A



B



**Figure S1**. **A)** Amino acid frequencies shown for the individual and pooled libraries compared to the yeast proteome. Residue colors are the same as in **Fig 2**. **B)** Distribution of AD scores obtained when a dipeptide coefficient is swapped to one of the 399 others (STAR Methods and text). Reverse: WD instead of DW etc. *Related to **Figs 1 and 2**.*

## A
### Ino2 AD1
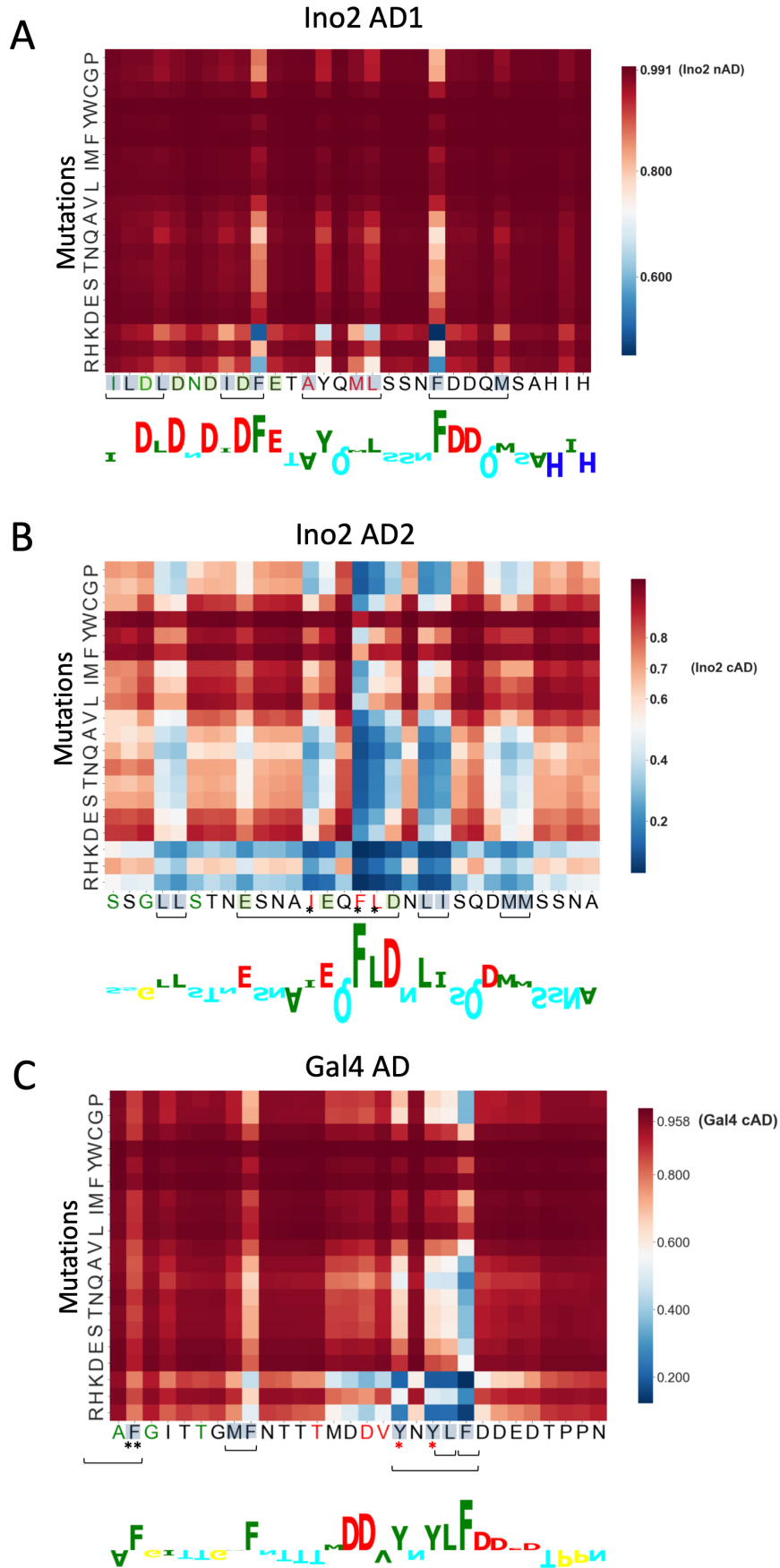


## B
### Ino2 AD2



## C
### Gal4 AD

**Figure S2**. Prediction of important residues within yeast ADs and comparison with in vivo analysis. *ADpred* scores predicting the probability of AD function for all possible single amino acid mutations for (**A**) activation domain 1 (AD1) of Ino2, (**B**) AD2 of Ino2, and (**C**) the AD of Gal4. Red indicates a high and blue a low *ADpred* probability for the in-silico mutation. Wild type *ADpred* scores are indicated in the colorbar. For comparison, results from an in vivo analysis where double or triple alanine substitutions were assayed for AD function (Pacheco et al., 2018; Tuttle et al., 2019). Conserved hydrophobic and acidic residues that were mutated are shown in blue and green, respectively. Double or triple alanine mutations resulting in less than ~ 50% AD function are marked with brackets below the x-label. For Ino2 AD1, conserved residues that ADpred predicts to be important but not tested experimentally are indicated by: *. For Gal4 mutations, residue F849 (marked with **) was mutated in conjunction with Y846 and this derivative has 47% WT activity. Red asterisk marks Gal4 residues Y865 and Y867, which have ≥75% WT function when individually mutated to Ala. *Related to* **Fig 4**.
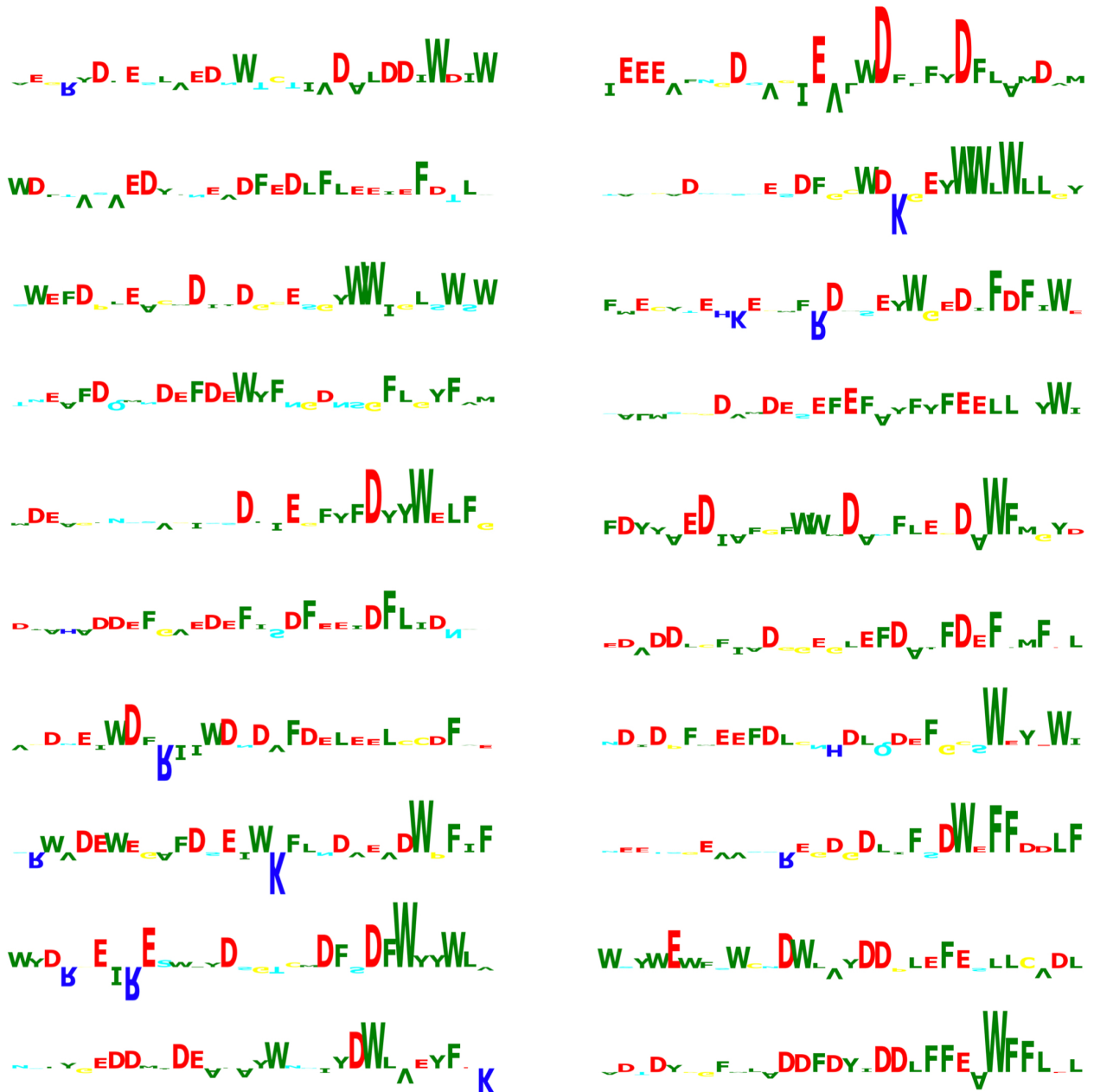
**Figure S3.** Prediction of functionally important residues in synthetic ADs. Shown is analysis for 20 high scoring synthetic ADs from the AD-positive set analyzed with the *Integrated Gradients* algorithm. Logos are drawn as in Fig 4C. *Related to **Fig 4**.*
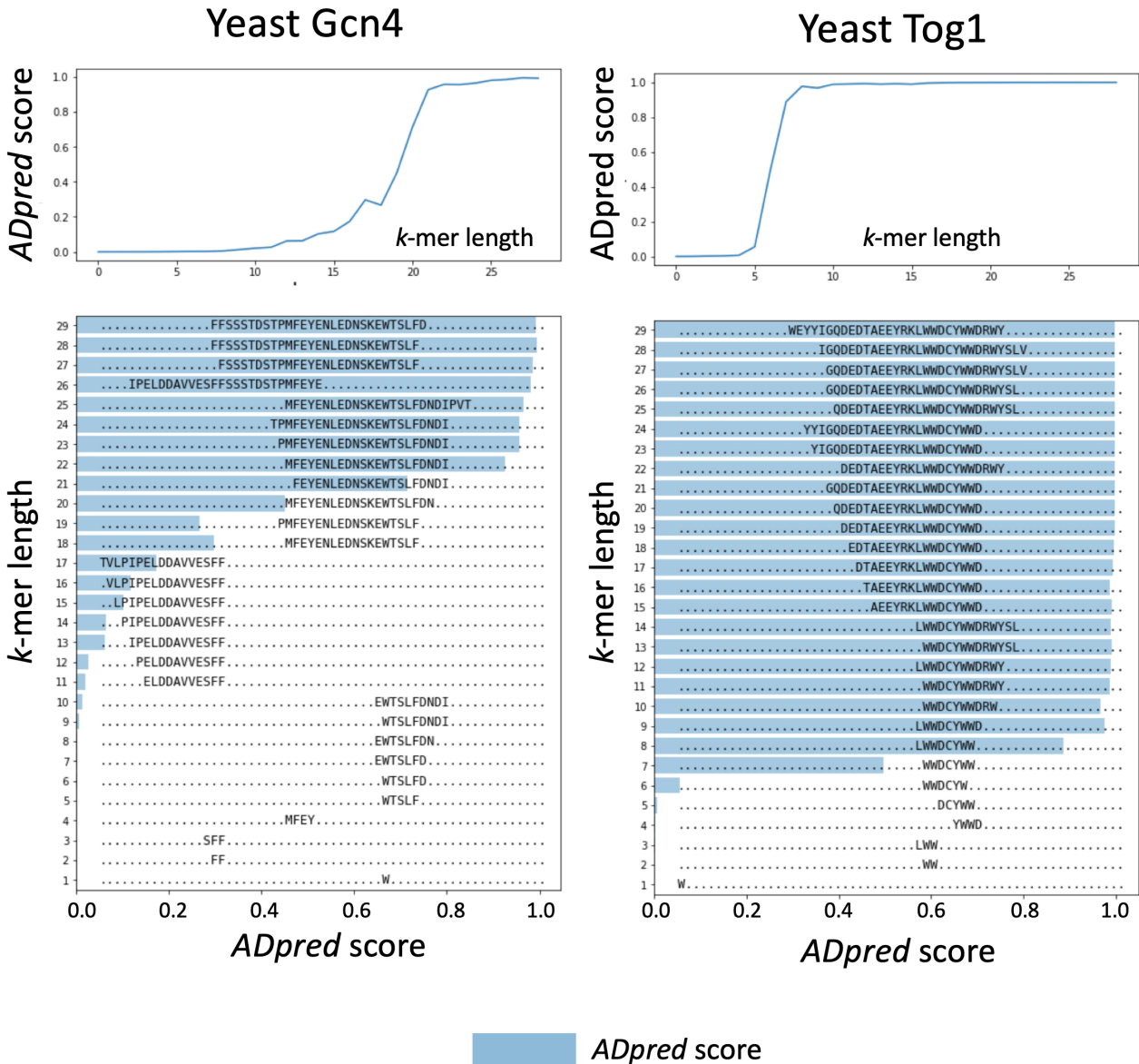
**Figure S4**. AD length determined by *k*-mer analysis of two yeast transcription factors. For each *k*-mer length between 1 and 29, we extracted each *k*-mer contained in the sequence of yeast transcription factors Gcn4 (left) and Tog1 (right) and computationally inserted them between 30 residue-long randomly generated flanking sequences that showed negligible ADpred scores: TNSANAANASASSQAGQQATQNQNTAQQNG (N-terminal) and GNGNQNQTTSTSNASANANSGSQGTGSSSQ (C-terminal). **Top**: For each length k, the *k*-mer with the maximum ADpred score is plotted. **Bottom**: sequences are aligned relative to the WT sequence with blue bars indicating the *ADpred* score for each individual peptide when inserted in the neutral flanking sequence. *Related to* **Fig 6.**
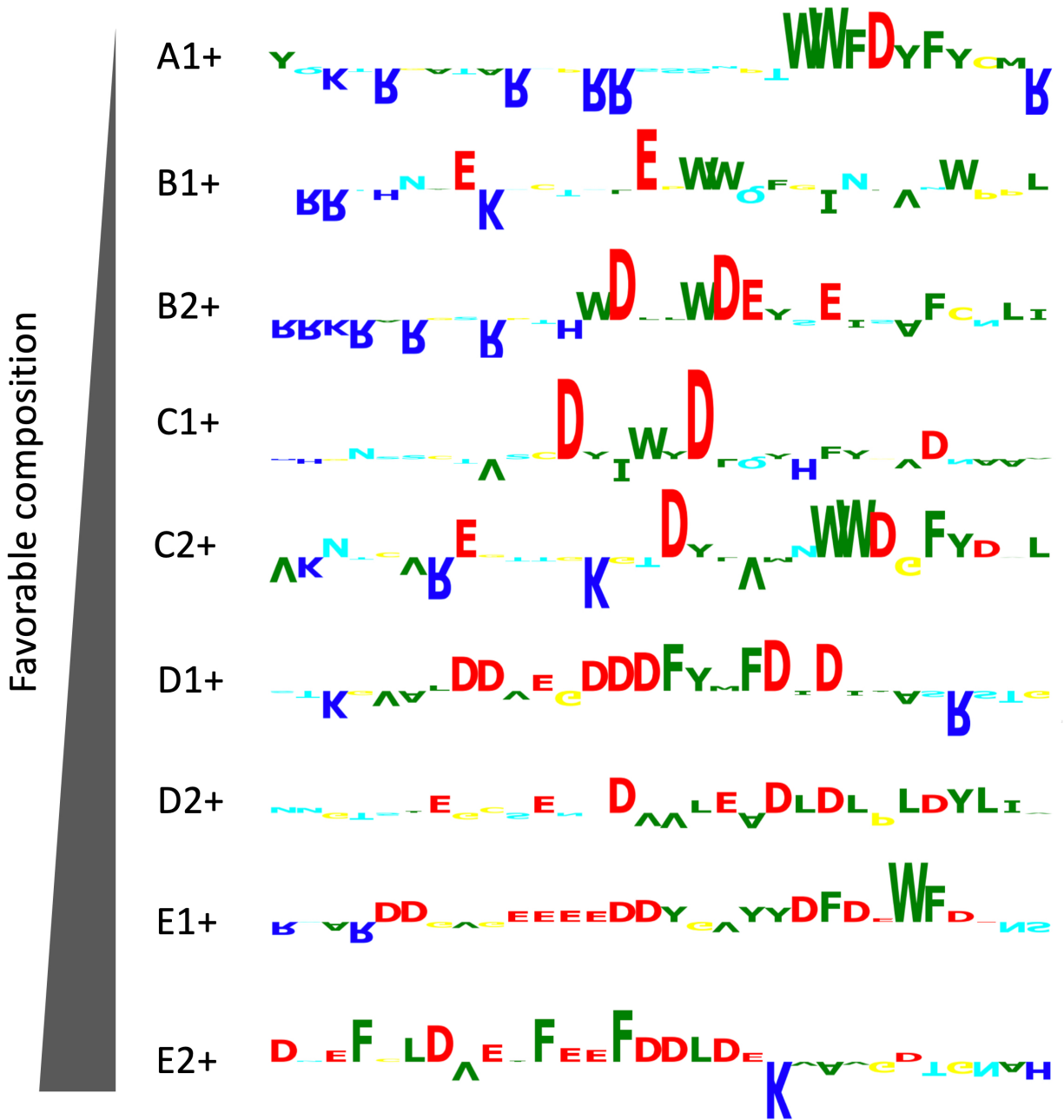
**Figure S5**. Shown is the *Integrated Gradient* analysis of predicted AD-positive peptides with variable amino acid composition from **Fig 6C.** The sequence names A-E indicate composition from very unfavorable (A) to very favorable (E). Logos are plotted as in **Figs 4 and S3**. *Related to Fig 6*.
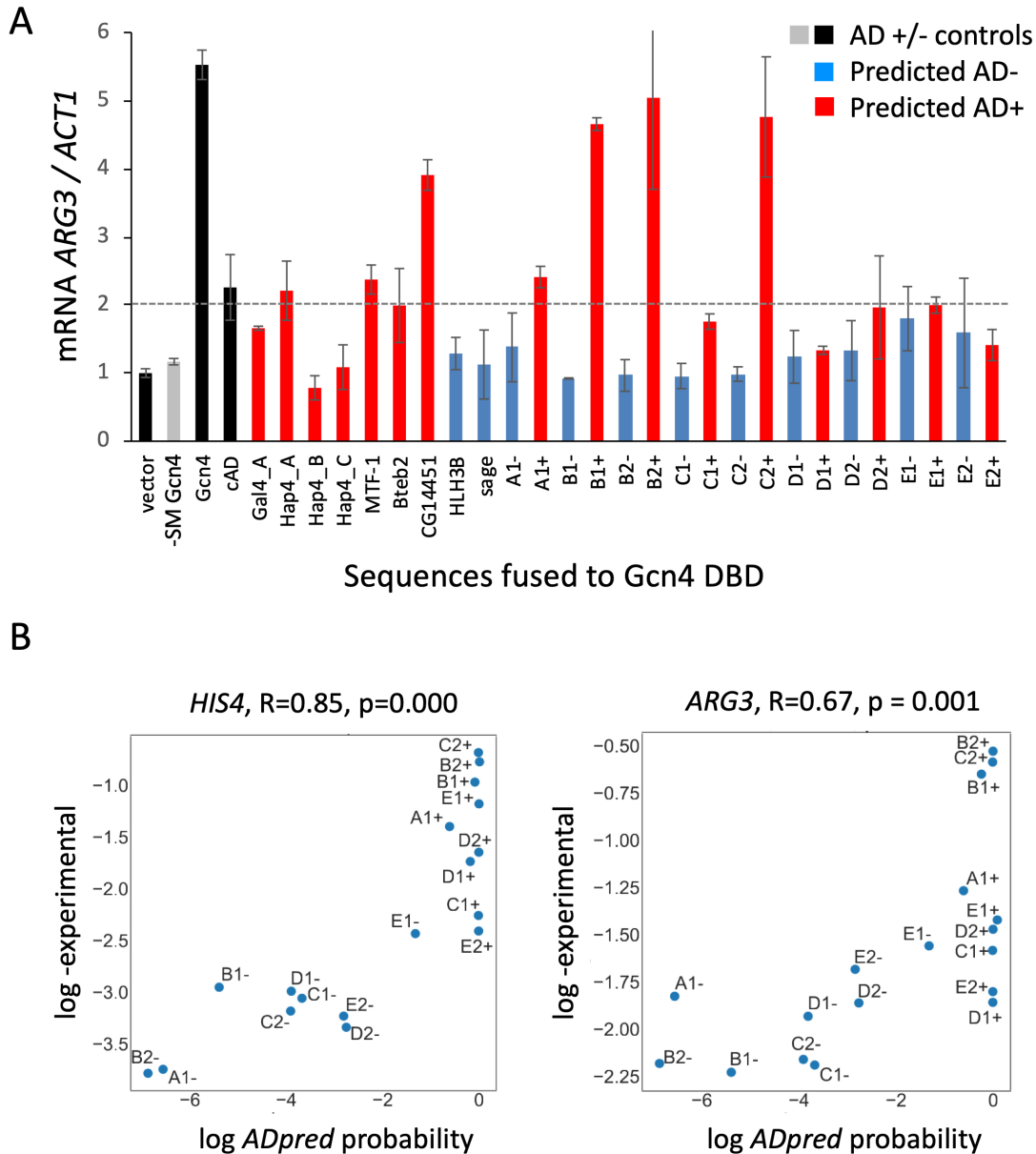
**Figure S6**. (**A**) Functional analysis of yeast and synthetic ADs. mRNA quantitation as in **Figs. 6A** and **6C** but with quantitation of mRNA at the Gcn4-dependent *ARG3* gene. Dotted line indicates 2-fold activation above cells lacking Gcn4 (vector). Red bars = sequences with high *ADpred* probability; blue bars = low *ADpred* probability. All samples were treated with SM unless otherwise indicated. Dotted horizontal line: level of SM-induced transcription in cells lacking Gcn4. (**B**) Scatter plot of the logarithm of *ADpred* probabilities versus log-experimental RT qPCR results obtained on *HIS4* and *ARG3* mRNAs. Pearson correlation and *p*-value for a two-sided hypothesis test (where the null hypothesis corresponds to slope=0) are indicated. *Related to **Fig 6.***
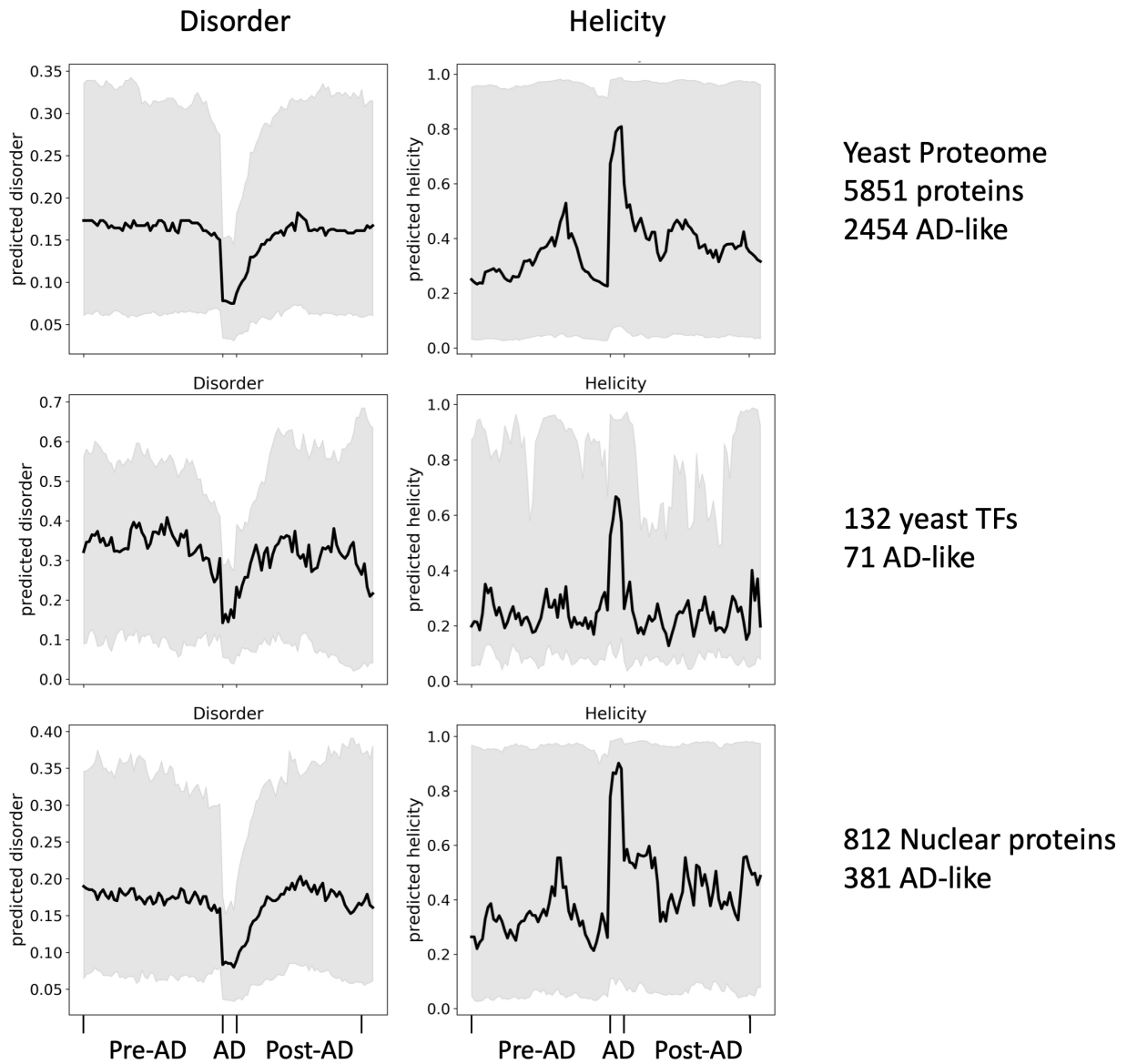
**Disorder** · **Helicity**



Yeast Proteome
5851 proteins
2454 AD-like

132 yeast TFs
71 AD-like

812 Nuclear proteins
381 AD-like

Pre-AD  AD  Post-AD

**Figure S7**. Structural properties of regions surrounding predicted ADs in yeast transcription factors. Analysis was carried out as described for **Fig 7**. The upper plot represents analysis of the yeast proteome, the middle plot represents analysis of all yeast proteins classified as nuclear and the lower plot (reproduced from **Fig 7B**) is an analysis of 132 curated yeast transcription factors. *Related to **Fig 7**.*

**Supplementary Tables**

**Table S1** (recommend opening with text editor). *See **Fig 1**.*

Unsorted list of AD-positive and AD-negative sequences with:

- List of AD-positive and negative sequences
- Distribution of sequences in the background and the four FACS bins
- Calculated AD-enrichment score

**Table S2**

- RT qPCR data and results. See ***Fig 6** and **Fig S6***.
- Sequences of the natural and synthetic ADs tested in Fig 6.

**Table S3**

The sets of yeast, *Drosophila* and human transcription factors (TFs) used in the AD enrichment analysis of **Fig 7A** and **Fig S7**. TFs are listed as UniProt IDs (Bateman et al., 2018). Yeast factors are a curated list combining data from mining the *Saccharomyces* Genome Database (Cherry et al., 2012), the set of TFs from Harbison (Harbison et al., 2004) and from manual inspection of known functional properties of each factor. Human and drosophila TF lists were obtained from factors (Stampfel et al., 2015; Vaquerizas et al., 2009). *See **Fig 7**.*

**Table S4.**

Performance metrics of the regression and deep learning models.

| Method | Feature | AUPRC | AUROC | Accuracy |
|---|---|---|---|---|
| **Regression** | **Single aa frequency** | 0.9337 ± 0.0024 | 0.9452 ± 0.0020 | 0.8830 ± 0.0032 |
| **Regression** | **Dipeptide frequency** | 0.9418 ± 0.0018 | 0.9508 ± 0.0017 | 0.8915 ± 0.0039 |
| **Deep NN** | **Seq.** | 0.9741 ± 0.0007 | 0.9762 ± 0.0004 | 0.9303 ± 0.0008 |
| **Deep NN** | **Seq._Dis.** | 0.9726 ± 0.0008 | 0.9747 ± 0.0005 | 0.9268 ± 0.0010 |
| **Deep NN** | **Seq._SS. (ADpred)** | 0.9750 ± 0.0007 | 0.9768 ± 0.0005 | 0.9324 ± 0.0013 |
| **Deep NN** | **Seq._SS._Dis.** | 0.9729 ± 0.0006 | 0.9750 ± 0.0005 | 0.9285 ± 0.0011 |