Nucleic Acids Research, 2026, **54**, D184–D193
https://doi.org/10.1093/nar/gkaf1209
Advance access publication date: 2 December 2025
Database issue

OXFORD

JASPAR 2026: expansion of transcription factor binding profiles and integration of deep learning models

Damla Ovek Baydar[1,†], Ieva Rauluseviciute[1,†], Dina R. Aronsen[1,‡], Romain Blanc-Mathieu[2,‡], Ine Bonthuis[1,‡], Herman de Beukelaer[3,4,‡], Katalin Ferenc[1,‡], Alice Jegou[2,‡], Vipin Kumar[1,‡], Roza Berhanu Lemma[1,‡], Jérémy Lucas[2,‡], Mathis Pochon[2,‡], Chang M. Yun[5,‡], Vivekanandan Ramalingam[6,‡], Salil Sanjay Deshpande[7,‡], Aman Patel[8], Georgi K. Marinov[6], Austin T. Wang[8], Alejandro Aguirre[9,10], Jaime A. Castro-Mondragon[1,11], Damir Baranasic[12,13,14], Jeanne Chèneby[15], Sveinung Gundersen[15], Morten Johansen[15], Aziz Khan[16], Marieke L. Kuijjer[1,17,18], Eivind Hovig[15], Boris Lenhard[13,14,*], Albin Sandelin[19,*], Klaas Vandepoele[3,4,20], Wyeth W. Wasserman[9,10,*], François Parcy[2,*], Anshul Kundaje[6,8,*], Anthony Mathelier[1,21,22,*]

[1]Norwegian Centre for Molecular Biosciences and Medicine (NCMBM), Nordic EMBL Partnership, University of Oslo, Oslo 0318, Norway
[2]Laboratoire Physiologie Cellulaire et Végétale, Univ. Grenoble Alpes, CNRS, CEA, INRAE, IRIG-DBSCI-LPCV, Grenoble F-38054 17 avenue des martyrs, France
[3]Department of Plant Biotechnology and Bioinformatics, Ghent University, 9051, Ghent, Belgium
[4]Center for Plant Systems Biology, VIB 9051 Ghent, Belgium
[5]Department of Chemical Engineering, Stanford University, Stanford, CA 94305, United States
[6]Department of Genetics, Stanford University, Stanford, CA 94305, United States
[7]Institute for Computational and Mathematical Engineering (ICME), Stanford University, Stanford, CA 94305, United States
[8]Department of Computer Science, Stanford University, Stanford, CA 94305, United States
[9]Department of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z3, Canada
[10]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children’s Hospital Research Institute, University of British Columbia, BC V5Z 4H4 Vancouver 950 W 28th Ave, Canada
[11]Akershus University Hospital, Department of Clinical Molecular Biology, Unit for Precision Medicine, Lørenskog,1478, Norway
[12]Division of Electronics, Ruđer Bošković Institute, 10000 Zagreb Bijenička cesta, Croatia
[13]MRC Laboratory of Medical Sciences, London W12 0NN Du Cane Road, United Kingdom
[14]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Hammersmith Hospital Campus, London W12 0NNDu Cane Road, United Kingdom
[15]Department of Biosciences, University of Oslo, Oslo 0316, Norway
[16]Department of Computational Biology, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE
[17]iCAN Flagship in Digital Precision Cancer Medicine, University of Helsinki, Helsinki, 00014, Finland
[18]Department of Biochemistry and Developmental Biology, University of Helsinki, Helsinki, 00014,Finland
[19]Department of Biology and Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen DK2200 N, Denmark
[20]Center for AI & Computational Biology, VIB, Ghent 9051, Belgium
[21]Department of Medical Genetics, Institute of Clinical Medicine, Oslo University Hospital and University of Oslo, Oslo 0318, Norway
[22]Bioinformatics in Life Science (BiLS) initiative, Department of Pharmacy, University of Oslo, Oslo 0316, Norway

*To whom correspondence should be addressed: Email: anthony.mathelier@ncmbm.uio.no
Correspondence may also be addressed to François Parcy. Email: francois.parcy@cea.fr
Correspondence may also be addressed to Wyeth W. Wasserman. Email: wyeth@cmmt.ubc.ca
Correspondence may also be addressed to Anshul Kundaje. Email: akundaje@stanford.edu
Correspondence may also be addressed to Albin Sandelin. Email: albin@binf.ku.dk
Correspondence may also be addressed to Boris Lenhard. Email: b.lenhard@imperial.ac.uk
[†]The first two authors should be regarded as Joint First Authors
[‡]These authors contributed equally to this work.

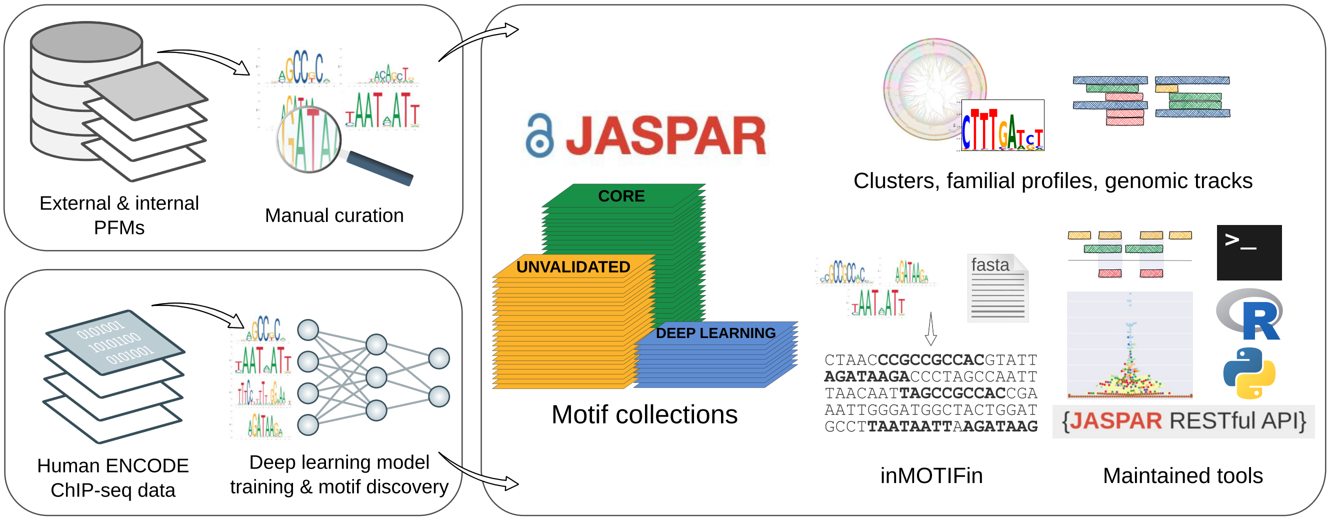
Abstract

JASPAR (https://jaspar.elixir.no/) is an open-access database that has provided high-quality, manually curated, and non-redundant DNA binding profiles for transcription factors (TFs) as position frequency matrices (PFMs) for over 20 years. We expanded the CORE (306 new profiles, 12% increase) and UNVALIDATED (433, 60% increase) collections with new PFMs and updated 13 existing profiles. We updated the TF binding site predictions and genome tracks for eight species. TF binding profile clusters and familial TF binding sites were updated accordingly. We integrate the inMOTIFin software to easily simulate regulatory sequences using JASPAR PFMs. To enrich TFs’ annotations, we provide scientific literature-

Received: September 15, 2025. **Revised:** October 14, 2025. **Accepted:** October 15, 2025

based human TF target information. Notably, this release features a deep learning (DL) collection, providing a paradigm shift in modeling and characterizing TF–DNA interactions with 1259 BPNet models trained on *Homo sapiens* ENCODE chromatin immunoprecipitation followed by sequencing (ChIP-seq) datasets from 240 TFs and interpreted to reveal predictive motif patterns for the models. The motifs associated with the same TF were clustered to provide a summary of the binding properties, resulting in 240 primary and 113 alternative motif patterns in the DL collection. The JASPAR 2026 collections lay a foundation for future endeavors in genomic research, serving the scientific community in uncovering the mechanisms of gene regulation.

Graphical abstract



Introduction

Transcription factors (TFs) are regulatory proteins that control gene transcription through *cis*-regulatory elements (CREs) such as promoters and enhancers. Although different types of TFs exist, this paper focuses on those that bind DNA in a sequence-specific manner [1]. For simplicity, we refer to them as TFs. The sequence-specific DNA binding of TFs is achieved through their DNA-binding domains (DBDs), which interact with the DNA at TF binding sites (TFBSs) [1]. While TFs recognize short DNA sequences, their binding activity in complex organisms is highly context-dependent [2]. In addition to the genomic sequence patterns recognized by TFs, their DNA occupancy is modulated by local chromatin accessibility, nucleosome positioning, DNA shape features, binding of other TFs, and the cooperation with other protein co-factors [3]. In many cases, TFs do not act in isolation but form cooperative complexes, stabilizing each other's binding event [2, 4]. This combinatorial binding enables precise and dynamic regulation of gene expression across cell types and conditions.

Position weight matrices (PWMs), derived from position frequency matrices (PFMs), are the most common computational representations modeling how TFs interact with DNA. PWMs are quantitative summaries of a TF's DNA-binding preferences, created by tallying the frequency, or log-likelihood score, of each nucleotide at every position within a set of experimentally observed TF–DNA interactions. These matrix models offer significant utility in various computational analyses, including the assessment of TFBS enrichment in regulatory regions [5], prediction of the impact of mutations in CREs, and guidance for *in vitro* mutagenesis experiments [6, 7]. Several open-access databases (e.g. CIS-BP [8], HOCOMOCO [9, 10], and JASPAR [11]) collect and store PFMs and PWMs.

JASPAR is an open-access database that provides manually curated, non-redundant TF binding profiles, primarily as PFMs across various taxonomic groups. Since its initial release in 2004 [12], JASPAR has become a standard resource in computational regulatory genomics due to its commitment to high-quality, accessible data and continuous expansion of content and tools, including a focus on open science and ease of use.

Despite their widespread use, PFMs have well-known limitations. They assume nucleotides at each position within TFBSs contribute independently to binding, even though interactions between adjacent or distant bases can be critical [13]. They also do not inherently consider genomic context (e.g. cooperativity, nucleosome positioning, co-factor binding) [3]. We and others have developed machine learning approaches, including Markov models [13, 14], variable-order Bayesian trees [15], support vector machines [16], and gradient boosting of decision trees [17], to improve TF–DNA interaction predictions by detecting more complex patterns.

Artificial intelligence, particularly the use of deep learning (DL) convolutional neural networks (CNNs), has led to new models that provide a shift in TF–DNA interaction modeling [18–22]. DL models are becoming an established method for decoding the *cis*-regulatory grammar of genomes [21, 22]. These models excel at autonomously discerning intricate regulatory patterns, facilitating context-specific and precise predictions [21, 22].

As such, DL models are trained in a supervised manner, and researchers aim to interpret the sequence patterns that have been captured. The interpretability of the models is performed using explainable artificial intelligence methods [20]. For instance, some methods, such as DeepLIFT [23], assign contribution scores to input nucleotides to pinpoint the ones with predictive power for the model. Then, tools like TF-MoDISco use contribution scores to derive predictive DNA motifs, thereby enabling the interpretability of the models [24]. Combining base-pair resolution predictive accuracy of experimental TF binding patterns (e.g. chromatin immunoprecipitation followed by sequencing, ChIP-seq) with motif-level interpretability, DL models, such as BPNet [25], have become transformative for studying condition-specific gene regulation by capturing complex features, such as the spacing and orientation of TFBSs and TF cooperativity.

To complement the classical PFM representation for modeling of TF–DNA binding, we now introduce a "Deep Learning" (DL) collection of BPNet-trained models on *Homo sapiens* TF ChIP-seq data from the ENCODE [26, 27], providing high-resolution, curated TF binding predictions and interpreted models visualized as logos. In this release, we have integrated 1259 BPNet models, identifying 353 primary and alternative motif patterns for 240 TFs in the DL collection.

Additionally, we have updated and expanded our CORE and UNVALIDATED collections. We have added 686 new profiles (265 in the CORE collection and 421 in the UNVALIDATED). Forty-one profiles from the UNVALIDATED collection of the previous release now have orthogonal support and were promoted to the CORE collection. Finally, we have updated 13 profiles with new matrices and also updated the metadata of 62 profiles.

Moreover, we have updated our tools and integrated new software into JASPAR. We introduce a simulation tool, inMOTIFin, that enables users to generate motifs and create or modify regulatory sequences by inserting motif instances into sequences with precise control over their frequencies, positions, and co-occurrences [28]. We have updated the pyJASPAR and Bioconductor packages. Finally, we introduce TF targets derived from a dedicated large language model that extracts TF-target relationships from the scientific literature.

Results

Expansion and update of the classical TF binding profiles

We performed manual curation of TF binding profiles using PFMs and position probability matrices from public resources (HOCOMOCO [10], ModERN [29], CIS-BP [30], Codebook/GRECO-BIT [31], and CAP-SELEX [32]). The collected set of profiles was complemented with PFMs that we generated through *de novo* motif enrichment analysis from ChIP-seq data from KRABopedia [33] and ModERN [29], as well as SMiLE-Seq, ChIP-seq, and GHT-SELEX from Codebook & GRECO-BIT [31] (see Supplementary Text for methodological details). Finally, we downloaded and processed CUT&RUN, ChIP-seq, DAP-seq, ampDAP-seq, and ChEC-seq data stored in GEO from individual studies (see Supplementary Table S1 for a complete list). JASPAR expert curators manually evaluated 11 565 profiles and selected the PFMs supported by orthogonal validation from the literature to add them or update former TF binding profiles in the JASPAR CORE collection. PFMs with high quality, but for which the curators did not find orthogonal support in the literature, were added to the JASPAR UNVALIDATED collection. In the current release, the JASPAR CORE collection has been complemented with 265 new binding profiles for four taxa: plants, 125 profiles (18% increase from the previous plants CORE collection), vertebrates, 121 profiles (16% increase), insects, four profiles (3.5% increase), and fungi, 15 profiles (8% increase) (Table 1 and Fig. 1). Furthermore, we promoted 41 profiles previously stored in the UNVALIDATED collection to the CORE collection after identifying orthogonal support from the literature (Table 1 and Fig. 1). Similarly, we complemented the UNVALIDATED collection with 871 new profiles for five taxa (Supplementary Table S2). Due to either profile redundancy or the underlying protein not being considered as a specific DNA-binding TF, we removed 11 profiles (seven from CORE and four from UNVALIDATED). After reevaluation of the JASPAR profiles, we downgraded 12 profiles from CORE to UNVALIDATED due to insufficient literature support. Subsequently, we updated 13 profiles from the CORE collection with new, higher-quality PFMs. In addition to updating the matrices, we also updated existing profile metadata for 62 profiles in both collections.

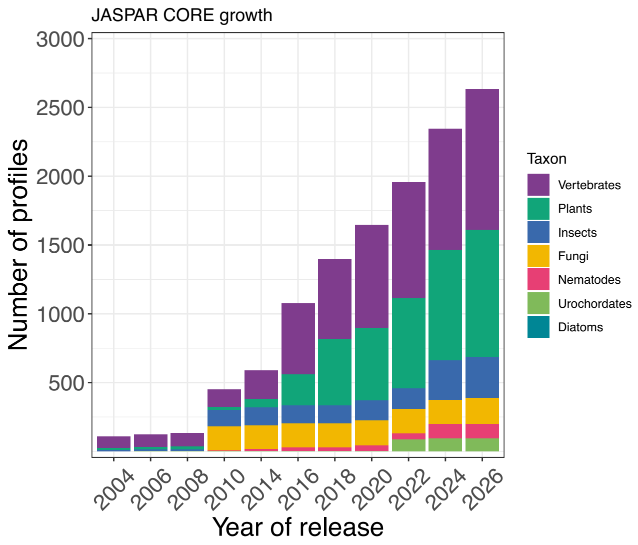


Figure 1. Overview of the growth of the number of profiles in the JASPAR CORE collection from the initiation of the database in 2004 to the latest 2026 release.

The current JASPAR 2026 release provides a total of 2633 and 1231 non-redundant TF DNA-binding profiles in the CORE and UNVALIDATED collections, respectively (Table 1, Fig. 1, Supplementary Table S2, and Supplementary Fig. S1).

TF binding profile clusters, familial binding profiles, word clouds, and genomic tracks

JASPAR has provided PFM collections for over 20 years. Still, in addition to the collections, we include complementary features that enable users to interact with the data and gain insights into the characteristics of TF–DNA interactions and transcriptional regulation. We provide a clustering of TF binding profiles for each taxonomic group, and these groupings can be visualized and downloaded for the CORE collection and the combined CORE and UNVALIDATED collections. Specifically, users can inspect the similarity between TF binding profiles through radial and linear trees. As previously described, the TF binding profiles for TFs belonging to the same structural family or class are often very similar [1]. Therefore, we provide summaries of familial binding profiles obtained from a hierarchical clustering applied to the CORE collection in six main taxonomic groups. These familial binding profiles summarize similar TF binding profiles with a single PFM. In JASPAR 2026, we constructed 504 familial profiles using PFMs from the CORE collection (233

Table 1. Summary of the JASPAR 2026 CORE collection update compared to the previous release

Taxonomic group	Non-redundant PFMs in JASPAR 2024	New non-redundant PFMs	Removed PFMs	Promoted PFMs (from UNVALIDATED to CORE)	Downgraded PFMs (from CORE to UNVALIDATED	Updated PFMs	Total non-redundant PFMs in JASPAR 2026
Plants	805	125	7	16	12	5	927
Vertebrates	879	121	–	19	–	7	1019
Urochordata	94	–	–	–	–	–	94
Insects	286	4	–	6	–	1	296
Nematodes	103	–	–	–	–	–	103
Fungi	178	15	–	–	–		193
Diatoms	1	–	–	–	–	–	1
CORE total	2346	265	7	41	12	13	2633

for vertebrates, 85 for insects, 69 for fungi, 55 for plants, 43 for nematodes, and 19 for urochordates). Users can access cluster and familial binding profile summaries at https://jaspar.elixir.no/matrix-clusters.

To supplement the information provided for each TF in JASPAR, we provide word clouds that summarize biological information explicitly associated with the TFs in the abstracts of scientific literature stored in PubMed. Since their introduction in 2022, we have updated the collection of word clouds and generated new ones for newly added profiles.

Lastly, we scanned the latest genome versions of *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Ciona intestinalis*, *Danio rerio*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, and *Saccharomyces cerevisiae* with TF binding profiles from the JASPAR 2026 CORE collection for the corresponding taxonomic group to predict potential TFBSs in the genomes. Using the predicted TFBSs and the familial binding profiles, we produced genomic tracks for familial binding sites by grouping TFBSs for TFs belonging to the same familial binding profile. Both TFBSs and familial binding site genomic tracks are available in multiple formats for users to visualize and interpret. Moreover, the TFBSs predicted in the mouse and human genomes with CORE vertebrate PFMs are available as native tracks in the UCSC Genome Browser [34].

Large language model-based TF–target gene associations

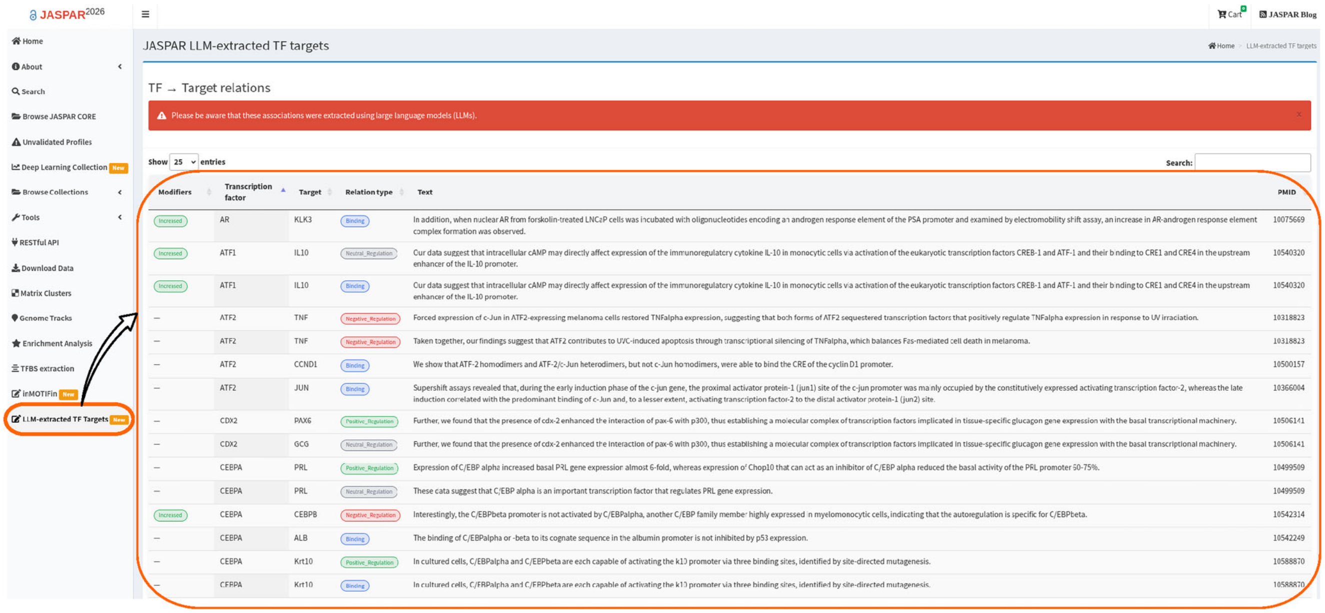
TF–target gene (TF–TG) regulatory relationships are described in the literature. To facilitate the systematic identification of these interactions from published studies and complement the available TF-centric information provided in JASPAR, we designed a prompt and corresponding JSON output schema for ChatGPT-5 to extract TF–TG relationships from a collection of text-mined sentences in the ExTRI resource [35]. Manual curation of a set of 192 ExTRI-derived sentences specified the text referring to TFs and TGs, and a set of 350 relationships between the entities was generated automatically. Annotated TF–TG relationships were determined across four categories: positive regulation, negative regulation, neutral regulation (where no direction can be inferred from the text), and binding (where a TF binds to a DNA element associated with the target gene).

In addition, we captured TF modifiers, which provide contextual information about TF expression or activity in a given sentence. Five non-exclusive categories of modifiers were defined: mutant (the TF carries a mutation or is described as a non-wild-type form), increased expression or activity (e.g. overexpression *in vitro*), reduced expression or activity (e.g. down-regulation, inhibition, or knockdown), absent (e.g. knockout, complete depletion, loss of expression), and altered function (the TF is expressed in a non-canonical context such as an atypical tissue or cell type or acquires a novel DNA binding site). These modifiers provide essential context for interpreting TF–TG relationships (see Supplementary Text for details). Figure 2 illustrates how these annotated relationships are visualized within the JASPAR web interface.

The initial set of relationships was manually reviewed for accuracy (as were the modifiers described earlier), but the future expansions of the relationships produced will be fully automated. Therefore, we have indicated in the interface that these relationships are LLM-extracted; errors may occur. The accuracy of the LLM-generated annotations was 88% (308/350), with the most common mistakes being attributed to TF–TG relations that are influenced by the Increased, Absent, and Reduced TF modifiers, as well as relations that are described in hypothetical statements. Only the 308 correct TF–TG relationships were included in the online interface.

Deep learning collection

We retrieved *Homo sapiens* TF ChIP-seq datasets from ENCODE [26]. We trained a specific BPNet [25] model for each ChIP-seq dataset to predict the ChIP-seq genomic tracks at base-pair resolution from input DNA sequences (see Supplementary Fig. S2 for model performance metrics). We revealed the motifs most contributing to the accuracy of the models using DeepLIFT [23] and TF-MoDISco [24] (Fig. 3 and Supplementary Text). Next, we used the MotifCompendium tool (https://github.com/kundajelab/MotifCompendium) [36] to cluster all the discovered TF-MoDISco motifs per TF, thereby constructing TF binding profiles that summarize the most critical canonical binding pattern(s) across datasets for each TF. We applied several quality-control metrics to ensure that the identified motif patterns corresponding to the cognate TF binding profiles are supported by *in vitro* orthogonal evidence (Fig. 3 and Supplementary Text). These processing steps culminated in motifs for 240 TFs. As multiple motif patterns can be identified for the same TF, we provide them in dedicated summary profile pages (one per TF); each summary profile page is assigned an identifier DLXXXXXX.Y, where XXXXXX is the summary ID and Y is the version number (Fig. 3B). For each TF, we provide the primary motif



JASPAR LLM-extracted TF targets

TF → Target relations

Please be aware that these associations were extracted using large language models (LLMs).

Modifiers	Transcription factor	Target	Relation type	Text	PMID
Increased	AR	KLK3	Binding	In addition, when nuclear AR from forskolin-treated LNCaP cells was incubated with oligonucleotides encoding an androgen response element of the PSA promoter and examined by electromobility shift assay, an increase in AR-androgen response element complex formation was observed.	10075669
Increased	ATF1	IL10	Neutral_Regulation	Our data suggest that intracellular cAMP may directly affect expression of the immunoregulatory cytokine IL-10 in monocytic cells via activation of the eukaryotic transcription factors CREB-1 and ATF-1 and their binding to CRE1 and CRE4 in the upstream enhancer of the IL-10 promoter.	10540320
Increased	ATF1	IL10	Binding	Our data suggest that intracellular cAMP may directly affect expression of the immunoregulatory cytokine IL-10 in monocytic cells via activation of the eukaryotic transcription factors CREB-1 and ATF-1 and their binding to CRE1 and CRE4 in the upstream enhancer of the IL-10 promoter.	10540320
—	ATF2	TNF	Negative_Regulation	Forced expression of c-Jun in ATF2-expressing melanoma cells restored TNFalpha expression, suggesting that both forms of ATF2 sequestered transcription factors that positively regulate TNFalpha expression in response to UV irradiation.	10318823
—	ATF2	TNF	Negative_Regulation	Taken together, our findings suggest that ATF2 contributes to UVC-induced apoptosis through transcriptional silencing of TNFalpha, which balances Fas-mediated cell death in melanoma.	10318823
—	ATF2	CCND1	Binding	We show that ATF-2 homodimers and ATF-2/c-Jun heterodimers, but not c-Jun homodimers, were able to bind the CRE of the cyclin D1 promoter.	10500157
—	ATF2	JUN	Binding	Supershift assays revealed that, during the early induction phase of the c-jun gene, the proximal activator protein-1 (jun1) site of the c-jun promoter was mainly occupied by the constitutively expressed activating transcription factor-2, whereas the late induction correlated with the predominant binding of c-Jun and, to a lesser extent, activating transcription factor-2 to the distal activator protein-1 (jun2) site.	10366004
—	CDX2	PAX6	Positive_Regulation	Further, we found that the presence of cdx-2 enhanced the interaction of pax-6 with p300, thus establishing a molecular complex of transcription factors implicated in tissue-specific glucagon gene expression with the basal transcriptional machinery.	10506141
—	CDX2	GCG	Neutral_Regulation	Further, we found that the presence of cdx-2 enhanced the interaction of pax-6 with p300, thus establishing a molecular complex of transcription factors implicated in tissue-specific glucagon gene expression with the basal transcriptional machinery.	10506141
—	CEBPA	PRL	Positive_Regulation	Expression of C/EBP alpha increased basal PRL gene expression almost 6-fold, whereas expression of Chop10 that can act as an inhibitor of C/EBP alpha reduced the basal activity of the PRL promoter 50-75%.	10499509
—	CEBPA	PRL	Neutral_Regulation	These data suggest that C/EBP alpha is an important transcription factor that regulates PRL gene expression.	10499509
Increased	CEBPA	CEBPB	Negative_Regulation	Interestingly, the C/EBPbeta promoter is not activated by C/EBPalpha, another C/EBP family member highly expressed in myelomonocytic cells, indicating that the autoregulation is specific for C/EBPbeta.	10542314
—	CEBPA	ALB	Binding	The binding of C/EBPalpha or -beta to its cognate sequence in the albumin promoter is not inhibited by p53 expression.	10542249
—	CEBPA	Krt10	Positive_Regulation	In cultured cells, C/EBPalpha and C/EBPbeta are each capable of activating the k10 promoter via three binding sites, identified by site-directed mutagenesis.	10588870
—	CEBPA	Krt10	Binding	In cultured cells, C/EBPalpha and C/EBPbeta are each capable of activating the k10 promoter via three binding sites, identified by site-directed mutagenesis.	10588870

Figure 2. Web interface showing a table of annotated TF–TG relationships, including associated TF modifiers, source sentences, and PMIDs where the relationships were identified.

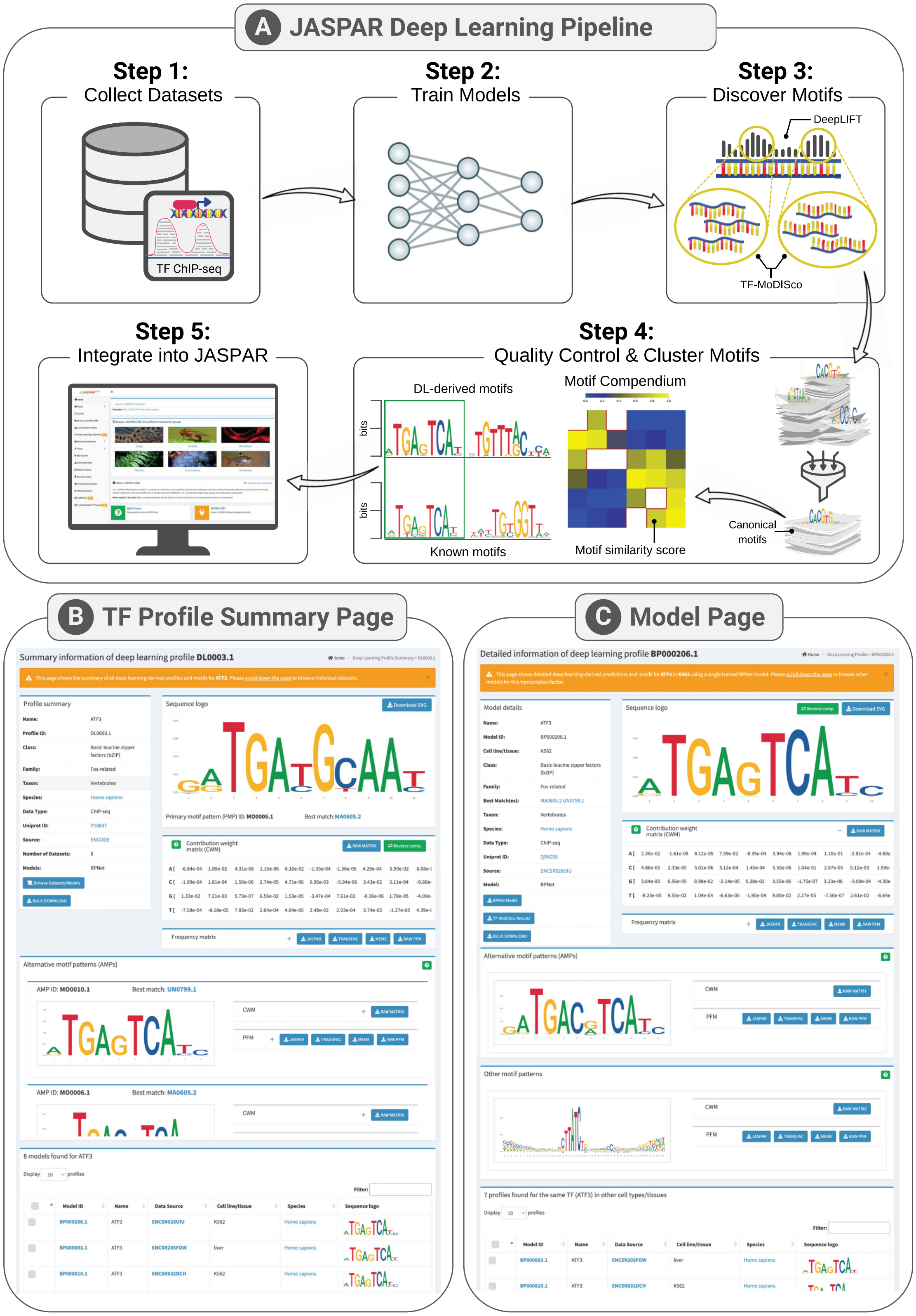
pattern (PMP) and the alternative motif patterns (AMPs) (see Supplementary Text), culminating in 353 motif patterns with their best JASPAR 2026 PFM match (Fig. 4) for the 240 TFs (from 1 to 7 motif patterns per TF, Supplementary Fig. S3). All PMPs and AMPs are labeled as MOXXXXXX.Y, where XXXXXX is the motif pattern ID and Y is the version number, as above. The summary profiles are linked to 1259 BPNet models, each specific to an individual ChIP-seq dataset (Supplementary Table S3). The specific BPNet models are labeled BPXXXXXX.Y, where XXXXXX is the model ID and Y is the version number, as above (Fig. 3C). Similarly to the summary profiles, we provide all motif patterns revealed by TF-MoDISco and clustered by MotifCompendium for each dataset. All trained BPnet models are available to users, and we provide the TF binding profiles as PFMs and contribution weight matrices (CWMs), which are similar to PFMs but capture contribution scores to the model prediction aggregated across sequences.

We provide all the profiles and models as part of the JASPAR Deep Learning (DL) collection, which can be accessed through the left sidebar of the JASPAR website. This opens an interactive interface that displays a searchable list of TF binding profiles, along with advanced filtering options. Users can switch to a different view to explore the list of deep learning models. We provide a functionality to select models to scan their motifs against input sequences using the *tangermeme* framework [37]. Each TF summary profile has a dedicated page presenting metadata, a visualization of the PMPs and AMPs, the corresponding CWMs and PFMs, and a list of all models trained to predict the summary profile (Fig. 3B). Complementarily, we provide each BPnet model on a dedicated page, which includes metadata, the PMP, AMPs, and other motifs found by TF-MoDISco with logo visualization and matrix representations, and links to download the model and TF-MoDISco results for further analysis (Fig. 3C).

JASPAR-associated tools

Incorporation of the regulatory sequence and motif simulation tool, inMOTIFin

When developing or evaluating computational methods to investigate the *cis*-regulatory grammar of genomes, it is critical to consider simulated data where the ground truth is known. To support such tasks, we developed inMOTIFin, a lightweight PFM and regulatory sequence simulation package [28]. In a nutshell, inMOTIFin can create PFMs with user-defined characteristics and generate DNA sequences with specific regulatory rules. To generate regulatory sequences, inMOTIFin inserts TFBSs in user-provided or random DNA sequences under flexible rules such as their positions, co-occurrences, and spacing. These features enable the design of synthetic data for benchmarking, analysis of TF binding cooperativity, and interpretation of computational models. We made it available on the JASPAR website to allow direct insertion of selected TF binding profiles. The user can set the number and length of random sequences to be generated or provide background sequences by uploading a FASTA file. When used from the website, each selected TF binding profile is inserted into the center of a randomly selected background by default. The user can also set the total number of motif instances per sequence. The output can be downloaded as a FASTA file, where the sequence headers include the background sequence and motif IDs, and a supporting BED file detailing the instances and positions of TFBS insertions. The package can also be installed locally and operated either through a command line or a Python interface (https://inmotifin.readthedocs.io/en/latest/, https://bitbucket.org/CBGR/inmotifin/src/main/). The standalone package has various additional features for tighter control over sequence simulations and modification. For seamless integration with JASPAR, the package supports the direct import of JASPAR profiles via the pyJASPAR module, given a user-provided set of matrix IDs.



Summary information of deep learning profile **DL0003.1**

Home > Deep Learning Profile Summary > DL0003.1

This page shows the summary of all deep learning-derived profiles and motifs for **ATF3**. Please scroll down the page to browse individual datasets.

Profile summary

Name:	ATF3
Profile ID:	DL0003.1
Class:	Basic leucine zipper factors (bZIP)
Family:	Fos-related
Taxon:	Vertebrates
Species:	Homo sapiens
Data Type:	ChIP-seq
Uniprot ID:	P18847
Source:	ENCODE
Number of Datasets:	8
Models:	BPNet

Primary motif pattern (PMP) ID: **MO0005.1** Best match: MA0605.2

Contribution weight matrix (CWM)

```
A [  -6.84e-04  1.89e-02  4.31e-06  1.15e-06  6.10e-02  -1.35e-04  -1.36e-05  4.29e-04  5.90e-02  6.08e-(
C [  -1.99e-04  1.81e-04  1.50e-06  2.74e-05  4.71e-06  6.95e-03  -5.94e-06  3.43e-02  3.11e-04  -9.80e-
G [  1.33e-02  7.21e-03  5.73e-07  6.56e-02  1.53e-05  -3.47e-04  7.61e-02  -8.36e-06  1.78e-05  -4.09e-
T [  -7.58e-04  -8.18e-05  7.83e-02  2.64e-04  4.64e-05  2.48e-02  2.03e-04  3.74e-03  -1.27e-05  4.39e-(
```

Frequency matrix

Alternative motif patterns (AMPs)

AMP ID: **MO0010.1** Best match: **UN0799.1**

AMP ID: **MO0006.1** Best match: **MA0605.2**

8 models found for ATF3

Display 10 profiles

Filter:

	Model ID	Name	Data Source	Cell line/tissue	Species	Sequence logo
□	BP000206.1	ATF3	ENCSR028UIU	K562	Homo sapiens	TGAGTCA
□	BP000693.1	ATF3	ENCSR205FOW	liver	Homo sapiens	TGAGTCA
□	BP000810.1	ATF3	ENCSR632DCH	K562	Homo sapiens	TGAGTCA

Detailed information of deep learning profile **BP000206.1**

Home > Deep Learning Profile > BP000206.1

This page shows detailed deep learning-derived predictions and motifs for **ATF3** in **K562** using a single trained BPNet model. Please scroll down the page to browse other models for this transcription factor.

Model details

Name:	ATF3
Model ID:	BP000206.1
Cell line/tissue:	K562
Class:	Basic leucine zipper factors (bZIP)
Family:	Fos-related
Best Match(es):	MA0605.2 UN0799.1
Taxon:	Vertebrates
Species:	Homo sapiens
Data Type:	ChIP-seq
Uniprot ID:	Q9XZS8
Source:	ENCSR028UIU
Model:	BPNet

Contribution weight matrix (CWM)

```
A [  2.35e-02  -1.61e-05  8.12e-05  7.59e-02  -8.35e-04  3.94e-06  1.99e-04  1.10e-01  -2.81e-04  -4.60e
C [  4.86e-05  2.33e-05  5.02e-06  3.12e-04  1.45e-04  5.55e-06  1.04e-01  2.67e-05  5.12e-03  1.59e-
G [  3.84e-03  6.56e-05  8.98e-02  -2.14e-05  5.28e-02  3.55e-06  -1.75e-07  3.23e-06  -3.00e-04  -4.30e
T [  -8.23e-05  9.55e-02  1.04e-04  -6.63e-05  -1.90e-04  9.80e-02  2.27e-05  -7.50e-07  2.61e-02  -6.64e
```

Frequency matrix

Alternative motif patterns (AMPs)

Other motif patterns

7 profiles found for the same TF (ATF3) in other cell types/tissues

Display 10 profiles

Filter:

	Model ID	Name	Data Source	Cell line/tissue	Species	Sequence logo
□	BP000693.1	ATF3	ENCSR205FOW	liver	Homo sapiens	TGAGTCA
□	BP000810.1	ATF3	ENCSR632DCH	K562	Homo sapiens	

Figure 3. JASPAR 2026 introduces a deep learning collection. The top panel illustrates the comprehensive workflow. The bottom panels present screenshots of the TF summary profile page (left) and the model page (right).

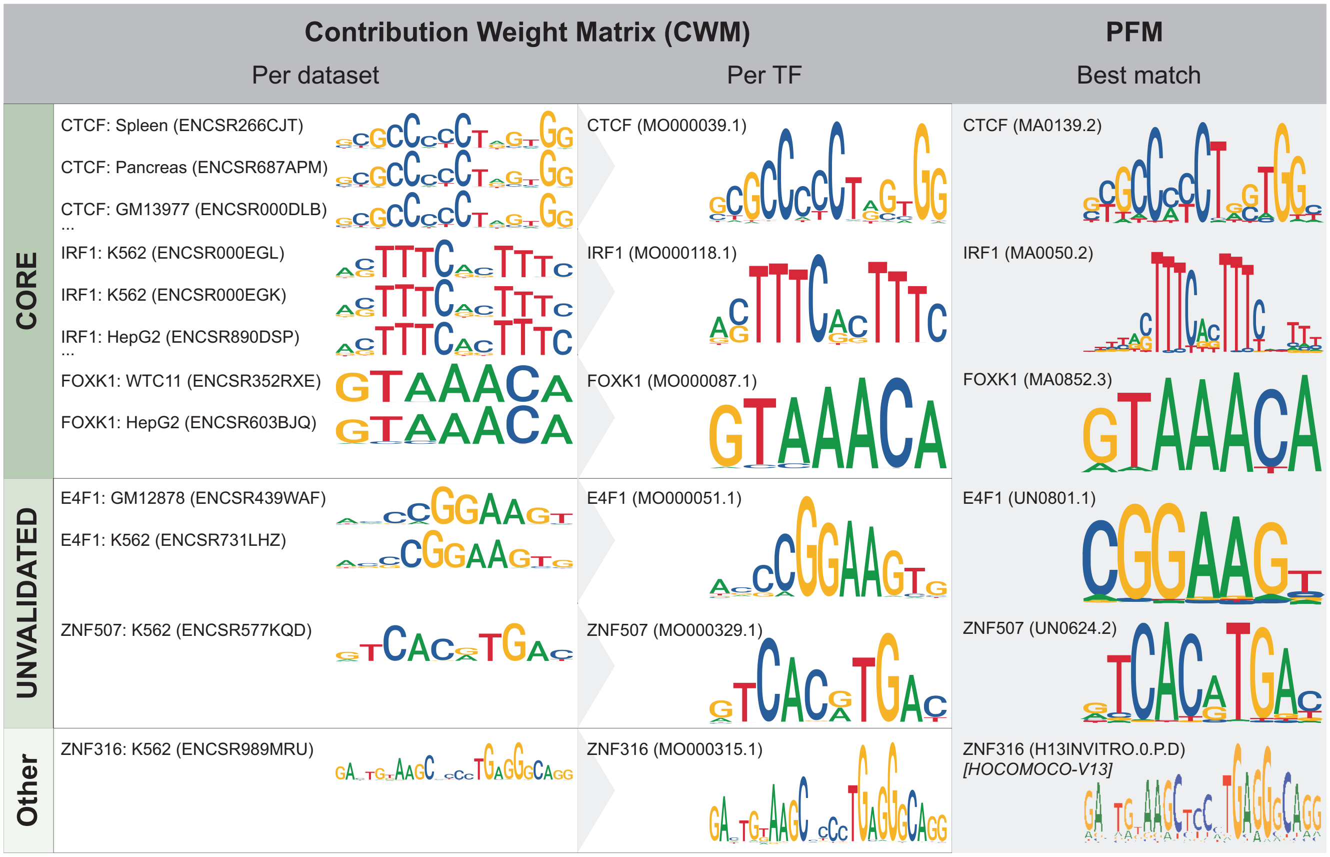


Figure 4. Examples of DL motif discovery, aggregation, and matching for CTCF, IRF1, FOXK1, E4F1, ZNF507, and ZNF316. (Left) Each CWM is derived from a BPNet model, each trained on an ENCODE TF ChIP-seq dataset, and revealed using TF-MoDISco. (Center) Individual CWMs are clustered and combined into aggregate CWMs per TF, using MotifCompendium. (Right) The closest matching PFM from JASPAR 2026, or other *in vitro*-derived PFM, was identified using MotifCompendium. (Top) CWMs that matched with a JASPAR 2026 CORE PFM. (Middle) CWMs that matched with a JASPAR 2026 UNVALIDATED PFM. (Bottom) CWMs that did not match with any JASPAR 2026 PFM, but matched with other *in vitro*-derived PFMs.

pyJASPAR and R/bioconductor packages for JASPAR data access

The 2026 release of the JASPAR database can be accessed through its web interface athttps://jaspar.elixir.no and its RESTful API (https://jaspar.elixir.no/api/) [38].In addition, we maintain and update the pyJASPAR Python package (https://github.com/asntech/pyjaspar; https://doi.org/10.5281/zenodo.4485856) [39] and the JASPAR R/Bioconductor data package, which includes the new JASPAR2026 release (https://github.com/da-bar/JASPAR). They have both been updated to fetch the most up-to-date JASPAR data. The R/Bioconductor package now stores the latest and the two previous JASPAR releases. All JASPAR versions included in the package are now accessible through the AnnotationHub [40]. Additionally, pyJASPAR has been updated to enable the retrieval of models from the Deep Learning collection. This allows seamless integration of JASPAR data when working within Python or R.

Conclusions and perspectives

As we present the 11th update of the JASPAR database, we expanded the JASPAR CORE collection by 12% (306 added or upgraded profiles; Table 1, Fig. 1). We have manually curated 11 565 profiles, derived from various databases and publications (Supplementary Table S1). It is becoming increasingly challenging to get big data sources to prepare motifs for manual curation, as we have nearly exhausted primary resources, such as GTRD or CIS-BP. The Codebook & GRECO-BIT consortium's [31] most recent effort focused on less studied TFs and their interactions with the DNA. This significant effort provided an important data source for this JASPAR release with 2425 motifs derived from ChIP-seq, PBM, SMiLE-seq, HT-SELEX, and GHT-SELEX, which we considered for manual curation. Many TFs were assayed using all or multiple of these techniques. Therefore, we were able to use this resource to provide orthogonal support for many motifs across technologies. Efforts, such as Codebook & GRECO-BIT, are essential for the community and have enabled us to enrich our JASPAR collections, especially with the motifs of less studied TFs. This puts us a step forward toward having a complete list of curated profiles for human TFs.

We continued our efforts to add high-quality profiles, even if we were unable to find orthogonal support. The UNVALIDATED collection expanded by 60% (433 profiles added; Supplementary Fig. S1 and Supplementary Table S2). Notably, the majority of these profiles were CAP-SELEX-derived PFMs for TFs binding DNA as dimers [32]. We were able to find orthogonal confirmation for only a few of these dimer motifs, indicating the lack of literature and data investigating cooperativity between TFs with dedicated assays. We hope that, in the future, more such studies will be published to delve deeper into the more complex regulatory mechanisms that better capture how TFs cooperate to regulate transcription. Deep learning-based models will provide a paradigm shift in modeling TF cooperativity [25].

When preparing a new release of the JASPAR PFM collections, we aim to refine the already existing profiles. For the current release, we systematically revised 638 profiles in the UNVALIDATED collection and promoted 41 of them to the CORE collection. Finding orthogonal support in the literature remains a manual effort that can be time-consuming. The emerging development of large language models (LLMs) will likely ease and automate this process for curators in the near future. With JASPAR 2026, we introduced LLM-extracted TF targets, exemplifying how such tools can parse the scientific literature to extract deeper insights into complex transcriptional regulation by TFs.

With this update, we are pleased to fulfill a long-awaited promise and launch the JASPAR Deep Learning (DL) Collection. This release features 353 primary and alternative motif patterns summarizing the binding properties captured by deep learning models for 240 TFs. Importantly, JASPAR also provides the underlying 1259 BPNet models that were trained on specific ChIP-seq datasets. As such, the JASPAR DL collection provides state-of-the-art models predicting TF binding ChIP-seq signals at base pair resolution from DNA sequences. It complements the PFMs by providing unprecedented means to decode the *cis*-regulatory code of genomes. Nevertheless, we recognize that the binding profiles extracted from the trained models were validated by matching them with existing PFMs, thereby strengthening the importance of establishing high-quality resources that store TF binding profiles as PFMs. Indeed, high-quality, manually curated resources play a key role in training large models and should be maintained alongside more expressive representations [41, 42].

It is expected that interpreting deep learning models will leverage important characteristics such as the surrounding genomic context, TF cooperativity, and nucleosome positioning, which cannot be achieved by traditional motif discovery approaches [43–45]. Moreover, DL models can infer the impact of sequence variations on binding affinities through *in silico* mutagenesis. This approach has the potential to further our understanding of the molecular mechanisms driving diseases. With emerging tools like *tangermeme* and *ledidi* [37, 46], the community now has access to an ecosystem enabling the large-scale use of DL models to perform a multitude of tasks, such as interpreting the models, performing *in silico* mutagenesis, predicting TFBSs through hit calling, and generating regulatory sequences.

Integrating deep learning models into JASPAR opens new opportunities for future research endeavors. In this release, we have begun integrating models trained on *Homo sapiens* ChIP-seq datasets, providing a robust foundation for understanding human TF binding. We plan to expand the DL collection in the future by including other organisms and taxa, enhancing the depth and diversity of this collection.

Concurrently, the scientific community has increasingly focused on making deep learning models more accessible by collaborating worldwide on open-source projects. This update to JASPAR supports this community-driven initiative by integrating BPNet models directly into the database and providing seamless and direct access to them.

From its initial release, JASPAR has consistently provided the community with a high-quality, user-friendly resource that promotes open science. In summarizing this latest database update, we emphasize our unwavering commitment to evolving in tandem with technological advancements and scientific discoveries in the field. While expanding our data coverage and embracing new methodologies, we ensure the high quality, non-redundancy, and ease of use of the JASPAR database. This update reinforces our ongoing commitment to providing a resource at the forefront of transcription factor binding research.

Acknowledgements

To prepare each JASPAR release, we mine and process publicly available data. As proud "research parasites" [47], we would like to thank all researchers for making this data available. We thank the IT team at NCMBM for their IT support, the members of the Kuijjer and Mathelier groups for their insightful discussions, and all of ELIXIR Norway, Rein Aasland, and Trine Ballestad Rounge for their support in the development. We also acknowledge the Norwegian Research and Education Cloud for hosting the web interface.

Author contributions: We follow the Contributor Roles Taxonomy (CRediT) [48]. Conceptualisation: D.O.B., I.R., A.S., B.L., W.W.W., K.V., F.P., A.Ku., A.M.; Data curation: D.O.B., I.R., D.R.A., R.B.M., I.B., H.D.B., V.H., A.J., V.K., R.B.L., J.L., M.P., V.R., S.S.D., C.M.Y., J.C.M.; Formal analysis: D.O.B., I.R. K.F., R.B.L., V.R., S.S.D., C.M.Y., A.A.; Funding acquisition: A.Kh., K.V., M.L.K., W.W.W., F.P., A.Ku., A.M.; Investigation: D.O.B., I.R., D.R.A., R.B.M., I.B., H.D.B., V.H., A.J., V.K., R.B.L., J.L., M.P., V.R., S.S.D., C.M.Y., A.A., J.C.M.; Methodology: D.O.B., I.R. K.F., R.B.L., V.R., S.S.D., C.M.Y.; Project administration: I.R., W.W.W., K.V., F.P., A.Ku., A.M.; Resources: E.H., W.W.W., K.V., F.P., A.Ku., A.M.; Software: I.R., K.F., D.O.B., A.Kh., C.M.Y., V.R., R.B.L.; Supervision: E.H., B.L., A.S., K.V., M.L.K., W.W.W., J.L., A.Ku., A.M.; Visualisation: D.O.B., I.R., J.C.M., C.M.Y, V.R.; Writing—original draft: D.O.B., I.R., H.D.B., K.F., R.B.L., A.M.; Writing—review and editing: D.O.B., I.R., D.R.A., R.B.M., I.B., H.D.B., K.F., A.J., V.K., R.B.L., J.L., M.P., C.M.Y., V.R., S.S.D., A.P., G.K.M., A.T.W., A.A., J.C.M., D.B., J.C., S.G., M.J., A.Kh., E.H., B.L., A.S., K.V., W.W.W., F.P., A.Ku., A.M.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

A. Kundaje is on the scientific advisory board of SerImmune, TensorBio, is a consultant with Arcardia Science, Inari, Precede Biosciences, Bristol Myers Squibb, and has a financial stake in DeepGenomics, Illumina, Immunai, SerImmune, TensorBio, and Freenome.

Funding

This work is supported by funding from the Research Council of Norway [187615], the Helse Sør-Øst, and the University of Oslo through the Norwegian Centre for Molecular Biosciences and Medicine (NCMBM; formerly NCMM) [to Mathelier and Kuijjer groups], the Norwegian Cancer Society [215027, 272930 to Mathelier group], the Norwegian Cancer Society [273592 to M.L.K. and I.B.], Familien Blix Fond [to M.L.K. and I.B.], the UiO:Life Science Internationalization support [to D.O.B.], the GRAL Labex financed within the University Grenoble Alpes graduate school (Ecoles Universitaires de Recherche) CBH-EUR-GS(ANR-17-EURE-0003), the ANR-23-CE20-0027 TRANSINET, and ANR-22-PEAE-0005 PEPR Agroécologie et Numérique [to Parcy group], the Research Council of Norway [322392 to ELIXIR Norway], the startup fund from Mohamed bin Zayed University of Artificial Intelligence [848087 to A.Kh.], the NIH grants U24HG012343 and U01HG012069 as well as compute support from the Stanford Sherlock Cluster [Kundaje group], the Stanford BioX fellowship [to A.T.W.], the European Union: NextGenerationEU (NPOO.C3.2.R2-I1.06.0060) [to DB], the Novo Nordisk Foundation [NNF20OC0059951, to A.S.] and the Danish cancer society [R325-A18868, to A.S.]. Funding to pay the Open Access publication charges for this article was provided by Research Council of Norway [187615].

Data availability

JASPAR is an open-access database available at https://jaspar.elixir.no/.

References

1. Lambert SA, Jolma A, Campitelli LF *et al.* The human transcription factors. *Cell* 2018;172:650–65. https://doi.org/10.1016/j.cell.2018.01.029
2. Nie Y, Shu C, Sun X. Cooperative binding of transcription factors in the human genome. *Genomics* 2020;112:3427–34. https://doi.org/10.1016/j.ygeno.2020.06.029
3. Kribelbauer JF, Loker RE, Feng S *et al.* Context-dependent gene regulation by homeodomain transcription factor complexes revealed by shape-readout deficient proteins. *Mol Cell* 2020;78:152–67. https://doi.org/10.1016/j.molcel.2020.01.027
4. Yan J, Enge M, Whitington T *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* 2013;154:801–13. https://doi.org/10.1016/j.cell.2013.07.034
5. Imrichová H, Hulselmans G, Atak ZK *et al.* i-cisTarget 2015 update: generalized *cis*-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Res* 2015;43:W57–64. https://doi.org/10.1093/nar/gkv395
6. Fornes O, Gheorghe M, Richmond PA *et al.* MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations. *Sci Data* 2018;5:180141. https://doi.org/10.1038/sdata.2018.141
7. Fu Y, Liu Z, Lou S *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* 2014;15:480. https://doi.org/10.1186/s13059-014-0480-5
8. Weirauch MT, Yang A, Albu M *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 2014;158:1431–43. https://doi.org/10.1016/j.cell.2014.08.009
9. Kulakovskiy IV, Vorontsov IE, Yevshin IS *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis. *Nucleic Acids Res* 2018;46:D252–9. https://doi.org/10.1093/nar/gkx1106
10. Vorontsov IE, Eliseeva IA, Zinkevich A *et al.* HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res* 2024;52:D154–63. https://doi.org/10.1093/nar/gkad1077
11. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R *et al.* JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2024;52:D174–82. https://doi.org/10.1093/nar/gkad1059
12. Sandelin A, Alkema W, Engström P *et al.* JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 2004;32:D91–4. https://doi.org/10.1093/nar/gkh012
13. Mathelier A, Wasserman WW. The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.* 2013;9:e1003214. https://doi.org/10.1371/journal.pcbi.1003214
14. Eggeling R, Roos T, Myllymäki P *et al.* Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics* 2015;16:375. https://doi.org/10.1186/s12859-015-0797-4
15. Grau J, Ben-Gal I, Posch S *et al.* VOMBAT: prediction of transcription factor binding sites using variable order bayesian trees. *Nucleic Acids Res* 2006;34:W529–33. https://doi.org/10.1093/nar/gkl212
16. Ghandi M, Lee D, Mohammad-Noori M *et al.* Enhanced regulatory sequence prediction using gapped *k*-mer features. *PLoS Comput. Biol* 2014;10:e1003711. https://doi.org/10.1371/journal.pcbi.1003711
17. Mathelier A, Xin B, Chiu T-P *et al.* DNA shape features improve transcription factor binding site predictions *in vivo*. *Cell Syst* 2016;3:278–86. https://doi.org/10.1016/j.cels.2016.07.001
18. Eraslan G, Avsec Ž, Gagneur J *et al.* Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019;20:389–403. https://doi.org/10.1038/s41576-019-0122-6
19. Chen L, Capra JA. Learning and interpreting the gene regulatory grammar in a deep learning framework. *PLoS Comput Biol* 2020;16:e1008334. https://doi.org/10.1371/journal.pcbi.1008334
20. Novakovsky G, Dexter N, Libbrecht MW *et al.* Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat Rev Genet* 2023;24:125–37. https://doi.org/10.1038/s41576-022-00532-2
21. Zeitlinger J, Roy S, Ay F *et al.* Perspective on recent developments and challenges in regulatory and systems genomics. *Bioinform Adv* 2025;5:vbaf106. https://doi.org/10.1093/bioadv/vbaf106
22. Perkel JM. Beyond AlphaFold: how AI is decoding the grammar of the genome. *Nature* 2025;644:829–32. https://doi.org/10.1038/d41586-025-02621-8
23. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. arXiv, https://arxiv.org/abs/1704.02685, 12 Oct 2019, preprint: not peer reviewed
24. Shrikumar A, Tian K, Avsec Ž *et al.* Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5, arXiv, https://arxiv.org/abs/1811.00416, 30 Apr 2020, preprint: not peer reviewed
25. Avsec Ž, Weilert M, Shrikumar A *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* 2021;53:354–66. https://doi.org/10.1038/s41588-021-00782-6
26. Luo Y, Hitz BC, Gabdank I *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* 2020;48:D882–9. https://doi.org/10.1093/nar/gkz1062
27. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74. https://doi.org/10.1038/nature11247
28. Ferenc K, Martini L, Rauluseviciute I *et al.* inMOTIFin: a lightweight end-to-end simulation software for regulatory sequences. 2025. https://arxiv.org/abs/2506.20769
29. Kudron MM, Victorsen A, Gevirtzman L *et al.* The ModERN resource: genome-wide binding profiles for hundreds of *Drosophila* and *Caenorhabditis elegans* transcription factors. *Genetics* 2018;208:937–49. https://doi.org/10.1534/genetics.117.300657
30. Lambert SA, Yang AWH, Sasse A *et al.* Similarity regression predicts evolution of transcription factor sequence specificity. *Nat Genet* 2019;51:981–9. https://doi.org/10.1038/s41588-019-0411-1

31. Jolma A, Laverty KU, Fathi A *et al.* bioRxiv, https://doi.org/10.1101/2024.11.11.622097, 12 November 2024, preprint: not peer reviewed
32. Xie Z, Sokolov I, Osmala M *et al.* DNA-guided transcription factor interactions extend human gene regulatory code. *Nature* 2025;641:1329–38. https://doi.org/10.1038/s41586-025-08844-z
33. de Tribolet-Hardy J, Thorball CW, Forey R *et al.* Genetic features and genomic targets of human KRAB-zinc finger proteins. *Genome Res* 2023;33:1409–23. https://doi.org/10.1101/gr.277722.123
34. Navarro Gonzalez J, Zweig AS, Speir ML *et al.* The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res* 2021;49:D1046–57. https://doi.org/10.1093/nar/gkaa1070
35. Vazquez M, Krallinger M, Leitner F *et al.* ExTRI: extraction of transcription regulation interactions from literature. *Biochim Biophys Acta Gene Regul Mech* 2022;1865:194778. https://doi.org/10.1016/j.bbagrm.2021.194778
36. Deshpande S, Yun CM, Ramalingam V *et al.* A unified lexicon of predictive DNA sequence motifs from ENCODE transcription factor binding and chromatin accessibility assays. 2025. https://zenodo.org/records/17179111
37. Schreiber J. tangermeme: a toolkit for understanding *cis*-regulatory logic using deep learning models. bioRxiv, https://doi.org/10.1101/2025.08.08.669296, 12 August 2025, preprint: not peer reviewed
38. Khan A, Mathelier A. JASPAR RESTful API: accessing JASPAR data from any programming language. *Bioinformatics* 2018;34:1612–4. https://doi.org/10.1093/bioinformatics/btx804
39. Khan A. pyJASPAR: a Pythonic interface to JASPAR transcription factor motifs. *Zenodo*, 2024. https://zenodo.org/records/17416190
40. Morgan M, Shepherd L. AnnotationHub. 2025. https://bioconductor.org/packages/3.9/bioc/html/AnnotationHub.html
41. Shumailov I, Shumaylov Z, Zhao Y *et al.* AI models collapse when trained on recursively generated data. *Nature* 2024;631:755–9. https://doi.org/10.1038/s41586-024-07566-y
42. Stroe O. *Case study: AlphaFold uses open data and AI to discover the 3D protein universe*. *European Molecular Biology Laboratory*, 2023.
43. Antikainen AA, Heinonen M, Lähdesmäki H. Modeling binding specificities of transcription factor pairs with random forests. *BMC Bioinformatics* 2022;23:212. https://doi.org/10.1186/s12859-022-04734-7
44. Luo H, Tang L, Zeng M *et al.* BertSNR: an interpretable deep learning framework for single-nucleotide resolution identification of transcription factor binding sites based on DNA language model. *Bioinformatics* 2024;40:btae461. https://doi.org/10.1093/bioinformatics/btae461
45. Novakovsky G, Fornes O, Saraswat M *et al.* ExplaiNN: interpretable and transparent neural networks for genomics. *Genome Biol* 2023;24:154. https://doi.org/10.1186/s13059-023-02985-y
46. Schreiber J, Lu YY, Noble WS. Ledidi: designing genomic edits that induce functional activity. bioRxiv, https://doi.org/10.1101/2020.05.21.109686, 25 May 2020, preprint: not peer reviewed
47. Longo DL, Drazen JM. Data sharing. *N Engl J Med* 2016;374:276–7. https://doi.org/10.1056/NEJMe1516564
48. Brand A, Allen L, Altman M *et al.* Beyond authorship: attribution, contribution, collaboration, and credit. *Learn Publ* 2015;28:151–5. https://doi.org/10.1087/20150211

Received: September 15, 2025. **Revised:** October 14, 2025. **Accepted:** October 15, 2025