. *et al.* The minimum information about a proteomics experiment (MIAPE). *Nat.*
682 *Biotechnol.* **25**, 887–893 (2007).
- 683 13. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and
684 stewardship. *Sci. Data* **3**, 160018 (2016).
- 685 14. Orchard, S. *et al.* The minimum information required for reporting a molecular interaction
686 experiment (MIMIx). *Nat. Biotechnol.* **25**, 894–898 (2007).
- 687 **Minimum information standards for the molecular interaction experiment that provided the**
688 **framework for the development of the MIADE guidelines**
- 689 15. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*
690 **49**, D480–D489 (2021).
- 691 16. Giglio, M. *et al.* ECO, the Evidence & Conclusion Ontology: community standard for evidence
692 information. *Nucleic Acids Res.* **47**, D1186–D1194 (2019).
- 693 17. Bruford, E. A. *et al.* Guidelines for human gene nomenclature. *Nat. Genet.* **52**, 754–758 (2020).
- 694 18. A one-letter notation for amino acid sequences. *Pure Appl. Chem.* **31**, 641–645 (1972).
- 695 19. Moses, D. *et al.* Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to their
696 Chemical Environment. *J. Phys. Chem. Lett.* **11**, 10131–10136 (2020).
- 697 20. Speer, S. L. *et al.* The intracellular environment affects protein-protein interactions. *Proc Natl*
698 *Acad Sci USA* **118**, (2021).
- 699 21. Sivade Dumousseau, M. *et al.* Encompassing new use cases - level 3.0 of the HUPO-PSI format
700 for molecular interactions. *BMC Bioinformatics* **19**, 134 (2018).
- 701 22. Montecchi-Palazzi, L. *et al.* The PSI-MOD community standard for representation of protein
702 modification data. *Nat. Biotechnol.* **26**, 864–866 (2008).
- 703 23. Nadendla, S. *et al.* ECO: the Evidence and Conclusion Ontology, an update for 2022. *Nucleic*
704 *Acids Res.* **50**, D1515–D1521 (2022).
- 705 24. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites.
706 *Nucleic Acids Res.* **44**, D1214-9 (2016).
- 707 25. Leinonen, R. *et al.* The european nucleotide archive. *Nucleic Acids Res.* **39**, D28-31 (2011).

- 708 26. RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence
709 search and new member databases. *Nucleic Acids Res.* **49**, D212–D220 (2021).
- 710 27. Kiss, R., Kovács, D., Tompa, P. & Perczel, A. Local structural preferences of calpastatin, the
711 intrinsically unstructured protein inhibitor of calpain. *Biochemistry* **47**, 6936–6945 (2008).
- 712 28. Krois, A. S., Dyson, H. J. & Wright, P. E. Long-range regulation of p53 DNA binding by its
713 intrinsically disordered N-terminal transactivation domain. *Proc Natl Acad Sci USA* **115**,
714 E11302–E11310 (2018).
- 715 29. Lalli, D., Rosa, C., Allegrozzi, M. & Turano, P. Distal unfolding of ferricytochrome c induced by
716 the F82K mutation. *Int. J. Mol. Sci.* **21**, (2020).
- 717 30. Thielges, M. C., Chung, J. K., Axup, J. Y. & Fayer, M. D. Influence of histidine tag attachment
718 on picosecond protein dynamics. *Biochemistry* **50**, 5799–5805 (2011).
- 719 31. Yamaguchi, Y. *et al.* NMR characterization of intramolecular interaction of osteopontin, an
720 intrinsically disordered protein with cryptic integrin-binding motifs. *Biochem. Biophys. Res.*
721 *Commun.* **393**, 487–491 (2010).
- 722 32. Scharnagl, C., Reif, M. & Friedrich, J. Stability of proteins: temperature, pressure and the role of
723 the solvent. *Biochim. Biophys. Acta* **1749**, 187–213 (2005).
- 724 33. Padan, E. *et al.* NhaA of Escherichia coli, as a model of a pH-regulated Na⁺/H⁺ antiporter.
725 *Biochim. Biophys. Acta* **1658**, 2–13 (2004).
- 726 34. Padan, E., Bibi, E., Ito, M. & Krulwich, T. A. Alkaline pH homeostasis in bacteria: new insights.
727 *Biochim. Biophys. Acta* **1717**, 67–88 (2005).
- 728 35. Haslbeck, M. *et al.* Hsp26: a temperature-regulated chaperone. *EMBO J.* **18**, 6744–6751 (1999).
- 729 36. Stromer, T., Fischer, E., Richter, K., Haslbeck, M. & Buchner, J. Analysis of the regulation of
730 the molecular chaperone Hsp26 by temperature-induced dissociation: the N-terminal domain is
731 important for oligomer assembly and the binding of unfolding proteins. *J. Biol. Chem.* **279**,
732 11222–11228 (2004).
- 733 37. Silva, J. L., Foguel, D. & Royer, C. A. Pressure provides new insights into protein folding,
734 dynamics and structure. *Trends Biochem. Sci.* **26**, 612–618 (2001).
- 735 38. Peng, X., Jonas, J. & Silva, J. L. Molten-globule conformation of Arc repressor monomers

- 736 determined by high-pressure ¹H NMR spectroscopy. *Proc Natl Acad Sci USA* **90**, 1776–1780
737 (1993).
- 738 39. Greene, D. N. *et al.* Single-molecule force spectroscopy reveals a stepwise unfolding of
739 *Caenorhabditis elegans* giant protein kinase domains. *Biophys. J.* **95**, 1360–1370 (2008).
- 740 40. Wood, M. J., Storz, G. & Tjandra, N. Structural basis for redox regulation of Yap1 transcription
741 factor localization. *Nature* **430**, 917–921 (2004).
- 742 41. Rubinstenn, G. *et al.* Structural and dynamic changes of photoactive yellow protein during its
743 photocycle in solution. *Nat. Struct. Biol.* **5**, 568–570 (1998).
- 744 42. Kar, M., Posey, A. E., Dar, F., Hyman, A. A. & Pappu, R. V. Glycine-Rich Peptides from FUS
745 Have an Intrinsic Ability to Self-Assemble into Fibers and Networked Fibrils. *Biochemistry* **60**,
746 3213–3222 (2021).
- 747 43. Vagenende, V., Yap, M. G. S. & Trout, B. L. Mechanisms of protein stabilization and prevention
748 of protein aggregation by glycerol. *Biochemistry* **48**, 11084–11096 (2009).
- 749 44. Ribeiro, S., Ebbinghaus, S. & Marcos, J. C. Protein folding and quinary interactions: creating
750 cellular organisation through functional disorder. *FEBS Lett.* **592**, 3040–3053 (2018).
- 751 45. Wingfield, P. T. Use of protein folding reagents. *Curr. Protoc. Protein Sci.* **Appendix 3**,
752 Appendix 3A (2001).
- 753 46. Theillet, F.-X. *et al.* Structural disorder of monomeric α -synuclein persists in mammalian cells.
754 *Nature* **530**, 45–50 (2016).
- 755 47. Fonseca-Ornelas, L. *et al.* Altered conformation of α -synuclein drives dysfunction of synaptic
756 vesicles in a synaptosomal model of Parkinson's disease. *Cell Rep.* **36**, 109333 (2021).
- 757 48. Lacy, E. R. *et al.* p27 binds cyclin-CDK complexes through a sequential mechanism involving
758 binding-induced protein folding. *Nat. Struct. Mol. Biol.* **11**, 358–364 (2004).
- 759 49. Huth, J. R. *et al.* The solution structure of an HMG-I(Y)-DNA complex defines a new
760 architectural minor groove binding motif. *Nat. Struct. Biol.* **4**, 657–665 (1997).
- 761 50. Léger, A. *et al.* PiP2 favors an α -helical structure of non-recombinant Hsp12 of *Saccharomyces*
762 *cerevisiae*. *Protein Expr. Purif.* **181**, 105830 (2021).
- 763 51. Wang, N. *et al.* Ordering a dynamic protein via a small-molecule stabilizer. *J. Am. Chem. Soc.*

- 764 **135**, 3363–3366 (2013).
- 765 52. O’Brien, D. P. *et al.* Calcium-dependent disorder-to-order transitions are central to the secretion
766 and folding of the CyaA toxin of *Bordetella pertussis*, the causative agent of whooping cough.
767 *Toxicon* **149**, 37–44 (2018).
- 768 53. Ulrich, E. L. *et al.* BioMagResBank. *Nucleic Acids Res.* **36**, D402-8 (2008).
- 769 54. Ramalli, S. G., Miles, A. J., Janes, R. W. & Wallace, B. A. The PCDDDB (protein circular
770 dichroism data bank): A bioinformatics resource for protein characterisations and methods
771 development. *J. Mol. Biol.* 167441 (2022) doi:10.1016/j.jmb.2022.167441.
- 772 55. Kikhney, A. G., Borges, C. R., Molodenskiy, D. S., Jeffries, C. M. & Svergun, D. I. SASBDB:
773 Towards an automatically curated and validated repository for biological scattering data. *Protein*
774 *Sci.* **29**, 66–75 (2020).
- 775 56. Lazar, T. *et al.* PED in 2021: a major update of the protein ensemble database for intrinsically
776 disordered proteins. *Nucleic Acids Res.* **49**, D404–D411 (2021).

