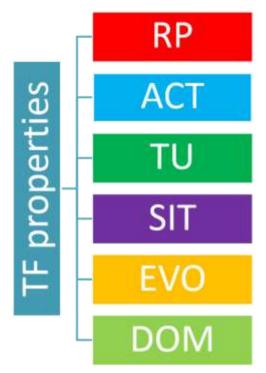
Supplementary material

Table S1. Specific relevant information associated to each TF property tagged in sentences of manual summaries

Property	Specific information manually tagged	Tag		
ACT	Growth condition in negative regulation	ACTCONDN		
	Growth condition in positive regulation	ACTCONDP		
	Effector	ACTEFFE		
	Regulation of the TF activity	ACTREG		
	Active conformation of the TF	ACTCONF		
EVO	Domain position and percentage of domain identity	EVPIDT		
	Percentage of TF identity with other TFs	EVPIT		
SIT	Symmetry	SSM		
	Size	SSZ		
TU	Regulation of the TU	TUR		
	Organization of the TU	TURO		
	Localization of the TU	TUL		



NagC, N-acetylglucosamine, transcriptional dual regulator participates in regulating <RP> phosphotransferase system (PTS) [18067539]. Its function is to coordinate biosynsis of amino sugars, D-glucosamine (GlcN) and N-acetylglucosamine (GlcNAc) with ir catabolism</RP> [1766379, 11361067, 11139621, 15066032].

<a href="<"><ACTEFFE> specific inducer for NagC is GlcNAc-6-P/ACTEFFE>, product of GlcNAc transport by PTS [1766379, 11361067]. NagC is displaced from its DNA targets by interacting with GlcNAc-6-P [1766379].

Based on structure of MIc DNA-binding transcriptional repressor, models for threedimensional structure of NagC and for binding of GlcNAc-6-P were developed [18067539].

Nag regulon consists of two divergent operons, nagE and nagBACD. nagC encodes repressor of this regulon. <TUL>Its localization within inducible nagBACD operon is unusual, since most transcriptional dual regulators are located near loci y control</TUL> [1766379].

NagC is present at low levels within bacterial cell [1766379].

consensus sequence for NagC-binding site has been determined via combination of native site comparisons, operator mutagenesis, and selection of DNA-binding fragments, [11139621]. dual function of NagC requires it to bind simultaneously to two operators in order to form a DNA loop [11361067]. <SSA> strictly conserved TT/AA motif, located in both ends of center of symmetry, is probably major recognition point for NagC</SSA> [11139621].

NagC is a <EVF>member of ROK (repressor, ORFs, kinases) (NagC/XyIR) family of proteins</EVF>, which contains at least two distinct classes of proteins, xylose repressor (XyIR) and a series of glucose/fructose kinases [7952186, 12374829]. NagC is similar to MIc transcriptional dual regulator [15916829]. However, y have different specificities for select binding sites [11139621].

<DOM>NagC has a helix-turn-helix motif in its N-terminal part [18067539].

Figure S1. Example of tagged manual summary with specific relevant information.

Table S2. Biological information automatically tagged in manual summaries to enrich features for supervised learning. This information was also used to tag sentences from articles of *E. coli* and *Salmonella*. For *E. coli*, we indicate the source, and for *Salmonella* we indicate if the source was the same, the new source, or if the tag was not used

Property	Biological information	Source for <i>E. coli</i>	Source for Salmonella	Tag
ACT	Dictionary of TFs Dictionary of growth conditions	RegulonDB RegulonDB	NO SÉ The same	ACTTF ACTCOND

DOM	Dictionary of effectors Keywords of effectors Dictionary of regulatory verbs Keywords of regulation Dictionary of conformations Keywords of conformations	RegulonDB Manually collected Manually collected Manually collected RegulonDB Manually collected	The same The same The same The same Not used The same The same	ACTEFFE ACTREG ACTREG ACTCONF ACTCONF DEAM
DOM	Dictionary of structural domain families Dictionary of molecular functions Dictionary of structural motifs Dictionary of TFs	DBD: Transcription Factor Prediction Database GO's OBO file Interpro RegulonDB	The same The same NO SÉ	MF DMOT TF
EVO	Keywords of structural domains Dictionary of evolutionary families Keywords of percentage of TF identity with other TFs and percentage of domain identity Dictionary of structural domain position	Frequent words RegulonDB Manually collected Manually collected	The same Not used The same The same	EVPI EVDOM
RP	Dictionary of biological processes Keywords of regulated processes	GO Frequent words	The same The same	PRO FWRP
SIT	Keywords of symmetry Keywords of size Spatial arrangement	Manually collected Manually collected Regular expression	The same The same The same	SSM SSZ SSA
TU	Dictionary of genes Dictionary of transcription units Keywords of organization of the TU Keywords of regulation of the TU Keywords of localization of the TU	RegulonDB RegulonDB Manually collected Manually collected Manually collected	NO SÉ NO SÉ The same The same The same	TURO TURO TURO TUR TUL

Table S3. A general description of the experimental setup, including the tested values of the different aspects employed for training the six classifiers

Aspect	Values			
Combination of features	lemma POS, lemma tag, tag for lemma			
Vectorizer	Binary, TF-IDF, TF-IDF binary			
N-grams	1, 1-2, 1-3			
Dimensionality reduction with SVD	300, 200 dimensions			
Feature selection with x^2	1000, 800, 500 features			
Under-sampling technique	RandomUS, Tomek, IHT, and OSS			
SVM kernel	rbf, lineal, poly			
Class weighting	true, false			

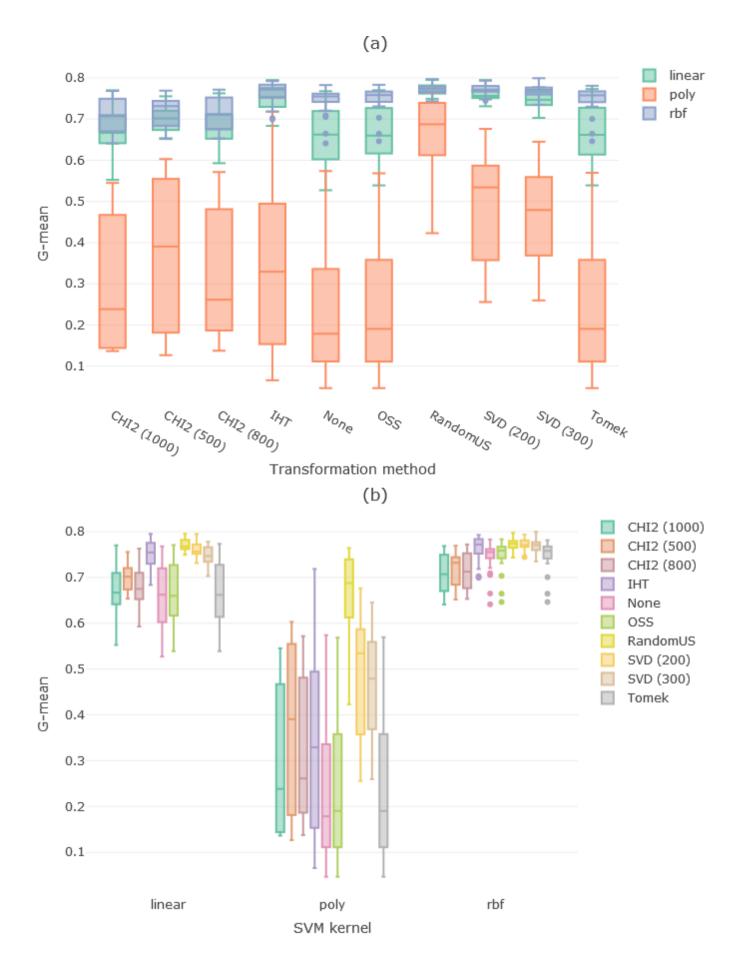


Figure S2. Distribution of performance in cross-validation of all trained predictive models for ACT property. (a) shows transformation methods in horizontal axis, CHI2 (χ^2) and SVD include number of final dimensions/features. (b) shows SVM kernel in horizontal axis.

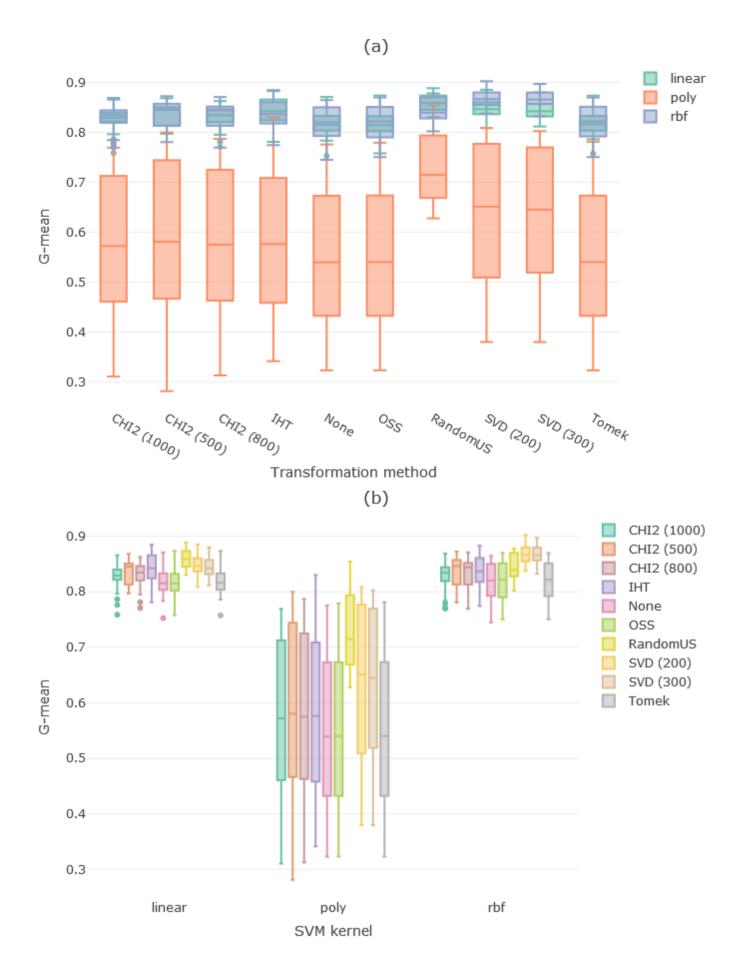


Figure S3. Distribution of performance in cross-validation of all trained predictive models for DOM property. (a) shows transformation methods in horizontal axis, CHI2 (χ^2) and SVD include number of final dimensions/features. (b) shows SVM kernel in horizontal axis.

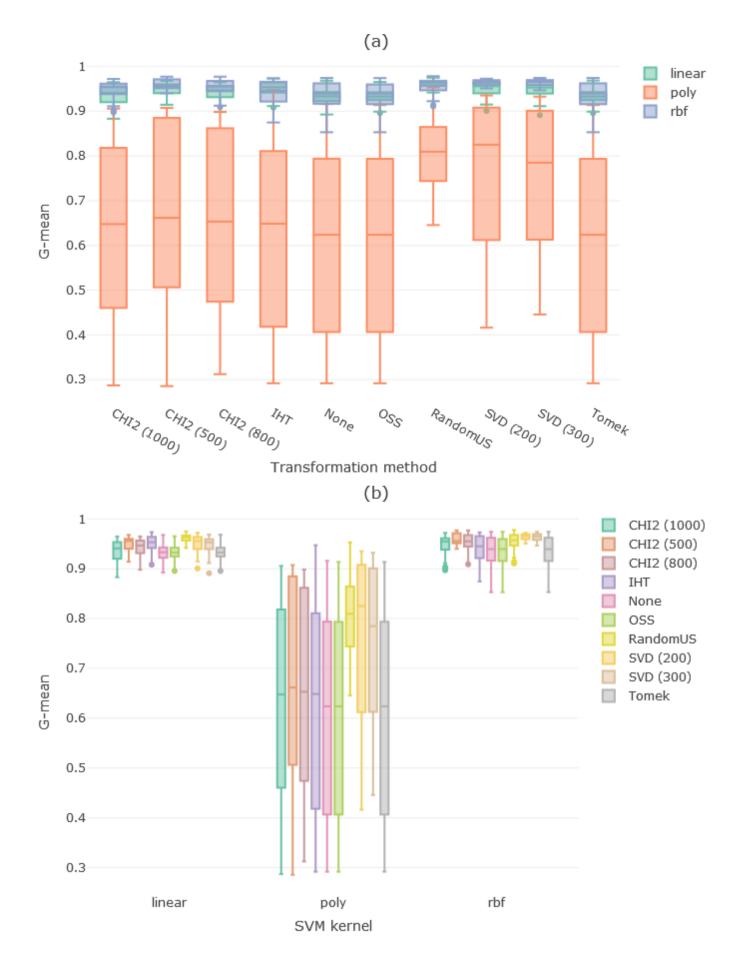


Figure S4. Distribution of performance in cross-validation of all trained predictive models for EVO property. (a) shows transformation methods in horizontal axis, CHI2 (χ^2) and SVD include number of final dimensions/features. (b) shows SVM kernel in horizontal axis.

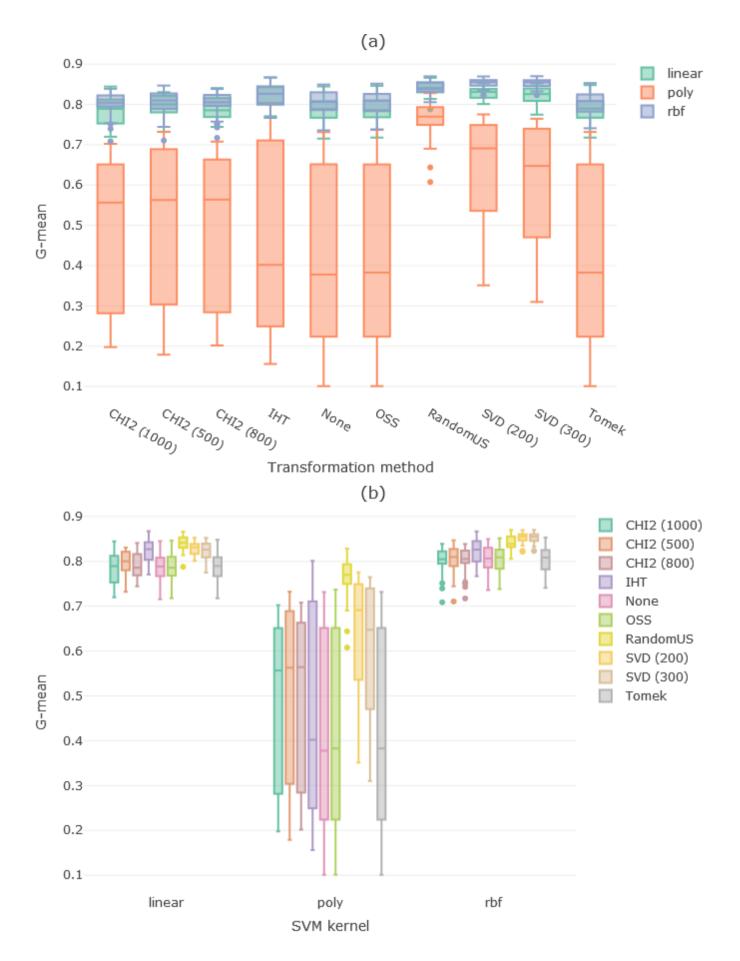


Figure S5. Distribution of performance in cross-validation of all trained predictive models for RP property. (a) shows transformation methods in horizontal axis, CHI2 (χ^2) and SVD include number of final dimensions/features. (b) shows SVM kernel in horizontal axis.

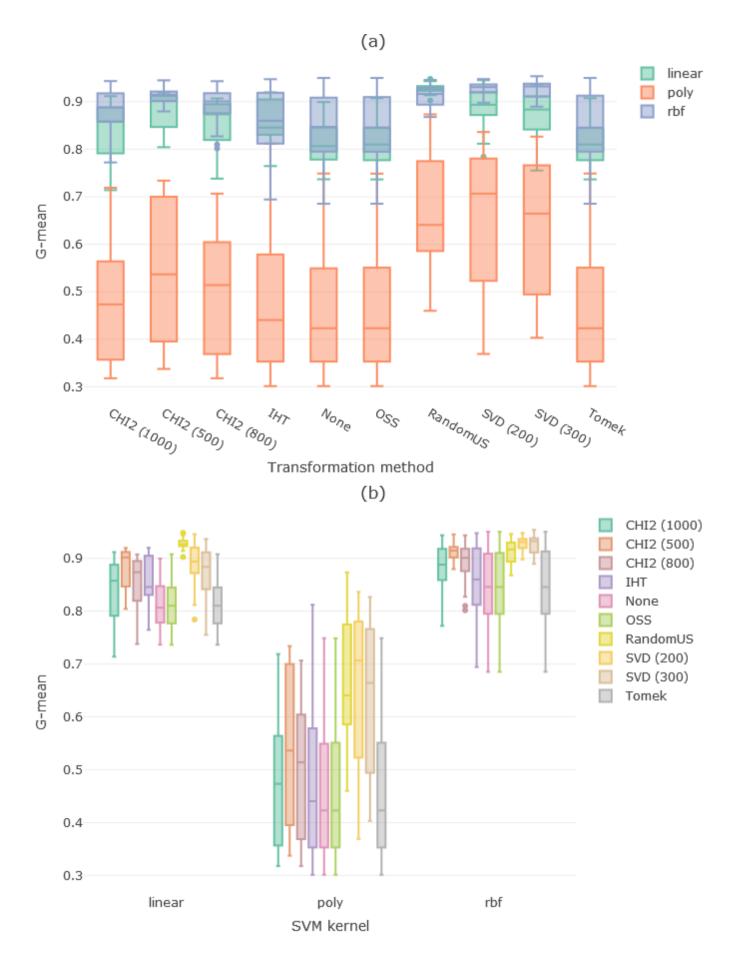


Figure S6. Distribution of performance in cross-validation of all trained predictive models for SIT property. (a) shows transformation methods in horizontal axis, CHI2 (χ^2) and SVD include number of final dimensions/features. (b) shows SVM kernel in horizontal axis.

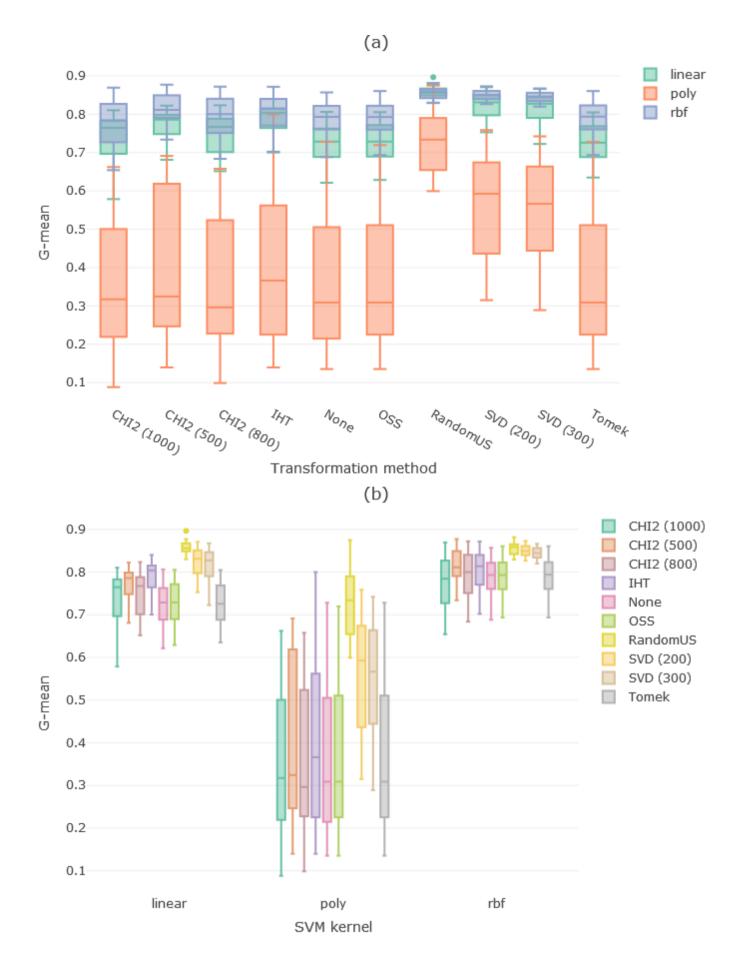


Figure S7. Distribution of performance in cross-validation of all trained predictive models for TU property. (a) shows transformation methods in horizontal axis, CHI2 (χ^2) and SVD include number of final dimensions/features. (b) shows SVM kernel in horizontal axis.

Table S4. Detailed characteristics of the best predictive model per TF property

Property	Vectorization N- grams		Final	SVM hyperparameters				G-mean	
			Transformation	dimensions	Kernel	С	Gamma	Class weight	score
ACT	TF-IDF	1,2	RandomUS		RBF	3.0	1.0	Balanced	0.80
DOM	TF-IDF	1	SVD	200	RBF	1.0	1.0	Balanced	0.90
EVO	TF-IDF binary	1	SVD	200	RBF	0.5	1.0	Balanced	0.97
RP	TF-IDF binary	1,2	SVD	200	RBF	1.0	1.0	Balanced	0.87
SIT	TF-IDF binary	1	SVD	200	RBF	0.5	1.0	Balanced	0.95
TU	TF-IDF binary	1	SVD	200	RBF	3.0	0.1	Balanced	0.87