Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Review

# Long-term evolution of regulatory DNA sequences. Part 1: simulations on global, biophysically-realistic genotype–phenotype maps

Elia Mascolo[1], Réka Borbély[1], Santiago Herrera-Álvarez[2], Calin C Guet[1], Justin Crocker[2] and Gašper Tkačik[1]

Promoters and enhancers are cis-regulatory elements (CREs), DNA sequences that bind transcription factor (TF) proteins to up- or down-regulate target genes. Decades-long efforts yielded TF-DNA interaction models that predict how strongly an individual TF binds arbitrary DNA sequences and how individual binding events on the CRE combine to affect gene expression. These insights can be synthesized into a global, biophysically realistic, and quantitative genotype–phenotype map for gene regulation, a 'holy grail' for the application of evolutionary theory. A global map provides a rare opportunity to simulate the long-term evolution of regulatory sequences and pose several fundamental questions: How long does it take to evolve CREs *de novo*? How many non-trivial regulatory functions exist in sequence space? How connected are they? For which regulatory architecture is CRE evolution most rapid and evolvable? In this article, the first of a two-part series, we briefly review the pertinent modeling and simulation efforts for a unique system that enables close, quantitative, and mechanistic links between biophysics, as well as systems, synthetic, and evolutionary biology.

**Addresses**
[1] Institute of Science and Technology Austria, Am Campus 1, Klosterneuburg AT-3400, Austria
[2] Developmental Biology Unit, European Molecular Biology Laboratory, Heidelberg DE-69117, Germany

Corresponding author: Tkačik, Gašper (Gasper.Tkacik@ist.ac.at)

**Current Opinion in Genetics & Development** 2026, **99**:102483

This review comes from a themed issue on **Genome Architecture and Expression**

Edited by **Luca Giorgetti** and **Daniel Jost**

Available online xxxx

https://doi.org/10.1016/j.gde.2026.102483



Much is understood about the power of evolutionary optimization. Assuming we know how the genetic program ('genotype') maps into observable organismal properties ('phenotype') and, subsequently, into fitness [1], a mathematically rigorous body of theory can predict evolutionary trajectories, diversity of outcomes, and adaptation rates across various population-genetic regimes [2]. Thus, the complete knowledge of a genotype–phenotype–fitness map implies a full knowledge of evolutionary change. In practice, however, our knowledge is remarkably incomplete: we know neither all relevant phenotypes nor how they map into fitness. Furthermore, the map typically depends on the environment in complex or unknown ways. Arguably, the biggest hurdle, however, is the jump from genotypes to phenotypes: the curse of dimensionality generally precludes assigning a phenotype to all $4^L$ possible genotypes even for modest sequence lengths ($L \gtrsim 10$ bp).

On the theory front, we typically circumvent this curse of dimensionality by adopting simplified models at the expense of biological realism. We study toy-model GP-fitness maps or 'fitness landscapes', where every possible genotype is mathematically assigned a fitness value. This exhaustive assignment to all $4^L$ possible genotypes defines a 'global map'. On such generic and stylized landscapes (often going by inventive or cryptic names such as the house-of-cards [3], Mount Fuji [4], NK [5,6], pairwise- or globally-epistatic landscapes [6–8], etc.), one can study global, arbitrarily long evolutionary trajectories. We refer to such approaches as addressing 'long-term evolution' if: evolution can start from any sequence (even fully random sequences, enabling simulation of *de novo* evolution); it can proceed indefinitely, with no imposed limit to how many mutations can be accumulated. However, using idealized maps sacrifices biological realism and a quantitative match to any real dataset: at best, we might capture generic properties of evolutionary dynamics; at worst, our toy model map may miss essential topological features of the real map, putting even qualitative predictions into doubt.

On the empirical front, in contrast, experiments are typically designed to measure putative fitness proxies using massive, controlled mutational libraries. For example, fluorescence has been measured for thousands of green fluorescent

protein mutants [9,10], yielding landscapes that enable guided design of new green fluorescent protein variants. Similarly, for transcriptional regulation, constitutive or regulated gene expression has been measured at scale using massively parallel assays [11,12]. In both cases, experiments provide a quantitative, system-specific map. The sacrifice here is the map's global nature: even the largest libraries probe only a vanishingly small fraction of the $4^L$ possible genotypes, confined to a local mutational neighborhood of the wild-type. We refer to such restricted assignments as 'local maps'. This limitation is not critical when considering point mutations over sufficiently short timescales. We refer to this regime as 'short-term evolution', characterized by two constraints: evolution starts from one or a few related sequences (typically the wild-type); it is limited to a handful of mutations ($\lesssim$10) exploring only the local mutational neighborhood around the initial sequences (the mutants for which experimental data is available). However, such approaches preclude more general predictions over arbitrarily long timescales (during which evolution may explore farther regions of genotype space) or modeling *de novo* evolution from non-functional sequences.

Genotype–phenotype (GP) maps that are both quantitative and global, and therefore uniquely suited to simulate long-term evolution, are few and far between. The established ones share a common theme: the phenotype depends on 'molecular recognition', the propensity of two molecules to interact with a strength set by their sequence-dependent structure. Because physicochemical laws strongly constrain the underlying interaction rules, the complexity of the GP map is drastically reduced. The prime example is the genetic code, where serial molecular recognition by tRNAs renders the GP map so simple that the phenotype (the protein sequence) can be decoded with a simple lookup table assigning 21 messages (20 amino acids and a *stop* message) to the $4^3$=64 codewords of 3 bp (codons). The success and centrality of this reconstruction arguably diverted attention from comparable efforts to decode non-protein-coding DNA.

For other, more complex GP maps, existing massively parallel experiments can be used to quantitatively calibrate theory-derived molecular interaction rules, which in turn generalize from measured genotypes to the entire sequence space. The first landmark success in this direction has been the prediction of secondary RNA structures from their sequence by the 'Vienna school' [13]. Subsequent efforts focused on molecular recognition in antigen-antibody interactions, emphasizing their specificity for pathogens and simultaneous avoidance of self-interactions, essential for the healthy functioning of immune systems [14]. Last but not least, interactions between transcription factors (TFs) and DNA that we focus on in this review series have also been a case-in-point for physically-informed quantitative GP maps that enable (semi-)realistic simulations of long-term evolution.

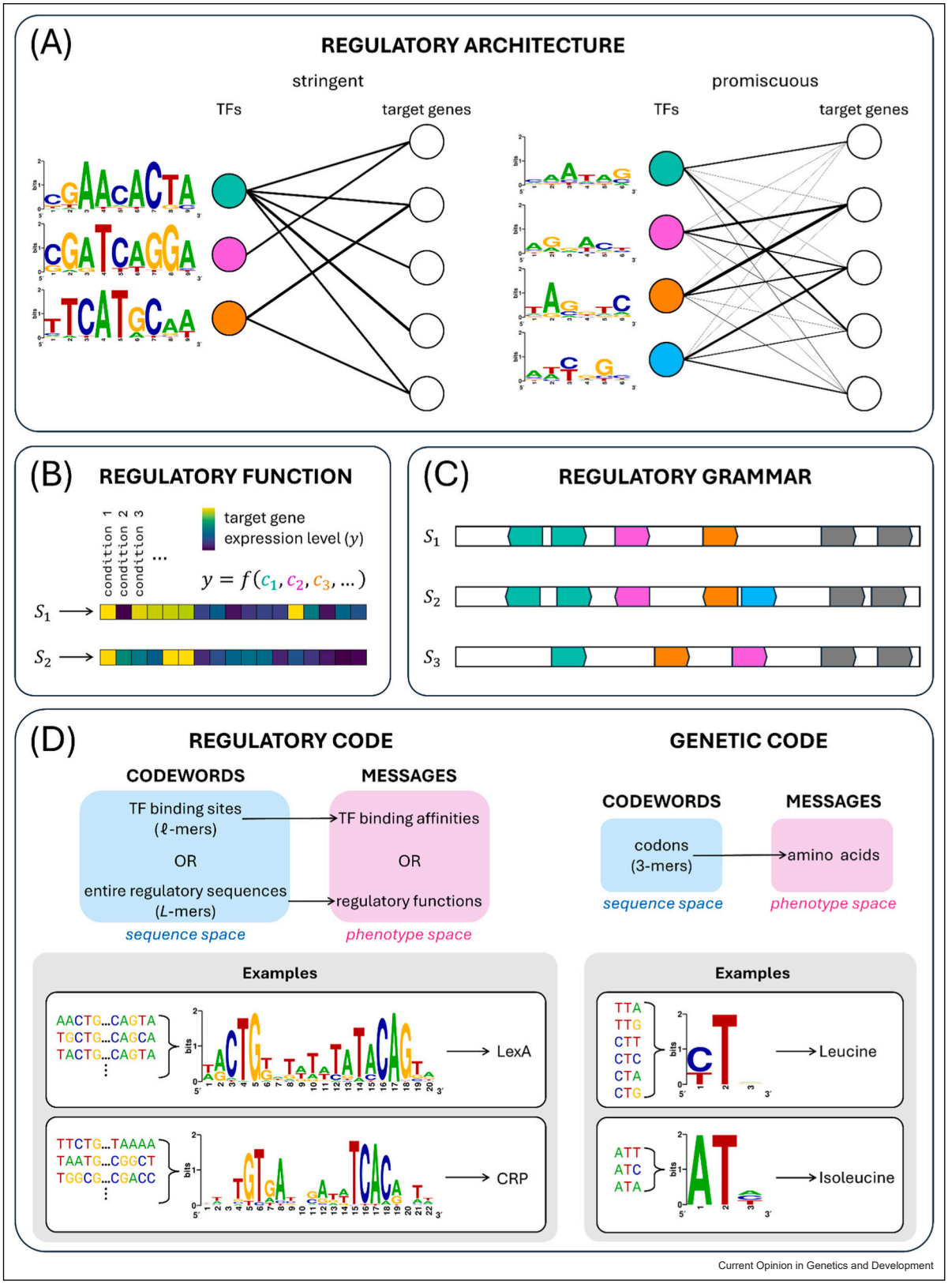
The scope of this review series is purposefully narrow: To evaluate our progress in simulating and theoretically understanding long-term regulatory sequence evolution. This goal closely interacts with several disciplines. From the evolutionary perspective, it presents a unique opportunity to apply and ultimately test population genetics — a mature mathematical theory — on questions inaccessible for most biological systems. For example, while we cannot expect the theory to answer "How long does it take to evolve an eye?", we may have a hope at "How long does it take to evolve a developmental enhancer?" From the systems and synthetic biology perspective, predicting gene expression from regulatory sequence is one of the field's defining questions, even though it is mostly pursued with minimal reference to evolutionary implications. Yet such pursued predictive models are literally the GP maps for gene regulation, when viewed through the evolutionary lens and assuming, importantly, that the molecular mechanisms of regulatory sequence readout do not change on cis-regulatory element (CRE) evolution timescales. Recent years have seen an explosion of successful predictive models combining large-scale experiments, biophysical constraints, and deep learning; we point the reader to several reviews on the topic [15–17].

To set the stage, *Genotype-phenotype maps for regulatory sequences* provides a rudimentary account of the essential building blocks of regulatory GP maps. Section II summarizes classic work on the evolution of individual binding sites, and *Evolution of entire regulatory sequences* focuses on more recent efforts to simulate the evolution of entire promoter and enhancer sequences. Part 2 of this review series [18] discusses evolutionary properties and candidate principles that may shape the evolution of gene regulatory architecture. An illustration of key concepts discussed in both Parts 1 and 2, namely *regulatory architecture*, *function*, *grammar*, and *code*, is provided by Figure 1.

## Genotype-phenotype maps for regulatory sequences

The basic building blocks of GP maps for regulatory sequence evolution comprise the physical mechanisms of regulatory protein-DNA sequence recognition, which are well understood [19,20]. TFs and other regulatory proteins typically recognize $\ell$ = 6–20 basepair motifs, with shorter lengths surprisingly reported in metazoans [21]. As an example, Figure 2a–d shows common representations of the binding preferences of the *Escherichia coli* RNA polymerase (RNAP)-σ70 complex. These representations allow prediction of protein-DNA binding. Building on that, constitutive expression driven by bacterial promoters is now largely predictable, not only with 'black-box' models [22], but also with interpretable ones based on protein-DNA interactions [11,23] (Figure 2e–g). For

**Figure 1**




*(caption on next page)*

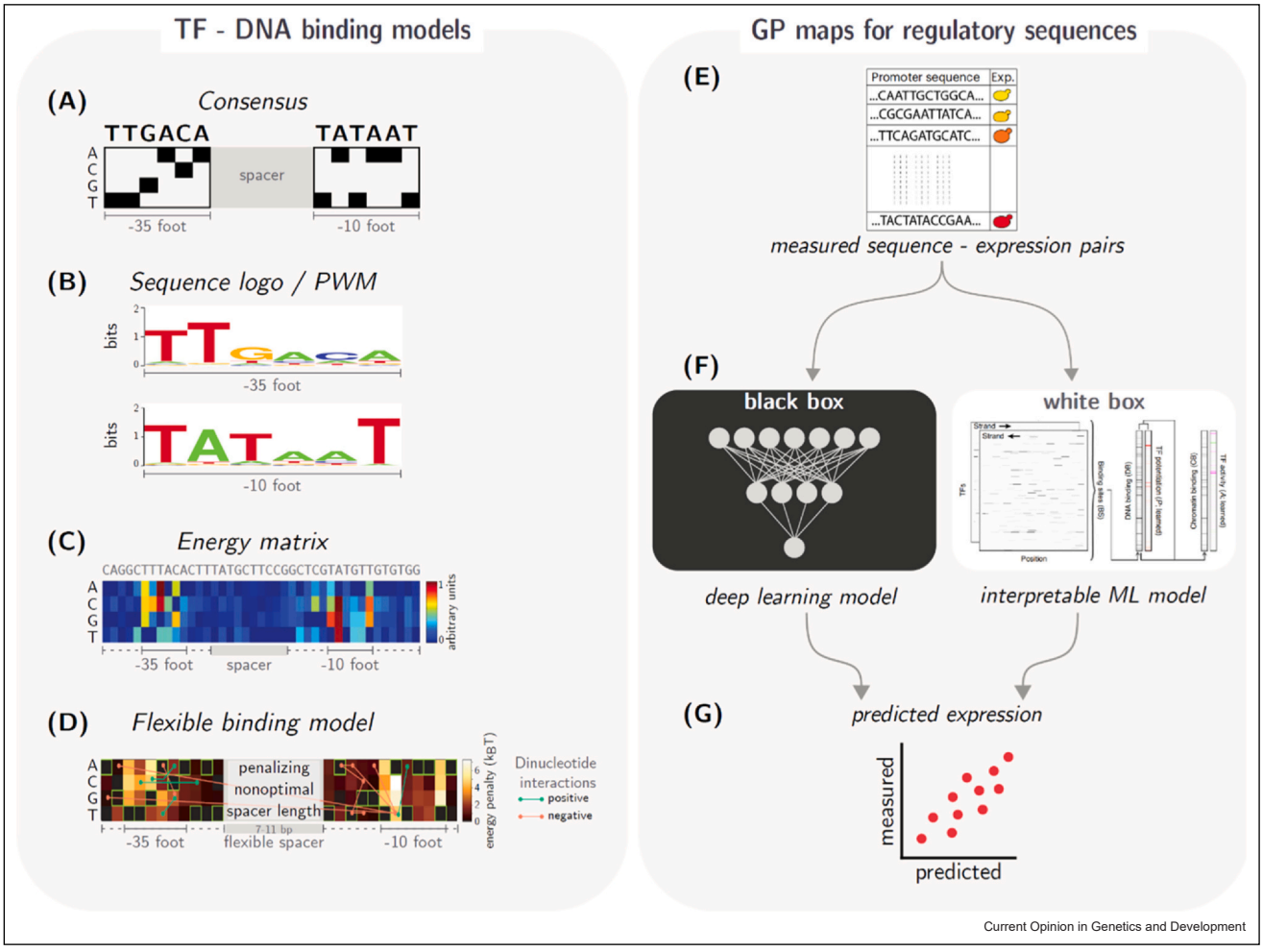
Illustration of key concepts. **(a)** *Regulatory architecture* refers to the large-scale organization of gene regulation: which genes are regulated by which TFs, whether regulation is layered or hierarchical, and whether interactions are additive or combinatorial. This notion is widely used in genetics, genomics, and systems biology. As an example, panel **(a)** compares two architectures: on the left, the targets of each TF are sharply discriminated from the non-targets (presence vs absence of a link), through motifs with high information content, with a very uneven number of targets per TF; on the right, the same target genes are instead regulated by a higher number of more promiscuous TFs, with the strength of each potential TF-target interaction occupying a continuum (links' widths), while the load of regulatory tasks is more homogeneously distributed across TFs. The contrasted features recapitulate certain well-known differences between prokaryotic and eukaryotic regulatory architectures. **(b)** *Regulatory function* is the quantitative mapping from regulatory inputs (encoded by TF concentrations) to gene expression output. Any given regulatory sequence implements a regulatory function that makes the expression of the cognate target gene context-dependent (i.e. a function of TF concentrations $c_i$). As input-output relationships, regulatory functions can be approximated by continuous or discrete (logic) mathematical functions. They are often considered in systems and synthetic biology. Panel **(b)** compares two distinct regulatory functions implemented by two alternative regulatory sequences ($S_1$ and $S_2$). Each condition entails specific concentrations of the TFs ($c_1$, $c_2$, $c_3$, ...), leading to a different expression level for the target gene in that condition. **(c)** *Regulatory grammar* describes the rules by which the number, identity, spacing, orientation, and interactions of TF binding sites determine the regulatory function. Grammar operates at an intermediate level of abstraction: binding sites are treated as functional units (akin to words), and emphasis is placed on how their combinatorial integration shapes gene expression. In this sense, grammar explains how regulatory sequences that contain similar or even identical 'vocabularies' of binding sites – like those illustrated in panel **(c)** — can give rise to different regulatory functions (illustrated in panel **(b)** for $S_1$ and $S_2$) when sites are arranged differently. This notion is common in genetics, evolutionary, and developmental biology. **(d)** *Regulatory code* is a map in which each possible DNA sequence is associated with a regulatory phenotype. This notion can be applied to cases in which the sequences are all the $\ell$-mers, and the phenotypes are their corresponding binding affinities for a given TF with an $\ell$-bp motif. Alternatively, it can be applied to cases in which the sequences are entire regulatory sequences (of $L$ bp in length), and the phenotypes are the regulatory functions they implement. Unlike grammar or function, the notion of (regulatory) code is inherently global and concerns the collective assignment of all (regulatory) *codewords* (i.e. DNA sequences) to (regulatory) *messages* (i.e. regulatory phenotypes). Panel **(d)** compares genetic and regulatory codes and provides examples of encoded messages for both. The *code* paradigm is extensively used for the genetic code (hence the term 'codon'), but often with minimal application of information theory. For regulatory codes, information-theoretic approaches are standard, but the sets of synonymous sequences are called 'motifs' and are seldom formally treated as codewords of an information-theoretic code.

eukaryotic promoters, predictive performance can be high but relies mostly on 'black-box' models, since the increased regulatory complexity (interactions of core, proximal, and distal elements, and chromatin context) makes interpretable models harder to construct [24,25]. Overall, these models can predict a substantial fraction, sometimes $\gtrsim$80%, of the variance in constitutive expression from completely random or designer DNA sequence libraries (sequences far from wildtype promoters!), thereby approaching *bona fide* global GP maps for *constitutive* expression.

To account for regulated — as opposed to constitutive — expression, two extensions to the standard GP map paradigm are necessary. First, the regulatory phenotype can no longer be a single scalar gene expression value [28]: regulation necessarily implies that expression is conditional on cell type or time, extracellular signals, or other causative factors manifesting as changes in TF concentrations and activities. These changes affect expression at the CRE locus through differential TF-binding. From the evolutionary perspective, all these effects are collectively grouped under the heading of the environment, forming the last leg of the genotype-phenotype-environment triad. The dependence of gene expression level on the environment (i.e. on the environment-specific combinations of TFs concentrations that encode the environment) is the 'regulatory function' (Figure 1b). Different regulatory sequences can implement different regulatory functions, and any putative fitness function must score how well gene expression approaches the optimal level across multiple such 'environments'.

Second, regulation by multiple TFs requires understanding how the larger regulatory sequence, typically ~200 bp for prokaryotic promoters or eukaryotic enhancers, affects gene expression. The essential simplification here is the convolutional nature of the transcriptional regulatory code: TFs recognize and bind short motifs within this longer sequence in a well-understood fashion, and our job is to devise a mathematical function, known as 'regulatory grammar' (Figure 1c), that describes how the arrangement of multiple such binding sites for various TFs, with different positions and orientations within the CRE, integrate into multi-valued regulatory phenotypes. Therefore, for a given set of motifs, the regulatory grammar is the mapping from the possible configurations of motif instances within L bp to regulatory output (e.g. expression levels across conditions). This is a drastic simplification of sequence space compared to an entirely unstructured GP map of dimension $4^L$ for L~200, which would be hopelessly out of reach. It also suggests a concrete experimental strategy for dissecting regulatory grammar: by randomly shuffling the positions, orientations, and combinations of predefined motifs and measuring the resulting expression profiles, one can empirically map how spatial organization integrates into regulatory function, as demonstrated by synthetic promoter shuffling approaches [29]. Thus, even though the space of possible motif configurations remains too large for exhaustive exploration, promoter shuffling and massively parallel reporter assays sample this space in a structured, high-throughput manner, and can reveal underlying grammar rules (e.g. that orientation matters for proximal elements but not much for distal enhancers [30]).

**Figure 2**




Models that map regulatory sequence to function. **(a–d)** Models of TF-DNA binding. In a common thermodynamic framework for gene regulation [26], stronger binding — and thus higher average occupancy of the bacterial promoter by the RNAP – leads to higher constitutive gene expression. **(a)** 'Consensus sequence' is a simple summary of RNAP-$\sigma^{70}$ preference to bind two specific sequence hexamers (two "feet") where the DNA is contacted at positions –10 and –35 relative to the transcription start site, separated by a canonical 17 bp spacer. The matrix immediately below the sequence shows a simplified 'mismatch energy model' often used in theoretical studies: any deviation from the consensus sequence at any position contributes a fixed additive penalty to RNAP-DNA binding energy of $\epsilon \sim 1 - 3\ k_BT$, set by the scale of hydrogen bonds that underpin molecular recognition. **(b)** Representation of the two feet as 'sequence logos', created from a curated and verified list of RNAP binding sites, displaying the information (in bits) carried by each basepair position as the height of the stack of letters. Different from the consensus sequence (and mismatch model), this representation takes into account the specificity for each of the four bases at each position, typically encoded as a 4×L *position weight matrix* (PWM). **(c)** Energy matrix representation for RNAP-DNA interaction, which can be calibrated to absolute energy scale, inferred from a massively parallel assay. **(d)** An extended biophysical model that includes a differential energy penalty for spacers between the two feet, and deviations from the additive model in terms of di-nucleotide interactions. **(e–g)** Use of massively parallel assays to infer GP maps for regulatory sequence. Typically, tens of thousands or more mutated versions of a CRE drive a fluorescent reporter (E), and such genotype-expression pairs serve as training samples for fully expressive ("black box") deep learning models (F left), or more interpretable, mechanistically-inspired models fit with modern machine learning tools (F right). The models are tested on withheld data or even mutational libraries with statistical properties that systematically differ from training samples **(g)**. **(c)** (reproduced from [27]). **(d)** (reproduced from [11]). **(e–g)** (F right, reproduced from [12]).

In prokaryotes, integration of binding into expression is based on the so-called 'thermodynamic models' [26], which represent an extensively tested and trusted paradigm [27,31]. In this paradigm, the thermodynamic equilibrium occupancy of all regulatory factors on the promoter stabilizes or destabilizes the binding of RNAP, whose occupancy monotonically maps into gene expression. Statistical physics provides the mathematical machinery to compute various occupancies given TF concentrations, and to systematically account for known complications such as binding cooperativity, steric occlusion between factors, DNA looping, etc. While some

measurements challenge certain thermodynamic models' assumptions [32], these models nevertheless remain a strong, identifiable, and mechanistically grounded baseline GP map for prokaryotic gene regulation [33].

The situation is more complicated in eukaryotes, especially in metazoans. Much is known about how their promoters activate and how signals are integrated across CREs [34,35]. Past groundbreaking efforts to understand eukaryotic regulatory grammar have been successful, especially in yeast [36] and in the context of developmental enhancers [35,37–39], boosted recently by massively parallel experiments and deep learning [12,40,41]. Despite this impressive progress, much remains unknown [42,43]. For example, a definite explanation of why eukaryotic gene regulation utilizes short TF binding sites that individually cannot confer sufficient specificity is still missing; this is called the 'specificity paradox' [44]. Similarly unexplained is the functional role of weak, low-affinity binding sites [45]. On the mechanistic side, key to understanding the regulatory grammar and thus the global regulatory GP map is the additivity vs cooperativity (or synergism) of TF binding, a subject of intense theoretical and experimental research [46–48]. Many additional, well-documented sequence-dependent mechanisms implicated in eukaryotic gene regulation (such as chromatin landscapes and accessibility, histone modifications, nucleosome positioning, methylation, loop extrusion and insulation, or non-equilibrium regulatory processes, etc.) are often studied in isolation and remain to be integrated into an overarching and predictive GP map for eukaryotic gene regulation. Current state-of-the-art either assumes that the effects of these mechanisms, as relevant for the evolutionary outcomes, will be successfully and automatically absorbed by expressive statistical models trained on rich datasets [49], or coarse-grains many such mechanisms into mathematical approximations, for example, by composing the GP map from successive yet understandable linear-nonlinear transformations [50], much like in computational neuroscience or neural networks. These approaches, schematized in Figure 2e–g, pursue from complementary directions the same question: which mechanistic features of gene regulation substantially affect CRE evolutionary trajectories, and which can be safely ignored? Much exciting work remains to be done on this front.

## Evolution of individual transcription factor binding sites

Models for single TF binding presented in Figure 2a–d allow us to study the emergence of TF binding sites *de novo* from a starting sequence ensemble, as well as their maintenance and turnover. The simplified mismatch model in Figure 2a enables analytical treatments that typically agree with simulations based on more detailed but analytically intractable models (Figure 2b–d). The simplest setup assumes directional selection for TF binding, either within a restricted sequence window of size $L=\ell$ bp (where $\ell$ is the length of the TF's motif) or the size of the entire CRE ($L\gg\ell$). In both cases, the sequence is typically short enough to neglect recombination on the relevant timescales. Two basic GP map features are key to the resulting evolutionary dynamics: (i) effects of individual basepairs first combine linearly (via the mismatch model or the energy matrix) into TF binding energy; this step captures the huge degeneracy of sequence space without intrinsic higher-order (epistatic) interactions; (ii) the resulting energy maps into binding probability, and thus fitness, via a sigmoid 'binding' nonlinearity, as dictated by the thermodynamics of TF-DNA interactions. This nonlinearity induces large neutral plateaus (where the sigmoid is flat) in sequence space.

Several approaches can be used to assess the extent to which selection can give rise to *de novo* TF binding sites and maintain them against the entropic forces imposed by mutation and drift. Explicitly modeling mutation rates leads to a mutation-selection balance, where deleterious mutations continually erode binding affinity and selection counteracts this loss. Even in the infinite-population limit (no drift), selection cannot fully concentrate probability mass on the consensus sequence because mutations continually redistribute it across sequence space [51]. Alternatively, one can work in the strong-selection, weak-mutation limit, often analyzed using the fixed-states approximation: the population is assumed fixed for a single genotype, and evolves via mutations that either fix or are lost according to Kimura's fixation probability. Here, mutation acts merely as a generator of variation, while the dominant entropic force opposing fitness maximization is genetic drift. The resulting evolutionary steady state reflects the drift-selection balance [52]. These dynamics map onto the energy-entropy tradeoff of statistical physics [53,54] and explain why mismatches between the consensus and functional binding sites should be expected. When applied to *de novo* emergence of binding sites, these models predict rapid evolution via point mutations for very short sites [55], but exponentially longer times as site length $\ell$ increases, which is difficult to reconcile with comparative genomics evidence for fast turnover at realistic $\ell$, unless selection is exceedingly strong [56]. Slow dynamics arise because random initial sequences are far from consensus, confining adaptation to a random walk over a vast, nearly flat fitness landscape with a vanishing selection gradient (due to the binding nonlinearity). This limitation is not easily circumvented, and active research is exploring mechanisms that could accelerate adaptation (such as noise [57]; sequence changes beyond point mutations [58,59]; promiscuity-inducing mutations decreasing TF specificity [60]; recognition beyond rigid motifs, as with flexible spacers [11,61], Figure 2d; and crosstalk-mediated TF co-option [62]).

A valuable counterpart to theoretically-driven studies are data-driven approaches that learn selection pressures or fitness landscape parameters from genomic data, using biophysical models (e.g. TF-DNA thermodynamic models). Effective fitness landscapes and selection pressures for RNAP and TF binding sites have been inferred from bioinformatic analyses in prokaryotes and yeast [63,64], and clever experimental designs, for example, using yeast hybrids, enable the dissection of specific mechanisms and evolutionary forces important for the evolution of gene regulation [65].

## Evolution of entire regulatory sequences

Simulating the evolution of entire CREs is challenging: not so much due to technicalities or simulation runtime, especially given today's computing power, but mainly due to the many structural and parametric assumptions needed to instantiate a quantitative and global GP map, especially in metazoans, as well as the difficulty of interpreting simulated outcomes. Despite these difficulties, some (though not many) such simulations have been attempted, as reviewed below, opening up opportunities for larger-scale exploration in the near future.

Early attempts to scale up from individual TF binding site evolution toward entire CREs focused on overlapping and competing TF binding sites and their functional importance within the CREs. The thermodynamic model was used to derive a GP map with overlapping and occluding sites, enabling general conclusions relevant to pro- and eukaryotes, while simplifying some aspects of the evolutionary process [66]. Further generic conclusions were pursued in a stylized biophysical model to ask about TF binding site turnover in a CRE while conserving regulatory function, focusing on selection strength and fitness effects of mutations over evolutionary time; importantly, this work was early to emphasize the importance of non-cognate binding that could give rise to deleterious regulatory crosstalk [67,68]. Another qualitatively new possible regulatory mechanism when considering the evolution of entire CREs is the summation of multiple, possibly overlapping, binding-site effects (as empirically demonstrated [45]), in particular when enabled by short tandem sequence repeats [69].

Moving from generic results to system-specific work, the evolution of constitutive bacterial promoters was experimentally studied and computationally simulated using an improved 'flexible RNAP binding model,' inferred from a massive mutagenesis assay (Figure 2d) [11]. GP maps derived from this model drastically increased the fraction of random sequences that drove significant expression, as reported previously [70], as well as the number of non-expressing sequences that are one mutation away from expressing; such sequence predictions were experimentally verified. Importantly, the flexible model predicts order-of-magnitude increases in the rate of *de novo* constitutive promoter evolution relative to previous models, rationalizing the existence of specific mechanisms that might appear unimportant without the evolutionary perspective. These findings served as a stepping stone toward understanding the evolution of not only constitutive, but also regulated bacterial promoters [28,71].

While thermodynamic models have proved their worth for prokaryotic promoters, their extension to eukaryotes is limited by the complex and long-range interactions mediated by enhancers and chromatin structure. Deep learning models, such as Enformer [24] and AlphaGenome [25], have made impressive progress in predicting regulatory activity from DNA sequences, but they learn within the fixed molecular context (e.g. which motifs the TFs recognize) of the extant genomes they are trained on. Moreover, predicting long-term evolution requires *global* GP maps that span sequence space far from wild-types, requiring complementary experimental and theoretical approaches. For example, statistical models inferred from massively parallel assays of yeast regulatory sequences [12] enabled intriguing evolutionary interrogation of how directional and stabilizing selection act on CREs [72]. The key questions relevant for any subsequent theory of regulatory sequence evolution (see Part 2 [18]) were addressed, including: systems-level constraints, diminishing returns epistasis as a consequence of a realistic GP map, regulatory phenotypes that are high-dimensional due to conditioning on the environment, and the importance of robustness and evolvability over long evolutionary timescales. Such approaches begin to chart a path toward predicting the long-term evolution of eukaryotic CREs.

In metazoans, the simulation of regulatory sequence evolution for developmental enhancers has a long history. Theoretical work demonstrated that *de novo* emergence of TF binding sites from truly random sequences is prohibitively slow [56], but left open the path to such emergence from non-random sequences that could contain the so-called 'pre-sites', whose importance has been suggested earlier [73]. The possible role of overlapping TF binding sites [74] or clusters of multiple (potentially weaker) sites [75] has also been highlighted. Several studies have tried to integrate known regulatory phenomenology and models of CREs involved in early *Drosophila* patterning to simulate regulatory sequence evolution [76], including focusing on adaptation timescales [77]. While promising, such sequence-level work remains to be integrated with a theory of how individual gene expression phenotypes and patterns contribute to organismal fitness, as suggested by normative theories [78]: this is one of the key open *Challenges* highlighted in Part 2 of this review series [18].

On the experimental front, the direction that might, perhaps surprisingly, interact most strongly with theory is the use of truly random mutational libraries for regulatory sequences. By 'truly random' we mean constructs that do not mutagenize wild-type promoters or enhancers, but explore sequence space in an unbiased fashion (in the extreme, by generating sequences with background A/C/G/T probabilities, without correlations between positions). A standard textbook viewpoint holds that a completely random sequence is typically inert, with no downstream gene expression, because the probability of containing a functional sequence is vanishingly small. Contrary to this expectation, many completely random promoter sequences drive significant expression in bacteria — around 10% of randomly generated sequences or more, depending on sequence length and other details — far exceeding expectations based on naive consensus sequence matching [11,70,79]. Related intriguing results are not limited to prokaryotes. In yeast, random promoter libraries have been proposed as a non-intuitive but productive means of GP map exploration [80], with strong indications that this extends to multicellular eukaryotes [81], including exciting recent results in the fruit fly [82]. What gives further credence to the random library approach is its strong connection to theory [28,50,83,84], as it provides empirical estimates of the neutral distribution of phenotypes across sequence space, which is a key element that explicitly appears in theoretical descriptions of evolutionary processes.

The empirical and modeling results reviewed here suggest that the information coding for regulatory phenotypes evolves in ways that depend on properties of the genotype–phenotype–fitness map, which can be characterized and quantified. Some of these properties are themselves genetically encoded, raising the possibility that not only regulatory sequences, but the regulatory code itself might evolve. This parallels decades of work on the genetic code, suggesting that, although extant protein-coding sequences share a (nearly) universal code, the architecture of the code itself might have evolved during pre-last universal common ancestor stages of life, favoring encoding strategies that jointly provide mutational robustness and evolvability [85,86]. In Part 2 of this review series [18], we discuss how developments at the interface of information theory and evolutionary theory can quantify the action of selection, mutation, and drift on regulatory sequences and codes, and outline future directions toward a unifying language for their emergence and organization.

## Data Availability

No data were used for the research described in the article.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We thank Nick Barton and Noa Ottilie Borst for essential contributions to this manuscript.

E.M. acknowledges support from the APART-USA fellowship, jointly funded by the Austrian Academy of Sciences (ÖAW) and the Institute of Science and Technology Austria (ISTA).

This study was supported by the European Molecular Biology Laboratory (J.C.); the European Molecular Biology Laboratory Interdisciplinary Postdoc Programme (EIPOD) under the Marie Skłodowska-Curie Actions cofund (S.H.A.).

## References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Gavrilets S: Fitness Landscapes and the Origin of Species (MPB-41) Princeton University Press; 2004, ⟨https://www.jstor.org/stable/j.ctv39x541⟩ (Available from).

2. Manrubia S, Cuesta JA, Aguirre J, Ahnert SE, Altenberg L, Cano AV, *et al.*: **From genotypes to organisms: state-of-the-art and perspectives of a cornerstone in evolutionary dynamics**. *Phys Life Rev* 2021, **38**:55-106, https://doi.org/10.1016/j.plrev.2021.03.004

3. Kingman JFC: **A simple model for the balance between selection and mutation**. *J Appl Probab* 1978, **15**:1-12, https://doi.org/10.2307/3213231

4. Aita T, Husimi Y: **Fitness spectrum among random mutants on Mt. Fuji-type fitness landscape**. *J Theor Biol* 1996, **182**:469-485, https://doi.org/10.1006/jtbi.1996.0189

5. Kauffman S, Levin S: **Towards a general theory of adaptive walks on rugged landscapes**. *J Theor Biol* 1987, **128**:11-45, https://doi.org/10.1016/S0022-5193(87)80029-2

6. Weinberger ED: **Local properties of Kauffman's N-k model: a tunably rugged energy landscape**. *Phys Rev A* 1991, **44**:6399-6413, https://doi.org/10.1103/PhysRevA.44.6399

7. Beerenwinkel N, Pachter L, Sturmfels B, Elena SF, Lenski RE: **Analysis of epistatic interactions and fitness landscapes using a new geometric approach**. *BMC Evol Biol* 2007, **7**:60, https://doi.org/10.1186/1471-2148-7-60

8. Otwinowski J, McCandlish DM, Plotkin JB: **Inferring the shape of global epistasis**. *Proc Natl Acad Sci* 2018, **115**:E7550-E7558, https://doi.org/10.1073/pnas.1804015115

9. Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, *et al.*: **Local fitness landscape of the green fluorescent protein**. *Nature* 2016, **533**:397-401, https://doi.org/10.1038/nature17995

10. Gonzalez Somermeyer L, Fleiss A, Mishin AS, Bozhanova NG, Igolkina AA, Meiler J, *et al.*: **Heterogeneity of the GFP fitness landscape and data-driven protein design**. *eLife* 2022, **11**:e75842, https://doi.org/10.7554/eLife.75842

11. •• Lagator M, Sarikas S, Steinrueck M, Toledo-Aparicio D, Bollback JP, Guet CC, *et al.*: **Predicting bacterial promoter function and evolution from random sequences**. *eLife* 2022, **11**, https://doi.org/10.7554/ELIFE.64543 PubMed PMID: 35080492.
This study uses random libraries to reveal and quantify the high evolvability of promoter function from random sequences, and shows how

accounting for the biophysical properties of the RNA polymerase can provide significant insights into *de novo* promoter evolution.

12. • de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A: **Deciphering eukaryotic gene-regulatory logic with 100 million random promoters**. *Nat Biotechnol* 2020, **38**:56-65, https://doi.org/10.1038/s41587-019-0315-8.
This work uses massive libraries of random promoters to uncover quantitative rules of eukaryotic gene regulation and shows that regulatory function is widespread in sequence space.

13. Fontana W: **Modelling 'evo-devo' with RNA**. *BioEssays* 2002, **24**:1164-1177, https://doi.org/10.1002/bies.10190

14. Adams RM, Kinney JB, Walczak AM, Mora T: **Epistasis in a fitness landscape defined by antibody-antigen binding free energy**. *Cell Syst* 2019, **8**:86-93, https://doi.org/10.1016/j.cels.2018.12.004 PubMed PMID: 30611676; PubMed Central PMCID: PMC6487650.

15. Levo M, Segal E: **pursuit of design principles of regulatory sequences**. *Nat Rev Genet* 2014, **15**:453-468, https://doi.org/10.1038/nrg3684

16. Sokolova K, Chen KM, Hao Y, Zhou J, Troyanskaya OG: **Deep learning sequence models for transcriptional regulation**. *Annu Rev Genom Hum Genet* 2024, **25**:105-122, https://doi.org/10.1146/annurev-genom-021623-024727 PubMed PMID: 38594933.

17. • Barbadilla-Martínez L, Klaassen N, van Steensel B, de Ridder J: **Predicting gene expression from DNA sequence using deep learning models**. *Nat Rev Genet* 2025, **26**:666-680, https://doi.org/10.1038/s41576-025-00841-2.
This article reviews how deep learning models trained on large-scale genomic data can accurately predict gene expression from sequence while implicitly learning regulatory grammar.

18. Mascolo E, Borbély R, Borst NO, Barton NH, Crocker J, Tkačik G: **Long-term evolution of regulatory DNA sequences. Part 2: Theory and future challenges**. *Curr Opin Genet Dev* 2026, **98**:102472, https://doi.org/10.1016/j.gde.2026.102472

19. von Hippel PH, Berg OG: **On the specificity of DNA-protein interactions**. *Proc Natl Acad Sci* 1986, **83**:1608-1612, https://doi.org/10.1073/pnas.83.6.1608

20. Maerkl SJ, Quake SR: **A systems approach to measuring the binding energy landscapes of transcription factors**. *Science* 2007, **315**:233-237, https://doi.org/10.1126/science.1131007

21. Wunderlich Z, Mirny LA: **Different gene regulation strategies revealed by analysis of binding motifs**. *Trends Genet* 2009, **25**:434-440, https://doi.org/10.1016/j.tig.2009.08.003

22. Wang X, Xu K, Huang Z, Lin Y, Zhou J, Zhou L, *et al*.: **Accelerating promoter identification and design by deep learning**. *Trends Biotechnol* 2025, **43**:3071-3087, https://doi.org/10.1016/j.tibtech.2025.05.008

23. LaFleur TL, Hossain A, Salis HM: **Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria**. *Nat Commun* 2022, **13**:5159, https://doi.org/10.1038/s41467-022-32829-5

24. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, *et al*.: **Effective gene expression prediction from sequence by integrating long-range interactions**. *Nat Methods* 2021, **18**:1196-1203, https://doi.org/10.1038/s41592-021-01252-x

25. Avsec Ž, Latysheva N, Cheng J, Novati G, Taylor KR, Ward T, *et al*.: **Advancing regulatory variant effect prediction with AlphaGenome**. *Nature* 2026, **649**:1206-1218, https://doi.org/10.1038/s41586-025-10014-0

26. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, *et al*.: **Transcriptional regulation by the numbers: models**. *Curr Opin Genet Dev* 2005, **15**:116-124, https://doi.org/10.1016/j.gde.2005.02.007 (Chromosomes and expression mechanisms).

27. Kinney JB, Murugan A, Callan CG, Cox EC: **Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence**. *Proc Natl Acad Sci* 2010, **107**:9158-9163, https://doi.org/10.1073/pnas.1004290107

28. • Grah R, Guet CC, Tkačik G, Lagator M: **Linking molecular mechanisms to their evolutionary consequences: a primer**. *Genetics* 2025, **229**:iyae191, https://doi.org/10.1093/genetics/iyae191.
This work provides a framework that connects biophysical mechanisms of molecular interactions to evolutionary dynamics and long-term evolutionary outcomes.

29. Kinkhabwala A, Guet CC: **Uncovering cis regulatory codes using synthetic promoter shuffling**. *PLoS One* 2008, **3**:e2030, https://doi.org/10.1371/journal.pone.0002030

30. Agarwal V, Inoue F, Schubach M, Penzar D, Martin BK, Dash PM, *et al*.: **Massively parallel characterization of transcriptional regulatory elements**. *Nature* 2025, **639**:411-420, https://doi.org/10.1038/s41586-024-08430-9 PubMed PMID: 39814889; PubMed Central PMCID: PMC11903340.

31. Jones DL, Brewster RC, Phillips R: **Promoter architecture dictates cell-to-cell variability in gene expression**. *Science* 2014, **346**:1533-1536, https://doi.org/10.1126/science.1255301

32. Garcia HG, Sanchez A, Boedicker JQ, Osborne M, Gelles J, Kondev J, *et al*.: **Operator sequence alters gene expression independently of transcription factor occupancy in bacteria**. *Cell Rep* 2012, **2**:150-161, https://doi.org/10.1016/j.celrep.2012.06.004 PubMed PMID: 22840405.

33. Barnes SL, Belliveau NM, Ireland WT, Kinney JB, Phillips R: **Mapping DNA sequence to transcription factor binding energy in vivo**. *PLoS Comput Biol* 2019, **15**:e1006226, https://doi.org/10.1371/journal.pcbi.1006226

34. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U: **Predicting expression patterns from regulatory sequence in Drosophila segmentation**. *Nature* 2008, **451**:535-540, https://doi.org/10.1038/nature06496

35. Spitz F, Furlong EEM: **Transcription factors: from enhancer binding to developmental control**. *Nat Rev Genet* 2012, **13**:613-626, https://doi.org/10.1038/nrg3207

36. Rajkumar AS, Dénervaud N, Maerkl SJ: **Mapping the fine structure of a eukaryotic promoter input-output function**. *Nat Genet* 2013, **45**:1207-1215, https://doi.org/10.1038/ng.2729

37. Fakhouri WD, Ay A, Sayal R, Dresch J, Dayringer E, Arnosti DN: **Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo**. *Mol Syst Biol* 2010, **6**:MSB200997, https://doi.org/10.1038/msb.2009.97

38. Small S, Arnosti DN: **Transcriptional enhancers in Drosophila**. *Genetics* 2020, **216**:1-26, https://doi.org/10.1534/genetics.120.301370

39. Reiter F, Almeida BP de, Stark A: **Enhancers display constrained sequence flexibility and context-specific modulation of motif function**. *Genome Res* 2023, **33**:346-358, https://doi.org/10.1101/gr.277246.122 PubMed PMID: 36941077.

40. •• Fuqua T, Jordan J, van Breugel ME, Halavatyi A, Tischer C, Polidoro P, *et al*.: **Dense and pleiotropic regulatory information in a developmental enhancer**. *Nature* 2020, **587**:235-239, https://doi.org/10.1038/s41586-020-2816-5.
This work reveals that developmental enhancers encode regulatory information densely and pleiotropically, with most point mutations affecting gene expression.

41. • de Almeida BP, Reiter F, Pagani M, Stark A: **DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers**. *Nat Genet* 2022, **54**:613-624, https://doi.org/10.1038/s41588-022-01048-5.
This work introduces a deep learning framework that predicts enhancer activity from sequence and enables rational design of synthetic enhancers with specified regulatory phenotypes.

42. Panigrahi A, O'Malley BW: **Mechanisms of enhancer action: the known and the unknown**. *Genome Biol* 2021, **22**:108, https://doi.org/10.1186/s13059-021-02322-1

43. Llères D, Cardozo Gizzi A, Nollmann M: **Redefining enhancer action: insights from structural, genomic, and single-molecule perspectives**. *Curr Opin Cell Biol* 2025, **95**:102527, https://doi.org/10.1016/j.ceb.2025.102527

44. Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS: **Low-affinity binding sites and the transcription factor specificity paradox in**

**eukaryotes**. *Annu Rev Cell Dev Biol* 2019, **35**:357-379, https://doi.org/10.1146/annurev-cellbio-100617-062719

45. •• Shahein A, López-Malo M, Istomin I, Olson EJ, Cheng S, Maerkl SJ: **Systematic analysis of low-affinity transcription factor binding site clusters in vitro and in vivo establishes their functional relevance**. *Nat Commun* 2022, **13**:5273, https://doi.org/10.1038/s41467-022-32971-0.
In this article, the authors demonstrate that clusters of low-affinity TF binding sites are biologically functional regulatory elements, emphasizing the importance of weak and distributed interactions in gene regulation.

46. Park J, Estrada J, Johnson G, Vincent BJ, Ricci-Tam C, Bragdon MD, *et al*.: **Dissecting the sharp response of a canonical developmental enhancer reveals multiple sources of cooperativity**. *eLife* 2019, **8**:e41266, https://doi.org/10.7554/eLife.41266

47. Grah R, Zoller B, Tkačik G: **Nonequilibrium models of optimal enhancer function**. *Proc Natl Acad Sci* 2020, **117**:31614-31622, https://doi.org/10.1073/pnas.2006731117

48. Rao S, Ahmad K, Ramachandran S: **Cooperative binding between distant transcription factors is a hallmark of active enhancers**. *Mol Cell* 2021, **81**:1651-1665.e4, https://doi.org/10.1016/j.molcel.2021.02.014

49. • Smith GD, Ching WH, Cornejo-Páramo P, Wong ES: **Decoding enhancer complexity with machine learning and high-throughput discovery**. *Genome Biol* 2023, **24**:116, https://doi.org/10.1186/s13059-023-02955-4.
This article reviews how the combination of high-throughput enhancer assays with machine learning can dissect the regulatory grammar underlying enhancer activity.

50. Borbély R, Tkačik G: **Regulatory architectures optimized for rapid evolution of gene expression**. *bioRxiv* 2025, https://doi.org/10.1101/2025.06.10.658850 [cited 2026 Jan 12]. p. 2025.06.10.658850. Available from: https://www.biorxiv.org/content/10.1101/2025.06.10.658850v1.

51. Gerland U, Hwa T: **On the selection and evolution of regulatory DNA motifs**. *J Mol Evol* 2002, **55**:386-400, https://doi.org/10.1007/s00239-002-2335-z

52. Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M: **Drift-barrier hypothesis and mutation-rate evolution**. *Proc Natl Acad Sci* 2012, **109**:18488-18492, https://doi.org/10.1073/pnas.1216223109

53. Iwasa Y: **Free fitness that always increases in evolution**. *J Theor Biol* 1988, **135**:265-281, https://doi.org/10.1016/S0022-5193(88)80243-1

54. Sella G, Hirsh AE: **The application of statistical physics to evolutionary biology**. *Proc Natl Acad Sci* 2005, **102**:9541-9546, https://doi.org/10.1073/pnas.0501865102

55. Berg J, Willmann S, Lässig M: **Adaptive evolution of transcription factor binding sites**. *BMC Evol Biol* 2004, **4**:42, https://doi.org/10.1186/1471-2148-4-42

56. Tuğrul M, Paixão T, Barton NH, Tkačik G: **Dynamics of transcription factor binding site evolution**. *PLoS Genet* 2015, **11**:e1005639, https://doi.org/10.1371/journal.pgen.1005639

57. Wolf L, Silander OK, van Nimwegen E: **Expression noise facilitates the evolution of gene regulation**. *eLife* 2015, **4**:e05856, https://doi.org/10.7554/eLife.05856

58. Tomanek I, Grah R, Lagator M, Andersson AMC, Bollback JP, Tkačik G, *et al*.: **Gene amplification as a form of population-level gene expression regulation**. *Nat Ecol Evol* 2020, **4**:612-625, https://doi.org/10.1038/s41559-020-1132-7

59. Steinrueck M, Guet CC: **Complex chromosomal neighborhood effects determine the adaptive potential of a gene under selection**. *eLife* 2017, **6**:e25100, https://doi.org/10.7554/eLife.25100

60. Friedlander T, Prizak R, Barton NH, Tkačik G: **Evolution of new regulatory functions on biophysically realistic fitness landscapes**. *Nat Commun* 2017, **8**:216, https://doi.org/10.1038/s41467-017-00238-8

61. Mascolo E, Erill I: **Information theory of composite sequence motifs: mutational and biophysical determinants of complex molecular recognition**. *bioRxiv* 2024, https://doi.org/10.1101/2024.11.11.623117 [cited 2026 Jan 12]. p. 2024.11.11.623117. Available from: https://www.biorxiv.org/content/10.1101/2024.11.11.623117v1.

62. Taylor T.B., Rice A.M. **Transcription Factor Promiscuity Drives Regulatory Rewiring and Evolvability in Gene Networks in Bacteria.** Adv Sci. n/a(n/a):e20406. doi:⟨10.1002/advs.202520406⟩.

63. Mustonen V, Lässig M: **Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies**. *Proc Natl Acad Sci* 2005, **102**:15936-15941, https://doi.org/10.1073/pnas.0505537102

64. Haldane A, Manhart M, Morozov AV: **Biophysical fitness landscapes for transcription factor binding sites**. *PLoS Comput Biol* 2014, **10**:e1003683, https://doi.org/10.1371/journal.pcbi.1003683

65. • Krieger G, Lupo O, Wittkopp P, Barkai N: **Evolution of transcription factor binding through sequence variations and turnover of binding sites**. *Genome Res* 2022, **32**:1099-1111, https://doi.org/10.1101/gr.276715.122 PubMed PMID: 35618416.
This article shows how transcription factor binding evolves through continuous sequence variation and can even maintain regulatory function across related species despite rapid motif divergence.

66. Hermsen R, Tans S, ten Wolde PR: **Transcriptional regulation by competing transcription factor modules**. *PLoS Comput Biol* 2006, **2**:e164, https://doi.org/10.1371/journal.pcbi.0020164 PubMed PMID: 17140283; PubMed Central PMCID: PMC1676028.

67. Bullaughey K: **Changes in selective effects over time facilitate turnover of enhancer sequences**. *Genetics* 2011, **187**:567-582, https://doi.org/10.1534/genetics.110.121590

68. Friedlander T, Prizak R, Guet CC, Barton NH, Tkačik G: **Intrinsic limits to gene regulation by global crosstalk**. *Nat Commun* 2016, **7**:12307, https://doi.org/10.1038/ncomms12307

69. Horton CA, Alexandari AM, Hayes MGB, Marklund E, Schaepe JM, Aditham AK, *et al*.: **Short tandem repeats bind transcription factors to tune eukaryotic gene expression**. *Science* 2023, **381**:eadd1250, https://doi.org/10.1126/science.add1250

70. • Yona AH, Alm EJ, Gore J: **Random sequences rapidly evolve into de novo promoters**. *Nat Commun* 2018, **9**:1530, https://doi.org/10.1038/s41467-018-04026-w.
This study pioneers the use of random libraries to study *de novo* promoter evolution, showing experimentally that random DNA sequences are at a short mutational distance from functional promoter sequences.

71. Razo-Mejia M, Boedicker JQ, Jones D, DeLuna A, Kinney JB, Phillips R: **Comparison of the theoretical and real-world evolutionary potential of a genetic circuit**. *Phys Biol* 2014, **11**:026005, https://doi.org/10.1088/1478-3975/11/2/026005

72. Vaishnav ED, de Boer CG, Molinet J, Yassour M, Fan L, Adiconis X, *et al*.: **The evolution, evolvability and engineering of gene regulatory DNA**. *Nature* 2022, **603**:455-463, https://doi.org/10.1038/s41586-022-04506-6

73. MacArthur S, Brookfield JFY: **Expected rates and modes of evolution of enhancer sequences**. *Mol Biol Evol* 2004, **21**:1064-1073, https://doi.org/10.1093/molbev/msh105

74. Lusk RW, Eisen MB: **Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers**. *PLoS Genet* 2010, **6**:e1000829, https://doi.org/10.1371/journal.pgen.1000829

75. He X, Duque TSPC, Sinha S: **Evolutionary origins of transcription factor binding site clusters**. *Mol Biol Evol* 2012, **29**:1059-1070, https://doi.org/10.1093/molbev/msr277

76. Duque T, Samee MdAH, Kazemian M, Pham HN, Brodsky MH, Sinha S: **Simulations of enhancer evolution provide mechanistic insights into gene regulation**. *Mol Biol Evol* 2014, **31**:184-200, https://doi.org/10.1093/molbev/mst170

77. Duque T, Sinha S: **What does it take to evolve an enhancer? A simulation-based study of factors influencing the emergence**

**of combinatorial regulation**. *Genome Biol Evol* 2015, **7**:1415-1431, https://doi.org/10.1093/gbe/evv080

78. Sokolowski TR, Gregor T, Bialek W, Tkačik G: **Deriving a genetic regulatory network from an optimization principle**. *Proc Natl Acad Sci* 2025, **122**:e2402925121, https://doi.org/10.1073/pnas.2402925121

79. Fuqua T, Sun Y, Wagner A: **The emergence and evolution of gene expression in genome regions replete with regulatory motifs**. *eLife* 2024, **13**:RP98654, https://doi.org/10.7554/eLife.98654

80. de Boer CG, Taipale J: **Hold out the genome: a roadmap to solving the cis-regulatory code**. *Nature* 2024, **625**:41-50, https://doi.org/10.1038/s41586-023-06661-w

81. Luthra I, Jensen C, Chen XE, Salaudeen AL, Rafi AM, de Boer CG: **Regulatory activity is the default DNA state in eukaryotes**. *Nat Struct Mol Biol* 2024, **31**:559-567, https://doi.org/10.1038/s41594-024-01235-4

82. • Galupa R, Alvarez-Canales G, Borst NO, Fuqua T, Gandara L, Misunou N, *et al*.: **Enhancer architecture and chromatin accessibility constrain phenotypic space during Drosophila development**. *Dev Cell* 2023, **58**:51-62, https://doi.org/10.1016/j.devcel.2022.12.003 PubMed PMID: 36626871.
This study shows that enhancer GP maps and chromatin accessibility jointly constrain the range of attainable regulatory phenotypes during development, strongly limiting the evolution of pre-existing enhancers while allowing *de novo* enhancer evolution to explore phenotype space much more freely.

83. Wagner A: **Information theory, evolutionary innovations and evolvability**. *Philos Trans R Soc B Biol Sci* 2017, **372**:20160416, https://doi.org/10.1098/rstb.2016.0416

84. Barton NH, Tkačik G: **Evolution and information content of optimal gene regulatory architectures**, [cited 2026 Jan 12]. p. 2025.06.10.657849. Available from: https://www.biorxiv.org/content/10.1101/2025.06.10.657849v1, *bioRxiv* 2025, https://doi.org/10.1101/2025.06.10.657849

85. Koonin EV, Novozhilov AS: **Origin and evolution of the genetic code: the universal enigma**. *IUBMB Life* 2009, **61**:99-111, https://doi.org/10.1002/iub.146

86. Rozhoňová H, Martí-Gómez C, McCandlish DM, Payne JL: **Robust genetic codes enhance protein evolvability**. *PLoS Biol* 2024, **22**:e3002594, https://doi.org/10.1371/journal.pbio.3002594