Evaluating the Effect of Disturbed Ensemble Distributions on SCFG Based Statistical Sampling of RNA Secondary Structures

Anika Scheid* and Markus E. Nebel
Department of Computer Science, University of Kaiserslautern,
P.O. Box 3049, D-67653 Kaiserslautern, Germany
{a_scheid,nebel}@cs.uni-kl.de

Abstract

Over the past years, statistical and Bayesian approaches have become increasingly appreciated to address the long-standing problem of computational RNA structure prediction. Recently, a novel probabilistic method towards the prediction of RNA secondary structures from a single sequence has been studied which is based on generating statistically representative and reproducible samples of the entire ensemble of feasible structures for a particular input sequence. This method actually samples the possible foldings from a distribution implied by a sophisticated (traditional or length-dependent) stochastic context-free grammar (SCFG) that mirrors the standard thermodynamic model applied in modern physics-based prediction algorithms. Specifically, that grammar represents an exact probabilistic counterpart to the energy model underlying the Sfold software, which employs on a sampling extension of the partition function (PF) approach to produce statistically representative subsets of the Boltzmann-weighted ensemble. Although both sampling approaches have the same worst-case time and space complexities, it has been indicated that they differ in performance (both with respect to prediction accuracy and quality of generated samples), where neither of these two competing approaches generally outperforms the other.

In this work, we will consider the SCFG based approach in order to perform an analysis on how the quality of generated sample sets and the corresponding prediction accuracy changes when different degrees of disturbances are incorporated into the needed sampling probabilities. This is motivated by the fact that if the results prove to behave resistant even with respect to large errors on the distinct sampling probabilities (compared to the exact ones), then it seems adequate to believe that these probabilities do not need to be computed in an exact way, but it may efficiently suffice to only approximate them. Thus, it might then be possible to decrease the worst-case time requirements of such an SCFG based sampling method without significant accuracy losses. If, on the other hand, the quality of sampled structures can be observed to strongly react on slight disturbances already, then there is little hope for improving the complexity by corresponding heuristic procedures.

1 Introduction

In computational structural biology, a well-established probabilistic methodology towards single sequence RNA secondary structure prediction is based on modeling secondary structures by *stochastic context-free grammars (SCFGs)*. In a sense, SCFGs can be seen as a generalization of *hidden Markov models (HMMs)*, which are widely and successfully used in the large field of bioinformatics. Briefly, SCFGs extend on traditional context-free grammars (CFGs) by additionally defining a (non-uniform) probability distribution on the generated structure class which is induced by the grammar parameters that can easily be derived from a given database of sample structures via maximum likelihood techniques. Notably, different SCFG designs can be used to model the same class of structures, where flexibility in model design comes from the fact that basically all distinct substructures can be distinguished and with increasing number of distinguished features, the resulting SCFG gains in both explicitness and complexity, which may result in a more realistic distribution on the modeled structure class.
Actually, by applying SCFGs for RNA structure prediction, the main focus of attention is laid on the typical structural composition that can be observed from a database of trusted secondary structures with annotated sequences. Hence, a SCFG model can easily be suited for a specific RNA type (given by the training data), but the performance of the corresponding prediction algorithms is strongly dependent on the availability of a rich training set. Traditionally, SCFG based prediction approaches are realized by

*Corresponding author. The research of this author has been supported by the Carl-Zeiss-Stiftung.

dynamic programming algorithms (DPAs) that require $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ storage for identifying the most probable folding for an input sequence of length n. Examples for successful applications of several lightweight (i.e. small and simple) SCFGs for RNA secondary structure prediction can be found in [DE04] and a popular SCFG based prediction tool is for instance given by the Pfold software [KH99, KH03].

However, for a very long time, the free energy minimization (MFE) paradigm has been the most common technique for predicting the secondary structure of a given RNA sequence. The respective methods are traditionally realized by DPAs that employ a particular thermodynamic model for the derivation of the corresponding recursions. They basically require $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ storage for identifying a set of candidate structures for an input sequence of length n. In fact, while early methods, like [NPGK78, NJ80, ZS81], computed only one structure (the MFE structure of the molecule), several more elaborate MFE based DPAs have been developed over the years for generating a set of suboptimal foldings (see, e.g., [WFHS99, Zuk89]). Some implementations are considered state-of-the-art tools for computational structure prediction from a single sequence, for instance the Mfold software [Zuk89, Zuk03] or the Vienna package [HFS$^+$94, Hof03].
One major drawback of these MFE approaches is that they generally build on the standard *Turner energy model* [XSB$^+$98, MSZT99], which still contains many imprecisions and uses the same experimentally derived parameters for all RNA types. Hence, their performance is strongly dependent on and thus limited by the applied thermodynamic model. Moreover, in the traceback steps of the corresponding DPAs, base pairs are successively generated according to the energy minimization principle, such that the predicted set of suboptimal foldings often contains many structures that are not significantly different (that have the same or very similar shapes and contain mostly the same actual base pairings).
To overcome these problems, several statistical sampling methods and clustering techniques have been invented over the last years that are based on the partition function (PF) approach for computing base pair probabilities as introduced in [McC90]. Briefly, these methods produce a statistical sample of the thermodynamic ensemble of suboptimal foldings and rely on a statistical representation of the Boltzmann-weighted ensemble of structures for a given sequence [DL03]. They are implemented in the widely used Sfold package [DCL04].

In fact, over the past years, statistical approaches to RNA secondary structure prediction have become an attractive alternative to the standard energy-based approach (which basically relies on several thousand experimentally-determined energy parameters). In principle, many of these approaches – in contrast to Sfold – rely on (thermodynamic) parameters estimated from growing databases of structural RNAs. For instance, the CONTRAfold tool [DWB06] is based on a *discriminative* statistical method and uses a simplified Mfold-like scoring scheme for the underlying *conditional log-linear model (CLLM)*. Briefly, CLLMs are flexible *discriminative* probabilistic models that generalize upon more intuitive *generative* probabilistic models (like vanilla SCFGs or HMMs), where any SCFG has an equivalent representation as an appropriately parameterized CLLM. The prime advantage of using discriminate instead of generative training is that more complex scoring schemes can be considered, whereas generative models are generally easier to train an use. Nevertheless, CONTRAfold in many cases manages to provide the highest single sequence prediction accuracy to date and eventually closes the performance gap between the best thermodynamic methods and the best (lightweight) SCFGs.
Notably, statistical methods for RNA folding have previously been chosen to be either purely physics-based (e.g., Sfold) or discriminative and implementing a thermodynamic model (e.g., CONTRAfold), not generative. This might have been due to the misconception that SCFGs could not easily be constructed to mirror energy-based models (as mentioned e.g. in [DWB06]), although it has been demonstrated lately that this is actually possible (see, e.g. [NS11]). In fact, a generative statistical method for predicting RNA secondary structure has recently been proposed [NSar]. This method builds on a novel probabilistic sampling approach for generating random candidate structures for a given input sequence that is based on a sophisticated SCFG design. Basically, it generates a statistical sample of possible foldings for the given sequence that is guaranteed to be representative with respect to the corresponding ensemble distribution implied by the parameters of the underlying SCFG. Particularly, conditional sampling probabilities for randomly creating unpaired bases and base pairs on actual sequence fragments are considered that are calculated by using only the grammar parameters and the corresponding inside and outside probabilities for the sequence. As the underlying elaborate SCFG mirrors the thermodynamic model employed in the Sfold software, this sampling algorithm represents a probabilistic counterpart to the sampling extension of the PF approach (as implemented in Sfold). In fact, the sole difference is that it incorporates only comprehensive structural features and additional information obtained from trusted databases of real-world RNA structures instead of the recent thermodynamic parameters.

Lately, in an attempt to improve the quality of generated sample sets, this probabilistic sampling approach has been extended to being capable of additionally incorporating *length-dependencies* [SN]. In particular, the employed (heavyweight) SCFG has been transformed into a corresponding *length-dependent stochastic context-free grammar (LSCFG)* and parts of the respective procedures have been modified accordingly (in order to deal with this grammar extension). LSCFGs have been formally introduced in [WN11], where the main difference to conventional SCFGs is that the lengths of generated substructures are taken into account when learning the grammar parameters, yielding a more explicit structure model induced by the resulting length-dependent probabilistic parameters. Note that in connection with problems related to RNA structure, the idea of considering computational methods that actually depend on the lengths of particular substructures is not only motivated by biological aspects but has also been discussed or applied by other authors (see, e.g., [Mai07, NE07]).

It remains to mention that although all three sampling approaches (PF, SCFG and LSCFG based variants) need $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ storage for the generation of a statistically representative sample for an input sequence of length n, they obviously use different ways to define a distribution on the ensemble of all feasible secondary structures for the sequence. Applications to structure prediction (with respect to sensitivity and PPV, as well as to the shapes of sampled structures and predictions) showed that none of these sampling variants generally yields the most realistic results. Actually, which one of them should be preferred seems to strongly depend on the RNA type of the input sequence, but most importantly on the quality of a corresponding training set and on the performance of the thermodynamic model on such RNAs. However, if the worst-case complexity of one of these variants could be improved without significant losses in sampling quality (that is, if any of them required less time or space than the others while it sacrificed only little predictive accuracy), then the corresponding method would be undoubtably the number one choice for RNA structure prediction, outperforming most if not all computational tools for predicting the secondary structure of a single sequence.
For these reasons, the main objective of this paper is given as follows: We will consider the (L)SCFG based statistical sampling approach from [NSar, SN] in order to perform a comprehensive experimental analysis on the influence of disturbances (in the considered conditional sampling distributions) on the quality of generated sample sets. Particularly, we want to explore to what extend the quality of produced secondary structure samples for a given input sequence and the corresponding predictive accuracy decreases when different degrees of disturbances are incorporated into the needed sampling probabilities. Note that some exemplary intuitive first results and corresponding observations have already been presented and discussed in [NS], where it is strongly suggested that a much more meaningful evaluation based on more substantial results (with respect to several reasonable applications that are of great interest in connection with sampling approaches) is needed to be able to draw reliable conclusions.
Actually, the prime motivation for such a disturbance analysis lies in the following facts: Suppose both the samples and predictive results are observed to behave rather resistant even with respect to large errors in the distinct sampling probabilities (compared to the exact values). Then it seems adequate to believe that the sampling procedure does not have to calculate these probabilities in the exact way, but it may efficiently suffice if they are only (adequately) approximated. Thus, in this case it might obviously be possible to employ an approximation algorithm (or at least a heuristic method) for sampling probability calculations in order to decrease the worst-case time (and maybe also space) requirements. Furthermore, to ensure that the quality of the generated sample sets remains sufficiently high, analysis results on the effects of different disturbance levels and types should be taken into account for the development of an appropriate approximation scheme (or heuristic). From the other perspective, suppose the quality of sampled structures seems to strongly react on rather slight disturbances already. In that case, there is obviously little hope that the the worst-case complexities of the sampling method can be improved by finding a suitable heuristic procedure for the computation of the needed sampling probabilities.

The rest of this paper is organized as follows: Section 2 introduces the formal framework, including the (L)SCFG model, definitions of various types and levels of disturbances and a corresponding recursive sampling strategy that will be considered within this article. A comprehensive disturbance analysis based on exemplary RNA data and the corresponding results will follow in Section 3, where both the quality of generated sample sets and their applicability to the problem of RNA structure prediction are investigated. Notably, we not only compare different ways for extracting predictions from generated samples in order to assess the predictive accuracy, but also present results on an the abstraction level of shapes that is of great interest and relevance for biologists. Finally, Section 4 concludes the paper.

2 Preliminaries

In this section, we provide all needed information and introduce the formal framework that will be used in the sequel. We start by a recap of the relevant details of the probabilistic sampling method from [NSar, SN] and proceed with formally defining how a number of different types and levels of disturbances can be incorporated into the corresponding (L)SCFG based statistical sampling variants. Last but not least, we present a modified version of the employed sampling strategy that (contrary to the original one) manages to deal with disturbed ensemble distributions.
Note that we assume the reader to be familiar with the notions and basic concepts regarding SCFGs. A fundamental introduction on stochastic context-free languages can be found in [HF71]. Moreover, since for the understanding of this paper, no additional information on length-dependent stochastic models is needed, we refer to [WN11] for details.

2.1 Sampling Based on (L)SCFG Model

In general, probabilistic sampling based on a suitable (L)SCFG has two basic steps: The first step (preprocessing) computes the inside and outside probabilities for all substrings of a given input sequence based on the considered (L)SCFG model. The second step (structure generation) takes the form of a recursive sampling algorithm to randomly draw a complete secondary structure by consecutively sampling substructures (defined by base pairs and unpaired bases) according to conditional sampling probabilities for particular sequence fragments that strongly depend on the inside and outside values derived in step one.

2.1.1 Step One – Preprocessing

According to the traditional DPA approach for predicting RNA structure via (L)SCFGs, a particular underlying grammar, say $\mathcal{G}_r$, must be constructed to generate all possible RNA sequences of any length (i.e., the language $\mathcal{L}_r$ of all non-empty strings over the alphabet $\Sigma_{\mathcal{G}_r} := \{A, C, G, U\}$), where any derivation tree for a particular sequence $r \in \mathcal{L}_r$ corresponds to one of the feasible secondary structures (according to certain structural constraints like for instance to absence of pseudoknots, as well as with respect to preliminary defined rules for base-pairing) for r. This means any such (inevitably ambiguous) grammar $\mathcal{G}_r$ basically relies on an appropriately designed (typically unambiguous) grammar $\mathcal{G}_s$ modeling the corresponding secondary structures (i.e., the language $\mathcal{L}_s$ of all corresponding words over $\Sigma_{\mathcal{G}_s} := \{(,), \circ\}$, where $(\ )$ and $\circ$ represents any of the possible base pairs and unpaired bases, respectively, see [VC85]). For our investigations, we decided to rely on a rather elaborate (L)SCFG design, namely the exact formal language counterpart to the thermodynamic model applied in the Sfold program, which is given as follows:

Definition 2.1 ([NSar, SN])**.** The (length-dependent) SCFG $\mathcal{G}_s$ generating exactly all secondary structures is given by $\mathcal{G}_s = (\mathcal{I}_{\mathcal{G}_s}, \Sigma_{\mathcal{G}_s}, \mathcal{R}_{\mathcal{G}_s}, S)$, where $\mathcal{I}_{\mathcal{G}_s} = \{S, T, C, A, P, L, F, H, G, B, M, O, N, U, Z\}$, $\Sigma_{\mathcal{G}_s} = \{(,), \circ\}$ and for $m_h := \min_{HL} \geq 1$ and $m_s := \min_{\text{hel}} \geq 1$, $\mathcal{R}_{\mathcal{G}_s}$ contains exactly the following rules:

$p_1 : S \to T$, $\rightsquigarrow$ initiate exterior loop

$p_2 : T \to C$, $p_3 : T \to A$, $p_4 : T \to CA$, $p_5 : T \to AT$, $p_6 : T \to CAT$, $\rightsquigarrow$ shape of exterior loop

$p_7 : C \to ZC$, $p_8 : C \to Z$, $\rightsquigarrow$ strands in exterior loop

$p_9 : A \to (^{m_s} L)^{m_s}$, $\rightsquigarrow$ initiate helix

$p_{10} : P \to (L)$, $\rightsquigarrow$ extend helix

$p_{11} : L \to F$, $p_{12} : L \to P$, $p_{13} : L \to G$, $p_{14} : L \to M$, $\rightsquigarrow$ initiate any loop

$p_{15} : F \to Z^{m_h - 1} H$, $\rightsquigarrow$ start hairpin loop

$p_{16} : H \to ZH$, $p_{17} : H \to Z$, $\rightsquigarrow$ extend hairpin loop

$p_{18} : G \to BA$, $p_{19} : G \to AB$, $p_{20} : G \to BAB$, $\rightsquigarrow$ shape of bulge/interior loop

$p_{21} : B \to ZB$, $p_{22} : B \to Z$, $\rightsquigarrow$ strands in bulge/interior loop

$p_{23} : M \to UAO$, $\rightsquigarrow$ first substructure of multiple loop

$p_{24} : O \to UAN$, $\rightsquigarrow$ second substructure of multiple loop

$p_{25} : N \to UAN$, $p_{26} : N \to U$, $\rightsquigarrow$ kth substructure of multiple loop, $k \geq 3$

$p_{27} : U \to ZU$, $p_{28} : U \to \epsilon$, $\rightsquigarrow$ strands in multiple loop

$p_{29} : Z \to \circ$. $\rightsquigarrow$ unpaired base

Note that $\mathcal{G}_s$ has been parameterized to impose two relevant restrictions on the class of all feasible structures: first, a minimum length of $\min_{HL}$ for hairpin loops and second, a minimum number of $\min_{\text{hel}}$ consecutive base pairs for helices, where common choices are $\min_{HL} \in \{1, 3\}$ and $\min_{\text{hel}} \in \{1, 2\}$. However, within this work we will only consider $\min_{HL} = \min_{\text{hel}} = 1$, which corresponds to the least restrictive (yet also most unrealistic) choice and usually yields the worst sampling results (see [NSar, SN]). Moreover, the needed grammar parameters (trained on a suitable RNA structure database) are splitted into a set of *transition probabilities* $\Pr_{tr}(rule)$ for $rule \in \mathcal{I}_{\mathcal{G}_s}$ and two sets of *emission probabilities* $\Pr_{em}(r_x)$ for $r_x \in \Sigma_{\mathcal{G}_r}$ and $\Pr_{em}(r_{x_1} r_{x_2})$ for $r_{x_1} r_{x_2} \in \Sigma^2_{\mathcal{G}_r}$, i.e. for the 4 unpaired bases and the 16 possible base pairings, respectively. It should be mentioned that in the length-dependent case, these probabilities depend on the length of the subwords generated, meaning we then have to use $\Pr_{tr}(rule, len = \text{len}(rule))$, where $\text{len}(rule)$ denotes the length of a specific application of *rule* in a parse tree, which is defined as the length of the (terminal) subword eventually generated from *rule*. Accordingly, we need to consider $\Pr_{em}(r_x, len = 1)$ and $\Pr_{em}(r_{x_1} r_{x_2}, len = x_2 - x_1 + 1)$, respectively. Note that for the sake of simplicity, we will omit the length (second parameter) in the sequel, hence using the same notations in either case (length-dependent or not).
However, according to [NSar, SN], the computation of all inside probabilities

$$\alpha_X(i, j) := \Pr(X \Rightarrow^*_{lm} r_i \dots r_j) \tag{1}$$

and all outside probabilities

$$\beta_X(i, j) := \Pr(S \Rightarrow^*_{lm} r_1 \dots r_{i-1}\ X\ r_{j+1} \dots r_n) \tag{2}$$

for a sequence r of size n, $X \in \mathcal{I}_{\mathcal{G}_s}$ and $1 \leq i, j \leq n$, can be done with a special variant of an Earley-style parser (such that the considered grammar does not need to be in *Chomsky normal form (CNF)*). Notably, both sampling variants (length-dependent or not) can be implemented to require $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory for this preprocessing step.

2.1.2 Step Two – Random Structure Generation

Once the preprocessing is finished, different strategies may be employed for realizing the recursive sampling step. In general, for any sampling decision (for example choice of a new base pair), a particular strategy relies on the respective set of all possible choices that might actually be formed on the currently considered fragment of the input sequence. Any of these sets contains exactly the mutually exclusive and exhaustive cases as defined by the alternative productions (of a particular intermediate symbol) of the underlying grammar. The corresponding random choice is then drawn according to the resulting conditional sampling distribution (for the considered sequence fragment). This means the respective sampling distributions are defined by the inside and outside values derived in step one (providing information on the distribution of all possible choices according to the actual input sequence) and the grammar parameters (transition probabilities).
In this work, we will only consider the well-established strategy from [NSar, SN], which is also implemented in the corresponding second step of the physics-based sampling algorithm underlying the popular Sfold tool. Basically, a secondary structure is sampled recursively by starting with the entire RNA sequence and consecutively computing the adjacent substructures (single-stranded regions and paired substructures) of the exterior loop (from left to right), where any paired substructure is completed by successively folding other loops. In fact, the base pairs and unpaired base(s) are successively sampled according to conditional probability distributions for the considered fragment, given a partially formed structure.

For example, suppose fragment $R_{i,j} := r_i \dots r_j$ of input sequence r, $1 \leq i, j \leq n = |r|$, is to be folded, where it is known that the resulting substructure on $R_{i,j}$ must correspond to a (valid) derivation of a particular intermediate symbol $X \in \mathcal{I}_{\mathcal{G}_s}$ (according to the partially formed structure). Then, the strategy considers the corresponding set $acX(i, j)$ of all choices for (valid) derivations of X on $R_{i,j}$, which actually correspond to all possible substructures on $R_{i,j}$ (the mutually exclusive and exhaustive cases for X on $R_{i,j}$). Under the assumption that the alternatives for intermediate symbol X are equal to $X \to Y$ and $X \to VW$, this set is defined as follows:

$$acX(i, j) := acX_Y(i, j) \cup acX_{VW}(i, j), \tag{3}$$

where

$$\begin{aligned} acX_Y(i,j) &:= \{prob \mid prob = \beta_X(i,j) \cdot \alpha_Y(i,j) \cdot \Pr_{tr}(X \to Y) \neq 0\} \\ &= \{\beta_X(i,j) \cdot prob \mid \beta_X(i,j) \neq 0 \text{ and } prob = \alpha_Y(i,j) \cdot \Pr_{tr}(X \to Y) \neq 0\} \end{aligned}$$

and

$$\begin{aligned} acX_{VW}(i,j) &:= \{\{k, prob\} \mid i \leq k \leq j \text{ and } prob = \beta_X(i,j) \cdot \alpha_V(i,k) \cdot \alpha_W(k+1,j) \cdot \Pr_{tr}(X \to VW) \neq 0\} \\ &= \{\{k, \beta_X(i,j) \cdot prob\} \mid i \leq k \leq j \text{ and } \beta_X(i,j) \neq 0 \text{ and} \\ &\qquad prob = \alpha_V(i,k) \cdot \alpha_W(k+1,j) \cdot \Pr_{tr}(X \to VW) \neq 0\}. \end{aligned}$$

Consequently, we have to sample from the corresponding conditional probability distribution induced by $acX(i,j)$, that is the random choice is drawn according to the following set of sampling probabilities:

$$\left\{ \frac{prob}{norm} \mid prob \in acX_Y(i,j) \text{ or } \{k, prob\} \in acX_{VW}(i,j) \right\}, \tag{4}$$

where obviously,

$$\sum_{prob \in acX_Y(i,j)} \frac{prob}{norm} + \sum_{\{k,prob\} \in acX_{VW}(i,j)} \frac{prob}{norm} = 1 \tag{5}$$

must hold, which can in general easily be guaranteed by using $norm = \beta_X(i,j) \cdot \alpha_X(i,j)$. However, if there may occur inconstancies in the distribution induced by the underlying grammar model (for example if a particular implementation faces problems that arise from numerical imprecisions or if the distribution has been deliberately disturbed as we intend to do in the sequel), we should instead use

$$\begin{aligned} norm &= \sum_{prob \in acX_Y(i,j)} prob + \sum_{\{k,prob\} \in acX_{VW}(i,j)} prob \\ &= \beta_X(i,j) \cdot \left(\sum_{\beta_X(i,j) \cdot prob \in acX_Y(i,j)} prob + \sum_{\{k, \beta_X(i,j) \cdot prob\} \in acX_{VW}(i,j)} prob \right) \\ &= \beta_X(i,j) \cdot \left(\alpha_Y(i,j) \cdot \Pr_{tr}(X \to Y) + \sum_{i \leq k \leq j} \alpha_V(i,k) \cdot \alpha_W(k+1,j) \cdot \Pr_{tr}(X \to VW) \right) \\ &= \beta_X(i,j) \cdot norm_\alpha, \end{aligned}$$

which then ensures that the corresponding sampling probabilities still sum up to unity, such that they indeed define a conditional probability distribution).

Note that the sampling strategy effectively works conform with the SCFG model, which means that it actually samples one of the possible parse tress of the given input sequence by randomly drawing one of the respective mutually exclusive and exhaustive cases (corresponding to the distinct grammar rules with same premise) at any point in the already partially constructed parse tree in order to generate one of the possible subtrees for the given input sequence (corresponding to one the possible substructures on the considered sequence fragment, which is currently being folded recursively).

Hence, according to the sampling process, we could have never gotten to a point where we have to consider all mutually exclusive and exhaustive cases for a particular premise $X \in \mathcal{I}_{\mathcal{G}_s}$ on an actual sequence fragment $R_{i,j}$, $1 \leq i,j \leq n$, if the grammar could not derive the sentential form $r_1 \dots r_{i-1} X r_{j+1} \dots r_n$ from the start symbol (axiom) $S \in \mathcal{I}_{\mathcal{G}_s}$, that is if the outside value $\beta_X(i,j)$ would be equal to 0. This in fact means that the respective probability distribution (conditioned on the considered fragment $R_{i,j}$) from which the strategy randomly samples one of the possible substructures (one valid subtree of the already partially constructed parse tree) is not influenced by the corresponding outside probability, due to the fact that $\beta_X(i,j) > 0$ indeed only represents a scaling factor common to all sampling probabilities for the relevant mutually exclusive and exhaustive cases. For this reason, we can obviously without loss of information remove the outside values from the definitions of the needed sampling probabilities.

The correctness of this simplification can easily be formally proven by considering the above defined set $acX(i,j)$ of all choices for possible derivations of intermediate symbol X on sequence fragment $R_{i,j}$. In fact, the sampling strategy randomly draws one of the elements from $acX(i,j)$ according to the corresponding distribution induced by normalizing the probabilities of the elements in $acX(i,j)$ such that they sum up to unity. Particularly, we have

$$1 = \sum_{\beta_X(i,j) \cdot prob \in acX_Y(i,j)} \frac{\beta_X(i,j) \cdot prob}{\beta_X(i,j) \cdot norm_\alpha} + \sum_{\{k, \beta_X(i,j) \cdot prob\} \in acX_{VW}(i,j)} \frac{\beta_X(i,j) \cdot prob}{\beta_X(i,j) \cdot norm_\alpha}$$

$$= \frac{1}{norm_\alpha} \cdot \left(\sum_{\beta_X(i,j)\cdot prob \in acX_Y(i,j)} prob + \sum_{\{k,\beta_X(i,j)\cdot prob\} \in acX_{VW}(i,j)} prob \right)$$
$$= \frac{1}{norm_\alpha} \cdot \left(\sum_{prob \in acX_Y(i,j)} \frac{prob}{\beta_X(i,j)} + \sum_{\{k,prob\} \in acX_{VW}(i,j)} \frac{prob}{\beta_X(i,j)} \right),$$

since $\beta_X(i,j) \neq 0$ holds (due to the definitions of $acX_Y(i,j)$ and $acX_{VW}(i,j)$).

Formal definitions of all corresponding sets $acX(i,j)$, $X \in \mathcal{I}_{\mathcal{G}_s}$ and $1 \leq i,j \leq n$, that are considered by the recursive sampling strategy for any input sequence of length n, including formulae for deriving the respective conditional sampling probabilities, can be found in Section Sm-I[1]. Notably, all those formulae only depend on some of the parameters of the underlying (L)SCFG model and the corresponding inside values, such that after a preprocessing of the given sequence (which includes the complete inside computation and needs $\mathcal{O}(n^3)$ time in the worst-case), a random candidate structure can be generated in $\mathcal{O}(n^2)$ time.

2.2 Considered Disturbance Types and Levels

Obviously, under the assumption of a particular (L)SCFG model (trained beforehand on arbitrary RNA data), the most straightforward way for improving the performance of the corresponding overall sampling algorithm seems to be by reducing the worst-case complexity of the inside calculations. Therefore, we decided to quantify to which extend the algorithm reacts to different types and degrees of disturbances incorporated into the considered inside probabilities in order the get evidence if it could actually be possible to find a corresponding approximation algorithm (or at least an appropriate heuristic method) that eventually requires less time but causes only acceptable losses in accuracy. In fact, with respect to developing a suitable heuristic method to be applied in practice, it is necessary to know about the effects of different disturbance levels and types to get an idea on how precisely the respective values need to be approximated in order to guarantee sufficiently good results and to find out which types of errors pose fundamental problems and which ones are negligible.
For these reasons, given an arbitrary input sequence r of length n, we decided to consider (more or less) skewed inside probabilities[2]

$$\widehat{\alpha}_X(i,j) := \max(\min(\alpha_X(i,j) + \alpha_X^{err}(i,j), 1), 0), \tag{6}$$

for $X \in \mathcal{I}_{\mathcal{G}_s}$ and $1 \leq i,j \leq n$, rather than the corresponding correct values $\alpha_X(i,j)$ (obtained in the preprocessing step for r) for defining the needed sampling probabilities. More precisely, we want to incorporate different stages of (more or less grave) randomly chosen errors into particular inside values for the given sequence, that is into preliminary chosen subsets of the set of all precomputed inside probabilities $\alpha_X(i,j)$, $X \in \mathcal{I}_{\mathcal{G}_s}$ and $1 \leq i,j \leq n$. Note that is actually suffices to consider $X \in \mathcal{I}_{\mathcal{G}_s}^{\alpha} := \{T, C, A, P, F, G, B, M, O, N, U\} \subset \mathcal{I}_{\mathcal{G}_s}$, since only those intermediate symbols are needed for defining the diverse sampling probabilities that are used by the employed sampling strategy for obtaining the distinct conditional distributions for drawing particular random choices.
However, to reach our previously declared goal, we decided to draw $\alpha_X^{err}(i,j)$ (uniformly) at random from either of the following sets:

$$func_{\mathcal{I}}^{win,op}(prob) := \begin{cases} Interval(func), & \text{if } X \in \mathcal{I} \subseteq \mathcal{I}_{\mathcal{G}_s}^{\alpha} \text{ and } [(j-i+1 > win \text{ and } op = +) \text{ or} \\ & \qquad (j-i+1 \leq win \text{ and } op = -)], \\ \{0\}, & \text{else,} \end{cases} \tag{7}$$

such that only inside values of particularly chosen intermediate symbols that lie outside ($op = +$) or within ($op = -$) a considered window of preliminary fixed size are actually disturbed, that is only for those values $\widehat{\alpha}_X(i,j) \neq \alpha_X(i,j)$ might result. Note that in the sequel, we will basically consider either

$$func^{win,op}(prob) := func_{\mathcal{I}_{\mathcal{G}_s}^{\alpha}}^{win,op}(prob) \tag{8}$$

[1] All references starting with Sm are references to the supplementary material available at `http:///wwwagak.cs.uni-kl.de/publications/`.

[2] Note that the function $\max(\min(x,1),0) = \min(\max(x,0),1)$ ensures that the resulting value is still a probability, i.e. a real value from $[0,1]$.

(i.e., disturbances only inside or outside fix-sized window, but for all intermediate symbols),

$$func_{\mathcal{I}}(prob) := func_{\mathcal{I}}^{n,+}(prob) = func_{\mathcal{I}}^{-1,-}(prob) \tag{9}$$

(i.e., errors for all subword lengths, but only for particular intermediate symbols), or simply

$$func(prob) := func_{\mathcal{I}_{\mathcal{G}_s}^{\alpha}}^{n,+}(prob) = func_{\mathcal{I}_{\mathcal{G}_s}^{\alpha}}^{-1,-}(prob) \tag{10}$$

(i.e., disturbances on all considered inside values).
Moreover, $func \in \{\text{mep}, \text{fep}, \text{mev}, \text{fev}\}$ denotes the actual disturbance type. Principally, we distinguish between two degrees of errors: relative and absolute ones. To generate relative errors, we might either use $func = \text{mep}$ (which stands for *maximum allowed error percentage*, with respect to the corresponding correct value) or $func = \text{fep}$ (for *fixed error percentage*, which is ought to force greater and hence more severe random errors). Formally, this means that either

$$Interval(\text{mep}) := [-prob \cdot \alpha_X(i,j), +prob \cdot \alpha_X(i,j)] \tag{11}$$

or

$$Interval(\text{fep}) := \{-prob \cdot \alpha_X(i,j), +prob \cdot \alpha_X(i,j)\} \tag{12}$$

might be employed for randomly drawing a relative error $\alpha_X^{err}(i,j)$, where $prob \in (0,1]$ indeed defines the desired percentage. In order to randomly choose an absolute error $\alpha_X^{err}(i,j)$ for obtaining a (potentially) disturbed probability $\widehat{\alpha}_X(i,j)$, we might equivalently consider either

$$Interval(\text{mev}) := [-prob, +prob] \tag{13}$$

or

$$Interval(\text{fev}) := \{-prob, +prob\}, \tag{14}$$

with $prob \in (0,1]$ being a preliminary fixed value. This means we may use $func = \text{mev}$ (which stands for *maximum allowed error value*, independent on the corresponding correct value) and $func = \text{fev}$ (for *fixed error value*, usually resulting in more grave disturbances) for causing absolute disturbances.
Note that random errors on all outside probabilities $\beta_X(i,j)$, $X \in \mathcal{I}_{\mathcal{G}_s}$ and $1 \leq i, j \leq n$, could be generated in basically the same way, but since those values can be deliberately excluded from the definition of sampling probabilities (according to the employed sampling strategy), this is obviously not necessary for the subsequent investigations.
Finally, it should be clear that for $func \in \{\text{mep}, \text{fep}\}$ (resulting in relative errors), only the magnitudes of the corresponding sampling probabilities (with respect to the implied skewed conditional sampling distributions) change, such that the exact same structures are possible as in the undisturbed case. Hence, we might expect that only the consideration of sufficiently large percentages $prob \in (0,1]$ for generating errors according to $func_{\mathcal{I}}^{win,op}(prob)$ can cause an actual shifting in the ensemble distribution, resulting in significant quality losses. The contrary holds for absolute errors created according to $func_{\mathcal{I}}^{win,op}(prob)$ with $func \in \{\text{mev}, \text{fev}\}$. In fact, since the (cardinalities of the) respective sets of relevant sampling choices implied by the skewed ensemble distribution generally differ (to a more or less severe extent) from the corresponding exact ones, it must be expected that only rather small fixed error values of $prob \in (0,1]$ are reasonable choices for our purpose. However, since for distinct subword lengths $j-i+1$, $1 \leq i, j \leq n$, the corresponding probabilities $\alpha_X(i,j)$ for any $X \in \mathcal{I}_{\mathcal{G}_s}^{\alpha}$ usually imply different orders of magnitudes[3], it seems practically impossible to tell how to find an appropriate fixed error value for creating absolute disturbances.

2.3 Resulting Modified Sampling Strategy

It should be clear that after the desired errors (according to any of the previously specified variants of either mep, fep, mev or fev) have been incorporated into the precomputed exact inside (and outside) values for a given sequence, the needed conditional sampling distributions (as considered by a particular strategy) are induced by the exact grammar parameters and the disturbed inside (and outside) probabilities for that sequence. This, however, might create the need to (slightly) modify the respective particularly employed sampling strategy such that it finally gets capable to deal with these skewed distributions.
As for this work, consider the previously sketched recursive sampling strategy from [NSar, SN]. Without any errors in the conditional probability distributions (i.e. by using the exact probabilistic parameters for

[3] In general, longer words tend to be generated with smaller probability since we have to apply more grammar rules, each implying a factor (typically) less than 1 to the probability.

the given input sequence, particularly the corresponding inside values), it always successfully generates the sampled loop type for a considered sequence fragment. For example, suppose the sampling procedure decides that base pair $r_i.r_j$ should close a multiloop, then the sequence fragment $R_{i+1,j-1} := r_{i+1} \dots r_{j-1}$ is guaranteed to be folded into an admissible multiloop that by definition contains at least two helical regions radiating out from this loop. However, by using disturbed sampling probabilities (given by the exact parameters of the underlying (L)SCFG model and disturbed inside values for input sequence r, derived by incorporating any sort of errors), the sampling algorithm may choose to form a particular substructure on the fragment $R_{i+1,j-1}$, although this would actually not be possible.
Therefore, we had to slightly modify the sampling procedure such that in any case where the chosen substructure type can not be successfully generated, it settles for the partially formed substructure. That is, it either leaves the complete fragment unpaired (if the desired base pairs could not be sampled at all), or else it for example only creates a bulge/interior loop although a multiloop should have been constructed (but only one helix has been successfully sampled). The resulting modified versions of the distinct sampling steps (in pseudocode) are given in Section Sm-I, Figure 1 gives a schematic overview of the overall sampling process.
Note that alternatively, the algorithm could have been modified to revise any decisions that lead to incompletely generated substructures, resulting in some sort of backtracking procedures that obviously would have to be applied in order to sample more realistic overall structures for a given RNA sequence. However, as this effectively results in much more complex modifications and eventually yields significant losses in performance, we opted for the simpler and more straightforward first variant to get rid of the described problem.

3 Analysis of the Influence of Disturbances

The aim of this section is to perform a comprehensive experimental analysis on the influence of disturbances (in the ensemble distribution for a given input sequence) on the quality of sample sets generated by the (L)SCFG based statistical sampling approach from [NSar, SN]. In fact, we want to explore to what extend the quality of produced secondary structure samples for a given input sequence and the corresponding predictive accuracy decreases when different degrees of errors are incorporated into the needed sampling probabilities.

3.1 RNA Structure Data

For our examinations, we decided to consider different sets of trusted RNA secondary structure data for which the (L)SCFG based sampling approach yields good quality results when no disturbances are included in the respective sampling distributions for a given sequence. Therefore, we took the same tRNA database (of 2163 distinct tRNA structures with lengths in $[64, 93]$ and about 76 on average, derived from [SHB+98]) and the identical 5S rRNA data set (of 1149 distinct sequences with lengths in $[102, 135]$ and about 119 on average, retrieved from [SBEB02]) as collected in [NSar]. These two rich data sets of trusted RNA secondary structures will be exclusively used as the basis for the following applications, such that the results can easily be opposed to the corresponding ones presented in [SN].

3.2 Probability Profiling for Specific Loop Types

A statistical sample of all possible secondary structures for a given RNA sequence can be used for sampling estimates of the probabilities of any structural motifs. Actually, *probability profiling* for unpaired bases within particular loop types can easily be applied for this purpose. In principle, for each nucleotide position i, $1 \leq i \leq n$, of a given sequence of length n, one computes the probabilities that i is an unpaired base within a specific loop type. These probabilities are given by the observed frequencies in a random sample set.
Since this application is rather intuitive, we decided to use it as a starting point for our disturbance analysis. Particularly, we derived a number of statistical samples for the well-known *Escherichia coli* tRNA^{Ala} sequence by applying the sampling strategy from Section 2.3 on the basis of diverse sets of probabilistic parameters (inside probabilities disturbed according to several particular variants as defined in Section 2.2) for that sequence and calculated corresponding probability profiles. All relevant results are displayed in Figures 5 to 18 of Section Sm-II. Some of the potentially most interesting ones are presented in Figures 2 to 4.

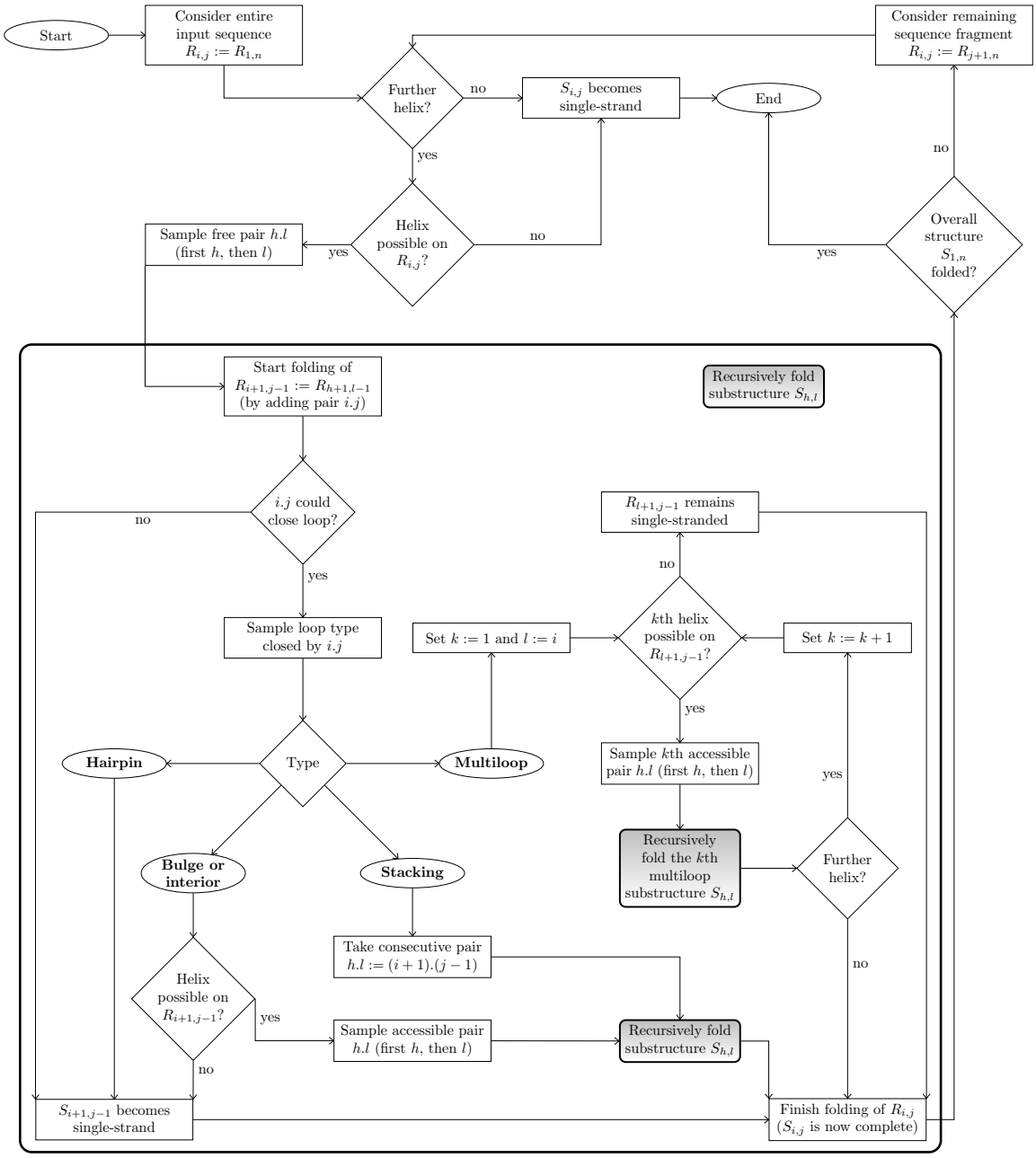
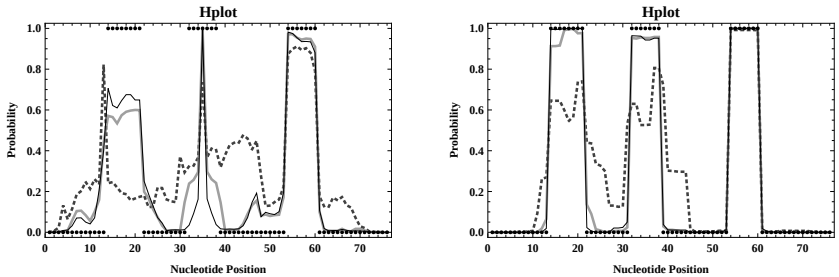


Figure 1: Flowchart for recursive sampling of an RNA secondary structure $S_{1,n}$ for a given input sequence r of length n according to an inherently controlled strategy with predetermined order, similar to that of [DL03, NSar]).

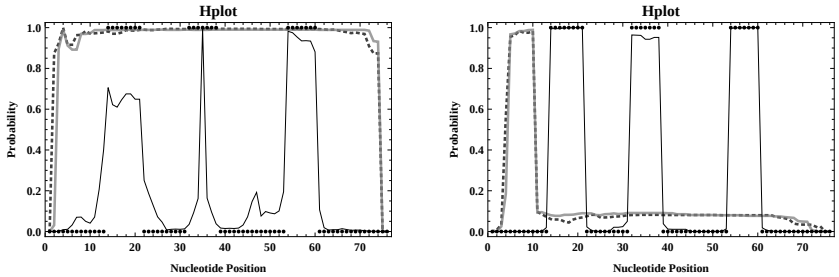
3.2.1 Errors on All Values

Let us first consider the profiles displayed in Figure 2 (and in Figures 5 and 6). Obviously, even if large relative errors on all inside probabilities and hence on the needed conditional sampling probabilities are generated, the sampled structures still exhibit the typical cloverleaf structure of tRNAs, especially for the length-dependent sampling approach where relative disturbances seem to have no significant negative effect on the sampling quality (see Figure 2a). However, Figure 2b perfectly demonstrates that if the disturbances have been created by adding absolute errors to all inside values, then – even for rather small absolute error values – the resulting samples obtained with both the SCFG and LSCFG approach are useless.

Note that for any given input sequence, it seems to be usually much more important for the employed sampling strategy to be able to identify which ones of the (combinatorially) possible substructures can actually be (validly) formed on the considered sequence fragment rather than to know their exact prob-

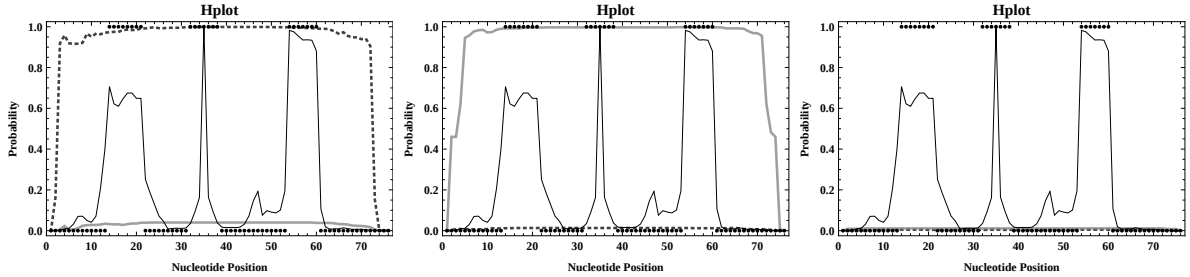


(a) Relative errors according to mep(*prob*) (thick gray lines) and fep(*prob*) (thick dotted darker gray lines), considering percentage $prob = 0.99$.

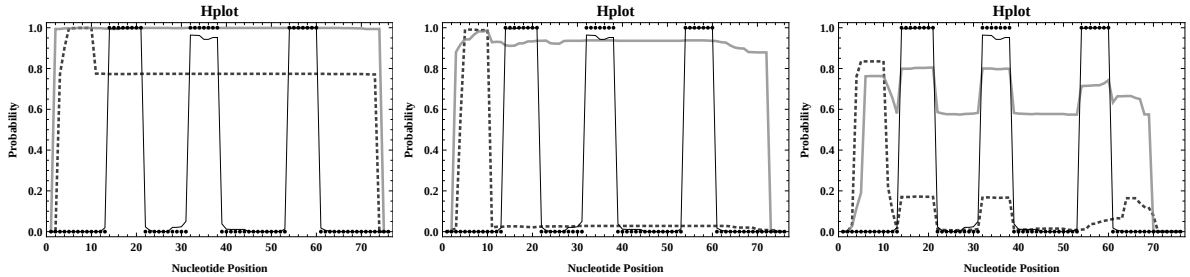


(b) Absolute errors according to mev(*prob*) (thick gray lines) and fev(*prob*) (thick dotted darker gray lines), using fixed value $prob = 10^{-9}$.

Figure 2: Hairpin loop profiles for *E.coli* tRNA^{Ala}, calculated from a random sample of 1000 structures generated with the SCFG (figures on the left) and LSCFG (figures on the right) approach, respectively (under the assumption of the less restrictive grammar parameters $\min_{\text{hel}} = 1$ and $\min_{HL} = 1$). The exact (undisturbed) results are displayed by the thin black lines, and the correct hairpin loops in *E.coli* tRNA^{Ala} are illustrated by the black points.



(a) Results for traditional SCFG model.

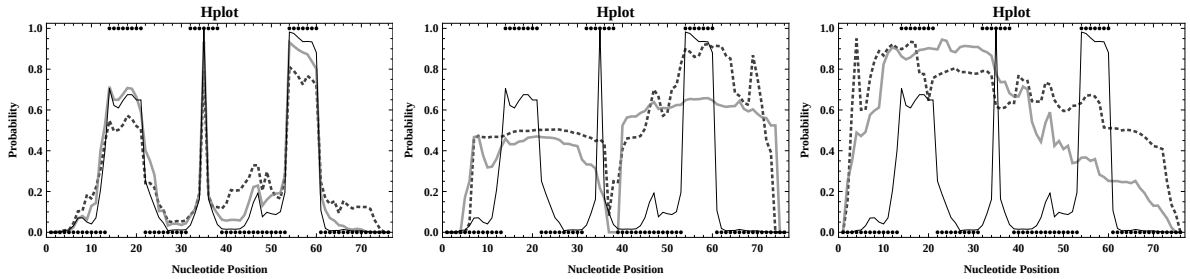


(b) Results for LSCFG model.

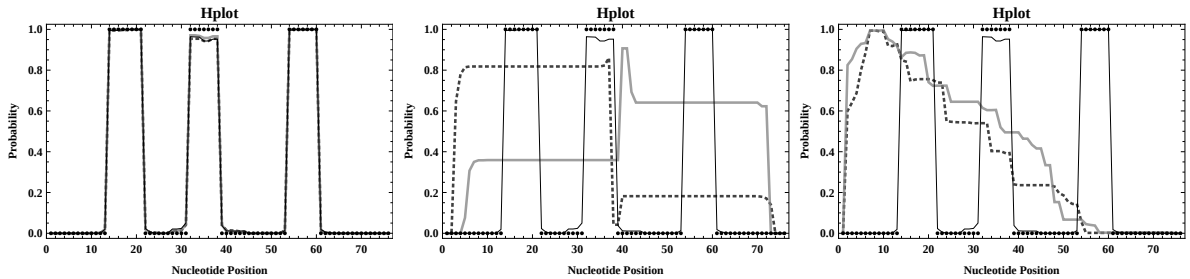
Figure 3: Hairpin loop profiles corresponding to those presented in Figure 2b, where absolute errors were derived according to mev$^{win,+}$(*prob*) (thick gray lines) and fev$^{win,+}$(*prob*) (thick dotted darker gray lines), respectively, with $prob = 10^{-9}$ and $win \in \{15, 38, 60\}$ (figures from left to right).

abilities (according to the conditional distribution for the respective fragment), for two contrary reasons: First, in order to avoid drawing practically impossible choices, which later forces it to leave the considered sequence fragment (at least partially) unpaired[4]. Second, for ensuring that none of the actually valid

[4]If those decisions are not revised by employing backtracking procedures, see the description of the modifications incor-



(a) Results for traditional SCFG model.



(b) Results for LSCFG model.

Figure 4: Hairpin loop profiles corresponding to those presented in Figure 2b, where absolute errors were derived according to $\mathrm{mev}^{win,-}(prob)$ (thick gray lines) and $\mathrm{fev}^{win,-}(prob)$ (thick dotted darker gray lines), respectively, with $prob = 10^{-9}$ and $win \in \{15, 38, 60\}$ (figures from left to right).

choices is prohibited during the folding process, such that the sampling procedure might inevitably prefer other (potentially even impossible) substructures.
Consequently, in order to prevent a decline in accuracy of generated structures and a reduction of the overall sampling quality, it seems to be of great importance that the sampling strategy is capable of distinguishing between inside values and especially sampling probabilities that are equal and unequal to zero according to the exact (undisturbed) ensemble distribution for the given input sequence. By adding absolute errors, however, inside or sampling probabilities being equal (unequal) to zero in the exact case might often become unequal (equal) to zero according to the resulting skewed (disturbed) distributions, whereas by incorporating relative errors, all considered inside and sampling probabilities obviously stay equal or unequal to zero (as in the exact case), which intuitively explains the basic observations made from Figure 2.

3.2.2 Relevant Sampling Probabilities

Nevertheless, in order to draw more detailed conclusions, we counted and compared the relevant (i.e., greater than zero) inside and sampling probabilities that were considered for obtaining the profiles presented in Figure 2. The results are collected in Tables 5 and 6 of Section Sm-II.
First, it seems obvious that due to the more explicit length-dependent version of the considered grammar parameters (length-dependently trained transition and emission probabilities), there should generally result a much smaller number of relevant inside values and sampling probabilities when applying the LSCFG model rather than the conventional one. Tables 5 and 6 exemplarily prove this intuitive assumption. Note that this effect might indeed be responsible for the observation that the LSCFG based sampling approach reacts considerably less to large relative errors than the conventional length-independent variant, as indicated by Figure 2a: less inside probabilities are effectively disturbed, such that the extend of the relative errors imposed on the corresponding sampling probabilities is inevitably smaller for the LSCFG variant than for the length-independent one.
Moreover, there are much more relevant exact inside and sampling probabilities than corresponding relevant disturbed values for basically any (intermediate) symbol when considering the traditional SCFG model, whereas for the LSCFG variant the contrary holds, that is generally way more inside and sampling probabilities are relevant in the disturbed cases than in the exact case. Actually, in both cases (length-dependent and not), the numbers of relevant disturbed inside values $\widehat{\alpha}_X(i,j)$, $1 \leq i,j \leq n$, are rather similar (for basically all $X \in \mathcal{I}^{\alpha}_{\mathcal{G}_s}$), in contrast to the numbers of relevant sampling probabilities

porated into the sampling algorithm in order to deal with such situations as given in Section 2.3.

(corresponding to valid choices for substructures) for the distinct sampling steps which are in general to a large extend greater when using the traditional SCFG approach than under the assumption of the corresponding LSCFG model. This behavior might be the reason for the fundamental differences in the resulting (albeit useless) loop profiles presented in Figure 2b.
Finally, it remains to mention that under the assumption of the conventional SCFG model, it happens that for any $X \in \mathcal{I}_{\mathcal{G}_s}^{\alpha}$, most inside values are relevant in both the exact and the disturbed case, whereas significantly less are relevant only in the exact case and very few are only relevant in the disturbed case (see Figure 5a). Considering the LSCFG variant, however, for any $X \in \mathcal{I}_{\mathcal{G}_s}^{\alpha}$ the least inside values are relevant only in the exact case, as indicated by Figure 5b. Obviously, this seems to be the natural consequence of the previously formulated observations.

3.2.3 Errors Only on Particular Values

Now, in an attempt to find out in which cases particular absolute errors have a very significant (negative) impact on the resulting sampling quality and to identify potentially existing situations where they barely influence the output of the applied statistical sampling algorithm, we want to consider some of the more specialized variants for generating absolute disturbances (as defined in Section 2.2). The corresponding profiles are basically shown in Figures 3 and 4 (as well as in Figures 7 to 18).
Notably, even if absolute disturbances may only occur for inside values $\alpha_X(i,j)$, $X \in \mathcal{I}_{\mathcal{G}_s}^{\alpha}$, with $j-i+1 > win$ (i.e., for substructure lengths greater than a particular fixed value win), the corresponding sampling results are of no practical use at all (see Figure 3). In fact, there seem to be no noticeable improvements when considering increasing values of win, which means that even if more inside values $\alpha_X(i,j)$, $X \in \mathcal{I}_{\mathcal{G}_s}^{\alpha}$, namely those satisfying $j-i+1 \leq win$, are guaranteed to be exact (contain no relative or absolute errors), the resulting samples might not be expected to gain in quality. This observation is actually unfortunate as regards the derivation of a corresponding heuristic version of the inside algorithm, since the inside computation starts by calculating the respective values for small sequence fragments and subsequently considers larger ones, meaning the straightforward approach of deriving all values $\alpha_X(i,j)$, $X \in \mathcal{I}_{\mathcal{G}_s}^{\alpha}$, with $j-i+1 \leq win$ in the exact way and approximating only the remaining ones (i.e., using a constant window size win for exact calculations) might not yield results of acceptable quality if absolute errors can not be ruled out (completely).
Nevertheless, as we can see from Figure 4, if absolute disturbances may only occur for inside values $\alpha_X(i,j)$, $X \in \mathcal{I}_{\mathcal{G}_s}^{\alpha}$, with $j-i+1 \leq win$ (i.e., for substructure lengths less than or equal to a particular fixed value win), the corresponding sampling results might actually be of acceptable quality, but seemingly only for rather small values of win. This means in order to obtain a practically applicable heuristic, it seems a good idea to consider a constant (small enough) window of size win and compute all values $\alpha_X(i,j)$, $X \in \mathcal{I}_{\mathcal{G}_s}^{\alpha}$, with $j-i+1 > win$ in the exact way, thus approximating only those satisfying $j-i+1 \leq win$. However, due to the contrary course of action of traditional inside calculations, this approach can obviously not be realized. Consequently, this observation does not contribute to developing an appropriate heuristic variant of the preprocessing step, but it actually motivates the construction of an innovative sampling strategy that takes on a reverse sampling direction (that constructs substructures in an inside-to-outside fashion, contrary to the generation of corresponding derivation trees according to the underlying grammar).
Finally, for the sake of completeness, it should be noted that by incorporating absolute errors (for all subword lengths) only for any of the distinct intermediate symbols $X \in \mathcal{I}_{\mathcal{G}_s}^{\alpha}$ at once (i.e., by disturbing only the inside values $\alpha_X(i,j)$, $1 \leq i,j \leq n$, for a particular $X \in \mathcal{I}_{\mathcal{G}_s}^{\alpha}$), we found out that some are more sensitive with respect to disturbances in the underlying ensemble distribution than others (see Figures 11 to 18 of Section Sm-II). In principle, the strongest (negative) reactions to the influence of the generated absolute errors were observed for symbols T, C, A, F (for the traditional SCFG model), G and U, whereas less severe quality losses basically resulted for intermediates M, O, N and P. Moreover, for two symbols, namely F (for the LSCFG model) and B, we recognized no noticeable impact of the caused disturbances to the accuracy of the generated sample sets.

3.3 Prediction Accuracy – Sensitivity and PPV

In connection with sampling approaches, there exist diverse (more or less) efficient well-defined principles for extracting a particular structure prediction from a generated set of candidate structures for a given input sequence. In fact, under the condition that a corresponding folding can be calculated in $\mathcal{O}(n^3)$ time and with $\mathcal{O}(n^2)$ storage (i.e., has the same worst-case complexities as the preprocessing step), the statistical sampling method considered in this work can easily be applied to single sequence secondary

structure prediction without significant losses in performance and its predictive power can easily be measured by means of *sensitivity* (Sens.) and *positive predictive value* (PPV)[5]. Briefly, these two common measures are widely used in order to quantify the accuracy of RNA secondary structure prediction methods and are usually defined as follows (see e.g. [BBC$^+$00]):

- Sens. is the relative frequency of correctly predicted pairs among all position pairs that are actually paired in a stem of native foldings, whereas
- PPV is defined as the relative frequency of correctly predicted pairs among all position pairs that were predicted to be paired with each other.

Formally, they are given by Sens. $= TP \cdot (TP + FN)^{-1}$ and PPV $= TP \cdot (TP + FP)^{-1}$, where TP is the number of correctly predicted base pairs (*true positives*), FN is the number of base pairs in the native structure that were not predicted (*false negatives*) and FP is the number of incorrectly predicted base pairs (*false positives*).

In order to investigate to what extend the accuracy of predicted foldings changes when different dimensions of relative disturbances are incorporated into the needed sampling probabilities, we decided to perform a series of cross-validation experiments based on the same partitions of the tRNA and 5S rRNA databases into 10 approximately equal-sized folds, respectively, as considered in [NSar, SN]. In particular, for each sequence, we generated several sample sets on the basis of different relative error types and values, where from each of the produced samples, we derived corresponding predictions according to a number of competing reasonable selection principles and construction schemes (which can all be applied to the respective sample set without increasing the worst-case complexity of the overall algorithm).
Briefly, we employed two different well-defined selection procedures in order to identify one particular structure from the produced sample as prediction: First, we picked the most likely secondary structure (i.e., the one with the highest probability among all feasible structures for the input sequence according to the induced (L)SCFG model), in strong analogy to traditional SCFG based probabilistic structure prediction methods. This choice will be denoted by *most probable (MP)* structure in the sequel. Additionally, as one of the most straightforward and reasonable choices for statistically representative samples of the overall structure ensemble, we took the most frequently sampled folding (i.e., the one with the highest number of occurrences among all candidate structures within the generated sample set), which will be named *most frequent (MF)* structure in the sequel.
Note that if the samples are indeed representative with respect to the underlying ensemble distribution (i.e., if a sufficiently large number of candidate foldings is randomly generated on the basis of the corresponding conditional probability distributions considered by the employed strategy), then these two predictions should be rather identical in most cases, at least if no disturbances are considered (i.e., under the condition that the exact inside probabilities are used for deriving the respective conditional sampling distributions). In fact, any representative set of candidate structures for a given input sequence obtained by (L)SCFG based statistical sampling obviously reflects the probability distribution on all feasible foldings of that sequence which strongly depends on the corresponding inside probabilities. Thus, if the preprocessed inside values contain any errors, then the MF structure of a particular statistically representative sample set corresponds to the most likely folding of the given sequence with respect to the skewed ensemble distribution induced by the disturbed inside values, whereas the MP structure of that sample is indeed equal to the most likely folding (among all generated candidate structures) with respect to the exact ensemble distribution[6]. Hence, the results for MP and MF structure predictions might differ in the disturbed cases, especially as the gravity of generated disturbances grows.
However, we decided to additionally apply two different commonly used construction schemes for computing a new structure as predicted folding, where the predicted structure itself must not necessarily be contained in the given sample. Particularly, we first determined a *maximum expected accuracy (MEA)* structure of the generated sample set as defined in [NSar], which maximizes the number of correctly unpaired and paired positions with respect to the true folding and is computed on the basis of the considered sample (rather than on the basis of the entire structure ensemble for the sequence as done for example in the Pfold [KH03] and CONTRAfold [DWB06] programs). Furthermore, we calculated the unique consensus structure of the produced sample, called the *centroid* structure, which effectively reflects the overall behavior of the sample set and is actually formed by all base pairs that occur in more than 50% of

[5]Note that the positive predictive value is often called *specificity*, although this measure formally obeys to a slightly different definition

[6]This is due to the fact that the probability of a particular folding of a given RNA sequence (i.e., the probability of the corresponding derivation tree) depends only on the considered set of grammar parameters (transition and emission probabilities).

the sampled structures (for details, see e.g. [DCL05]). Note that for similar reasons as discussed above for MF structure predictions, MEA and centroid structures obtained from statistically representative sample sets can only reflect the skewed ensemble distribution rather than the exact one in the disturbed case.
Last but not least, we derived two different sets of so-called $\gamma_{\mathrm{t-o}}$-MEA and $\gamma_{\mathrm{t-o}}$-centroid structures for the produced samples, respectively, as defined in [NSar] (in connection with sampling algorithms), where $\gamma_{\mathrm{t-o}} \in [0, \infty)$ is a trade-off parameter for controlling the sensitivity and PPV of the predicted foldings. Note that the default choice $\gamma_{\mathrm{t-o}} = 1$ serves as the neutral element with respect to the prediction, meaning the prediction is neither biased towards a better sensitivity nor to a better PPV and corresponds to the above described well-known MEA or unique centroid structure, respectively. Notably, by measuring the performance at several different settings of $\gamma_{\mathrm{t-o}}$ (i.e. by determining the (adjusted) sensitivity and PPV for various values of $\gamma_{\mathrm{t-o}}$), it becomes possible to derive a corresponding *receiver operating characteristic (ROC)* curve and to calculate the estimated *area under this curve (AUC)*, for both the MEA and the centroid prediction principle, respectively. This obviously allows for a much more informative and reliable comparison of the predictive powers of the different sampling variants than considering only the corresponding results for the default choice $\gamma_{\mathrm{t-o}} = 1$.
However, the (unadjusted) sensitivity and PPV measures obtained by considering the four different (unparameterized) prediction principles sketched above are listed in Tables 7a and 8a, where a few selected ones are presented in Table 1. The corresponding AUC values obtained by varying instances of $\gamma_{\mathrm{t-o}}$ are all collected in Tables 7b and 8b, some of them are presented in Table 2. Note that in accordance with [NSar, SN], we considered any value of $\gamma_{\mathrm{t-o}} \in \{1.25^k \mid -12 \leq k \leq -1\} \cup \{2^k \mid 0 \leq k \leq 12\}$ in order to obtain appropriate ROC curves and corresponding AUC values. Plots of some of the resulting curves can be found in Figures 19 to 22 of Section Sm-II.

Approach	Errors	MP struct.		MF struct.		MEA struct.		Centroid	
		Sens.	PPV	Sens.	PPV	Sens.	PPV	Sens.	PPV
SCFG	—	0.7818	0.8437	0.7792	0.8445	0.7324	0.8939	0.6754	0.9158
	mep(0.5)	0.7822	0.8447	0.7599	0.8370	0.7169	0.8927	0.6607	0.9140
	mep(0.99)	0.7590	0.8388	0.6768	0.8004	0.6414	0.8877	0.5817	0.9127
	fep(0.5)	0.7798	0.8440	0.7234	0.8184	0.6864	0.8896	0.6292	0.9134
	fep(0.99)	0.4101	0.7295	0.2864	0.5590	0.2532	0.7776	0.2157	0.8291
LSCFG	—	0.8545	0.9534	0.8542	0.9535	0.8335	0.9736	0.8250	0.9783
	mep(0.5)	0.8545	0.9534	0.8429	0.9524	0.8236	0.9731	0.8150	0.9773
	mep(0.99)	0.8519	0.9533	0.7988	0.9413	0.7833	0.9676	0.7735	0.9726
	fep(0.5)	0.8548	0.9536	0.8224	0.9486	0.8029	0.9707	0.7940	0.9758
	fep(0.99)	0.7530	0.9325	0.5769	0.8623	0.5668	0.9075	0.5567	0.9195

(a) For our tRNA database.

Approach	Errors	MP struct.		MF struct.		MEA struct.		Centroid	
		Sens.	PPV	Sens.	PPV	Sens.	PPV	Sens.	PPV
SCFG	—	0.4251	0.5372	0.4251	0.5363	0.3403	0.6967	0.2689	0.8044
	mep(0.5)	0.4143	0.5280	0.4160	0.5290	0.3334	0.6987	0.2643	0.8051
	mep(0.99)	0.3897	0.5227	0.3894	0.5216	0.2957	0.7069	0.2362	0.8072
	fep(0.5)	0.4055	0.5203	0.4049	0.5198	0.3209	0.7068	0.2532	0.8087
	fep(0.99)	0.2043	0.4410	0.1756	0.3788	0.1066	0.6867	0.0814	0.7666
LSCFG	—	0.8993	0.9412	0.8997	0.9409	0.8959	0.9513	0.8873	0.9574
	mep(0.5)	0.8993	0.9412	0.8909	0.9380	0.8903	0.9478	0.8819	0.9541
	mep(0.99)	0.8989	0.9414	0.8639	0.9269	0.8659	0.9408	0.8574	0.9482
	fep(0.5)	0.8993	0.9412	0.8796	0.9328	0.8798	0.9445	0.8716	0.9515
	fep(0.99)	0.8251	0.9052	0.7162	0.8375	0.7148	0.8661	0.6986	0.8879

(b) For our 5S rRNA database.

Table 1: Prediction results by means of sensitivity and PPV (computed by 10-fold cross-validation procedures, using sample size 1000 and $\min_{\mathrm{hel}} = \min_{HL} = 1$).

Let us first consider the results reported in Table 1. As we can see, the PPV is principally not affected by the different dimensions of disturbances caused according to mep(*prob*), as only in the case of MF structure prediction one can observe a slight change for the worse. However, with increasing value of

Approach	Errors	MEA struct.	Centroid
SCFG	—	0.828522	0.833894
	mep(0.5)	0.819658	0.823811
	mep(0.99)	0.786645	0.788478
	fep(0.5)	0.805999	0.807240
	fep(0.99)	0.440021	0.422778
LSCFG	—	0.936285	0.919736
	mep(0.5)	0.932121	0.916321
	mep(0.99)	0.916540	0.896024
	fep(0.5)	0.924191	0.908943
	fep(0.99)	0.752030	0.722737

(a) For our tRNA database.

Approach	Errors	MEA struct.	Centroid
SCFG	—	0.409278	0.408549
	mep(0.5)	0.401914	0.400515
	mep(0.99)	0.376683	0.375488
	fep(0.5)	0.400827	0.397566
	fep(0.99)	0.189628	0.182902
LSCFG	—	0.914801	0.918933
	mep(0.5)	0.911963	0.915503
	mep(0.99)	0.902330	0.905126
	fep(0.5)	0.906507	0.911063
	fep(0.99)	0.776239	0.777355

(b) For our 5S rRNA database.

Table 2: Prediction results by means of AUC values (computed by 10-fold cross-validation procedures, using sample size 1000 and $\min_{\text{hel}} = \min_{HL} = 1$).

mep, there results a moderate decline in sensitivity (with respect to all four prediction schemes) of up to about 10% for the traditional and 5% for the length-dependent sampling approach in the case of tRNAs, whereas for 5S rRNAs, the sensitivity values only decrease up to about 3% to 4% for both sampling variants. Unsurprisingly, for both RNA data, the change for the worse by means of measured sensitivity is less significant when considering MP structure predictions than when employing any of the other three principles, especially in the case of the LSCFG model. This is due to the fact that MP structures are always extracted by relying on the exact distribution (see discussion above). Altogether, these observations indicate that relative disturbances caused by mep do not have a significant negative effect on the predictive accuracy.

Moreover, Table 1 indicates that generating errors according to the fep(*prob*) variant (unsurprisingly) yields greater losses in the accuracies of selected predictions. In fact, as *prob* gets greater, there generally result considerably smaller PPV values for all four prediction schemes (mostly for MF structures) than in the corresponding undisturbed case. Furthermore, the respective sensitivity values degrade enormously, albeit again comparatively less in connection with MP structure predictions. However, these changes for the worse are obviously less significant when using the length-dependent sampling approach instead of the more general conventional variant, which matches the observations made above for disturbances caused by mep(*prob*). Nevertheless, errors produced according to fep(*prob*) for moderate percentages *prob* seem to generally have only a rather small influence on the resulting prediction accuracy. In most cases, only marginal losses in performance can be expected when disturbances are generated by fep(*prob*) with values *prob* of up to about 0.5, whereas for percentages of up to about 0.75, there should usually still result an acceptable accuracy of selected predictions (according to any of the four considered extraction principles).

Finally, it should be mentioned that all these observations and conclusions are actually affirmed by comparing the more reliable AUC results given in Table 2, which draw a rather similar picture of the behavior of both sampling approaches under the influence of the considered types and dimensions of relative disturbances in the underlying ensemble distribution.

3.3.1 Sampling Quality – Specific Values Related to Shapes

Obviously, the sensitivity and PPV measures used in the last section for assessing the accuracy of predicted foldings depend only on the numbers of correctly and incorrectly predicted base pairs (compared to the trusted database structure). For biologists, however, it is usually much more important to get the correct *shape* of the native folding. This is due to the fact that a predicted set of suboptimal foldings calculated by modern computational structure prediction methods generally contains lots of similar foldings but for biologists, only those with significant structural differences are of interest. According to these aspects, the concept of *abstract shapes* was introduced [GVR04, SVR$^+$06, JRG08], which are defined as morphic images of secondary structures such that each shape comprises a class of analogical foldings. Notably, there are five different shape levels which have been proven to gradually increase abstraction by disregarding certain unpaired regions or combining nested helices (see e.g. [NS09]), where secondary structures can accordingly be considered level 0 shapes.
For these reasons, we decided to complete our analysis of the influence of disturbances to the quality of probabilistic statistical sampling by considering the following meaningful specific values related to the shapes of predictions and sampled structures as defined in [NSar, SN]:

- Frequency of prediction of correct structure ($\mathrm{CSP}_{\mathrm{freq}}$): In how many cases is the predicted secondary structure (or its shape) equal to the correct structure (or the correct shape)?
- Frequency of correct shape occurring in a sample ($\mathrm{CSO}_{\mathrm{freq}}$): In how many cases can the correct shape (on different levels) be found in the generated sample set?
- Number of occurrences of correct shape in a sample ($\mathrm{CS}_{\mathrm{num}}$): How many times can the correct shape be found in the generated sample set?
- Number of different shapes in a sample ($\mathrm{DS}_{\mathrm{num}}$): How many different secondary structures (or shapes) can be found in the generated sample set?

We can easily compute the respective values from the predicted structures and the corresponding sample sets that were derived for the calculation of the sensitivity and PPV measures in the last section. The obtained results are collected in Tables 9a to 10g of Section Sm-II. Some of the most interesting ones are recorded in Tables 3 and 4.

First, as regards tRNAs, we observe that for MP predictions, disturbances caused by mep(*prob*) do generally not have a noticeable negative impact on the frequency of correct structure predictions (see Table 9a), and for the three other extraction principles, such disturbances do at least not yield a significant decline of the corresponding $\mathrm{CSP}_{\mathrm{freq}}$ value for shape levels 2 to 5 and under the assumption of the LSCFG approach, where for MF structures, there indeed results a slightly higher $\mathrm{CSP}_{\mathrm{freq}}$ value with increasing relative error percentage *prob* (see Tables 9b to 9d). When the more intensive variant as defined by fep(*prob*) is used for incorporating random errors into the considered sampling probabilities, the LSCFG based sampling algorithm still yields acceptable results with respect to $\mathrm{CSP}_{\mathrm{freq}}$ on abstraction levels 2 to 5, where for MP and MF structure predictions it obviously behaves quite resistant to the imposed distributions even for large values of *prob*.
Similar results are observed for 5S rRNAs (see Tables 10a to 10d, where for all four prediction selecting principles, the $\mathrm{CSP}_{\mathrm{freq}}$ values (for all shape levels in case of MP predictions and at least for shape levels 1 to 5 for all other prediction types) generally do not get significantly worse when applying the LSCFG sampling approach with inside values disturbed according to mep(*prob*) for any percentage $prob \in (0, 1)$ or according to the more intense relative disturbance variant fep(*prob*) for moderate values $prob \in (0, 1)$ (of up to about $prob = 0.75$).
Moreover, comparing the discussed $\mathrm{CSP}_{\mathrm{freq}}$ results for the LSCFG variant to the corresponding ones for the conventional SCFG approach, we get additional evidence that the length-independent sampling method reacts stronger to relative disturbances in the underlying ensemble distribution for a given sequence than its length-dependent counterpart. As already mentioned, this is due to the fact that the ensemble distribution considered in the length-dependent case is much more centered due to the more explicit (length-dependently trained) grammar parameters, such that randomly generated errors on particular probabilities carry less weight.

Now, let us consider the three remaining specific values $\mathrm{CSO}_{\mathrm{freq}}$, $\mathrm{CS}_{\mathrm{num}}$ and $\mathrm{DS}_{\mathrm{num}}$ that can eventually be used to assess the overall quality of generated sample sets rather than the accuracy of corresponding selected predictions. Basically, the obtained $\mathrm{CSO}_{\mathrm{freq}}$ and $\mathrm{CS}_{\mathrm{num}}$ results for tRNAs and 5S rRNAs (as

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.2413	0.4082	0.5548	0.5548	0.5552	0.6278
	mep(0.5)	0.2409	0.4068	0.5548	0.5548	0.5552	0.6265
	mep(0.99)	0.1877	0.3551	0.5382	0.5382	0.5386	0.6075
	fep(0.5)	0.2339	0.4017	0.5511	0.5511	0.5516	0.6269
	fep(0.99)	0.0014	0.0384	0.1979	0.1979	0.1984	0.2326
LSCFG	—	0.3324	0.4956	0.6574	0.6574	0.6579	0.7351
	mep(0.5)	0.3324	0.4956	0.6574	0.6574	0.6579	0.7351
	mep(0.99)	0.3236	0.4892	0.6560	0.6560	0.6565	0.7332
	fep(0.5)	0.3324	0.4966	0.6588	0.6588	0.6593	0.7369
	fep(0.99)	0.0624	0.2626	0.6246	0.6250	0.6250	0.6967

(a) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MP struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.2099	0.3699	0.5594	0.5594	0.5599	0.6302
	mep(0.5)	0.1683	0.3301	0.5372	0.5372	0.5377	0.6047
	mep(0.99)	0.0522	0.1822	0.4517	0.4517	0.4517	0.5215
	fep(0.5)	0.1049	0.2547	0.5155	0.5155	0.5160	0.5793
	fep(0.99)	0.0000	0.0125	0.1110	0.1110	0.1119	0.2062
LSCFG	—	0.3269	0.4892	0.6560	0.6565	0.6565	0.7337
	mep(0.5)	0.2534	0.4235	0.6708	0.6708	0.6713	0.7485
	mep(0.99)	0.1137	0.2954	0.6801	0.6801	0.6801	0.7568
	fep(0.5)	0.1794	0.3653	0.6704	0.6704	0.6709	0.7531
	fep(0.99)	0.0023	0.1262	0.6334	0.6334	0.6357	0.7240

(b) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MF struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0555	0.2094	0.4193	0.4193	0.4207	0.4679
	mep(0.5)	0.0416	0.1817	0.4045	0.4045	0.4055	0.4489
	mep(0.99)	0.0125	0.0989	0.3112	0.3112	0.3126	0.3570
	fep(0.5)	0.0245	0.1364	0.3662	0.3662	0.3666	0.4059
	fep(0.99)	0.0000	0.0014	0.0245	0.0245	0.0250	0.0546
LSCFG	—	0.1854	0.3574	0.4919	0.4919	0.4919	0.5465
	mep(0.5)	0.1405	0.3056	0.4998	0.4998	0.4998	0.5567
	mep(0.99)	0.0730	0.2191	0.4753	0.4753	0.4753	0.5284
	fep(0.5)	0.1003	0.2556	0.4836	0.4836	0.4836	0.5409
	fep(0.99)	0.0009	0.0781	0.3902	0.3902	0.3921	0.4508

(c) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MEA struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0374	0.1276	0.2973	0.2973	0.2977	0.3130
	mep(0.5)	0.0273	0.1045	0.2779	0.2779	0.2783	0.2908
	mep(0.99)	0.0083	0.0541	0.2007	0.2007	0.2007	0.2173
	fep(0.5)	0.0134	0.0795	0.2473	0.2473	0.2473	0.2603
	fep(0.99)	0.0000	0.0009	0.0120	0.0120	0.0120	0.0227
LSCFG	—	0.1729	0.3158	0.4300	0.4300	0.4300	0.4762
	mep(0.5)	0.1322	0.2728	0.4374	0.4374	0.4374	0.4859
	mep(0.99)	0.0693	0.1914	0.4101	0.4101	0.4101	0.4558
	fep(0.5)	0.0957	0.2261	0.4207	0.4207	0.4207	0.4642
	fep(0.99)	0.0009	0.0633	0.3264	0.3264	0.3269	0.3648

(d) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle Centroid).

Table 3: Specific values related to shapes of predictions and sampled structures, obtained from our tRNA database (by 10-fold cross-validation procedures, using sample size 1000 and $\min_{\mathrm{hel}} = \min_{HL} = 1$).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0000	0.0026	0.0052	0.0131	0.0366	0.7110
	mep(0.5)	0.0000	0.0009	0.0026	0.0113	0.0287	0.7128
	mep(0.99)	0.0000	0.0026	0.0044	0.0095	0.0227	0.6919
	fep(0.5)	0.0000	0.0017	0.0043	0.0113	0.0374	0.6954
	fep(0.99)	0.0000	0.0000	0.0000	0.0017	0.0096	0.5474
LSCFG	—	0.2141	0.4256	0.4744	0.4900	0.9408	0.9843
	mep(0.5)	0.2141	0.4256	0.4744	0.4900	0.9408	0.9843
	mep(0.99)	0.1941	0.4221	0.4761	0.4892	0.9452	0.9852
	fep(0.5)	0.2124	0.4248	0.4726	0.4883	0.9417	0.9852
	fep(0.99)	0.0209	0.3029	0.3725	0.4186	0.8529	0.9809

(a) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MP struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0000	0.0026	0.0052	0.0131	0.0357	0.7128
	mep(0.5)	0.0000	0.0009	0.0026	0.0122	0.0305	0.7180
	mep(0.99)	0.0000	0.0026	0.0044	0.0105	0.0235	0.6902
	fep(0.5)	0.0000	0.0017	0.0043	0.0113	0.0383	0.6971
	fep(0.99)	0.0000	0.0000	0.0000	0.0035	0.0200	0.5439
LSCFG	—	0.2002	0.4256	0.4700	0.4866	0.9417	0.9861
	mep(0.5)	0.1332	0.3960	0.4439	0.4587	0.9434	0.9869
	mep(0.99)	0.0365	0.3630	0.4308	0.4491	0.9304	0.9861
	fep(0.5)	0.0801	0.3847	0.4404	0.4561	0.9400	0.9861
	fep(0.99)	0.0035	0.1497	0.2106	0.3325	0.5440	0.9730

(b) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MF struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0000	0.0000	0.0000	0.0000	0.0261	0.3821
	mep(0.5)	0.0000	0.0000	0.0000	0.0000	0.0209	0.3698
	mep(0.99)	0.0000	0.0000	0.0000	0.0000	0.0122	0.3003
	fep(0.5)	0.0000	0.0000	0.0000	0.0000	0.0252	0.3438
	fep(0.99)	0.0000	0.0000	0.0000	0.0000	0.0026	0.0444
LSCFG	—	0.1062	0.3891	0.4291	0.4378	0.9051	0.9835
	mep(0.5)	0.1010	0.3751	0.4134	0.4239	0.8921	0.9782
	mep(0.99)	0.0392	0.3429	0.3986	0.4213	0.8712	0.9791
	fep(0.5)	0.0740	0.3839	0.4239	0.4387	0.8877	0.9791
	fep(0.99)	0.0017	0.1358	0.1863	0.2942	0.4970	0.9634

(c) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MEA struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0000	0.0000	0.0000	0.0000	0.0104	0.1097
	mep(0.5)	0.0000	0.0000	0.0000	0.0000	0.0104	0.1062
	mep(0.99)	0.0000	0.0000	0.0000	0.0000	0.0078	0.0827
	fep(0.5)	0.0000	0.0000	0.0000	0.0000	0.0061	0.0932
	fep(0.99)	0.0000	0.0000	0.0000	0.0000	0.0009	0.0078
LSCFG	—	0.0966	0.2916	0.3238	0.3316	0.8703	0.9686
	mep(0.5)	0.0879	0.3142	0.3516	0.3621	0.8625	0.9686
	mep(0.99)	0.0322	0.2924	0.3377	0.3595	0.8294	0.9651
	fep(0.5)	0.0662	0.3194	0.3551	0.3638	0.8512	0.9695
	fep(0.99)	0.0017	0.1053	0.1471	0.2219	0.4831	0.9339

(d) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle Centroid).

Table 4: Specific values related to shapes of predictions and sampled structures, obtained from our 5S rRNA database (by 10-fold cross-validation procedures, using sample size 1000 and $\min_{\mathrm{hel}} = \min_{HL} = 1$).

reported in Tables 9e to 9f and Tables 10e to 10f), respectively, show a similar picture and thus yield similar conclusions as the corresponding CSP_{freq} values discussed above. As a consequence to the fact that for larger relative error percentages *prob*, for the less intensive disturbance variant defined by mep(*prob*) and especially for the more grave version implied by fep(*prob*), the resulting values for CSO_{freq} and CS_{num} usually get smaller, the corresponding DS_{num} values inevitably increase with growing disturbance influences imposed by mep(*prob*) and especially fep(*prob*) (see Tables 9g and 10g). This actually means that the diversity within the generated sample sets generally gets greater as the overall sampling quality (with respect to occurrences of the correct structure in the sample) decreases, which could be fully expected.

4 Conclusion and Future Work

In this article, we performed a comprehensive experimental analysis on the effect of disturbances in the ensemble distribution for a given sequence to the quality of corresponding sets of candidate structures generated with the (L)SCFG based statistical sampling method studied in [NSar, SN]. Basically, two different levels of errors were considered for randomly creating disturbances on all inside values for a given input sequence according to the underlying grammar model: relative and absolute ones.
During our analysis (on the basis of trusted sets of tRNA and 5S rRNA data), we immediately observed that even incorporating only rather small absolute errors into (all or particular instances of the) inside values causes problematic disturbances of the resulting sampling probabilities that generally lead to the generation of useless sample sets. This can be assumed to be due to the fact that the installation of absolute errors usually makes it impossible for the employed sampling strategy to identify which ones of the considered inside probabilities for a given input sequence must originally (i.e., in the exact case) have been equal or unequal to zero, which inevitably results in a misguided behavior of the strategy, as it is no longer ensured that it creates only reasonable substructures for a considered sequence fragment.
However, both SCFG approaches (length-dependent and traditional one) behave rather resistant to disturbances of the needed conditional sampling probabilities that are caused by generating (moderate) relative errors on all (and also only on particular) inside values for a given input sequence. In general, even large relative errors seem to have no enormous negative impact on both the predictive accuracy and the overall quality of generated sample sets. That is, the reaction of the (L)SCFG based statistical sampling algorithm to the relative disturbances is fair enough to still obtain meaningful structure predictions (especially if the most likely structure of the sample is selected as predicted folding, in strong analogy to conventional SCFG based DPAs), and the overall quality of the resulting sample sets is still acceptable such that they might often also be used for further applications (like, e.g. probability profiling for specific loop types).
Consequently, it seems reasonable to believe that the needed sampling probabilities do not necessarily have to be computed in the exact way, but it may probably suffice to only (adequately) approximate them. In fact, the worst-case time complexity of any particular (L)SCFG based sampling method could potentially be reduced by developing a suitable approximation procedure (or at least an adequate heuristic method) for the computation of the needed sampling probabilities, where an appropriate approximation ratio (or at least an acceptable ratio of correctly and incorrectly computed zero values) should be attempted to ensure that the sampling quality remains sufficiently high, as indicated by the experimental disturbance analysis results discussed within this article.

References

[BBC+00] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[DCL04] Y. Ding, C. Y. Chan, and C. E. Lawrence. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Research*, 32:W135–W141, 2004.

[DCL05] Ye Ding, Chi Yu Chan, and Charles E. Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, 11:1157–1166, 2005.

[DE04] Robin D. Dowell and Sean R. Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 5:71, 2004.

[DL03] Ye Ding and Charles E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, 2003.

[DWB06] Chuong B. Do, Daniel A. Woods, and Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.

[GVR04] Robert Giegerich, Björn Voß, and Marc Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Research*, 32(16):4843–4851, 2004.

[HF71] T. Huang and K. S. Fu. On stochastic context-free languages. *Information Sciences*, 3:201–224, 1971.

[HFS+94] IvoL. Hofacker, Walter Fontana, PeterF. Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures (the Vienna RNA package). *Monatsh Chem.*, 125(2):167–188, 1994.

[Hof03] Ivo L. Hofacker. The vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, 2003.

[JRG08] Stefan Janssen, Jens Reeder, and Robert Giegerich. Shape based indexing for faster search of RNA family databases. *BMC Bioinformatics*, 9(131), 2008.

[KH99] B. Knudsen and J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–454, 1999.

[KH03] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 31(13):3423–3428, 2003.

[Mai07] Robert S. Maier. Parametrized stochastic grammars for RNA secondary structure prediction. *Information Theory and Applications Workshop*, pages 256–260, 2007.

[McC90] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[MSZT99] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[NE07] Eric P. Nawrocki and Sean R. Eddy. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Computational Biology*, 3(3):e56, 2007.

[NJ80] R. Nussinov and A. B. Jacobson. Fast algorithms for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Science of the USA*, 77(11):6309–6313, 1980.

[NPGK78] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35:68–82, 1978.

[NS] Markus E. Nebel and Anika Scheid. A n^2 RNA secondary structure prediction algorithm. Accepted for Bioinformatics 2012, International Conference on Bioinformatics Models, Methods and Algorithms.

[NS09] Markus E. Nebel and Anika Scheid. On quantitative effects of RNA shape abstraction. *Theory in Biosciences*, 128(4):211–225, 2009.

[NS11] Markus E. Nebel and Anika Scheid. Analysis of the free energy in a stochastic RNA secondary structure model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(6):1468–1482, 2011.

[NSar] Markus E. Nebel and Anika Scheid. Evaluation of a sophisticated SCFG design for RNA secondary structure prediction. *Theory in Biosciences*, to appear.

[SBEB02] Maciej Szymanski, Miroslawa Z. Barciszewska, Volker A. Erdmann, and Jan Barciszewski. 5s ribosomal RNA database. *Nucleic Acids Res.*, 30:176–178, 2002.

[SHB+98] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, 26:148–153, 1998.

[SN] Anika Scheid and Markus E. Nebel. Statistical RNA secondary structure sampling based on a length-dependent SCFG model. Submitted.

[SVR+06] Peter Steffen, Björn Voß, Marc Rehmsmeier, Jens Reeder, and Robert Giegerich. RNAshapes 2.1.1 manual, February 2006.

[VC85] G. Viennot and M. Vauchaussade De Chaumont. Enumeration of RNA secondary structures by complexity. *Mathematics in medicine and biology, Lecture Notes in Biomathematics*, 57:360–365, 1985.

[WFHS99] S. Wuchty, W. Fontana, I. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–165, 1999.

[WN11] Frank Weinberg and Markus E. Nebel. Applying length-dependent stochastic context-free grammars to RNA secondary structure prediction. *Algorithms*, 4(4):223–238, 2011.

[XSB+98] T. Xia, J. SantaLucia Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–14735, 1998.

[ZS81] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.

[Zuk89] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

[Zuk03] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, 2003.

Supplementary Material

Sm-I Formal Description of the Sampling Process

In the sequel, given an RNA molecule r consisting of n nucleotides, we denote the corresponding sequence fragment from position i to position j, $1 \leq i \leq j \leq n$, by $R_{i,j} = r_i r_{i+1} \dots r_{j-1} r_j$. Additionally, by $S_{i,j}$ we denote a structure on the sequence fragment $R_{i,j}$ that meets all the constraints of our definition of RNA secondary structures.
Briefly, according to [NSar, SN], a complete secondary structure $S_{1,n}$ for a given input sequence r of length n can be sampled in the following recursive way: Start with the entire RNA sequence $R_{1,n}$ and consecutively compute the adjacent substructures (single-stranded regions and paired substructures) of the exterior loop (from left to right). Any (paired) substructure on fragment $R_{i,j}$, $1 \leq i < j \leq n$, is folded by recursively constructing substructures (hairpins, stacked pairs, bulges, interior and multibranched loops) on smaller fragments $R_{l,h}$, $i \leq l < h \leq j$.
Note that this sampling process is similar to the traceback algorithm employed in MFE based dynamic programming algorithms. Actually, the main difference is that in those algorithms, base pairings are selected by the minimum free energy principle for the fragments $R_{i,j}$, $1 \leq i, j \leq n$ whereas here, base pairs are randomly sampled according to conditional probability distributions for the corresponding fragments. These distributions are derived from definitions of probabilities for particular choices (such as paired and unpaired bases or specific loop types). Notably, they only depend on the precomputed (skewed) inside probabilities $\widehat{\alpha}_X(i,j)$ for input sequence r, the thereof additionally precalculated probabilities

$$\widehat{\alpha}_{AT}(h,j) := \sum\nolimits_{l=(h-1)+\min_{\mathrm{ps}}}^{(j-1)} \widehat{\alpha}_A(h,l) \cdot \widehat{\alpha}_T(l+1,j), \tag{15}$$

$$\widehat{\alpha}_{AB}(h,j) := \sum\nolimits_{l=(h-1)+\min_{\mathrm{ps}}}^{(j-2)} \widehat{\alpha}_A(h,l) \cdot \widehat{\alpha}_B(l+1,j-1), \tag{16}$$

$$\widehat{\alpha}_{AO}(h,j) := \sum\nolimits_{l=(h-1)+\min_{\mathrm{ps}}}^{(j-1)-\min_{\mathrm{ps}}} \widehat{\alpha}_A(h,l) \cdot \widehat{\alpha}_O(l+1,j-1), \tag{17}$$

$$\widehat{\alpha}_{AN}(h,j) := \sum\nolimits_{l=(h-1)+\min_{\mathrm{ps}}}^{(j-1)} \widehat{\alpha}_A(h,l) \cdot \widehat{\alpha}_N(l+1,j-1), \tag{18}$$

corresponding to inside values for combined intermediate symbols, where $i \leq h \leq j$, and of course the trained grammar parameters (transition probabilities only).
Algorithms 1 to 4 formally describe how the sampling strategy works. Note that the type (or shape) and actual composition (of accessible base pairs and unpaired bases) of a particular substructure (corresponding to a valid derivation tree) on a given fragment $R_{i,j}$ are randomly drawn according to the conditional probability distributions induced by the respective sets of all (valid) choices for the unique intermediate symbol of the grammar that generates such substructures (that represents the root of the corresponding subtree). Principally, each of the presented algorithms describing the employed sampling strategy relies on a moderate number of formal set definitions for the respective mutually exclusive and exhaustive cases in order to perform the needed random choices, which basically all obey to the same scheme.
Particularly, for sampling shape and actual composition (free base pairs and unpaired bases) of the exterior loop, Algorithm 1 considers the following sets:

$$acT(i,j) := \{\{x, prob\} \mid x \in \{C, A, CA, AT, CAT\} \text{ and } prob = \sum\nolimits_{\{y,pr\} \in acT_x(i,j)} pr \neq 0\}, \tag{19}$$

where

$$acT_C(i,j) := \{\{0, prob\} \mid prob = \widehat{\alpha}_C(i,j) \cdot \Pr_{tr}(T \to C) \neq 0\}, \tag{20}$$

$$acT_A(i,j) := \{\{0, prob\} \mid prob = \widehat{\alpha}_A(i,j) \cdot \Pr_{tr}(T \to A) \neq 0\}, \tag{21}$$

$$\begin{aligned} acT_{CA}(i,j) := \{\{h, prob\} \mid\ & (i+1) \leq h \leq (j+1) - \min_{\mathrm{ps}} \text{ and} \\ & prob = \widehat{\alpha}_C(i,h-1) \cdot \widehat{\alpha}_A(h,j) \cdot \Pr_{tr}(T \to CA) \neq 0\}, \end{aligned} \tag{22}$$

$$\begin{aligned} acT_{AT}(i,j) := \{\{l, prob\} \mid\ & (i-1) + \min_{\mathrm{ps}} \leq l \leq (j-1) \text{ and} \\ & prob = \widehat{\alpha}_A(i,l) \cdot \widehat{\alpha}_T(l+1,j) \cdot \Pr_{tr}(T \to AT) \neq 0\}, \end{aligned} \tag{23}$$

$$\begin{aligned} acT_{CAT}(i,j) := \{\{h, prob\} \mid\ & (i+1) \leq h \leq j - \min_{\mathrm{ps}} \text{ and} \\ & prob = \widehat{\alpha}_C(i,h-1) \cdot \widehat{\alpha}_{AT}(h,j) \cdot \Pr_{tr}(T \to CAT) \neq 0\}, \end{aligned} \tag{24}$$

Algorithm 1 Sampling an entire secondary structure

Input: RNA sequence r of length $n \geq 1$,
trained transition probabilities $\Pr_{tr}(rule)$, for $rule \in \mathcal{R}_{\mathcal{G}_s}$,
precomputed inside probabilities $\widehat{\alpha}_X(i,j)$, for $X \in \mathcal{I}_{\mathcal{G}_s} \cup \{AT, AB, AO, AN\}$ and $1 \leq i, j \leq n$.

Output: $helices = \big\{\{i,j,k\} \mid 1 \leq i < j \leq n$ and $k \geq \min_{\text{hel}}$ and
$i.j, (i+1).(j-1), \ldots, (i+(k-1)).(j-(k-1))$ are consecutive base pairs$\big\}$.

procedure ComputeRandomExteriorLoop()
$helices = \emptyset$
$i = 1$, $j = n$
while $(j - i + 1) \neq 0$ **do**
/*Sample next substructure on $R_{i,j}$ according to $acT(i,j)$, i.e. construct paired substructure starting with free base pair $h.l$, for $i \leq h < l \leq j$, or leave $R_{i,j}$ unpaired:*/
$extLoopType =$ Sample exterior loop substructure type for $R_{i,j}$ according to $acT(i,j)$
if $extLoopType = C$ **then**
/*$R_{i,j}$ becomes single-stranded:*/
return *helices*
else if $extLoopType = A$ **then**
/*$R_{i,j}$ becomes paired structure:*/
$h = i$, $l = j$
else if $extLoopType = CA$ **then**
/*$R_{i,j}$ becomes paired structure preceded by single-strand:*/
Sample h according to $acT_{CA}(i,j)$
$l = j$
else if $extLoopType = AT$ **then**
/*$R_{i,j}$ becomes paired structure followed by further structure(s):*/
$h = i$
Sample l according to $acT_{AT}(i,j)$
else if $extLoopType = CAT$ **then**
/*$R_{i,j}$ becomes paired structure preceded by single-strand and followed by further structure(s):*/
Sample h according to $acT_{CAT}(i,j)$
Sample l according to $ac^*_{AT}(h,j)$
end if
if $extLoopType \in \{A, CA, AT, CAT\}$ **and** $h.l$ successfully sampled **then**
/*Recursively fold substructures on $R_{h,l}$:*/
$helices = helices \cup \{\{h, l, \min_{\text{hel}}\}\}$
$helices =$ ComputeRandomLoop$(h + (\min_{\text{hel}} - 1), l - (\min_{\text{hel}} - 1), helices)$
/*Consider the remaining fragment $R_{(l+1),j}$:*/
$i = l + 1$
else
/*Sampling failed (as there exist no valid choices), so stop folding the loop (such that $R_{i,j}$ becomes single-stranded):*/
return *helices*
end if
end while
return *helices*
end procedure

Algorithm 2 Sampling any substructure

procedure ComputeRandomLoop($i, j, helices$)
$loopType$ = Sample loop type closed by $i.j$ according to $acL(i, j)$
if $loopType = F$ **then**
/*Pair $i.j$ closes hairpin loop:*/
return $helices$
else if $loopType = P$ **then**
/*Pair $i.j$ closes stacked pair:*/
$helices[-1, 3] = helices[-1, 3] + 1$ /*increments length of last added helix*/
$helices$ = ComputeRandomLoop($i + 1, j - 1, helices$)
else if $loopType = G$ **then**
/*Pair $i.j$ closes bulge or interior loop:*/
$helices$ = ComputeRandomBulgeInteriorLoop($i, j, helices$)
else if $loopType = M$ **then**
/*Pair $i.j$ closes multiloop:*/
$helices$ = ComputeRandomMultiLoop($i, j, helices$)
else
/*Sampling failed (as there exist no valid choices), so stop folding the loop (such that $R_{i+1,j-1}$ becomes single-stranded hairpin loop):*/
return $helices$
end if
return $helices$
end procedure

Algorithm 3 Sampling a particular bulge or interior loop

procedure ComputeRandomBulgeInteriorLoop($i, j, helices$)
/*Note that the following allows $\max_{bulge} = \infty$ (then no restrictions are applied):*/
$loopType$ = Sample bulge or interior loop type on $R_{i+1,j-1}$ according to $acG(i, j)$
if $loopType = BA$ **then**
/*Bulge on the left:*/
Sample h according to $acG_{BA}(i, j)$
$l = j$
else if $loopType = AB$ **then**
/*Bulge on the right:*/
$h = i$
Sample l according to $acG_{AB}(i, j)$
else if $loopType = BAB$ **then**
/*Interior loop:*/
Sample h according to $acG_{BAB}(i, j)$
Sample l according to $ac^*_{AB}(h, j)$
end if
if $loopType \in \{BA, AB, BAB\}$ **and** $h.l$ successfully sampled **then**
/*Recursively fold substructures on $R_{h,l}$:*/
$helices = helices \cup \{\{h, l, \min_{\text{hel}}\}\}$
$helices$ = ComputeRandomLoop($h + (\min_{\text{hel}} - 1), l - (\min_{\text{hel}} - 1), helices$)
else
/*Sampling failed (as there exist no valid choices), so stop folding the loop (such that $R_{i+1,j-1}$ becomes single-stranded hairpin loop):*/
return $helices$
end if
return $helices$
end procedure

Algorithm 4 Sampling a complete multiloop

procedure ComputeRandomMultiLoop($i, j, helices$)
$k = 0$, $l_k = i$
while $(j - l_k - 1) \geq \min_{\text{ps}}$ **do**
/*Create $(k+1)$th paired substructure on $R_{l_k+1,j-1}$, starting with accessible base pair $h_{k+1}.l_{k+1}$, for $l_k < h_{k+1} < l_{k+1} < j$:*/
if $(k+1) = 1$ **then**
Sample h according to $acM_{UAO}(l_k, j)$
Sample l according to $ac^*_{AO}(h, j)$
else if $(k+1) = 2$ **then**
Sample h according to $acO_{UAN}(l_k, j)$
Sample l according to $ac^*_{AN}(h, j)$
else if $(k+1) \geq 3$ **then**
Sample h according to $acN_{UAN}(l_k, j)$
Sample l according to $ac^*_{AN}(h, j)$
end if
if $h.l$ successfully sampled **then**
$h_{k+1} = h$, $l_{k+1} = l$
/*Recursively fold substructures on $R_{h_{k+1},l_{k+1}}$:*/
$helices = helices \cup \{\{h_{k+1}, l_{k+1}, \min_{\text{hel}}\}\}$
$helices = \text{ComputeRandomLoop}(h_{k+1} + (\min_{\text{hel}} - 1), l_{k+1} - (\min_{\text{hel}} - 1), helices)$
/*Decide whether to leave the remaining fragment $R_{l_{k+1}+1,j-1}$ unpaired or not:*/
if $(k+1) \geq 2$ **then**
Uniformly draw real value $random \in (0, 1]$
if $random \in (0, dec_U(l_{k+1}, j)]$ **then**
/*No additional base pairs:*/
return *helices*
else if $random \in (dec_U(l_{k+1}, j), 1]$ **then**
/*At least one more paired substructure:*/
$k = k + 1$
end if
end if
else
/*Sampling failed (as there exist no valid choices), so stop folding the loop (such that $R_{l_k+1,j-1}$ becomes single-stranded):*/
return *helices*
end if
end while
return *helices*
end procedure

and

$$ac^*_{AT}(h,j) := \{\{l, prob\} \mid (h-1) + \min_{\text{ps}} \leq l \leq (j-1) \text{ and } prob = \widehat{\alpha}_A(h,l) \cdot \widehat{\alpha}_T(l+1,j) \neq 0\}. \quad (25)$$

For sampling the type of the loop closed by a given base pair $i.j$, Algorithm 2 relies on

$$acL(i,j) := \{\{x, prob\} \mid x \in \{F,P,G,M\} \text{ and } prob = \widehat{\alpha}_x(i+1,j-1) \cdot \Pr_{tr}(L \to x) \neq 0\}. \quad (26)$$

Algorithm 3 employs the following sets in order to sample a particular bulge or interior loop (closed by a given base pair $i.j$) on the considered sequence fragment $R_{i+1,j-1}$:

$$acG(i,j) := \{\{x, prob\} \mid x \in \{BA, AB, BAB\} \text{ and } prob = \sum\nolimits_{\{y,pr\} \in acG_x(i,j)} pr \neq 0\}, \quad (27)$$

where

$$\begin{aligned} acG_{BA}(i,j) := \{\{h, prob\} \mid\ & (i+2) \leq h \leq j - \min_{\text{ps}} \text{ and} \\ & prob = \widehat{\alpha}_B(i+1,h-1) \cdot \widehat{\alpha}_A(h,j-1) \cdot \Pr_{tr}(G \to BA) \neq 0\}, \quad (28) \\ acG_{AB}(i,j) := \{\{l, prob\} \mid\ & i + \min_{\text{ps}} \leq l \leq (j-2) \text{ and} \\ & prob = \widehat{\alpha}_A(i+1,l) \cdot \widehat{\alpha}_B(l+1,j-1) \cdot \Pr_{tr}(G \to AB) \neq 0\}, \quad (29) \\ acG_{BAB}(i,j) := \{\{h, prob\} \mid\ & (i+2) \leq h \leq j - \min_{\text{ps}} - 1 \text{ and} \\ & prob = \widehat{\alpha}_B(i+1,h-1) \cdot \widehat{\alpha}_{AB}(h,j) \cdot \Pr_{tr}(G \to BAB) \neq 0\}, \quad (30) \end{aligned}$$

and

$$ac^*_{AB}(h,j) := \{\{h, prob\} \mid (h-1) + \min_{\text{ps}} \leq l \leq (j-2) \text{ and } prob = \widehat{\alpha}_A(h,l) \cdot \widehat{\alpha}_B(l+1,j-1) \neq 0\}. \quad (31)$$

Finally, for sampling a complete multiloop (closed by a given base pair $i.j$) on the considered sequence fragment $R_{i+1,j-1}$, the following formal definitions are used by Algorithm 4:

$$\begin{aligned} acM_{UAO}(i,j) := \{\{h, prob\} \mid\ & (i+1) \leq h \leq j - 2 \cdot \min_{\text{ps}} \text{ and} \\ & prob = \widehat{\alpha}_U(i+1,h-1) \cdot \widehat{\alpha}_{AO}(h,j) \cdot \Pr_{tr}(M \to UAO) \neq 0\}, \quad (32) \\ acO_{UAN}(l_k,j) := \{\{h, prob\} \mid\ & (l_k+1) \leq h \leq j - \min_{\text{ps}} \text{ and} \\ & prob = \widehat{\alpha}_U(l_k+1,h-1) \cdot \widehat{\alpha}_{AN}(h,j) \cdot \Pr_{tr}(O \to UAN) \neq 0\}, \quad (33) \\ acN_{UAN}(l_k,j) := \{\{h, prob\} \mid\ & (l_k+1) \leq h \leq j - \min_{\text{ps}} \text{ and} \\ & prob = \widehat{\alpha}_U(l_k+1,h-1) \cdot \widehat{\alpha}_{AN}(h,j) \cdot \Pr_{tr}(N \to UAN) \neq 0\}, \quad (34) \end{aligned}$$

as well as

$$ac^*_{AO}(h,j) := \{\{l, prob\} \mid (h-1) + \min_{\text{ps}} \leq l \leq (j-1) - \min_{\text{ps}} \text{ and } prob = \widehat{\alpha}_A(h,l) \cdot \widehat{\alpha}_O(l+1,j-1) \neq 0\}, \quad (35)$$

$$ac^*_{AN}(h,j) := \{\{l, prob\} \mid (h-1) + \min_{\text{ps}} \leq l \leq (j-1) \text{ and } prob = \widehat{\alpha}_A(h,l) \cdot \widehat{\alpha}_N(l+1,j-1) \neq 0\}, \quad (36)$$

and finally (for deciding whether an additional substructure should be added or not),

$$dec_U(l_{k+1},j) := \frac{\widehat{\alpha}_U(l_{k+1}+1,j-1) \cdot \Pr_{tr}(N \to U)}{\widehat{\alpha}_U(l_{k+1}+1,j-1) \cdot \Pr_{tr}(N \to U) + \sum_{\{h,prob\} \in acN(l_{k+1},j)} prob}. \quad (37)$$

It remains to mention that after a preprocessing of the given input sequence (including the complete dynamic programming method for deriving all inside probabilities $\widehat{\alpha}_X(i,j)$, for $X \in \mathcal{I}_{\mathcal{G}_s}$ and $1 \leq i,j \leq n$, as well as the subsequent calculation of the additionally needed probabilities $\widehat{\alpha}_x(h,j)$, for $x \in \{AT, AB, AO, AN\}$ and $1 \leq h,j \leq n$, which both take $\mathcal{O}(n^3)$ time and require $\mathcal{O}(n^2)$ storage [7]), each of the probabilities *prob* defined for a particular choice of a paired base (h or l) in the respective subset ($acX_y(i,j)$ or $ac^*_z(h,j)$) of all possible choices can be derived in constant time. Furthermore, according

[7] Note that if we modify the considered SCFG $\mathcal{G}_s$ such that each occurrence of any pattern $x \in \{AT, AB, AO, AN\}$ (in the conclusions of the production rules of $\mathcal{G}_s$) is replaced by a new intermediate symbol $Y \notin \mathcal{I}_{\mathcal{G}_s}$ corresponding to the respective pattern x, then $\widehat{\alpha}_x(i,j)$, $1 \leq i,j \leq n$, is equal to the inside probability $\widehat{\alpha}_Y(i,j)$ of this new intermediate symbol Y and is automatically derived during the inside value computations.

to their definitions, none of these subsets contains more than n choices for a particular paired base in the worst-case, that is $\texttt{card}(acX_y(i,j)) \in \mathcal{O}(n)$ and $\texttt{card}(ac^*_z(h,j)) \in \mathcal{O}(n)$. Hence, the sampling strategy needs $\mathcal{O}(n)$ time for deriving the respective probability distribution and drawing a corresponding random choice.
Additionally, due to the cardinalities of $\mathcal{O}(n)$ for each of the $\mathcal{O}(1)$ distinct subsets $acX_y(i,j)$ (corresponding to production rules $X \to y$) for any main set $acX(i,j)$ (for premise X), each of the probabilities defined for a particular choice of the shape of a random substructure (corresponding to one of its subsets and hence to the respective production $X \to y$ applied from the considered intermediate symbol X in order to generate that shape) can be computed in $\mathcal{O}(n)$ time (since for each of the $\mathcal{O}(1)$ rules $X \to y$, we have to compute the sum of $\texttt{card}(acX_y(i,j)) \in \mathcal{O}(n)$ terms, where each term is obtained in constant time, see above). Then, the respective probability distribution employed for (shape or loop type) sampling can be derived in constant time (as $\texttt{card}(acX(i,j)) \in \mathcal{O}(1)$). For example, the distribution for sampling the exterior loop substructure type according to $acT(i,j)$ can be derived in $2 \cdot \mathcal{O}(1) + 3 \cdot \mathcal{O}(n)$ time.
Altogether, there obviously results $\mathcal{O}(n)$ time complexity for sampling a random base pair $h.l$ (by first sampling the substructure type (if needed), then the leftmost base h and finally the rightmost base l) on $R_{i,j}$, $1 \leq i \leq h < l \leq j \leq n$. Thus, since any structure of size n can have at most $\lfloor \frac{n - \min_{HL}}{2} \rfloor \in \mathcal{O}(n)$ base pairs and any base pair can be sampled in linear time, the time requirements of the sampling strategy for constructing a complete secondary structure $S_{1,n}$ is bounded by $\mathcal{O}(n^2)$.

Sm-II Tables and Figures

X	$\texttt{card}(\mathcal{X}^e)$	$\texttt{card}(\mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^e \cap \mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^e \setminus \mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^d \setminus \mathcal{X}^e)$	
		mev	fev	mev	fev	mev	fev	mev	fev
A	2649	1733	1698	1547	1529	1102	1120	186	169
B	2926	1704	1741	1660	1709	1266	1217	44	32
C	2926	1847	1847	1806	1808	1120	1118	41	39
F	2926	1873	1891	1840	1849	1086	1077	33	42
G	2696	1777	1781	1616	1617	1080	1079	161	164
M	2548	1597	1573	1357	1338	1191	1210	240	235
N	3002	1957	1945	1957	1945	1045	1057	0	0
O	2770	1838	1818	1727	1692	1043	1078	111	126
P	2649	1721	1745	1553	1563	1096	1086	168	182
T	2926	1865	1905	1822	1869	1104	1057	43	36
U	3002	1938	1913	1938	1913	1064	1089	0	0
AT	2697	2699	2698	2697	2697	0	0	2	1
AB	2552	2554	2553	2552	2552	0	0	2	1
AO	2478	2482	2481	2478	2478	0	0	4	3
AN	2697	2699	2698	2697	2697	0	0	2	1

(a) Traditional SCFG model.

X	$\texttt{card}(\mathcal{X}^e)$	$\texttt{card}(\mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^e \cap \mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^e \setminus \mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^d \setminus \mathcal{X}^e)$	
		mev	fev	mev	fev	mev	fev	mev	fev
A	469	1587	1630	325	326	144	143	1262	1304
B	1651	1996	1980	1337	1310	314	341	659	670
C	1096	2001	1987	1053	1025	43	71	948	962
F	871	1888	1850	819	801	52	70	1069	1049
G	729	1603	1583	457	457	272	272	1146	1126
M	359	1517	1525	170	184	189	175	1347	1341
N	1331	1601	1626	786	758	545	573	815	868
O	690	1524	1527	357	355	333	335	1167	1172
P	435	1612	1565	312	306	123	129	1300	1259
T	708	1772	1752	614	594	94	114	1158	1158
U	1571	2038	2059	1323	1322	248	249	715	737
AT	1394	2630	2613	1394	1394	0	0	1236	1219
AB	1829	2485	2469	1829	1829	0	0	656	640
AO	499	2308	2291	499	499	0	0	1809	1792
AN	1832	2620	2602	1831	1832	1	0	789	770

(b) LSCFG model.

Table 5: Numbers of relevant inside values (inside probabilities being greater than zero) that were considered for obtaining the profiles presented in Figure 2b (and Figure 6), where $\mathcal{X}^e := \{\{i,j\} \mid 1 \leq i,j \leq n \text{ and } \alpha_X(i,j) \neq 0\}$ and $\mathcal{X}^d := \{\{i,j\} \mid 1 \leq i,j \leq n \text{ and } \widehat{\alpha}_X(i,j) \neq 0\}$.

$\mathcal{X}$	$\texttt{card}(\mathcal{X}^e)$	$\texttt{card}(\mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^e \cap \mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^e \setminus \mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^d \setminus \mathcal{X}^e)$	
		mev	fev	mev	fev	mev	fev	mev	fev
$\mathcal{T}_C$	76	36	39	36	39	40	37	0	0
$\mathcal{T}_A$	54	33	41	22	32	32	22	11	9
$\mathcal{T}_{CA}$	1829	766	856	480	673	1349	1156	286	183
$\mathcal{T}_{AT}$	2595	999	961	966	924	1629	1671	33	37
$\mathcal{T}_{CAT}$	2628	1644	1646	1644	1646	984	982	0	0
$\mathcal{AT}$	62102	25675	25693	24940	24671	37162	37431	735	1022
$\mathcal{L}_F$	2775	1750	1777	1750	1777	1025	998	0	0
$\mathcal{L}_P$	2522	1533	1542	1487	1486	1035	1036	46	56
$\mathcal{L}_G$	2552	1543	1539	1540	1536	1012	1016	3	3
$\mathcal{L}_M$	2408	1290	1265	1288	1261	1120	1147	2	4
$\mathcal{G}_{BA}$	59580	23057	23288	22390	22547	37190	37033	667	741
$\mathcal{G}_{AB}$	0	0	0	0	0	0	0	0	0
$\mathcal{G}_{BAB}$	59476	36453	37479	36428	37461	23048	22015	25	18
$\mathcal{AB}$	1041908	454132	457662	441016	442856	600892	599052	13116	14806
$\mathcal{M}_{UAO}$	56901	41296	40054	41208	40010	15693	16891	88	44
$\mathcal{AO}$	980735	488660	456896	473742	442002	506993	538733	14918	14894
$\mathcal{O}_{UAN}$	56999	41352	40087	41329	40066	15670	16933	23	21
$\mathcal{N}_{UAN}$	49970	36636	35377	36615	35356	13355	14614	21	21
$\mathcal{AN}$	985172	511715	490277	497364	474716	487808	510456	14351	15561

(a) Traditional SCFG model.

$\mathcal{X}$	$\texttt{card}(\mathcal{X}^e)$	$\texttt{card}(\mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^e \cap \mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^e \setminus \mathcal{X}^d)$		$\texttt{card}(\mathcal{X}^d \setminus \mathcal{X}^e)$	
		mev	fev	mev	fev	mev	fev	mev	fev
$\mathcal{T}_C$	7	7	7	7	7	0	0	0	0
$\mathcal{T}_A$	1	3	1	0	0	1	1	3	1
$\mathcal{T}_{CA}$	33	280	198	13	7	20	26	267	191
$\mathcal{T}_{AT}$	55	256	298	33	25	22	30	223	273
$\mathcal{T}_{CAT}$	161	477	461	157	153	4	8	320	308
$\mathcal{AT}$	2936	17032	26734	1916	1870	1020	1066	15116	24864
$\mathcal{L}_F$	845	795	780	795	780	50	65	0	0
$\mathcal{L}_P$	409	603	581	295	292	114	117	308	289
$\mathcal{L}_G$	669	431	429	428	423	241	246	3	6
$\mathcal{L}_M$	308	152	162	152	162	156	146	0	0
$\mathcal{G}_{BA}$	401	844	881	351	355	50	46	493	526
$\mathcal{G}_{AB}$	0	0	0	0	0	0	0	0	0
$\mathcal{G}_{BAB}$	5074	11964	11884	4002	3916	1072	1158	7962	7968
$\mathcal{AB}$	173376	457279	487255	109429	110855	63947	62521	347850	376400
$\mathcal{M}_{UAO}$	4229	10068	10201	3926	3939	303	290	6142	6262
$\mathcal{AO}$	19149	279648	298214	8090	8203	11059	10946	271558	290011
$\mathcal{O}_{UAN}$	11284	16633	16787	9773	9863	1511	1421	6860	6924
$\mathcal{N}_{UAN}$	11880	18444	18496	10324	10491	1556	1389	8120	8005
$\mathcal{AN}$	89494	306125	329250	45081	44257	44413	45237	261044	284993

(b) LSCFG model.

Table 6: Numbers of relevant sampling probabilities (being greater than zero) that were considered for obtaining the profiles presented in Figure 2b (and Figure 6), where $\mathcal{X}_y^z := \bigcup_{1 \leq i,j \leq n} acX_y(i,j)$ and $\mathcal{Y}^z := \bigcup_{1 \leq i \leq h \leq j \leq n} ac_Y^*(h,j)$, with $z = e$ and $z = d$ denoting the exact and disturbed values, respectively.

Approach	Errors	MP struct.		MF struct.		MEA struct.		Centroid	
		Sens.	PPV	Sens.	PPV	Sens.	PPV	Sens.	PPV
SCFG	—	0.7818	0.8437	0.7792	0.8445	0.7324	0.8939	0.6754	0.9158
	mep(0.5)	0.7822	0.8447	0.7599	0.8370	0.7169	0.8927	0.6607	0.9140
	mep(0.75)	0.7793	0.8431	0.7303	0.8217	0.6935	0.8917	0.6356	0.9123
	mep(0.9)	0.7699	0.8409	0.7075	0.8117	0.6715	0.8893	0.6097	0.9115
	mep(0.99)	0.7590	0.8388	0.6768	0.8004	0.6414	0.8877	0.5817	0.9127
	fep(0.5)	0.7798	0.8440	0.7234	0.8184	0.6864	0.8896	0.6292	0.9134
	fep(0.75)	0.7442	0.8313	0.6414	0.7736	0.6066	0.8802	0.5507	0.9032
	fep(0.9)	0.6644	0.8106	0.5257	0.7229	0.4934	0.8652	0.4375	0.8952
	fep(0.99)	0.4101	0.7295	0.2864	0.5590	0.2532	0.7776	0.2157	0.8291
LSCFG	—	0.8545	0.9534	0.8542	0.9535	0.8335	0.9736	0.8250	0.9783
	mep(0.5)	0.8545	0.9534	0.8429	0.9524	0.8236	0.9731	0.8150	0.9773
	mep(0.75)	0.8542	0.9533	0.8281	0.9485	0.8098	0.9709	0.8018	0.9758
	mep(0.9)	0.8546	0.9539	0.8104	0.9425	0.7978	0.9697	0.7889	0.9744
	mep(0.99)	0.8519	0.9533	0.7988	0.9413	0.7833	0.9676	0.7735	0.9726
	fep(0.5)	0.8548	0.9536	0.8224	0.9486	0.8029	0.9707	0.7940	0.9758
	fep(0.75)	0.8524	0.9532	0.7763	0.9323	0.7674	0.9620	0.7589	0.9687
	fep(0.9)	0.8315	0.9492	0.7223	0.9162	0.7131	0.9523	0.7038	0.9601
	fep(0.99)	0.7530	0.9325	0.5769	0.8623	0.5668	0.9075	0.5567	0.9195

(a) Sensitivity and PPV.

Approach	Errors	MEA struct.	Centroid
SCFG	—	0.828522	0.833894
	mep(0.5)	0.819658	0.823811
	mep(0.75)	0.810331	0.813818
	mep(0.9)	0.801393	0.801842
	mep(0.99)	0.786645	0.788478
	fep(0.5)	0.805999	0.807240
	fep(0.75)	0.761806	0.759493
	fep(0.9)	0.682057	0.676879
	fep(0.99)	0.440021	0.422778
LSCFG	—	0.936285	0.919736
	mep(0.5)	0.932121	0.916321
	mep(0.75)	0.925639	0.907926
	mep(0.9)	0.919747	0.900505
	mep(0.99)	0.916540	0.896024
	fep(0.5)	0.924191	0.908943
	fep(0.75)	0.900592	0.884400
	fep(0.9)	0.872742	0.848190
	fep(0.99)	0.752030	0.722737

(b) AUC values.

Table 7: Prediction results for our tRNA database (computed by 10-fold cross-validation procedures, using sample size 1000 and $\min_{\text{hel}} = \min_{HL} = 1$).

Approach	Errors	MP struct.		MF struct.		MEA struct.		Centroid	
		Sens.	PPV	Sens.	PPV	Sens.	PPV	Sens.	PPV
SCFG	—	0.4251	0.5372	0.4251	0.5363	0.3403	0.6967	0.2689	0.8044
	mep(0.5)	0.4143	0.5280	0.4160	0.5290	0.3334	0.6987	0.2643	0.8051
	mep(0.75)	0.4113	0.5303	0.4105	0.5289	0.3234	0.7031	0.2566	0.8098
	mep(0.9)	0.4071	0.5311	0.4064	0.5297	0.3120	0.7007	0.2466	0.8050
	mep(0.99)	0.3897	0.5227	0.3894	0.5216	0.2957	0.7069	0.2362	0.8072
	fep(0.5)	0.4055	0.5203	0.4049	0.5198	0.3209	0.7068	0.2532	0.8087
	fep(0.75)	0.3713	0.5070	0.3708	0.5050	0.2795	0.7121	0.2247	0.8183
	fep(0.9)	0.3321	0.4953	0.3261	0.4858	0.2296	0.7344	0.1829	0.8161
	fep(0.99)	0.2043	0.4410	0.1756	0.3788	0.1066	0.6867	0.0814	0.7666
LSCFG	—	0.8993	0.9412	0.8997	0.9409	0.8959	0.9513	0.8873	0.9574
	mep(0.5)	0.8993	0.9412	0.8909	0.9380	0.8903	0.9478	0.8819	0.9541
	mep(0.75)	0.8993	0.9411	0.8816	0.9348	0.8822	0.9459	0.8746	0.9528
	mep(0.9)	0.8993	0.9414	0.8745	0.9323	0.8739	0.9438	0.8666	0.9500
	mep(0.99)	0.8989	0.9414	0.8639	0.9269	0.8659	0.9408	0.8574	0.9482
	fep(0.5)	0.8993	0.9412	0.8796	0.9328	0.8798	0.9445	0.8716	0.9515
	fep(0.75)	0.8963	0.9400	0.8548	0.9217	0.8560	0.9346	0.8480	0.9432
	fep(0.9)	0.8854	0.9353	0.8240	0.9065	0.8260	0.9234	0.8170	0.9338
	fep(0.99)	0.8251	0.9052	0.7162	0.8375	0.7148	0.8661	0.6986	0.8879

(a) Sensitivity and PPV.

Approach	Errors	MEA struct.	Centroid
SCFG	—	0.409278	0.408549
	mep(0.5)	0.401914	0.400515
	mep(0.75)	0.397622	0.396770
	mep(0.9)	0.383750	0.383935
	mep(0.99)	0.376683	0.375488
	fep(0.5)	0.400827	0.397566
	fep(0.75)	0.363824	0.363257
	fep(0.9)	0.326873	0.325467
	fep(0.99)	0.189628	0.182902
LSCFG	—	0.914801	0.918933
	mep(0.5)	0.911963	0.915503
	mep(0.75)	0.908958	0.911579
	mep(0.9)	0.905646	0.908203
	mep(0.99)	0.902330	0.905126
	fep(0.5)	0.906507	0.911063
	fep(0.75)	0.893417	0.895371
	fep(0.9)	0.875529	0.877256
	fep(0.99)	0.776239	0.777355

(b) AUC values.

Table 8: Prediction results for our 5S rRNA database (computed by 10-fold cross-validation procedures, using sample size 1000 and $\min_{\text{hel}} = \min_{HL} = 1$).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.2413	0.4082	0.5548	0.5548	0.5552	0.6278
	mep(0.5)	0.2409	0.4068	0.5548	0.5548	0.5552	0.6265
	mep(0.75)	0.2335	0.3990	0.5506	0.5506	0.5511	0.6246
	mep(0.9)	0.2159	0.3809	0.5446	0.5446	0.5451	0.6135
	mep(0.99)	0.1877	0.3551	0.5382	0.5382	0.5386	0.6075
	fep(0.5)	0.2339	0.4017	0.5511	0.5511	0.5516	0.6269
	fep(0.75)	0.1586	0.3269	0.5257	0.5257	0.5261	0.5908
	fep(0.9)	0.0564	0.1979	0.4401	0.4401	0.4401	0.4952
	fep(0.99)	0.0014	0.0384	0.1979	0.1979	0.1984	0.2326
LSCFG	—	0.3324	0.4956	0.6574	0.6574	0.6579	0.7351
	mep(0.5)	0.3324	0.4956	0.6574	0.6574	0.6579	0.7351
	mep(0.75)	0.3329	0.4952	0.6579	0.6579	0.6584	0.7351
	mep(0.9)	0.3315	0.4901	0.6574	0.6574	0.6579	0.7351
	mep(0.99)	0.3236	0.4892	0.6560	0.6560	0.6565	0.7332
	fep(0.5)	0.3324	0.4966	0.6588	0.6588	0.6593	0.7369
	fep(0.75)	0.3232	0.4827	0.6551	0.6551	0.6556	0.7341
	fep(0.9)	0.2358	0.4055	0.6394	0.6399	0.6399	0.7166
	fep(0.99)	0.0624	0.2626	0.6246	0.6250	0.6250	0.6967

(a) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MP struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.2099	0.3699	0.5594	0.5594	0.5599	0.6302
	mep(0.5)	0.1683	0.3301	0.5372	0.5372	0.5377	0.6047
	mep(0.75)	0.1128	0.2700	0.5062	0.5062	0.5067	0.5682
	mep(0.9)	0.0712	0.2173	0.4808	0.4808	0.4813	0.5511
	mep(0.99)	0.0522	0.1822	0.4517	0.4517	0.4517	0.5215
	fep(0.5)	0.1049	0.2547	0.5155	0.5155	0.5160	0.5793
	fep(0.75)	0.0231	0.1317	0.4087	0.4087	0.4092	0.4623
	fep(0.9)	0.0032	0.0518	0.2918	0.2918	0.2918	0.3505
	fep(0.99)	0.0000	0.0125	0.1110	0.1110	0.1119	0.2062
LSCFG	—	0.3269	0.4892	0.6560	0.6565	0.6565	0.7337
	mep(0.5)	0.2534	0.4235	0.6708	0.6708	0.6713	0.7485
	mep(0.75)	0.1872	0.3666	0.6741	0.6741	0.6745	0.7550
	mep(0.9)	0.1502	0.3384	0.6694	0.6694	0.6699	0.7545
	mep(0.99)	0.1137	0.2954	0.6801	0.6801	0.6801	0.7568
	fep(0.5)	0.1794	0.3653	0.6704	0.6704	0.6709	0.7531
	fep(0.75)	0.0726	0.2492	0.6708	0.6717	0.6713	0.7596
	fep(0.9)	0.0301	0.1933	0.6847	0.6852	0.6857	0.7688
	fep(0.99)	0.0023	0.1262	0.6334	0.6334	0.6357	0.7240

(b) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MF struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0555	0.2094	0.4193	0.4193	0.4207	0.4679
	mep(0.5)	0.0416	0.1817	0.4045	0.4045	0.4055	0.4489
	mep(0.75)	0.0222	0.1456	0.3694	0.3699	0.3703	0.4147
	mep(0.9)	0.0148	0.1179	0.3555	0.3560	0.3583	0.4031
	mep(0.99)	0.0125	0.0989	0.3112	0.3112	0.3126	0.3570
	fep(0.5)	0.0245	0.1364	0.3662	0.3662	0.3666	0.4059
	fep(0.75)	0.0069	0.0712	0.2682	0.2686	0.2705	0.3070
	fep(0.9)	0.0005	0.0240	0.1655	0.1655	0.1669	0.2006
	fep(0.99)	0.0000	0.0014	0.0245	0.0245	0.0250	0.0546
LSCFG	—	0.1854	0.3574	0.4919	0.4919	0.4919	0.5465
	mep(0.5)	0.1405	0.3056	0.4998	0.4998	0.4998	0.5567
	mep(0.75)	0.1128	0.2760	0.4864	0.4873	0.4864	0.5432
	mep(0.9)	0.0924	0.2478	0.4827	0.4827	0.4827	0.5377
	mep(0.99)	0.0730	0.2191	0.4753	0.4753	0.4753	0.5284
	fep(0.5)	0.1003	0.2556	0.4836	0.4836	0.4836	0.5409
	fep(0.75)	0.0532	0.2011	0.4771	0.4776	0.4771	0.5423
	fep(0.9)	0.0213	0.1341	0.4508	0.4517	0.4508	0.5095
	fep(0.99)	0.0009	0.0781	0.3902	0.3902	0.3921	0.4508

(c) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MEA struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0374	0.1276	0.2973	0.2973	0.2977	0.3130
	mep(0.5)	0.0273	0.1045	0.2779	0.2779	0.2783	0.2908
	mep(0.75)	0.0139	0.0716	0.2362	0.2362	0.2362	0.2520
	mep(0.9)	0.0074	0.0656	0.2354	0.2354	0.2354	0.2502
	mep(0.99)	0.0083	0.0541	0.2007	0.2007	0.2007	0.2173
	fep(0.5)	0.0134	0.0795	0.2473	0.2473	0.2473	0.2603
	fep(0.75)	0.0037	0.0360	0.1609	0.1609	0.1609	0.1734
	fep(0.9)	0.0000	0.0069	0.0865	0.0865	0.0869	0.0939
	fep(0.99)	0.0000	0.0009	0.0120	0.0120	0.0120	0.0227
LSCFG	—	0.1729	0.3158	0.4300	0.4300	0.4300	0.4762
	mep(0.5)	0.1322	0.2728	0.4374	0.4374	0.4374	0.4859
	mep(0.75)	0.1100	0.2469	0.4258	0.4258	0.4258	0.4748
	mep(0.9)	0.0874	0.2140	0.4189	0.4189	0.4189	0.4660
	mep(0.99)	0.0693	0.1914	0.4101	0.4101	0.4101	0.4558
	fep(0.5)	0.0957	0.2261	0.4207	0.4207	0.4207	0.4642
	fep(0.75)	0.0481	0.1688	0.4046	0.4046	0.4046	0.4559
	fep(0.9)	0.0199	0.1146	0.3828	0.3833	0.3828	0.4262
	fep(0.99)	0.0009	0.0633	0.3264	0.3264	0.3269	0.3648

(d) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle Centroid).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.6838	0.9459	0.9903	0.9903	0.9908	0.9995
	mep(0.5)	0.6274	0.9376	0.9880	0.9884	0.9889	0.9995
	mep(0.75)	0.5724	0.9274	0.9898	0.9908	0.9908	1.0000
	mep(0.9)	0.4707	0.9219	0.9866	0.9871	0.9875	1.0000
	mep(0.99)	0.3837	0.9057	0.9898	0.9903	0.9908	0.9995
	fep(0.5)	0.5534	0.9293	0.9884	0.9889	0.9889	0.9995
	fep(0.75)	0.2903	0.8849	0.9852	0.9857	0.9861	0.9995
	fep(0.9)	0.0883	0.8077	0.9838	0.9843	0.9843	0.9995
	fep(0.99)	0.0018	0.4808	0.9556	0.9575	0.9603	0.9931
LSCFG	—	0.8234	0.9288	0.9723	0.9750	0.9727	0.9986
	mep(0.5)	0.8169	0.9311	0.9658	0.9681	0.9663	0.9986
	mep(0.75)	0.7827	0.9260	0.9732	0.9760	0.9737	0.9986
	mep(0.9)	0.7291	0.9191	0.9718	0.9732	0.9723	0.9986
	mep(0.99)	0.6653	0.9122	0.9709	0.9746	0.9713	0.9986
	fep(0.5)	0.7735	0.9173	0.9704	0.9741	0.9709	0.9986
	fep(0.75)	0.6191	0.9048	0.9686	0.9713	0.9690	0.9986
	fep(0.9)	0.3777	0.8604	0.9732	0.9750	0.9736	0.9986
	fep(0.99)	0.0763	0.7106	0.9663	0.9686	0.9667	0.9986

(e) $\mathrm{CSO}_{\mathrm{freq}}$ values.

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	16.202	98.357	327.26	327.27	327.51	418.80
	mep(0.5)	13.511	90.408	314.52	314.53	314.83	405.33
	mep(0.75)	9.9097	77.641	295.10	295.12	295.47	387.67
	mep(0.9)	7.1723	66.885	278.33	278.35	278.81	373.45
	mep(0.99)	5.2356	56.709	255.51	255.54	256.20	354.90
	fep(0.5)	9.7356	77.552	294.89	294.91	295.31	387.38
	fep(0.75)	3.0058	48.215	239.50	239.53	240.34	333.77
	fep(0.9)	0.5193	22.122	171.61	171.72	173.31	270.84
	fep(0.99)	0.0028	5.3030	62.460	62.587	66.606	168.64
LSCFG	—	101.69	326.26	708.52	708.94	709.42	805.87
	mep(0.5)	90.408	307.25	712.29	712.73	713.23	810.52
	mep(0.75)	75.220	288.57	710.49	710.94	711.54	810.28
	mep(0.9)	62.276	270.87	708.82	709.20	709.93	809.72
	mep(0.99)	51.262	252.51	708.00	708.35	709.20	807.00
	fep(0.5)	70.493	281.00	710.79	711.18	711.75	810.24
	fep(0.75)	40.142	229.20	704.79	705.33	706.27	807.32
	fep(0.9)	20.373	193.27	695.77	696.29	698.20	802.86
	fep(0.99)	1.6771	118.55	612.73	612.89	620.77	752.68

(f) $\mathrm{CS}_{\mathrm{num}}$ values.

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	802.27	244.52	60.504	60.030	59.916	28.764
	mep(0.5)	805.91	250.25	63.247	62.668	62.520	29.780
	mep(0.75)	812.80	259.32	67.778	67.072	66.880	31.417
	mep(0.9)	812.01	267.01	72.572	71.611	71.323	32.755
	mep(0.99)	822.99	285.70	81.360	80.006	79.564	35.462
	fep(0.5)	813.01	261.18	68.849	68.053	67.839	31.848
	fep(0.75)	820.06	289.39	86.049	84.434	83.880	37.363
	fep(0.9)	823.37	338.60	120.77	116.73	114.89	47.031
	fep(0.99)	787.70	437.67	225.13	209.08	198.18	68.691
LSCFG	—	238.30	15.045	5.6854	5.4122	5.1806	3.2274
	mep(0.5)	237.12	15.061	5.7478	5.4543	5.2231	3.1970
	mep(0.75)	234.02	15.304	5.9289	5.6123	5.3695	3.2496
	mep(0.9)	230.18	15.572	6.0783	5.7631	5.5047	3.2811
	mep(0.99)	226.25	16.097	6.3746	6.0086	5.7673	3.3176
	fep(0.5)	234.80	15.367	6.0292	5.7213	5.4755	3.2501
	fep(0.75)	215.02	15.921	6.4449	6.0982	5.8435	3.2944
	fep(0.9)	199.10	17.503	7.5125	7.0207	6.7294	3.4531
	fep(0.99)	164.88	22.047	10.309	9.5385	9.1512	3.7212

(g) $\mathrm{DS}_{\mathrm{num}}$ values.

Table 9: Specific values related to shapes of predictions and sampled structures, obtained from our tRNA database (by 10-fold cross-validation procedures, using sample size 1000 and $\min_{\mathrm{hel}} = \min_{HL} = 1$).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0000	0.0026	0.0052	0.0131	0.0366	0.7110
	mep(0.5)	0.0000	0.0009	0.0026	0.0113	0.0287	0.7128
	mep(0.75)	0.0000	0.0017	0.0035	0.0105	0.0322	0.7050
	mep(0.9)	0.0000	0.0009	0.0017	0.0078	0.0331	0.7180
	mep(0.99)	0.0000	0.0026	0.0044	0.0095	0.0227	0.6919
	fep(0.5)	0.0000	0.0017	0.0043	0.0113	0.0374	0.6954
	fep(0.75)	0.0000	0.0000	0.0009	0.0113	0.0321	0.6710
	fep(0.9)	0.0000	0.0009	0.0009	0.0052	0.0261	0.6536
	fep(0.99)	0.0000	0.0000	0.0000	0.0017	0.0096	0.5474
LSCFG	—	0.2141	0.4256	0.4744	0.4900	0.9408	0.9843
	mep(0.5)	0.2141	0.4256	0.4744	0.4900	0.9408	0.9843
	mep(0.75)	0.2141	0.4248	0.4726	0.4892	0.9399	0.9843
	mep(0.9)	0.2089	0.4274	0.4761	0.4926	0.9399	0.9843
	mep(0.99)	0.1941	0.4221	0.4761	0.4892	0.9452	0.9852
	fep(0.5)	0.2124	0.4248	0.4726	0.4883	0.9417	0.9852
	fep(0.75)	0.1898	0.4213	0.4674	0.4831	0.9408	0.9843
	fep(0.9)	0.1314	0.4013	0.4518	0.4726	0.9321	0.9869
	fep(0.99)	0.0209	0.3029	0.3725	0.4186	0.8529	0.9809

(a) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MP struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0000	0.0026	0.0052	0.0131	0.0357	0.7128
	mep(0.5)	0.0000	0.0009	0.0026	0.0122	0.0305	0.7180
	mep(0.75)	0.0000	0.0017	0.0043	0.0113	0.0331	0.7067
	mep(0.9)	0.0000	0.0009	0.0017	0.0078	0.0357	0.7215
	mep(0.99)	0.0000	0.0026	0.0044	0.0105	0.0235	0.6902
	fep(0.5)	0.0000	0.0017	0.0043	0.0113	0.0383	0.6971
	fep(0.75)	0.0000	0.0000	0.0009	0.0113	0.0296	0.6745
	fep(0.9)	0.0000	0.0000	0.0000	0.0035	0.0261	0.6631
	fep(0.99)	0.0000	0.0000	0.0000	0.0035	0.0200	0.5439
LSCFG	—	0.2002	0.4256	0.4700	0.4866	0.9417	0.9861
	mep(0.5)	0.1332	0.3960	0.4439	0.4587	0.9434	0.9869
	mep(0.75)	0.0923	0.3847	0.4448	0.4639	0.9356	0.9861
	mep(0.9)	0.0575	0.3508	0.4135	0.4352	0.9373	0.9887
	mep(0.99)	0.0365	0.3630	0.4308	0.4491	0.9304	0.9861
	fep(0.5)	0.0801	0.3847	0.4404	0.4561	0.9400	0.9861
	fep(0.75)	0.0339	0.3630	0.4230	0.4430	0.9208	0.9843
	fep(0.9)	0.0131	0.3160	0.3743	0.4204	0.8442	0.9843
	fep(0.99)	0.0035	0.1497	0.2106	0.3325	0.5440	0.9730

(b) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MF struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0000	0.0000	0.0000	0.0000	0.0261	0.3821
	mep(0.5)	0.0000	0.0000	0.0000	0.0000	0.0209	0.3698
	mep(0.75)	0.0000	0.0000	0.0000	0.0000	0.0209	0.3559
	mep(0.9)	0.0000	0.0000	0.0000	0.0000	0.0131	0.3290
	mep(0.99)	0.0000	0.0000	0.0000	0.0000	0.0122	0.3003
	fep(0.5)	0.0000	0.0000	0.0000	0.0000	0.0252	0.3438
	fep(0.75)	0.0000	0.0000	0.0000	0.0000	0.0139	0.2463
	fep(0.9)	0.0000	0.0000	0.0000	0.0000	0.0070	0.1619
	fep(0.99)	0.0000	0.0000	0.0000	0.0000	0.0026	0.0444
LSCFG	—	0.1062	0.3891	0.4291	0.4378	0.9051	0.9835
	mep(0.5)	0.1010	0.3751	0.4134	0.4239	0.8921	0.9782
	mep(0.75)	0.0749	0.3647	0.4047	0.4282	0.8894	0.9774
	mep(0.9)	0.0470	0.3290	0.3769	0.3917	0.8834	0.9817
	mep(0.99)	0.0392	0.3429	0.3986	0.4213	0.8712	0.9791
	fep(0.5)	0.0740	0.3839	0.4239	0.4387	0.8877	0.9791
	fep(0.75)	0.0287	0.3516	0.3943	0.4134	0.8616	0.9713
	fep(0.9)	0.0139	0.2968	0.3490	0.3855	0.8120	0.9739
	fep(0.99)	0.0017	0.1358	0.1863	0.2942	0.4970	0.9634

(c) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle MEA struct.).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0000	0.0000	0.0000	0.0000	0.0104	0.1097
	mep(0.5)	0.0000	0.0000	0.0000	0.0000	0.0104	0.1062
	mep(0.75)	0.0000	0.0000	0.0000	0.0000	0.0078	0.0923
	mep(0.9)	0.0000	0.0000	0.0000	0.0000	0.0044	0.0896
	mep(0.99)	0.0000	0.0000	0.0000	0.0000	0.0078	0.0827
	fep(0.5)	0.0000	0.0000	0.0000	0.0000	0.0061	0.0932
	fep(0.75)	0.0000	0.0000	0.0000	0.0000	0.0026	0.0696
	fep(0.9)	0.0000	0.0000	0.0000	0.0000	0.0017	0.0479
	fep(0.99)	0.0000	0.0000	0.0000	0.0000	0.0009	0.0078
LSCFG	—	0.0966	0.2916	0.3238	0.3316	0.8703	0.9686
	mep(0.5)	0.0879	0.3142	0.3516	0.3621	0.8625	0.9686
	mep(0.75)	0.0644	0.3029	0.3403	0.3551	0.8407	0.9678
	mep(0.9)	0.0427	0.2829	0.3194	0.3299	0.8451	0.9678
	mep(0.99)	0.0322	0.2924	0.3377	0.3595	0.8294	0.9651
	fep(0.5)	0.0662	0.3194	0.3551	0.3638	0.8512	0.9695
	fep(0.75)	0.0261	0.2907	0.3255	0.3516	0.8103	0.9608
	fep(0.9)	0.0113	0.2411	0.2872	0.3194	0.7650	0.9565
	fep(0.99)	0.0017	0.1053	0.1471	0.2219	0.4831	0.9339

(d) $\mathrm{CSP}_{\mathrm{freq}}$ values (for selection principle Centroid).

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0000	0.2855	0.4526	0.9852	0.9974	1.0000
	mep(0.5)	0.0000	0.2750	0.4256	0.9835	0.9991	1.0000
	mep(0.75)	0.0000	0.2367	0.3768	0.9774	0.9982	1.0000
	mep(0.9)	0.0000	0.2185	0.3394	0.9696	0.9965	1.0000
	mep(0.99)	0.0000	0.1715	0.2977	0.9756	0.9991	1.0000
	fep(0.5)	0.0000	0.2237	0.3543	0.9739	0.9991	1.0000
	fep(0.75)	0.0000	0.1584	0.2472	0.9574	0.9957	1.0000
	fep(0.9)	0.0000	0.0749	0.1497	0.9147	0.9930	1.0000
	fep(0.99)	0.0000	0.0174	0.0296	0.6763	0.9608	1.0000
LSCFG	—	0.6258	0.8912	0.9295	0.9504	0.9948	1.0000
	mep(0.5)	0.6084	0.8947	0.9286	0.9469	0.9948	1.0000
	mep(0.75)	0.5727	0.8886	0.9269	0.9521	0.9957	1.0000
	mep(0.9)	0.5231	0.8851	0.9252	0.9521	0.9948	1.0000
	mep(0.99)	0.4630	0.8894	0.9199	0.9452	0.9939	1.0000
	fep(0.5)	0.5553	0.8868	0.9234	0.9504	0.9948	1.0000
	fep(0.75)	0.4248	0.8894	0.9225	0.9521	0.9948	1.0000
	fep(0.9)	0.2393	0.8720	0.9077	0.9504	0.9957	1.0000
	fep(0.99)	0.0279	0.7580	0.8460	0.9617	0.9939	1.0000

(e) $\mathrm{CSO}_{\mathrm{freq}}$ values.

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	0.0000	0.5432	1.1811	20.640	51.834	573.72
	mep(0.5)	0.0000	0.4980	1.0481	19.498	49.614	566.13
	mep(0.75)	0.0000	0.3977	0.8859	18.385	47.128	556.17
	mep(0.9)	0.0000	0.3655	0.7353	16.331	43.735	544.22
	mep(0.99)	0.0000	0.2768	0.5850	14.689	40.062	527.12
	fep(0.5)	0.0000	0.3865	0.7982	17.401	45.868	552.82
	fep(0.75)	0.0000	0.2481	0.4961	13.092	38.088	507.97
	fep(0.9)	0.0000	0.0957	0.2141	8.7270	27.742	443.65
	fep(0.99)	0.0000	0.0191	0.0348	2.9269	12.064	285.20
LSCFG	—	42.599	347.33	421.29	455.78	881.11	983.88
	mep(0.5)	38.324	346.45	419.23	455.04	875.92	983.26
	mep(0.75)	31.890	338.26	411.90	451.93	865.28	983.84
	mep(0.9)	23.180	316.85	389.48	434.11	853.47	984.36
	mep(0.99)	17.873	312.64	386.51	436.20	832.96	983.29
	fep(0.5)	29.194	342.28	413.57	454.07	861.46	983.23
	fep(0.75)	14.829	304.49	376.16	430.85	811.76	980.03
	fep(0.9)	7.7946	250.74	312.98	391.21	713.62	980.94
	fep(0.99)	0.9219	93.694	133.47	260.11	418.20	970.01

(f) $\mathrm{CS}_{\mathrm{num}}$ values.

Approach	Errors	Shape Level					
		0	1	2	3	4	5
SCFG	—	999.67	941.77	866.98	336.69	167.10	16.476
	mep(0.5)	999.59	943.47	871.08	345.35	171.57	16.766
	mep(0.75)	999.61	946.99	878.58	358.32	179.50	17.508
	mep(0.9)	999.53	949.65	884.73	372.57	188.32	18.198
	mep(0.99)	999.49	953.90	894.21	393.84	201.28	18.948
	fep(0.5)	999.53	947.08	879.39	363.20	182.17	17.663
	fep(0.75)	999.39	955.12	898.94	414.68	213.39	20.622
	fep(0.9)	998.86	962.12	917.65	484.74	258.19	25.174
	fep(0.99)	996.37	966.76	933.73	632.35	367.71	40.976
LSCFG	—	318.99	24.878	19.283	8.2879	4.4246	1.2088
	mep(0.5)	320.15	25.352	19.707	8.3590	4.4759	1.2245
	mep(0.75)	318.55	26.391	20.555	8.5940	4.6395	1.2271
	mep(0.9)	320.83	27.964	21.834	8.8192	4.7788	1.2219
	mep(0.99)	326.13	30.176	23.552	9.1464	4.9729	1.2463
	fep(0.5)	321.32	26.848	21.000	8.6976	4.6169	1.2202
	fep(0.75)	324.45	31.466	24.610	9.4810	5.1226	1.2445
	fep(0.9)	336.16	41.060	32.527	11.069	6.0148	1.2880
	fep(0.99)	401.88	84.057	69.689	18.023	9.1200	1.3690

(g) $\mathrm{DS}_{\mathrm{num}}$ values.

Table 10: Specific values related to shapes of predictions and sampled structures, obtained from our 5S rRNA database (by 10-fold cross-validation procedures, using sample size 1000 and $\min_{\mathrm{hel}} = \min_{HL} = 1$).

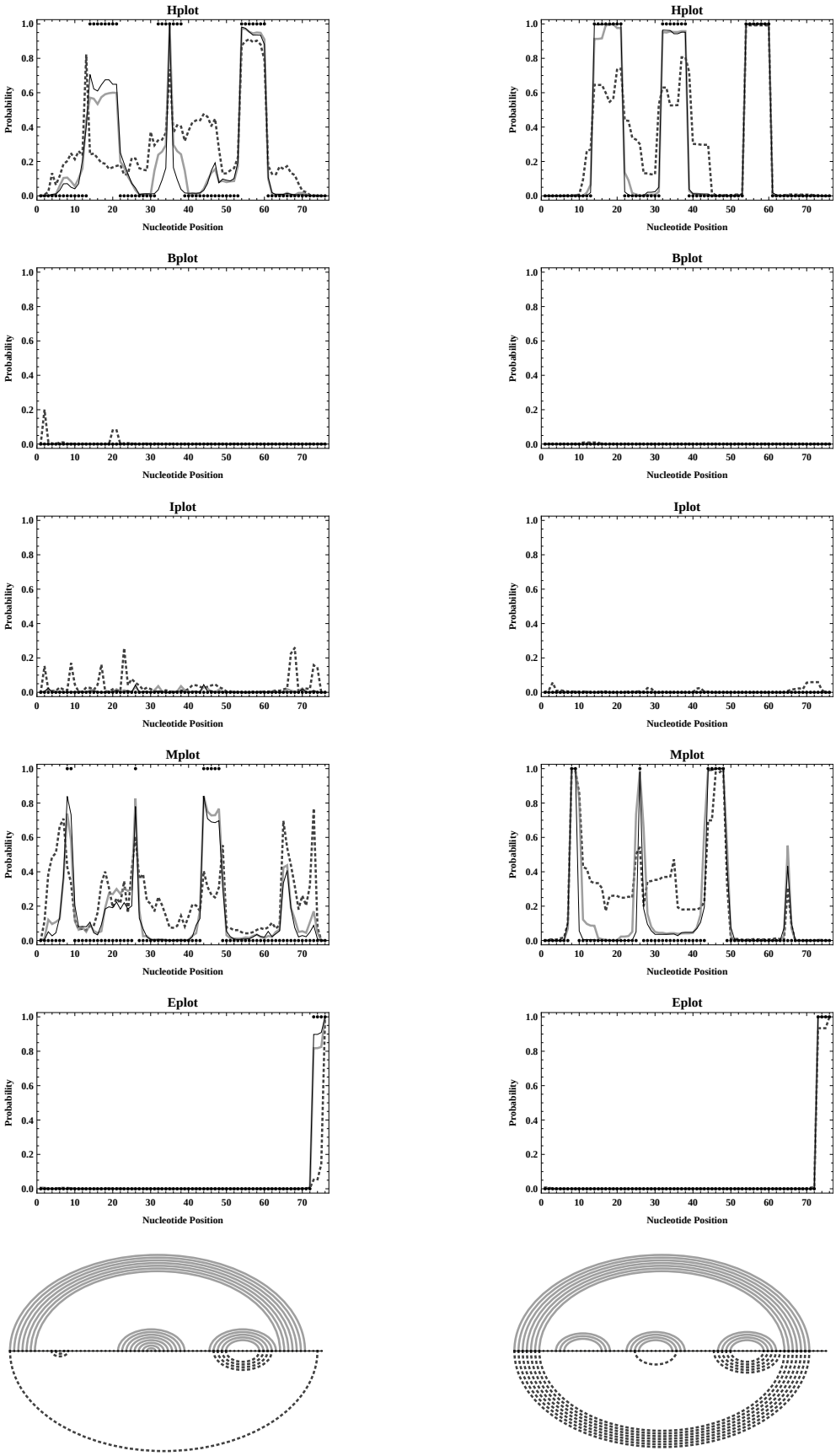


Figure 5: Loop profiles and centroid for *E.coli* tRNA^{Ala} derived according to mep(*prob*) (thick gray lines) and fep(*prob*) (thick dotted darker gray lines) under the assumption of the SCFG (figures on the left) and LSCFG (figures on the right) model, respectively, where percentage $prob = 0.99$ has been used for generating the relative errors. Hplot, Bplot, Iplot, Mplot and Eplot display the probability that an unpaired base lies in a hairpin, bulge, interior, multi-branched and exterior loop, respectively.

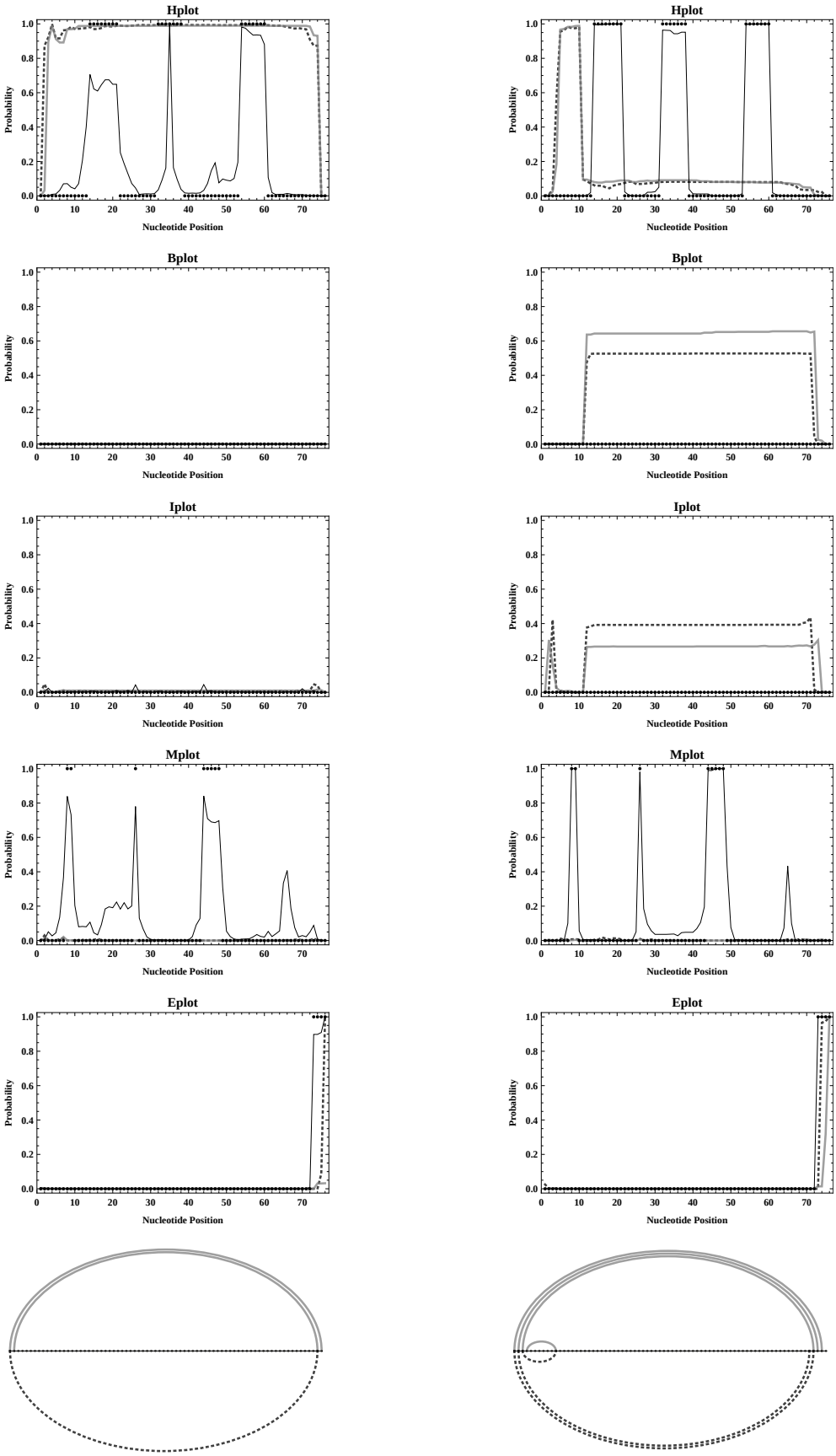


Figure 6: Loop profiles and centroid for *E.coli* tRNA^{Ala} derived according to mev(*prob*) (thick gray lines) and fev(*prob*) (thick dotted darker gray lines) under the assumption of the SCFG (figures on the left) and LSCFG (figures on the right) model, respectively, where fixed value $prob = 10^{-9}$ has been used for generating the absolute errors. Hplot, Bplot, Iplot, Mplot and Eplot display the probability that an unpaired base lies in a hairpin, bulge, interior, multi-branched and exterior loop, respectively.

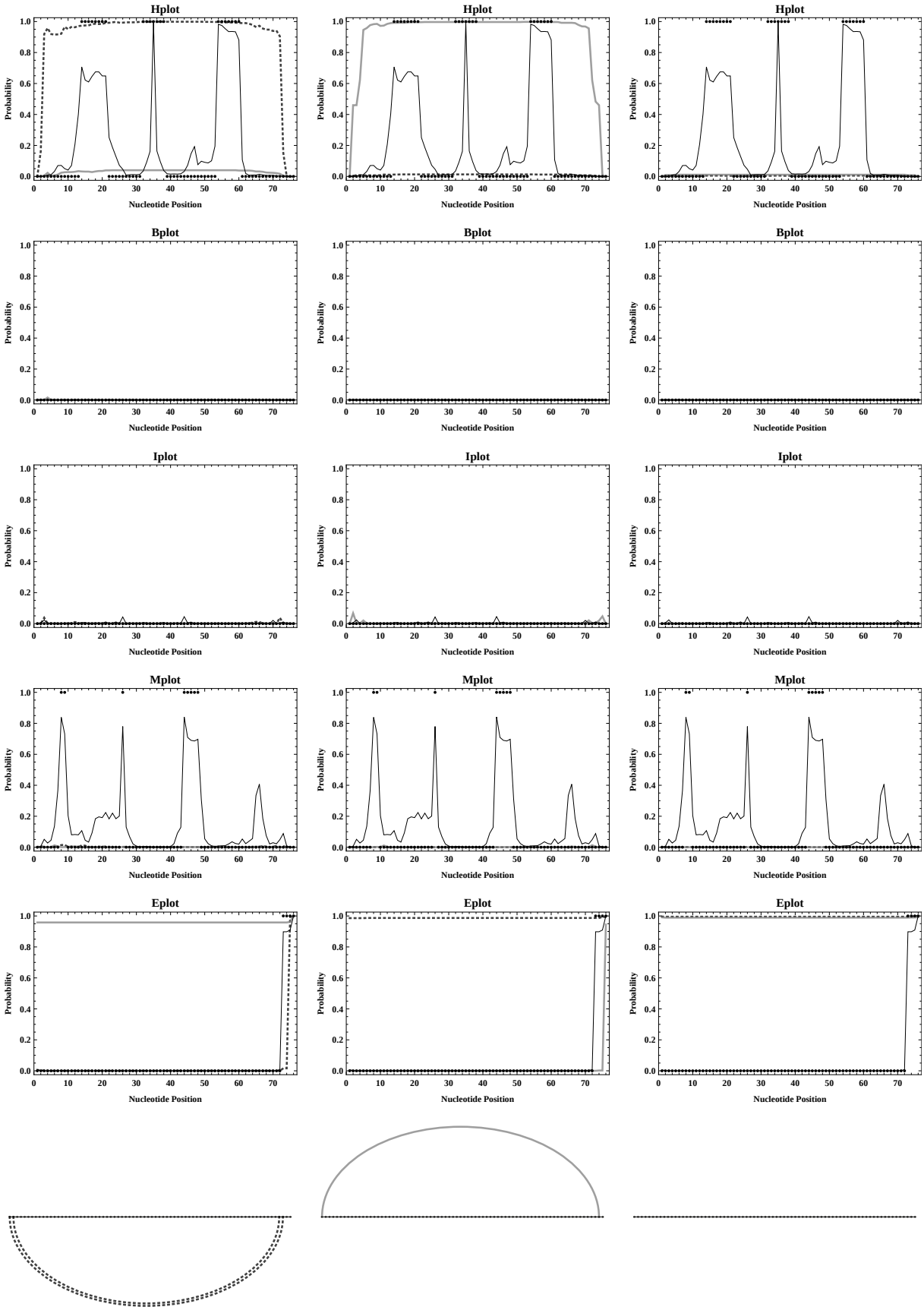


Figure 7: Loop profiles and centroid derived according to $\mathrm{mev}^{win,+}(prob)$ (thick gray lines) and $\mathrm{fev}^{win,+}(prob)$ (thick dotted darker gray lines) for the traditional SCFG model, respectively, where $prob = 10^{-9}$ and $win \in \{15, 38, 60\}$ (figures from left to right).

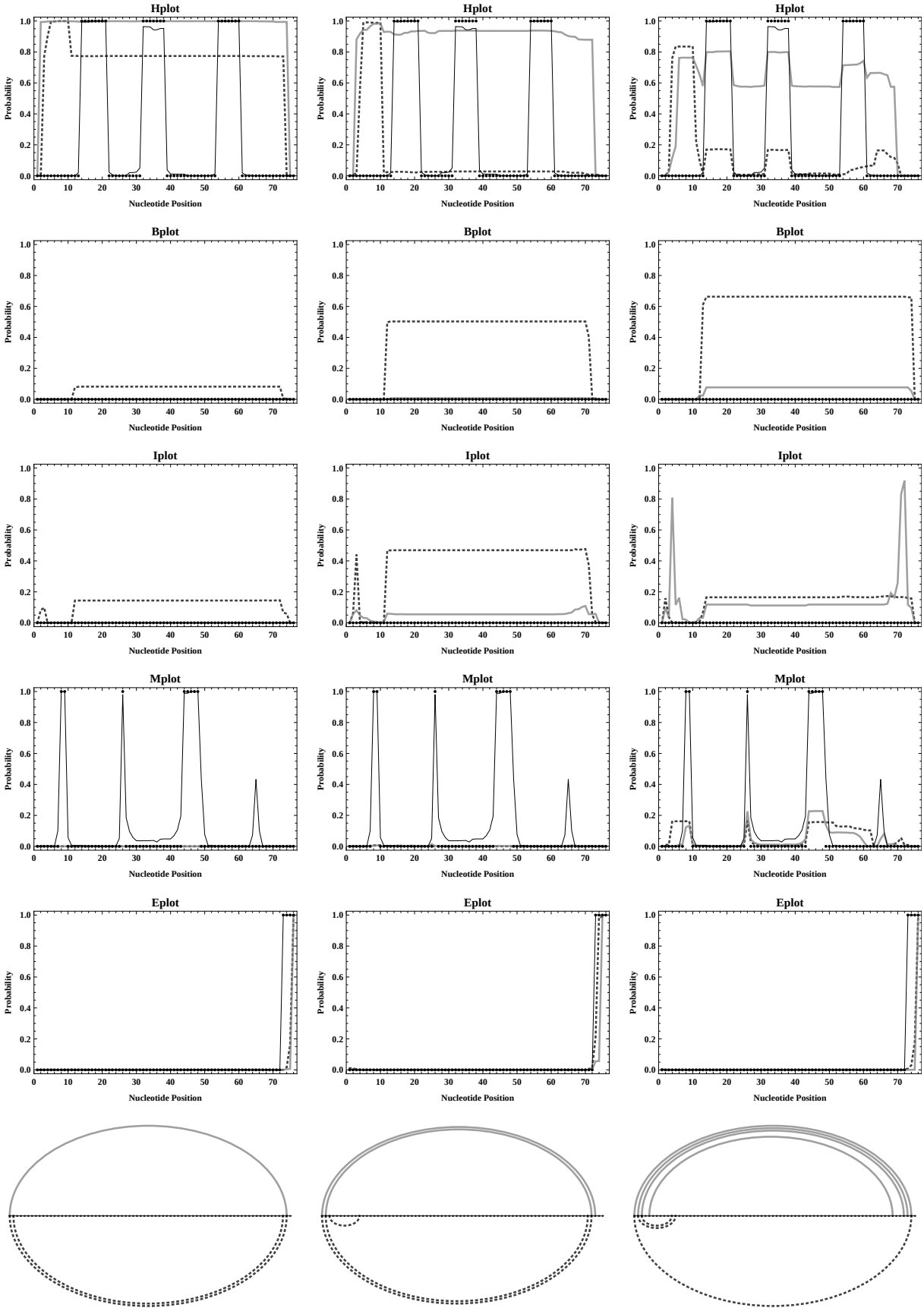


Figure 8: Loop profiles and centroid derived according to $\mathrm{mev}^{win,+}(prob)$ (thick gray lines) and $\mathrm{fev}^{win,+}(prob)$ (thick dotted darker gray lines) for the LSCFG model, respectively, where $prob = 10^{-9}$ and $win \in \{15, 38, 60\}$ (figures from left to right).

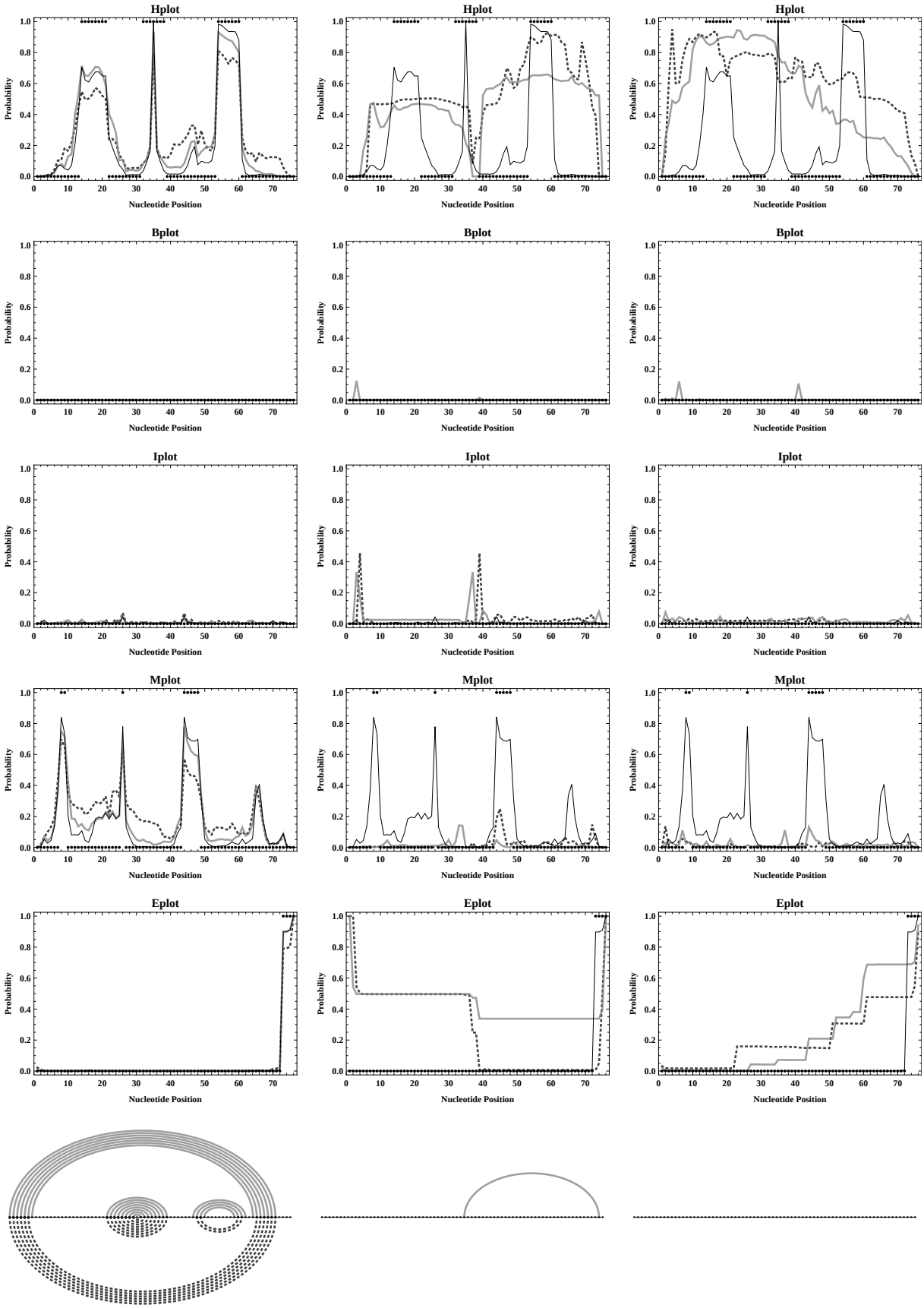


Figure 9: Loop profiles and centroid derived according to $\mathrm{mev}^{win,-}(prob)$ (thick gray lines) and $\mathrm{fev}^{win,-}(prob)$ (thick dotted darker gray lines) for the traditional SCFG model, respectively, where $prob = 10^{-9}$ and $win \in \{15, 38, 60\}$ (figures from left to right).

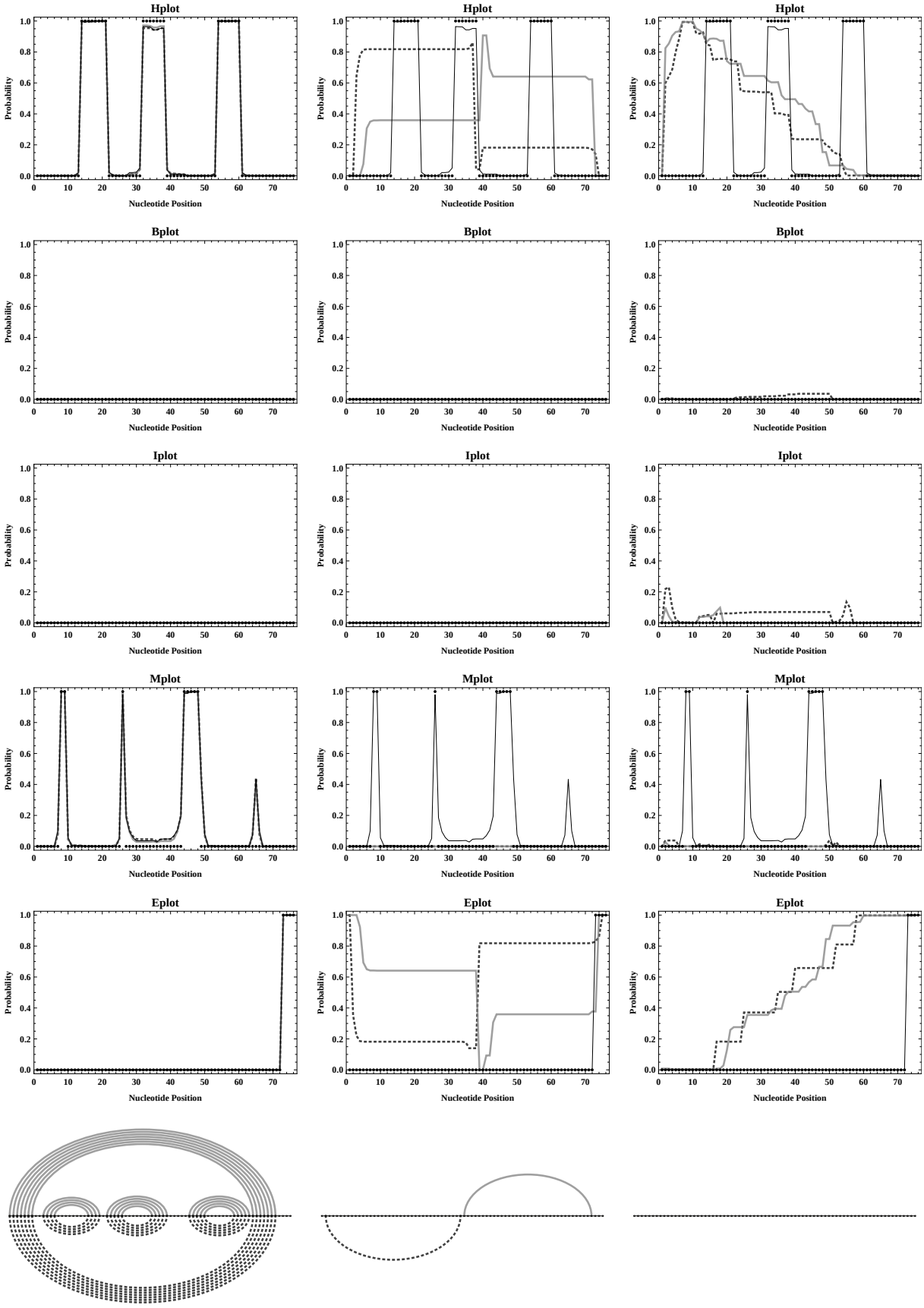


Figure 10: Loop profiles and centroid derived according to $\mathrm{mev}^{win,-}(prob)$ (thick gray lines) and $\mathrm{fev}^{win,-}(prob)$ (thick dotted darker gray lines) for the LSCFG model, respectively, where $prob = 10^{-9}$ and $win \in \{15, 38, 60\}$ (figures from left to right).

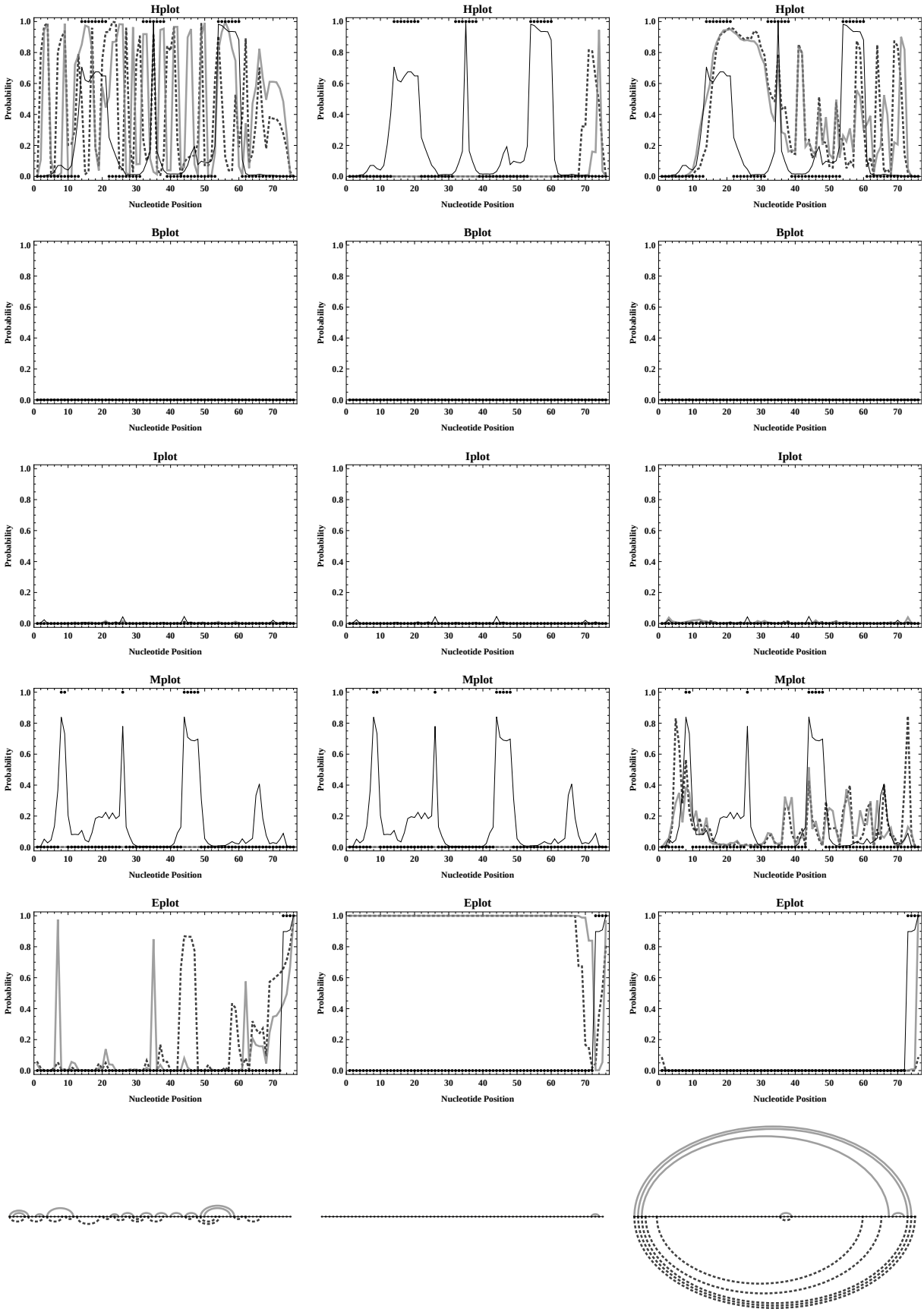


Figure 11: Loop profiles and centroid derived according to $\mathrm{mev}_{\mathcal{I}}(prob)$ (thick gray lines) and $\mathrm{fev}_{\mathcal{I}}(prob)$ (thick dotted darker gray lines) for the traditional SCFG model, respectively, where $prob = 10^{-9}$ and $\mathcal{I} \in \{\{T\}, \{C\}, \{A\}\}$ (figures from left to right).

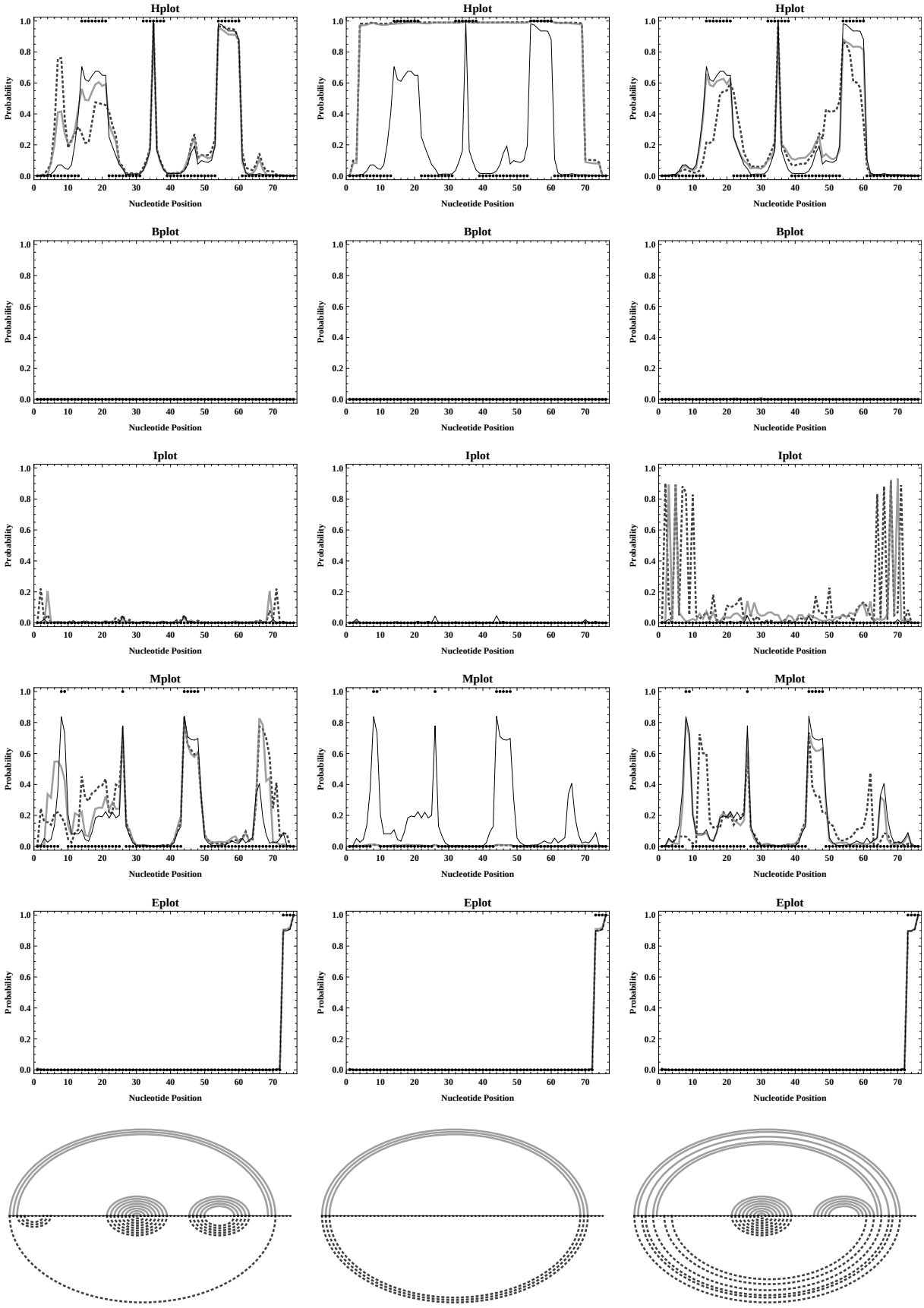


Figure 12: Loop profiles and centroid derived according to $\mathrm{mev}_{\mathcal{I}}(prob)$ (thick gray lines) and $\mathrm{fev}_{\mathcal{I}}(prob)$ (thick dotted darker gray lines) for the traditional SCFG model, respectively, where $prob = 10^{-9}$ and $\mathcal{I} \in \{\{P\}, \{F\}, \{G\}\}$ (figures from left to right).

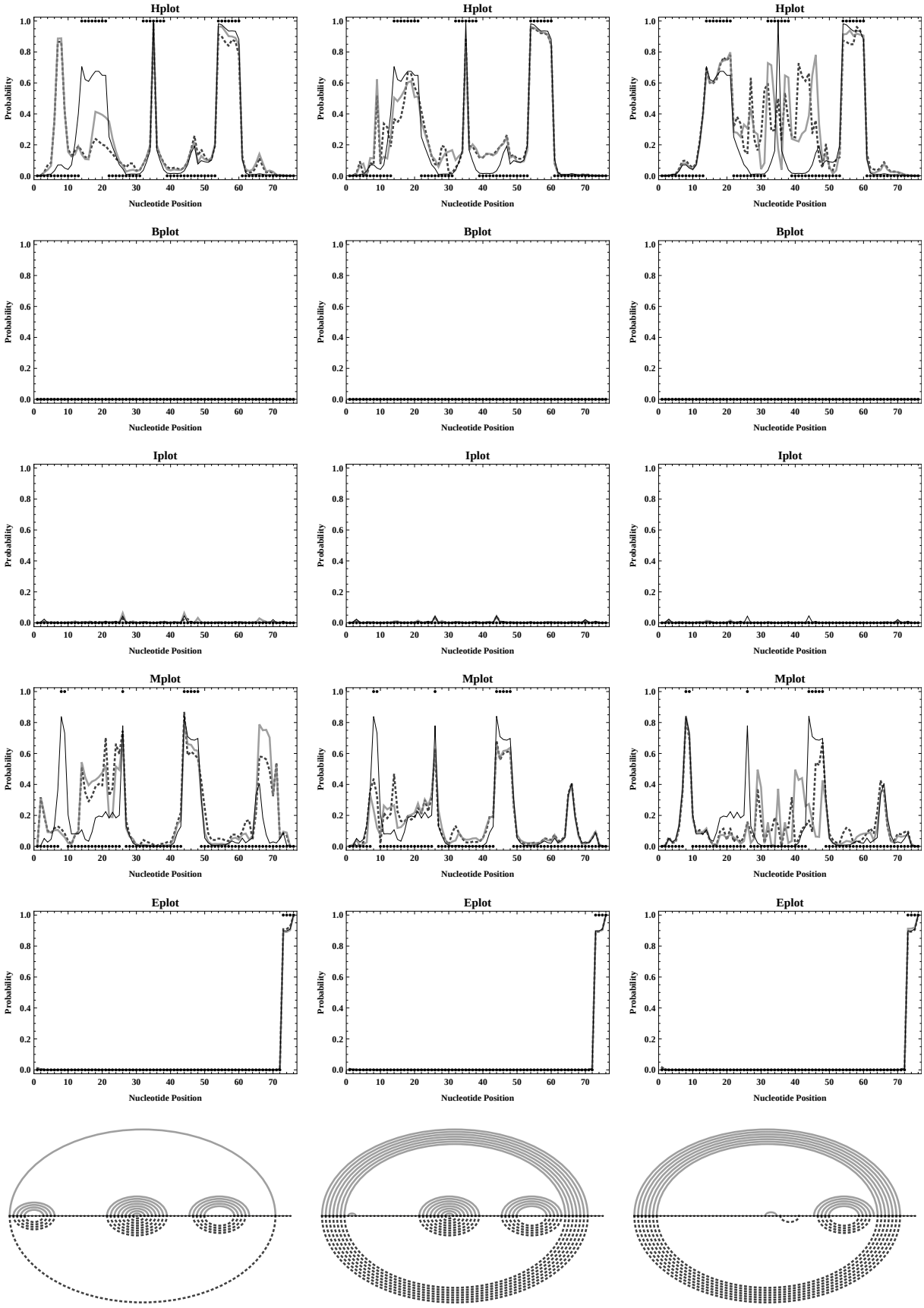


Figure 13: Loop profiles and centroid derived according to $\mathrm{mev}_{\mathcal{I}}(prob)$ (thick gray lines) and $\mathrm{fev}_{\mathcal{I}}(prob)$ (thick dotted darker gray lines) for the traditional SCFG model, respectively, where $prob = 10^{-9}$ and $\mathcal{I} \in \{\{M\}, \{O\}, \{N\}\}$ (figures from left to right).

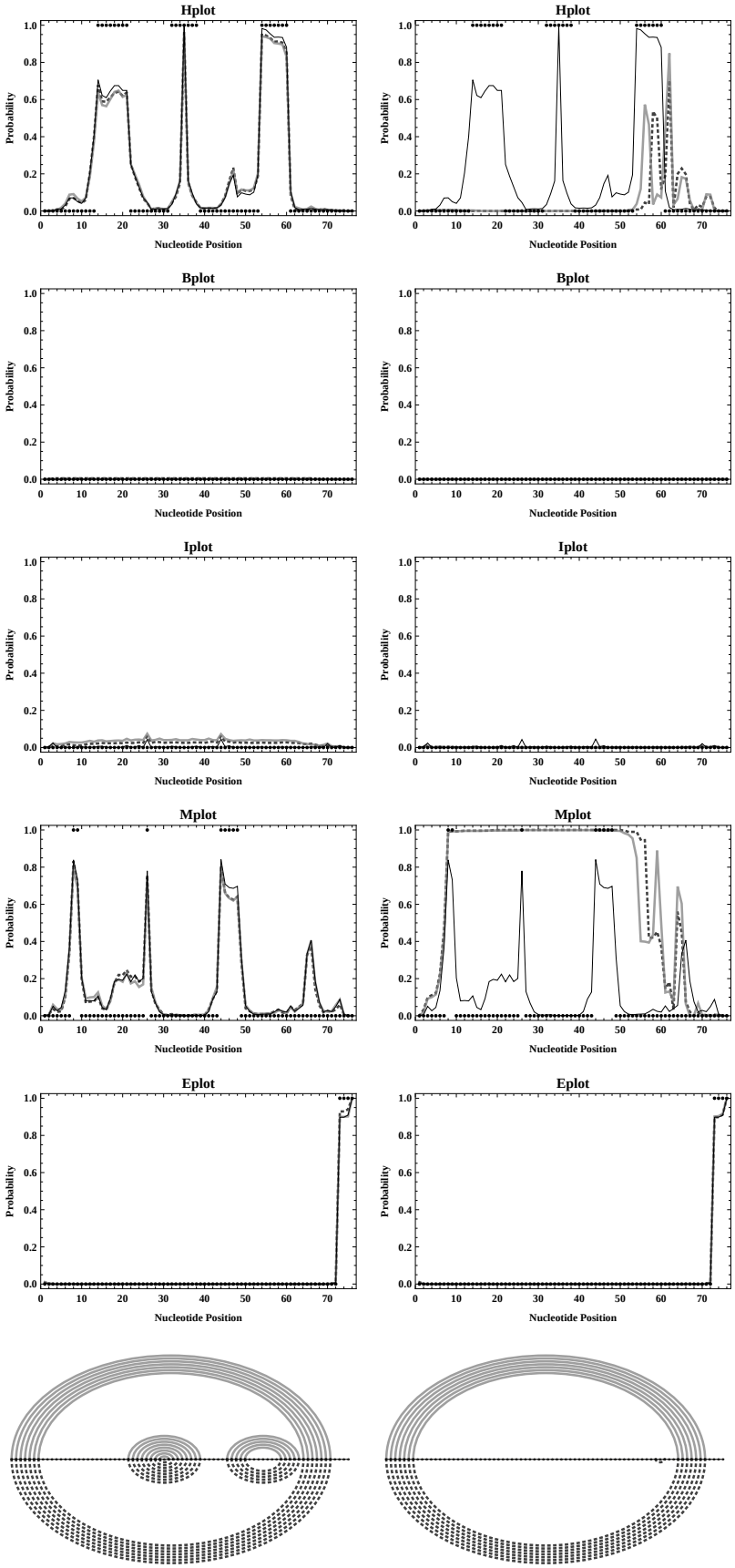


Figure 14: Loop profiles and centroid derived according to $\mathrm{mev}_{\mathcal{I}}(prob)$ (thick gray lines) and $\mathrm{fev}_{\mathcal{I}}(prob)$ (thick dotted darker gray lines) for the traditional SCFG model, respectively, where $prob = 10^{-9}$ and $\mathcal{I} \in \{\{B\}, \{U\}\}$ (figures from left to right).

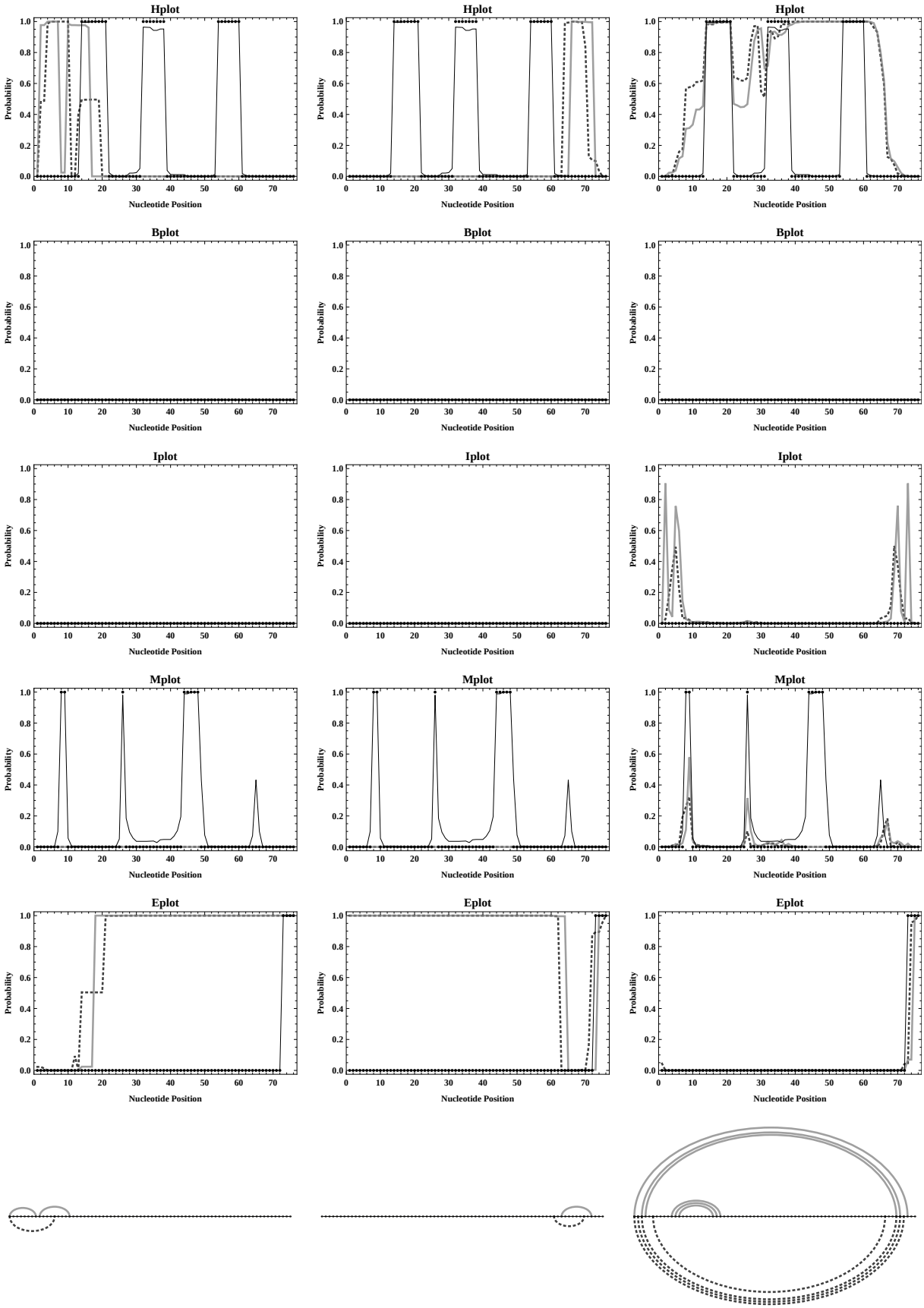


Figure 15: Loop profiles and centroid derived according to $\mathrm{mev}_{\mathcal{I}}(prob)$ (thick gray lines) and $\mathrm{fev}_{\mathcal{I}}(prob)$ (thick dotted darker gray lines) for the LSCFG model, respectively, where $prob = 10^{-9}$ and $\mathcal{I} \in \{\{T\}, \{C\}, \{A\}\}$ (figures from left to right).

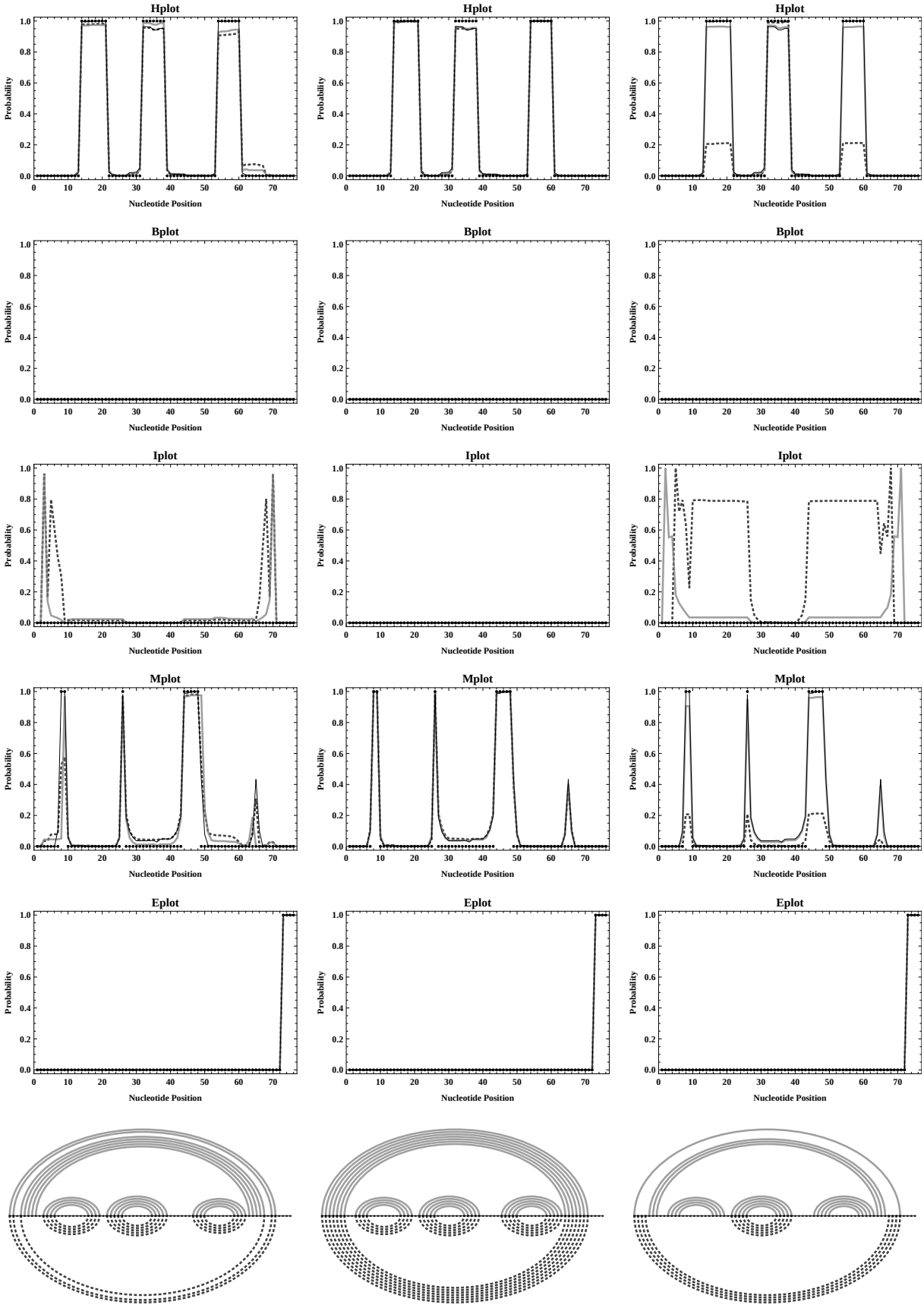


Figure 16: Loop profiles and centroid derived according to $\mathrm{mev}_{\mathcal{I}}(prob)$ (thick gray lines) and $\mathrm{fev}_{\mathcal{I}}(prob)$ (thick dotted darker gray lines) for the LSCFG model, respectively, where $prob = 10^{-9}$ and $\mathcal{I} \in \{\{P\}, \{F\}, \{G\}\}$ (figures from left to right).

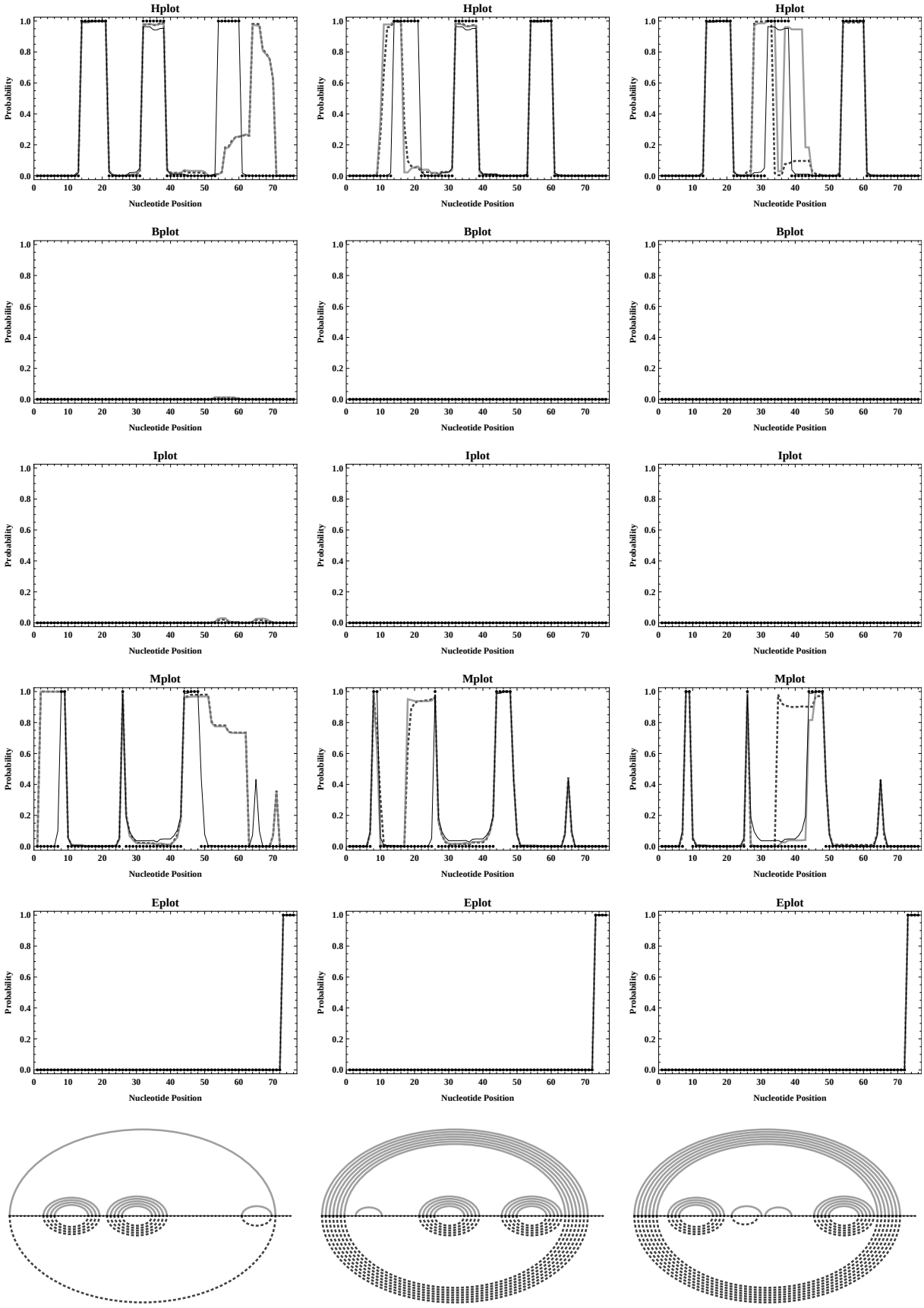


Figure 17: Loop profiles and centroid derived according to $\mathrm{mev}_{\mathcal{I}}(prob)$ (thick gray lines) and $\mathrm{fev}_{\mathcal{I}}(prob)$ (thick dotted darker gray lines) for the LSCFG model, respectively, where $prob = 10^{-9}$ and $\mathcal{I} \in \{\{M\}, \{O\}, \{N\}\}$ (figures from left to right).

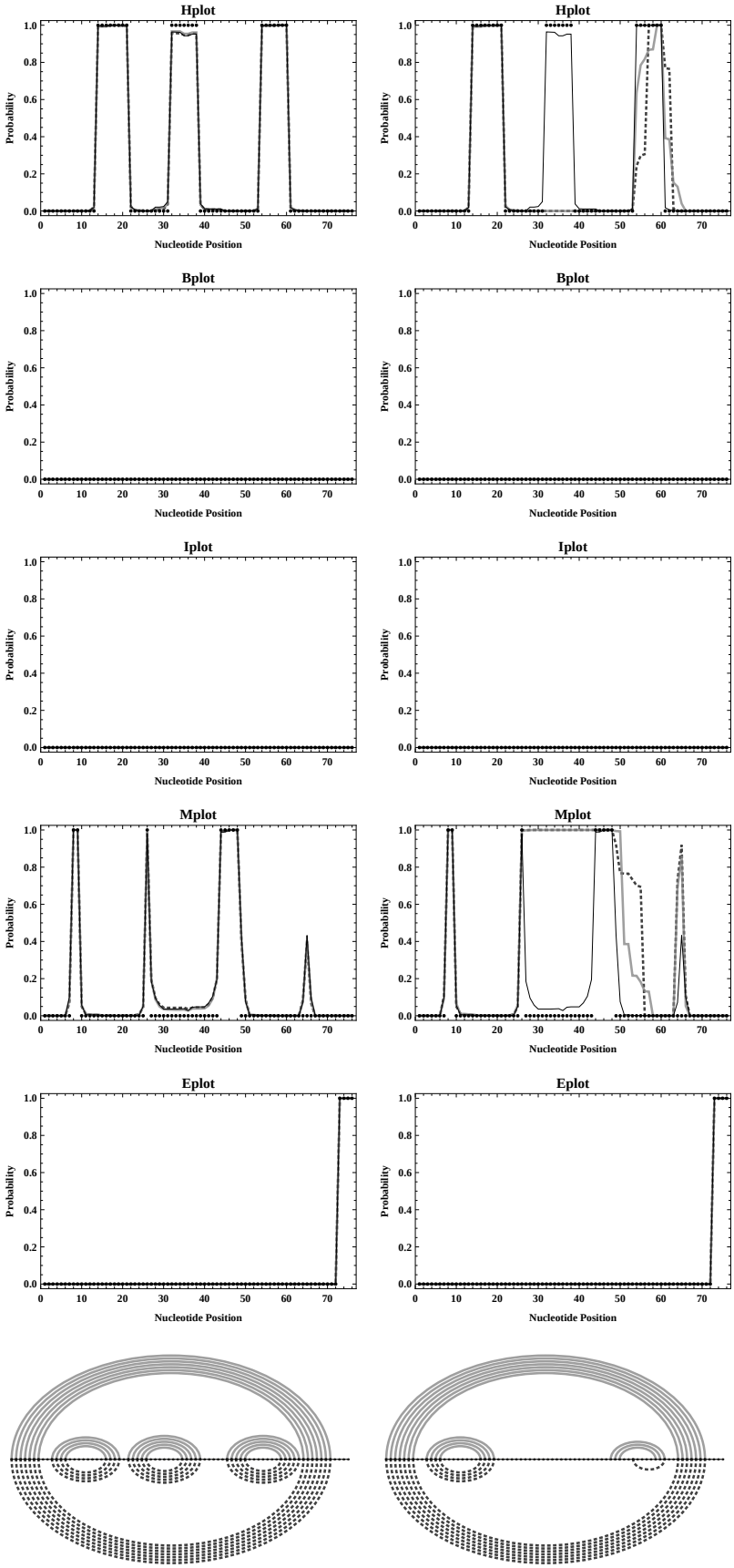


Figure 18: Loop profiles and centroid derived according to $\mathrm{mev}_{\mathcal{I}}(prob)$ (thick gray lines) and $\mathrm{fev}_{\mathcal{I}}(prob)$ (thick dotted darker gray lines) for the LSCFG model, respectively, where $prob = 10^{-9}$ and $\mathcal{I} \in \{\{B\}, \{U\}\}$ (figures from left to right).

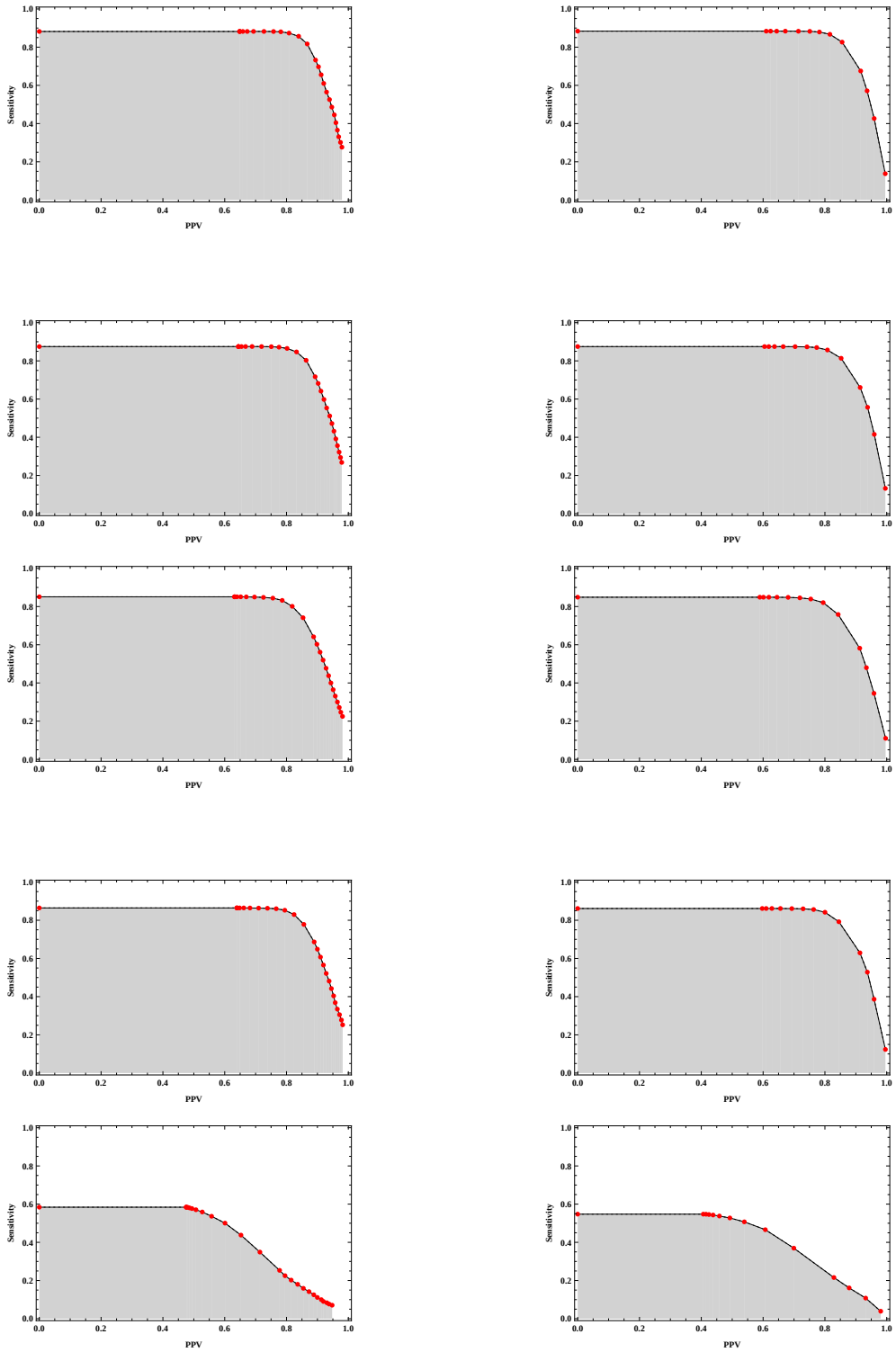


Figure 19: Comparison of the (areas under) ROC curves obtained for our tRNA database, derived without disturbances (top line) and by considering random relative disturbances according to mep(0.5), mep(0.99), fep(0.5) and fep(0.99) (from top to bottom line) under the assumption of the traditional SCFG model (for $\min_{\text{hel}} = 1$ and $\min_{HL} = 1$). For each preprocessing variant, corresponding ROC curves are shown for prediction principle MEA structure (figure on the left) and centroid (figure on the right), respectively.

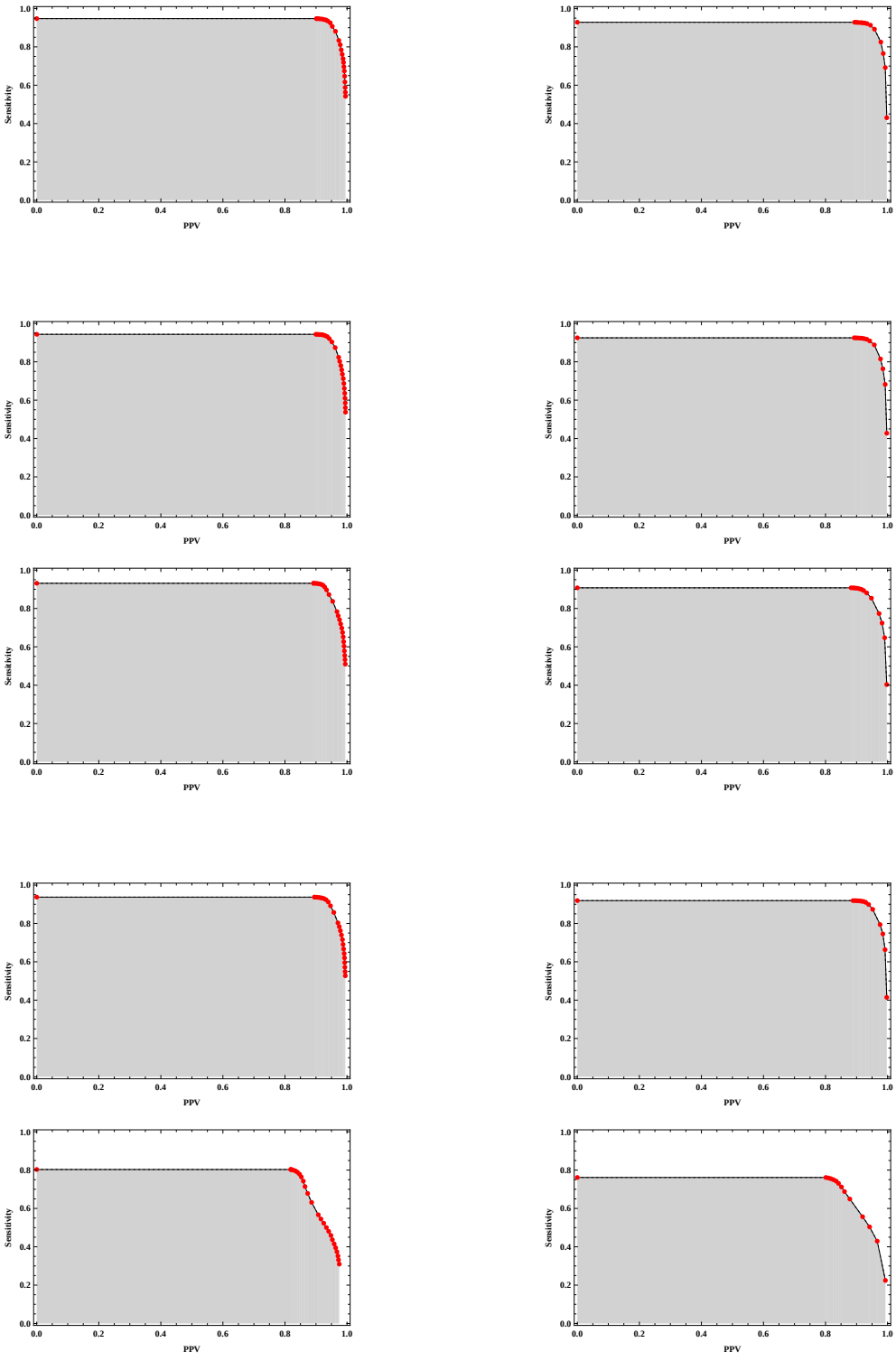


Figure 20: Results corresponding to those of Figure 19, derived under the assumption of the LSCFG model.

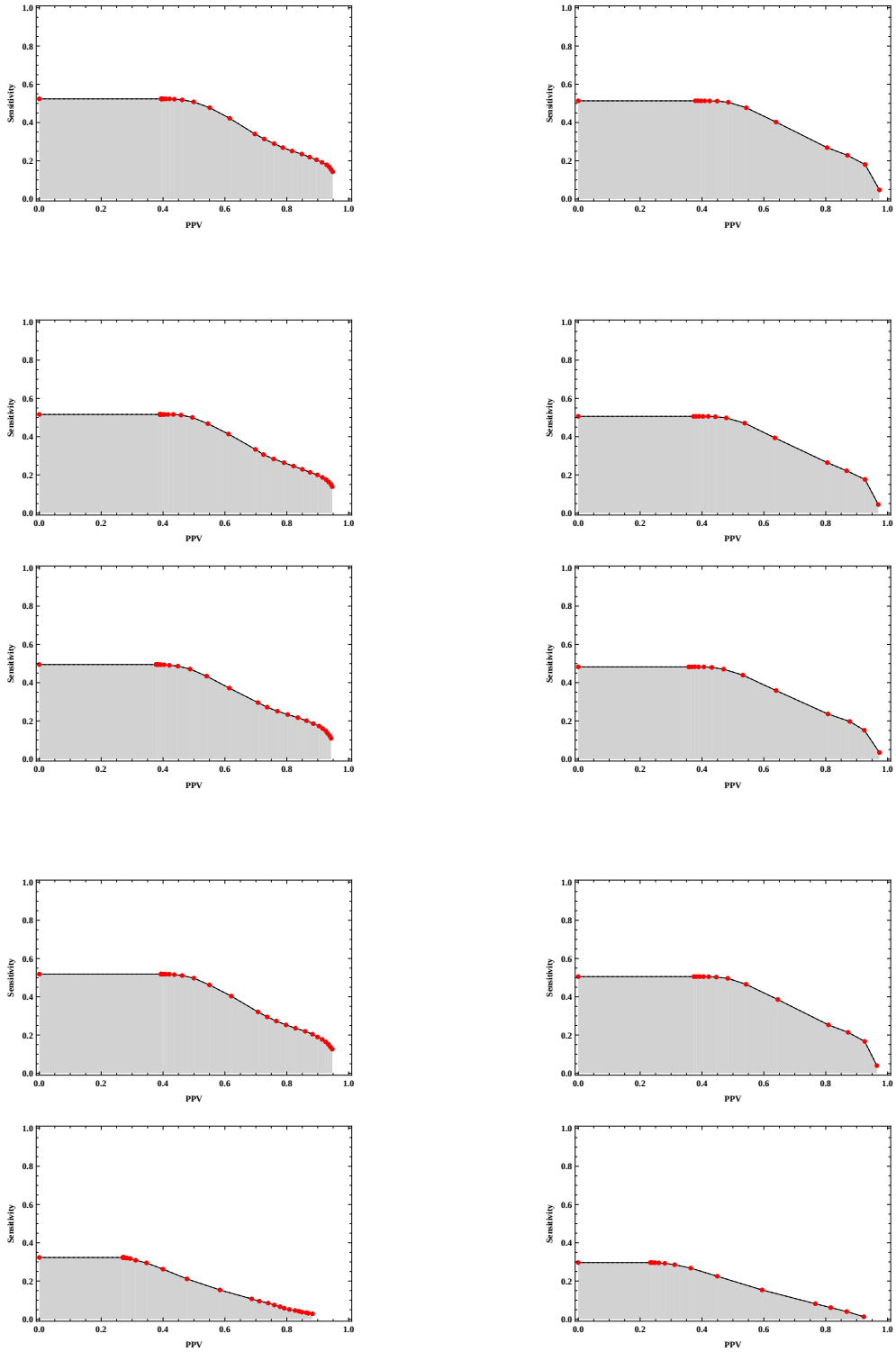


Figure 21: Comparison of the (areas under) ROC curves obtained for our 5S rRNA database, derived without disturbances (top line) and by considering random relative disturbances according to mep(0.5), mep(0.99), fep(0.5) and fep(0.99) (from top to bottom line) under the assumption of the traditional SCFG model (for $\min_{\text{hel}} = 1$ and $\min_{HL} = 1$). For each preprocessing variant, corresponding ROC curves are shown for prediction principle MEA structure (figure on the left) and centroid (figure on the right), respectively.

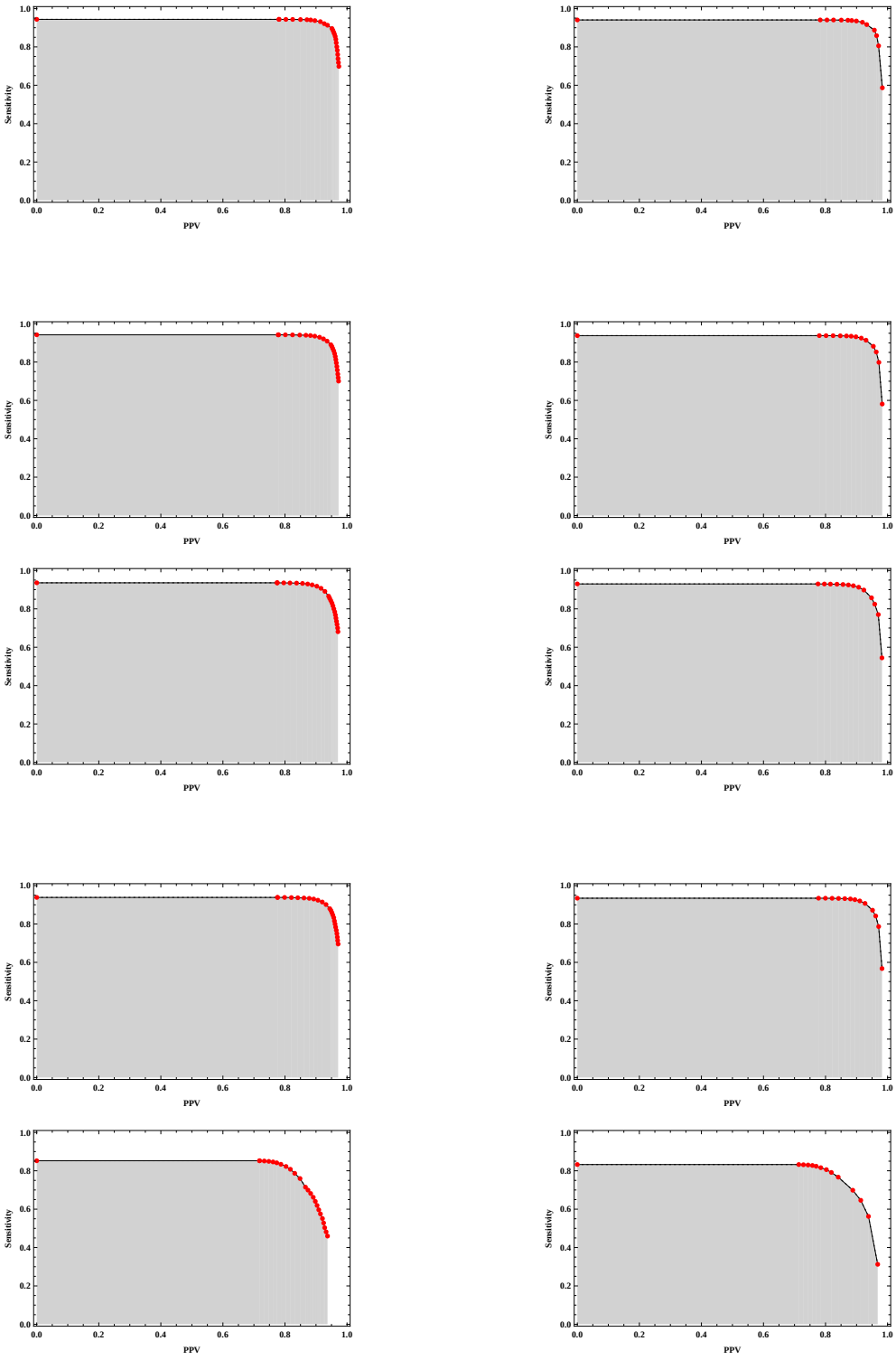


Figure 22: Results corresponding to those of Figure 21, derived under the assumption of the LSCFG model.