

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
- ☐ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

Universal Protein Binding Microarray (PBM) Analysis Suite (http://the_brain.bwh.harvard.edu/PBMAAnalysisSuite/indexSep2017.html)
python Dependencies: pandas, biopython
R Dependencies (packages at: <https://cran.r-project.org/>): caret (v6), glmnet (v 2.0-13), PRROC (v1.3), aphid (v 1.0.1), seqinr (v3.4.5)

Data analysis

The SR code, and examples, are made available on GitHub (<https://github.com/smlmbt/SimilarityRegression>).
hmmscan (v3.2.1, part of the HMMER toolkit, <http://hmm.org/>)
Affinity Regression (<https://bitbucket.org/leslielab/affreg>)
PreMoTF (<http://stormo.wustl.edu/PreMoTF/>)
ZFModels (<http://stormo.wustl.edu/ZFModels/>)
C2H2 SVM Method (<http://zf.princeton.edu/>)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

New PBM data and motifs are deposited in GEO (accession number: GSE121420), and Cis-BP (<http://cisbp.ccbr.utoronto.ca/>).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We successfully generated new PBM data for 340 TFs representing multiple eukaryotic kingdoms, with a particular focus on Cannabis sativa (a medicinal plant), Caenorhabditis briggsae (a nematode), Aspergillus nidulans and Neurospora crassa (model fungi), and also 15 human TFs. These TFs were selected on the basis of at least one of two different criteria: first, to increase the number of experimentally determined motifs for TFs in these species of interest, and second, to obtain novel motifs by analyzing proteins that are dissimilar to TFs with known motifs. For A. nidulans TF 350 novel DBDs were selected for analysis and 180 were successfully cloned into the expression vector (pTH6838) and validated by sequencing. Sample size was not predetermined, but is large enough to test the SR and other prediction algorithms on unseen data from relevant/highly-abundant TF families.
Data exclusions	No data passing our PBM success criteria was excluded from the analysis.
Replication	Successful replication of SR model accuracy was evaluated using the new PBM data, and motif similarity calculated from other motifs included in the human vs. fly analysis.
Randomization	Randomization was not necessary in this study as samples were allocated to different groups (TF similarity) based on uniform rules learned during cross-validation.
Blinding	Blinding was not necessary as SR models are developed automatically using labeled training data based on established thresholds. Performance metrics were calculated without human intervention.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials TF clones used for PBMs are available upon direct request to the corresponding author.