PNAS
www.pnas.org

Supplementary Information for

Efficient compression in color naming and its evolution

Noga Zaslavsky, Charles Kemp, Terry Regier and Naftali Tishby

Corresponding author: Noga Zaslavsky
E-mail: noga.zaslavsky@mail.huji.ac.il

This PDF file includes:

Supplementary text
Figs. S1 to S15
Tables S1 to S7
Captions for Movies S1 to S2
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Movies S1 to S2

Contents

Movie Captions

Movie S1. Evolution of the IB color naming systems. Left panel: Bifurcation diagram, similar to Fig.5. This diagram shows the full range of IB solutions, whereas Fig.5 shows only the range relevant for the languages in our data. The black line indicates the location in the diagram that corresponds to the value of β. **Right panel:** Visualization (as in Fig.4) of the IB system that corresponds to β. The IB systems evolve as β gradually increases from $\beta = 1$, where there is only one category, to $\beta = 2^{13}$, where each color is mapped deterministically to its own unique category. In between these two extremes, the IB systems induce soft color categories. Structural phase transitions occur at critical values of β along this trajectory of efficient solutions, in which new categories appear. Low-consensus regions often appear in systems near these phase transitions.

Movie S2. Languages achieve near-optimal compression. Left panel: The red dot traces along the optimal systems on the IB curve (theoretical limit), while the blue dot follows nearby, indicating the position of selected languages just below the curve in the information plane. A total of 23 representative languages are shown, which were selected to demonstrate the range of empirical variation accommodated by the IB model and the relation of that variation to languages' positions near the IB curve. **Right panel:** Contour plots of the language's naming distribution (top) and the IB encoder (bottom) that correspond to the blue and red dots on the left panel, respectively. The IB systems captures much of the structural variability in the data, and even languages that are less similar to the IB systems are still highly efficient, as seen on the left panel.

Supporting Information Text

1. Theoretical framework

1.1. Summary of notation. We use capital letters to denote random variables (e.g. M and U), calligraphic letters to denote their support (e.g. $\mathcal{M}$ and $\mathcal{U}$), and lower case letters to denote a specific realization (e.g. m and u). In our formulation we consider a finite set of distributions $\mathcal{M}$. Each element in this set (i.e., each $m \in \mathcal{M}$) is a distribution over the set $\mathcal{U}$. In other words, m is a function that takes u as an argument. We use the notation $m(u)$ when we wish to make explicit that m is a function of u, or when we wish to denote the probability of a specific u according to m. It may be helpful to think of $m(u)$ in terms of conditional probabilities, i.e., $m(u) = p(u|m)$. Table S1 summarizes the notation used in the IB framework (1), in our current formulation of IB, in the framework of RKK (2) and in the adjusted RKK model (RKK+) which we constructed as a baseline for evaluation. A detailed description of RKK+ appears in section 4.

Table S1. Summary of notation

	Component	IB (1999)	IB (current)	RKK+ (current)	RKK (2015)
Communication model	Target variable / universe	$y \in \mathcal{Y}$	$u \in \mathcal{U}$	$u \in \mathcal{U}$	$t \in \mathcal{U}$
	Source variable	$x \in \mathcal{X}$	$m \in \mathcal{M}$	$m \in \mathcal{M}$	-
	Speaker's intended meaning	$p(y\|x)$	$m(u)$	$m(u)$	$s(t)$
	Source distribution / need	$p(x)$	$p(m)$	$p(m)$	$n(t)$
	Cluster / word	$\hat{x} \in \hat{\mathcal{X}}$	$w \in \mathcal{W}$	$w \in \mathcal{W}$	$w \in \mathcal{W}$
	Encoder / naming distribution	$q(\hat{x}\|x)$	$q(w\|m)$	$q(w\|m)$	$t \mapsto w$ if $t \in \mathrm{cat}(w)$
	Decoder	$\hat{x} \mapsto q(y\|\hat{x})$	$q(\hat{m}\|w)$	$q(\hat{m}\|w)$	-
	Listener's interpreted meaning	$q(y\|\hat{x})$	$\hat{m}_w(u)$	$\hat{m}_w(u)$	$l(t)$
Optimization principle	Complexity	$I_q(X;\hat{X})$	$I_q(M;W)$	$\log K$	$K = \|\mathcal{W}\|$
	Distortion / communicative cost	$D[p(y\|x)\|q(y\|\hat{x})]$	$D[m\|\hat{m}]$	$D[m\|\hat{m}]$	$D[s\|l]$
	Accuracy	$I_q(\hat{X};Y)$	$I_q(W;U)$	$I_q(W;U)$	-
	Tradeoff parameter	β	β	-	-

1.2. Bayesian listener. We show that the ideal listener with respect to a given speaker is an optimal Bayesian decision maker. In our case, this means that we can assume an ideal listener that always decodes w deterministically by interpreting it as meaning $\hat{m}_w(u) = \sum_{m\in\mathcal{M}} q(m|w)m(u)$, where $q(m|w)$ is obtained by applying Bayes' rule,

$$q(m|w) = \frac{q(w|m)p(m)}{q(w)}\,, \qquad \text{[S1]}$$

where $q(w) = \sum_{m'} p(m')q(w|m')$. To show that this Bayesian listener is optimal, assume that the speaker's encoder is given by $q(w|m)$. The optimal listener for this speaker is defined by the decoder $q(\hat{m}|w)$ that minimizes

$$\mathcal{F}_\beta[q] = I_q(M;W) - \beta I_q(W;U) = I_q(M;W) - \beta\left(I(M;U) - \mathbb{E}_q\left[D[M\|\hat{M}]\right]\right), \tag{S2}$$

where the second equality follows from Eq. (5). Note that $I(M;U)$ is constant in q and $I_q(M;W)$ depends on the encoder but not on the decoder. Only the last term depends on the decoder, and it holds that

$$\mathbb{E}_q\left[D[M\|\hat{M}]\right] = \sum_{m,w,\hat{m}} p(m)q(w|m)q(\hat{m}|w)D\left[m\|\hat{m}\right] \tag{S3}$$

$$= \sum_{m,w,\hat{m}} q(w)q(m|w)q(\hat{m}|w)D\left[m\|\hat{m}\right] \tag{S4}$$

$$\geq \sum_{w} q(w) \operatorname*{argmin}_{\hat{m}'} \sum_{m} q(m|w)D\left[m\|\hat{m}'\right] \tag{S5}$$

Therefore, there is a deterministic decoder $q(\hat{m}|w)$ that minimizes Eq. (S2),

$$q(\hat{m}|w) = \begin{cases} 1 & \text{if } \hat{m} = \operatorname*{argmin}_{\hat{m}'} \mathbb{E}_{q(m|w)}\left[D\left[m\|\hat{m}'\right]\right] \\ 0 & \text{otherwise} \end{cases}. \tag{S6}$$

Differentiating $\mathbb{E}_{q(m|w)}\left[D\left[m\|\hat{m}'\right]\right]$ with respect to $\hat{m}'$ and equating to 0 gives that the minimum is attained at $\hat{m}_w$. Since $\sum_u \hat{m}_w(u) = 1$ we did not need to impose this normalization constraint on the optimization, and because the KL divergence is convex in both arguments $\hat{m}_w$ is indeed the minimum.

1.3. Relation between IB and rate distortion theory. It has been shown that IB can be considered a special type of rate distortion (RD) with a variable distortion measure (3), and that the IB distortion measure has unique properties that distinguish IB from other RD problems (4). Furthermore, it was shown in (4) that IB can be considered a standard RD problem over probability measures, where the reconstruction alphabet is continuous. This view is closely related to the interpretation of IB as distributional clustering (5), in contrast to many applications of IB in the context of supervised learning (6). The setting we consider in this paper corresponds to a RD problem where M is compressed into $\hat{M}$. Although we are explicitly interested in the compression of M into a codeword W and in the reconstruction of $\hat{M}$ from W, it can be shown that the two problems are equivalent under mild assumptions.

A formal proof of this statement is beyond the scope of this work, but the main idea is that we can assume w.l.o.g. that the decoder is information lossless, i.e., $I_q(M;W) = I_q(M;\hat{M})$. In this case, minimizing $\mathcal{F}_\beta[q]$ is equivalent to minimizing the RD objective $I_q(M;\hat{M}) + \beta\mathbb{E}_q[D[M\|\hat{M}]]$, under the constraint $q(\hat{m}|m) = \sum_w q(w|m)\mathbf{1}_{[\hat{m}=\hat{m}_w]}$. It is possible to show that, under mild assumptions, this additional constraint on $q(\hat{m}|m)$ would not change the optimum of the RD problem. However, here we will only justify the assumption that the decoder is information lossless. Let $\varphi(w) = \hat{m}_w$ with respect to some encoder q. The decoder is information lossless if $\varphi(w)$ is a one-to-one mapping over the support of q (i.e., over $Sup(q) = \{w \in \mathcal{W} : q(w) > 0\}$). We can assume that this property holds, because otherwise it is possible to construct q' for which this property holds and $\mathcal{F}_\beta[q'] \leq \mathcal{F}_\beta[q]$. Assume there are $w_1, w_2 \in Sup(q)$ such that $w_1 \neq w_2$ and $\varphi(w_1) = \varphi(w_2)$. Define q' by merging them, namely for all m let $q'(w_1|m) = q(w_1|m) + q(w_2|m)$, $q'(w_2|m) = 0$, and for all $w \neq w_1, w_2$ let $q'(w|m) = q(w|m)$. This does not change the expected distortion; however, $I_{q'}(M;W) \leq I_q(M;W)$.

1.4. The IB method and deterministic annealing. Given a value of β, the IB method (1) iteratively updates the following IB equations until convergence,

$$q_\beta(w|m) = \frac{q_\beta(w)}{Z(m;\beta)} \exp(-\beta D\left[m\|\hat{m}_w\right]) \tag{S7}$$

$$q_\beta(w) = \sum_{m\in\mathcal{M}} q_\beta(w|m)p(m) \tag{S8}$$

$$\hat{m}_w(u) = \sum_{m\in\mathcal{M}} m(u)q_\beta(m|w), \tag{S9}$$

where $Z(m;\beta)$ is the normalization factor. At the optimum, these equations are satisfied self-consistently. Because IB is a non-convex problem, the method of deterministic annealing (7) is often used to mitigate the problem of converging to sub-optimal fixed points of the IB equations (e.g. 5, 8). Deterministic annealing starts at a low value of β ($\beta = 1$ in IB) where the solution is trivial, and then gradually increases β. For each β, the IB method is initialized with the solution found for the previous value of β. In practice, for better convergence, we evaluated the IB curve by reverse deterministic annealing (9); i.e., starting at a very high value of β, where the solution is given by a one-to-one mapping from M to W, and then gradually decreasing β. We repeated this process with 1500 values of β in $[1, 2^{13}]$.

2. Least informative source

How to accurately model a cognitive process that generates meanings for the speaker is an open question that is beyond the scope of this work. Instead, we wish to estimate a source distribution that is more realistic than the uniform distribution, but does not require prior knowledge. In this work we propose a general approach for doing so, based on the following observation: if a source distribution exists, it should be reflected somehow in the way people speak, i.e., in the naming distribution. Therefore, it makes sense to try to infer the source distribution directly from the naming data. We do so without making assumptions about the cognitive source by building on the notion of least informative priors. Our approach is domain-general; however, for simplicity we present it here in terms our color naming model. In section 7 we discuss other approaches for estimating the source distribution, and show that our conclusions also hold under these alternative source distributions.

2.1. Definition for a given language. We begin by defining a least informative prior over color chips, with respect to a given naming distribution $q_l(w|c)$. Because we assumed that each chip c is associated with a unique meaning m_c, any prior $p(c)$ induces a source distribution by setting $p(m_c) = p(c)$. One common approach for obtaining uninformative priors is by invoking the maximum entropy principle. However, in our case the maximum entropy distribution over color chips is simply the uniform distribution. Another natural approach in our setting is to find a distribution that maximizes the entropy of c while minimizing the expected uncertainty over c give a term w in the language. That is,

$$p_l(c) = \underset{p(c)}{\operatorname{argmax}}\; H(C) - H_q(C|W) \qquad \text{[S10]}$$

where $H_q(C|W) = -\sum_{c,w} p(c)q(w|c)\log\frac{q(c|w)}{p(c)}$ is the conditional entropy, and $q(c|w) = \frac{q(w|c)p(c)}{q(w)}$ is the posterior distribution of c given w.

This definition has two interesting interpretations, in addition to being a constrained maximum entropy distribution. First, note that

$$I_q(W;C) = \underset{p(c)}{\operatorname{argmax}}\; H(C) - H_q(C|W)\,, \qquad \text{[S11]}$$

which implies that $p_l(c)$ maximizes the mutual information between colors and words. This type of prior distribution is also called a capacity achieving prior, and can be evaluated using the Blahut-Arimoto algorithm (10, 11). Note that in the IB model, a language l would be maximally complex if the source distribution were defined from $p_l(c)$. This contrasts with the IB principle, which aims to minimize complexity. Second, $p_l(c)$ is considered the least informative prior over c in the sense that it minimizes information about the posterior $q(c|w)$ by maximizing the KL divergence between the prior and posterior. This interpretation follows from the identity

$$I_q(W;C) = \sum_w q(w)D[q(c|w)\|p(c)]\,, \qquad \text{[S12]}$$

and it is closely related to the notion of reference priors in Bayesian inference (12). Reference priors are considered objective priors in the sense that they depend solely on the given distribution $q(w|c)$, but not on other assumptions that may reflect subjective prior beliefs.

2.2. Estimation across languages. Our approach for estimating a LI source can be applied on a language-specific basis. However, we leave this language-specific analysis for future research and instead focus on estimating a single source distribution for all languages. We obtain this universal LI source by averaging across the language-specific LI priors, namely

$$p_{\mathrm{LI}}(m_c) = \frac{1}{L}\sum_{l=1}^{L} p_l(c)\,. \qquad \text{[S13]}$$

To control for overfitting and to test the ability of our model to generalize to languages which are not used for estimating the source, we performed 5-fold cross-validation over the languages that contribute to the average in Eq. (S13). Fifteen WCS languages were excluded from this process, to ensure that the naming probabilities for each language were estimated from at least 5 responses for every chip. This regularization process is further explained in section 4.1, and the excluded 15 languages are listed in section 10. Section 10 contains the results for all 111 languages.

The full LI source, estimated by averaging over 96 languages, is shown in Fig.S1. This source distribution is non-uniform; however, it still has relatively high entropy, $H[p(m_c)] \approx 7.41$, compared to the maximal entropy $\log_2(330) \approx 8.36$. This means that the KL divergence between the LI source and the uniform source is roughly 1 bit.

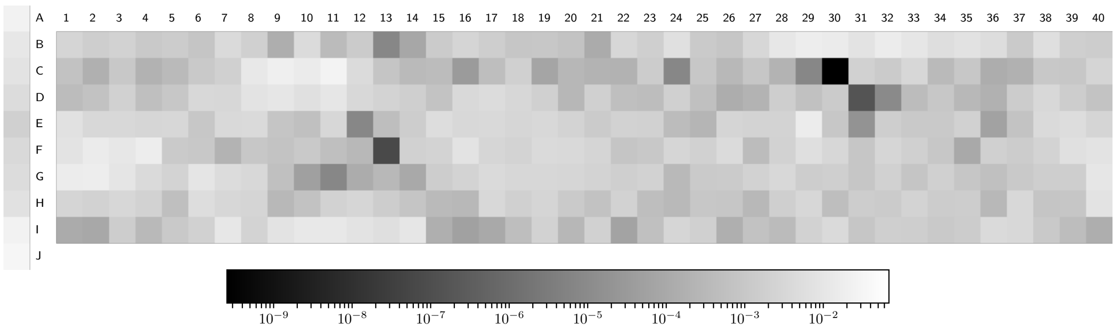


Fig. S1. The LI prior over the color chips obtained by averaging across the LI priors of 96 languages. Each chip is colored according to its probability mass, where black corresponds to probability 0 and white corresponds to probability 1. Gray colors are based on a logarithmic scale.

3. Dissimilarity measures

We compared different encoders, or color naming systems, by building on standard information-theoretic dissimilarity measures between clusterings (13). In our setting, these measures have an intuitive interpretation that relates them to the information between speakers of two languages.

Assume a language l_1 with lexicon $\mathcal{W}_1$ and an encoder $q_1(w_1|m)$, and a language l_2 with lexicon $\mathcal{W}_2$ and an encoder $q_2(w_1|m)$. In addition, assume that given a meaning $m \sim p(m)$, a speaker of l_1 produces a word $W_1 \sim q_1(w_1|m)$ and a speaker of l_2 independently produces a word $W_2 \sim q_2(w_2|m)$. The joint distribution of W_1 and W_2 is given by

$$q(w_1, w_2) = \sum_{m \in \mathcal{M}} p(m) q_1(w_1|m) q_2(w_2|m)\,. \qquad \text{[S14]}$$

Similarly, we can consider the joint distribution of two speakers of the same language that independently produce words W_i and W_i' given m,

$$q(w_i, w_i') = \sum_{m \in \mathcal{M}} p(m) q_i(w_i|m) q_i(w_i'|m)\,. \qquad \text{[S15]}$$

Intuitively, two languages are similar if the cross-language information $I(W_1; W_2)$ is large compared to the information within each language.

3.1. Normalized Information Distance (NID). The normalized information distance (NID 13) is defined by

$$\mathrm{NID}(W_1, W_2) = 1 - \frac{I(W_1; W_2)}{\max\{H(W_1), H(W_2)\}}\,. \qquad \text{[S16]}$$

NID has been defined in (13) for hard partitions; i.e., in the case where $q(w|m)$ is a deterministic distribution. In this case NID has several desirable properties (13): it is a metric, it is bounded in the interval $[0, 1]$, and it was shown to outperform other methods for measuring similarity between hard clusterings. Therefore, we measured the distance between the mode maps that correspond to q_1 and q_2 by the NID between them.

3.2. Generalization of NID to soft partitions (gNID). Although it is straightforward to apply the NID formula to soft partitions (soft-NID), we noticed that soft-NID is not sensitive enough to differences in the full probabilistic structure of the encoders. This can be seen in Fig.S2, which shows Dyimini for example. The soft-NID between Dyimini and different IB theoretical systems along the IB curve has a relatively flat part. This means that soft-NID can barely distinguish between these different IB systems. We therefore slightly modified soft-NID in a way that also generalized NID to soft partitions. We define this generalization by

$$\mathrm{gNID}(W_1, W_2) = 1 - \frac{I(W_1; W_2)}{\max\{I(W_1, W_1'), I(W_2, W_2')\}} . \qquad \text{[S17]}$$

If $q_1(w_1|m)$ and $q_2(w_2|m)$ are both deterministic conditional distributions (i.e., W_1 and W_2 are selected deterministically given m), then gNID reduces to NID. To see this, notice that $I(W_i; W_i') = H(W_i) - H(W_i|W_i')$ and $H(W_i|W_i') = 0$ in the deterministic case.

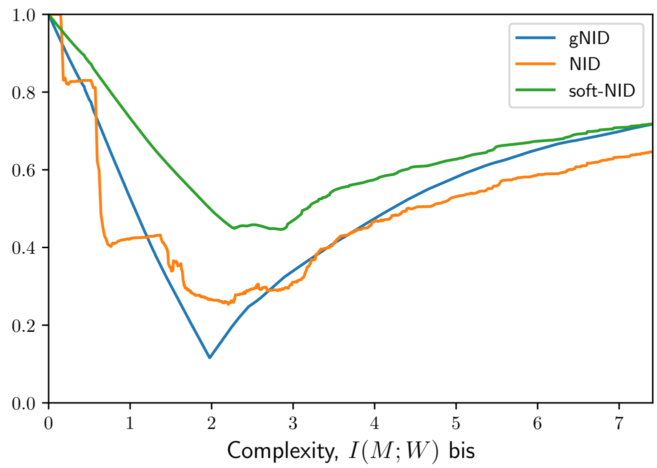


Fig. S2. Dissimilarity measures between the color naming system of Dyimini and the IB theoretical systems along the IB curve. gNID and soft-NID apply to the full distributions, whereas NID applies to their corresponding mode maps.

gNID has a few desirable properties. It holds that $\mathrm{gNID}(W_1, W_2) \leq 1$ because mutual information is non-negative, and $\mathrm{gNID}(W_1, W_2) = 1$ when W_1 and W_2 are independent because in that case $I(W_1; W_2) = 0$. When the two encoders are equivalent, then $\mathrm{gNID}(W_1, W_2) = 0$, as opposed to soft-NID which could be positive in this case. Although gNID in general is not necessarily non-negative, we did not encounter cases in which the gNID between a language's color naming distribution and an IB or RKK+ encoder was negative. In addition, for most languages gNID exhibits qualitatively similar behavior as seen for Dyimini (Fig.S2). That is, the gNID between the language and the IB systems follows a similar trend as NID and soft-NID; however, unlike NID and soft-NID, gNID is unimodal.

4. The RKK+ model

The RKK+ model is based on our communication model (Fig.1), but the definition of efficiency and the treatment of the data are derived from RKK's approach to color naming (2). Our communication model is very similar to RKK's communication model, although we relaxed a few assumptions made by RKK. In this section we discuss in detail the derivation of RKK+ from RKK's notion of efficiency, and explain the differences between the RKK+ model and RKK's color naming model. The mapping between our notation and RKK's notation is described in Table S1. For simplicity, we use here our notation for RKK+ and refer to the components of RKK's color naming model by the corresponding RKK+ notation.

4.1. Encoders based on major color terms. RKK's approach to the WCS color naming data relies on the notion of a major color term. According to RKK, w is a major color term in a given language, if it is the modal term for at least 10 color chips. Otherwise, w is considered a minor color term. For English, which was not included in RKK's color naming analysis, we set the threshold at 5 chips in order to obtain the 11 basic color terms in English. As in RKK's analysis, only data for major color terms is considered for the evaluation of the RKK+ model. That is, for each language l RKK+ considers a naming distribution $q_l^+(w|c)$ which is obtained from $q_l(w|c)$ by restricting it only to the major color terms in l. Restricting the data of a language to major terms may result in insufficient data for estimating the color naming distribution of that language. In 15 WCS languages some chips had fewer than 5 naming responses, and therefore we excluded these languages from the quantitative model evaluation and from the estimation of the LI source. These 15 languages are presented in section 10.

4.2. Relaxing RKK's assumptions.

Stochastic speaker. RKK made the simplifying assumption that the speaker chooses words deterministically, which induces hard partitions of the color space into named categories. For each color term w RKK defined $\mathrm{cat}(w)$ by the set of colors that are named by w. This corresponds to a deterministic encoder: $q(w|c) = 1$ if $c \in \mathrm{cat}(w)$ and 0 else. In RKK+ this assumption was relaxed because the encoder in our communication model can be stochastic.

Perceptual uncertainty. RKK assumed that the speaker has no perceptual uncertainty, which means that colors are represented by delta functions, i.e., $m_c(u) = \delta_{c,u}$. In our model we allow for perceptual uncertainty and instead assume that each color c is represented by the Gaussian m_c.

Bayesian listener. RKK assumed that the listener's interpreted meanings take the form

$$l_w(u) \propto \sum_{c \in \mathrm{cat}(w)} \exp\left(-\frac{1}{2\sigma^2}\|u - c\|^2\right) . \qquad \text{[S18]}$$

Although this form is justified (2), we show in section 4.4 that a similar form can emerge directly from the need for efficiency. Therefore, we waive this assumption and consider a listener who is adapted to the speaker without additional constraints, as in the IB model.

4.3. Efficiency according to RKK. RKK argued that theoretically efficient languages minimize a communicative cost for a given level of complexity. We next present their definitions of complexity and communicative cost, and discuss the specific form these measures take in RKK+.

Complexity. RKK's notion of complexity is derived from the minimum description length principle on a domain-specific basis. In the domain of color, RKK defined the complexity of a language by the number of major terms in that language, denoted here by K. In RKK+ we slightly adjust this complexity measure and consider instead $\log K$. This does not change the essence of the measure nor the structure of the theoretically optimal systems, but allows us to measure complexity in bits, as in IB.

Communicative cost. RKK defined the error between the speaker's intended meaning and listener's interpreted meaning by the KL divergence between these two distributions. This definition coincides with the distortion measure in IB. RKK's communicative cost is the expected error, as it corresponds to the expected distortion in IB. Following the same argument as in section 1.2, we obtain that the ideal listener in RKK+ takes the same form as in IB; i.e., it is given by $\hat{m}_w$. Therefore, in RKK+ the communicative cost of an encoder $q(w|m)$ is given by

$$D[q] = \sum_{m,w} p(m)q(w|m)D\left[m\|\hat{m}_w\right] . \qquad \text{[S19]}$$

This definition is the same as Eq. (S3), but in RKK+ it applies to q_l^+ instead of q_l. We can therefore apply Eq. (5) to inversely relate the communicative cost $D[q_l^+]$ to the accuracy of the language according to RKK+. The complexity-accuracy pairs of the languages we considered, according to RKK+, are shown in Fig.S3.

4.4. Structure of the solution. An optimal speaker-listener pair in RKK+ jointly minimizes the expected distortion between them, for a given K. The hard constraint on the number of major terms is enforced by only considering encoders $q(w|m)$ over K terms. We have already seen that optimizing this distortion with respect to the speaker's interpreted meanings, while fixing the speaker's encoder, gives $\hat{m}_w$. Now, fix $\hat{m}_w$ and consider the encoder that minimizes Eq. (S19). Since this objective is linear in $q(w|m)$ the minimum is attained at

$$q(w|m) = \begin{cases} 1 & \text{if } w = w_m, \quad \text{where } w_m = \operatorname*{argmin}_{w'} D\left[m\|\hat{m}_{w'}\right] \\ 0 & \text{otherwise} \end{cases} . \qquad \text{[S20]}$$

Formally, this can be shown by following a similar argument as in section 1.2. This means that even though we relaxed the assumption that the speaker is deterministic, RKK+ does not predict any advantage for non-deterministic speakers that induce soft categories, and the theoretically optimal RKK+ systems can be characterized by hard partitions of color space. We can therefore define $\mathrm{cat}(w)$ as in RKK's color naming model, namely $\mathrm{cat}(w) = \{m \in \mathcal{M} : w_m = w\}$. Plugging back this encoder into the formula of $\hat{m}_w$ (i.e., Eq. (1)) and substituting the structure of m_c, gives a similar form as RKK's assumed listener in Eq. (S18).

4.5. Evaluation of the RKK+ bounds. The RKK+ fixed points are characterized by the self-consistent $\text{cat}(w)$ and $\hat{m}_w$. This suggests an iterative algorithm for finding these fixed points, which can be considered as K-Means over distributions where the KL divergence is used as the dissimilarity function. Denote by $C_w^{(t)}$ the set $\text{cat}(w)$ obtained at the t-th iteration of the algorithm. The algorithm can be described as follows:

- Initialize $C_w^{(0)}$ for $w \in \{1, \ldots, K\}$
- For $t = 1, \ldots$ (until convergence) update:

$$\hat{m}_w^{(t)}(u) = \frac{1}{|C_w^{(t-1)}|} \sum_{m \in C_w} m(u) \qquad \text{[S21]}$$

$$C_w^{(t)} = \left\{ m \in \mathcal{M} : w = \underset{w'}{\text{argmin}}\, D[m \| \hat{m}_{w'}^{(t)}] \right\} \qquad \text{[S22]}$$

This is a non-convex optimization problem, and only convergence to a local optimum is guaranteed. Therefore, for each K we repeated this algorithm 300 times with random initializations, and selected the best result. We evaluated the RKK+ bounds for $K = 2, \ldots, 11$. These bounds are shown in Fig.S3 (orange bars). The number of major terms in the languages we considered varies between 3-11.

4.6. Relation to IB. RKK+ is equivalent to IB when the lexicon size is restricted to K terms, and when $\beta \to \infty$. To see this, notice that taking $\beta \to \infty$ means that the speaker and listener only care about accuracy, and therefore minimizing $\mathcal{F}_\beta$ amounts to only minimizing the expected distortion. For every K we can evaluate the IB solution for $0 \leq \beta < \infty$, where the hard constraint on the lexicon size is imposed by considering only encoders $q(w|m)$ over a lexicon of size K. While the optimal IB curve is estimated for $K = |\mathcal{M}|$ (see 4), for smaller values of K we can obtain sub-optimal IB curves. This means that the IB curve upper bounds the RKK+ bounds. This relation between IB and RKK+ can be seen in Fig.S3.

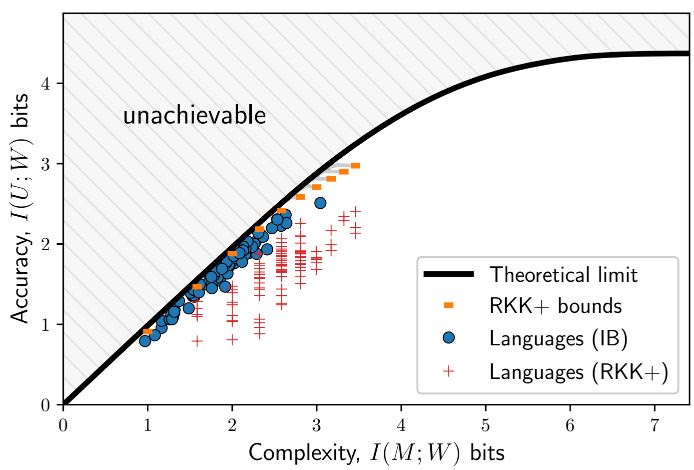


Fig. S3. Comparison between IB and RKK+. Complexity-accuracy values for all languages according to IB (blue dots) and RKK+ (red crosses). The IB curve (black) is evaluated for $K = 330$, and it defines the theoretical limit of achievable tradeoffs, including those achieved by the optimal systems according to RKK+. RKK+ bounds (orange bars) correspond to the deterministic limits of sub-optimal IB curves (gray curves) obtained by restricting the lexicon size to $K = 2, \ldots, 11$. The efficiency of the languages according to each model is evaluated with respect to the model's bounds.

5. Quantitative evaluation and variants of the IB model

Our goal in comparing IB with RKK+ is to test which principle can account better for the data, while holding all other elements of the model constant. Although IB and RKK+ are defined over the same communication model, there are two differences in the way these models treat the data: (1) RKK+ only considers major terms while IB considers the full set of naming responses, and (2) RKK+ evaluates each language against an optimal system with the same complexity, whereas in the IB model each language is evaluated against an optimal system at β_l which may have a different complexity than that of the language. We controlled for these differences by considering two variants of the IB model that match how RKK+ treats the data. We show here that the results in both cases are similar to our main evaluation, which suggests that these two differences are mainly technical and do not impact our conclusions.

We evaluate RKK+ in the same way we evaluate IB. Namely, we are interested in (A) whether color naming systems across languages are near-optimally efficient according to RKK+; and (B) how well a theoretically optimal RKK+ encoder for a given K_l can explain the structure of the color naming distribution q_l^+ in languages with K_l major color terms. We use the same quantitative measures for evaluating IB and RKK+, namely ε_l, gNID and NID, where ε_l is defined with respect to the objective in each model. Although there is no tradeoff parameter in RKK+, the definition of ε_l coincides with the definition of ε_l in IB, because in RKK+ the complexity term cancels out. Recall that for IB we defined

$$\varepsilon_l = \frac{1}{\beta_l}\left(\mathcal{F}_{\beta_l}[q_l] - \mathcal{F}^*_{\beta_l}\right) = \frac{1}{\beta_l}\Big(I_{q_l}(W;M) - I_{q_{\beta_l}}(W;M)\Big) - \Big(I_{q_l}(W;U) - I_{q_{\beta_l}}(W;U)\Big)\,. \qquad \text{[S23]}$$

From Eq. (5) we get that

$$\varepsilon_l = \frac{1}{\beta_l}\Big(I_{q_l}(W;M) - I_{q_{\beta_l}}(W;M)\Big) + \Big(D[q_l] - D[q_{\beta_l}]\Big)\,. \qquad \text{[S24]}$$

If q_l and q_{β_l} have the same complexity then we get that $\varepsilon_l = D[q_l] - D[q_{\beta_l}]$. In RKK+ we have $\varepsilon_l = D[q_l^+] - D[q_{K_l}]$, where q_{K_l} is an optimal RKK+ encoder for K_l.

5.1. IB with constrained complexity. We considered a variant of the IB model in which β_l is determined such that the complexity at β_l matches the language's complexity (IB-C). Formally, this means that in IB-C β_l is selected such that $I_{q_l}(W;M) = I_{q_{\beta_l}}(W;M)$ and therefore $\varepsilon_l = D[q_l] - D[q_{\beta_l}]$. Table S2 shows the results for IB-C, together with the results for IB and RKK+ that are reported in main text (Table 1). The differences between IB and IB-C are not substantial, both for the LI source and for the uniform source. Therefore, our conclusions hold even for IB-C.

Table S2. Quantitative evaluation via fivefold cross-validation (including IB-C)

Source	Model	ε_l	gNID	NID	β_l
LI	IB	0.18 (±0.07)	0.18 (±0.10)	0.31 (±0.07)	1.03 (±0.01)
	IB-C	0.18 (±0.07)	0.21 (±0.08)	0.31 (±0.08)	1.04 (±0.02)
	RKK+	0.70 (±0.23)	0.47 (±0.10)	0.32 (±0.10)	
U	IB	0.24 (±0.09)	0.39 (±0.12)	0.56 (±0.07)	1.06 (±0.01)
	IB-C	0.24 (±0.09)	0.40 (±0.10)	0.56 (±0.08)	1.07 (±0.02)
	RKK+	0.95 (±0.22)	0.65 (±0.08)	0.50 (±0.10)	

Averages over left-out languages ±1 SD for the least informative (LI) and uniform (U) source distributions. Lower values of ε_l, gNID and NID are better.

5.2. IB for major color terms. Applying RKK+ to both major and minor terms can only increase the gap between the performance of RKK+ and IB. This is because in some languages there are many low frequency terms which do not much affect the partition of color space, however the optimal RKK+ encoders are very much affected by K. IB is more robust to low frequency terms, because the informational complexity in IB takes this into account by considering the frequency of each term. Therefore, we considered a variant of IB and a variant of IB-C in which they are applied to the color naming distributions restricted to major terms, i.e., to q_l^+ instead of q_l. Table S3 shows that the results in this case are not substantially different from the results in Table S2, which correspond to our main evaluation. Therefore, our conclusions hold whether or not the data are restricted to major color terms.

Table S3. Quantitative evaluation via fivefold cross-validation (based only on major color terms)

Source	Model	ε_l	gNID	NID	β_l
LI	IB	0.14 (±0.06)	0.20 (±0.11)	0.31 (±0.07)	1.03 (±0.01)
	IB-C	0.14 (±0.06)	0.20 (±0.09)	0.31 (±0.08)	1.04 (±0.02)
	RKK+	0.70 (±0.23)	0.47 (±0.10)	0.32 (±0.10)	
U	IB	0.19 (±0.07)	0.42 (±0.12)	0.57 (±0.07)	1.06 (±0.01)
	IB-C	0.19 (±0.07)	0.40 (±0.10)	0.56 (±0.08)	1.07 (±0.02)
	RKK+	0.95 (±0.22)	0.65 (±0.08)	0.50 (±0.10)	

Averages over left-out languages ±1 SD for the least informative (LI) and uniform (U) source distributions. Lower values of ε_l, gNID and NID are better.

6. Foundational assumptions

In this section we examine the foundational assumptions of our communication model more closely, and discuss the robustness of our results to these assumptions.

6.1. Choice of color space. Our model is based on the assumption that colors are represented in CIELAB space. To test the robustness of our results to this assumption, we repeated our full analysis with colors that are represented in the CIELUV color space (similarly to (14)) instead of CIELAB. Apart from this, all the other assumptions and methods were kept fixed. Table S4 shows quantitatively that this analysis yields similar results as the main analysis which is based on the CIELAB assumption. In particular, in both cases IB with the LI source provides the best account of the data. This conclusion is also supported by the qualitative results shown in Fig.S4 and in Fig.S5A, which are very similar to the corresponding results based on the CIELAB space. The main difference appears to be in the bifurcation diagram (Fig.S5B), where a red category appears much earlier compared to the results based on CIELAB.

Table S4. Quantitative evaluation via fivefold cross-validation (based on CIELUV)

Source	Model	ε_l	gNID	NID	β_l
LI	IB	0.14 (±0.06)	0.19 (±0.10)	0.30 (±0.09)	1.02 (±0.01)
	RKK+	0.71 (±0.23)	0.45 (±0.10)	0.29 (±0.10)	
U	IB	0.19 (±0.08)	0.36 (±0.12)	0.54 (±0.11)	1.03 (±0.01)
	RKK+	0.97 (±0.24)	0.66 (±0.07)	0.51 (±0.09)	

Averages over left-out languages ±1 SD for the least informative (LI) and uniform (U) source distributions. Lower values of ε_l, gNID and NID are better.

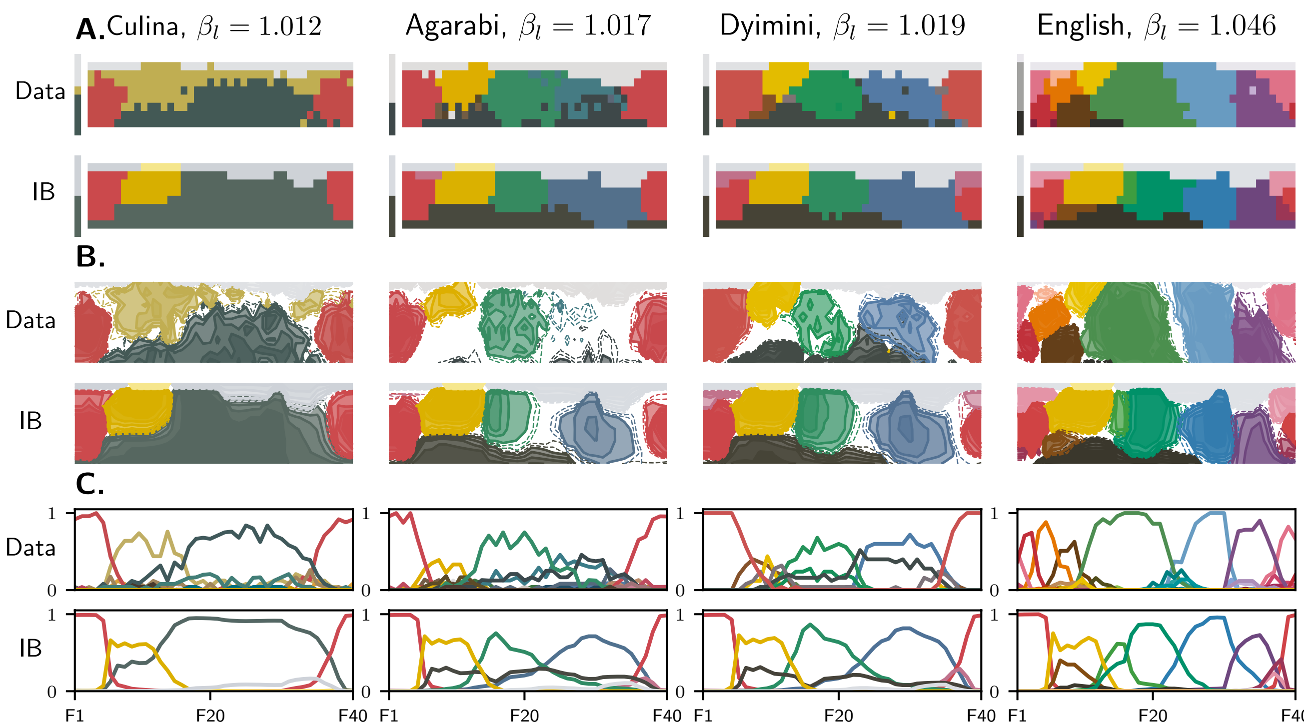


Fig. S4. CIELUV space. Mode maps (**A**), contour plots (**B**) and naming probabilities along row F (**C**), similar to Fig.4 in main text but based on the results for CIELUV instead of CIELAB.

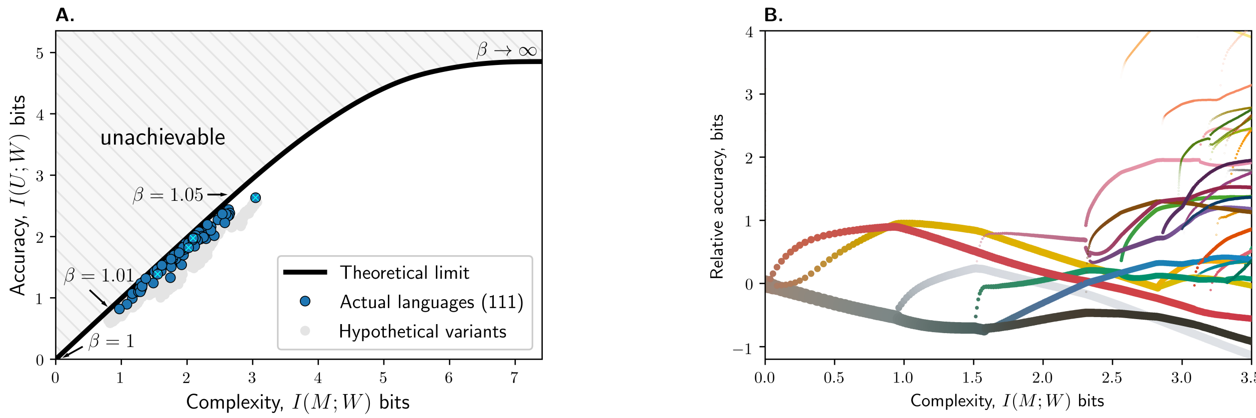


Fig. S5. CIELUV space. Information plane (**A**) and bifurcation diagram (**B**) for the full LI source. These figures are similar to Fig.3 and Fig.5 in main text, but they are based on the results for CIELUV instead of CIELAB.

6.2. Category effects and biological constraints. By grounding our model in a presumed universal perceptual color space such as CIELAB, we have implicitly assumed that this underlying representation is not affected by language. However, it is known that in fact there are lexical effects on the perceived similarity of colors (e.g. 15). While distances between colors in CIELAB may have been influenced to some extent by such category effects, we believe it is unlikely that this has introduced a substantial bias to our model. One reason for this belief is that our model is able to account for wide cross-language variation in color naming based on the same underlying perceptual space for all languages. Another reason is that category effects on color memory (e.g. 16, 17) have themselves been accounted for by assuming the same universal perceptual space, CIELAB, combined with knowledge of language-specific categories (18). These outcomes, which are consistent with a universal perceptual space, seem unlikely given a perceptual space that is instead strongly biased toward lexical categorization in one language, such as English.

It has recently been shown that pre-linguistic infants exhibit categorical distinctions that resemble common patterns in the WCS data (14), and this finding has been taken to suggest a pre-linguistic biological basis for color categorization. That conclusion is broadly consistent with our assumption of a universal color space, although our analysis is based solely on data from adults, and we do not attempt to directly engage the question of color categorization in infants.

6.3. Perceptual uncertainty. The color meaning space ($\mathcal{M}$) that we assumed has a free parameter, σ^2, that determines the speaker's level of perceptual uncertainty. We set $\sigma^2 = 64$ based on a result reported in (19) which suggested that this value corresponds to a distance over which two colors can be comfortably distinguished. To further justify this setting, we evaluated our IB model with a higher ($\sigma^2 = 500$) and lower ($\sigma^2 = 36$) level of perceptual uncertainty. The higher value, $\sigma^2 = 500$, corresponds to a level of perceptual uncertainty that has been used in previous studies (e.g. 2, 20). Table S5 shows the quantitative results for our IB model with different levels of perceptual uncertainty, and with respect to the full LI source. It can be seen that under higher uncertainty, the model is slightly worse on all three measures. Under lower uncertainty the model is slightly better in terms of ε_l but slightly worse in terms of gNID. This suggests that the value of σ^2 that we used is in a reasonable region; however, slightly lower values could perhaps improve the model. This remains a question for future work.

Table S5. Evaluation of IB with different levels of perceptual uncertainty.

	σ^2	ε_l	gNID	NID	β_l
Lower perceptual uncertainty	36	0.13 (±0.06)	0.23 (±0.11)	0.31 (±0.08)	1.01 (±0.01)
Baseline (main model)	64	0.18 (±0.07)	0.18 (±0.1)	0.31 (±0.07)	1.03 (±0.01)
Higher perceptual uncertainty	500	0.26 (±0.06)	0.31 (±0.12)	0.41 (±0.08)	1.77 (±0.20)

Numbers correspond to averages over languages ±1 SD. Lower values are better for ε_l, gNID and NID.

6.4. Validity of the WCS protocol. In the WCS protocol, field workers were instructed to encourage participants to provide short color terms. In practice, these instructions were not applied equally across languages, and in some languages this biased the free naming task towards frequently used terms. This raises a concern about the quality of the WCS data and questions results based on these data. Gibson et al. (21) addressed this issue by comparing color naming data they collected in a free naming task and in a fixed naming task, and showing that their results were robust to these two conditions. To assure that our results were also not influenced by this issue, we applied a similar approach to our analysis.

Specifically, we considered the English color naming data that were collected by Lindsey and Brown (LB, 22) in a free naming task. LB used an improved experimental protocol for this task, and therefore the quality of their data is irrefutable. We also considered a modified version of these data which is based only on major terms (MT data), as described in section 4.1. Fig.S6D shows that the complexity and accuracy values evaluated from the LB data and the MT data are very similar. In addition, Fig.S6A-Fig.S6C show that the naming distribution estimated from the LB data is fairly similar to the naming distribution estimated from the MT data, and that the IB predictions are also similar in both cases. This suggests that our information-theoretic analysis is robust to restricting the naming responses to major terms, and thus the WCS data can be considered reliable in our setting.

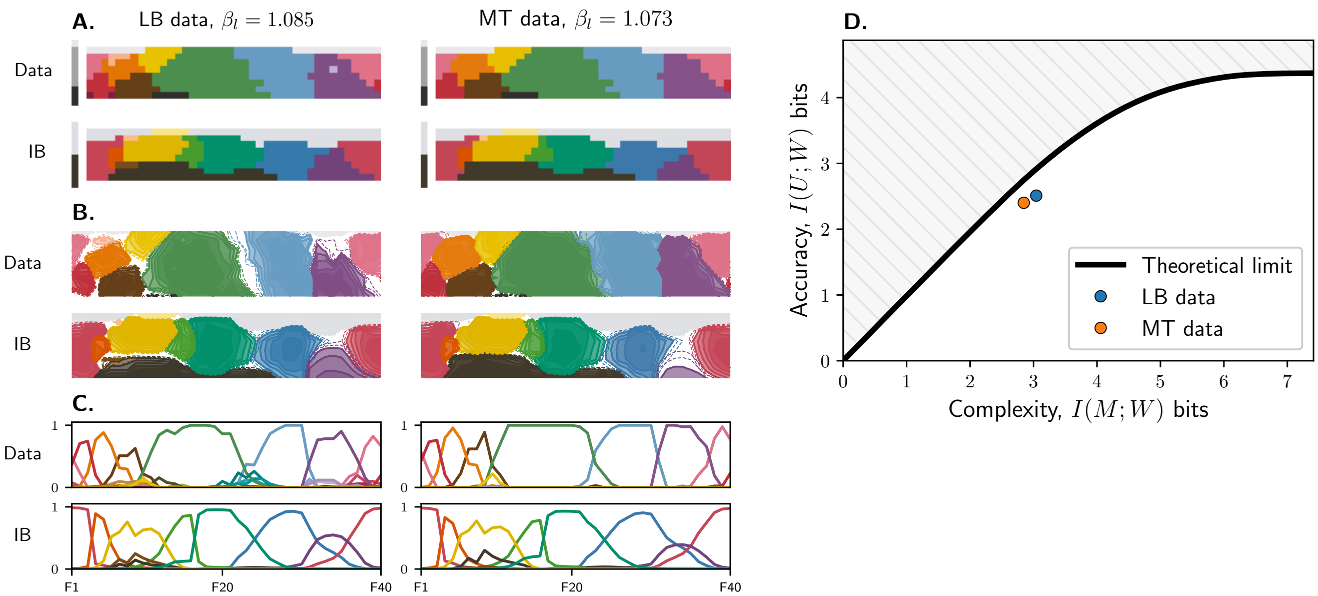


Fig. S6. English color naming data. Mode maps (**A**), contour plots (**B**) and naming probabilities along row F of the WCS palette (**C**), as in Fig.4. Data rows correspond to the English color naming distribution estimated from the LB data (left), which considers all color terms, and from the modified MT data (right), which was restricted to major color terms. **D.** Complexity and accuracy evaluated based on the LB data the modified MT data.

7. Alternative source distributions

In this section we examine two alternatives to the LI source – the uniform distribution, which we used as a baseline for evaluation, and another approach based on image statistics.

7.1. Uniform distribution. The quantitative results for the uniform source are reported in the main text. We complete this picture by presenting Fig.S7A, Fig.S7B and Fig.S8, which are analogous to Fig.3, Fig.5 and Fig.4 in the main text, but were evaluated for the uniform source. In this case, the languages in our data also lie near the theoretical limit (Fig.S7), although not as close as they do with the LI source (this can be seen by comparing ε_l for IB under the uniform and LI source in Table 1). In addition, although both IB and RKK+ capture some of the structure in the data even with the uniform source (Fig.S8), this fit does not look as good as the fit based on the LI source (Fig.4 and section 10). This is consistent with Table 1, which quantitatively shows that the LI source improves the similarity between each model and the data.

Note that since the uniform source does not take into account communicative needs, the IB model with this source only reflects properties of the perceptual CIELAB space that are extracted by IB. The bifurcation diagram (Fig.S7B) in this case reveals a similar yellow discrepancy as observed for the LI source, in which a yellow category emerges at the earliest stage. This suggests that the yellow discrepancy is directly related to the irregular distribution of stimulus colors in CIELAB space.

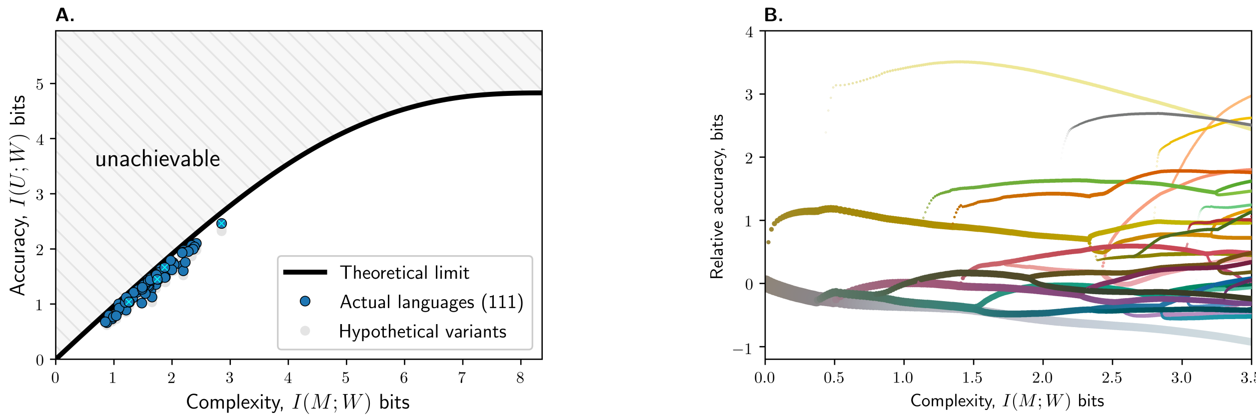


Fig. S7. Uniform source. Information plane (**A**) and bifurcation diagram (**B**) evaluated for the uniform source. For more details see captions of Fig.3 and Fig.5 in main text.

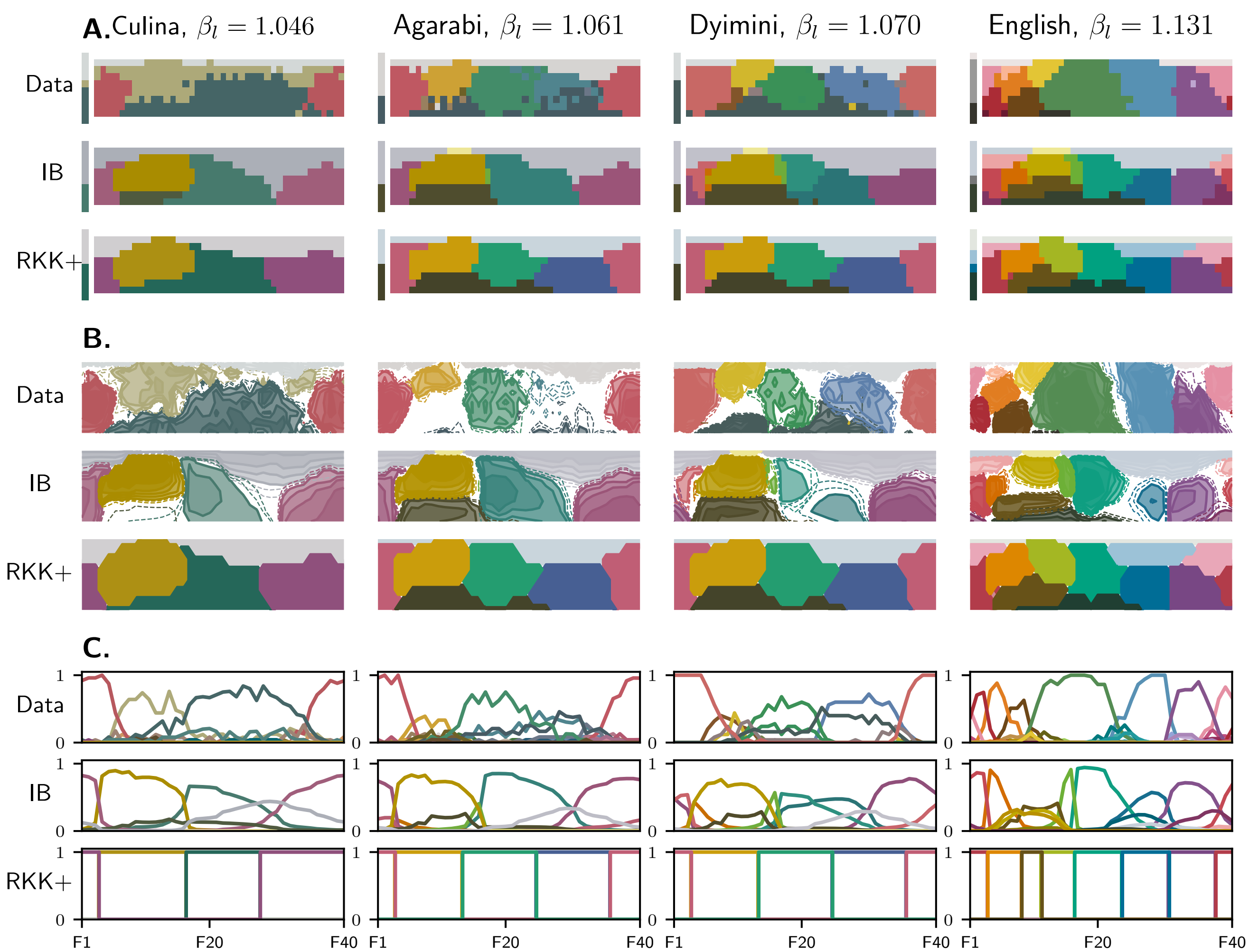


Fig. S8. Uniform source. Mode maps (**A**), contour plots (**B**) and naming probabilities along row F of the WCS palette (**C**), for the color naming distributions (data) and for the IB and RKK+ models. These plots are similar to Fig.4 in main text, where the only difference is that they were evaluated with respect to the uniform source.

7.2. Salience-weighted distribution. Another possible approach for estimating the source distribution is based on the frequencies of colors in natural images. We used the color salience data of Gibson et al. (21), in which the salience of a color is defined by the frequency with which it appears in objects in a large set of images, relative to its frequency either in objects or in backgrounds, under the assumption that foreground objects are more likely to be spoken about than backgrounds are. Gibson et al. estimated the salience of 80 out of the 320 chromatic chips in the WCS palette, and obtained a salience-weighted (SW) prior by taking the probability of each chip to be proportional to its salience.

In order to apply the SW approach to our setting, we first constructed a salience function over CIELAB space by interpolating Gibson et al.'s salience data. We used RBF interpolation with basis functions $\phi(x - x_i) = \sqrt{\frac{\|x-x_i\|^2}{2\sigma^2} + 1}$ and $\sigma^2 = 64$ as in our main analysis. Based on this interpolated function, we estimated the salience of all 330 WCS chips and constructed a SW prior over them (see Fig.S9). This prior corresponds to a SW source.

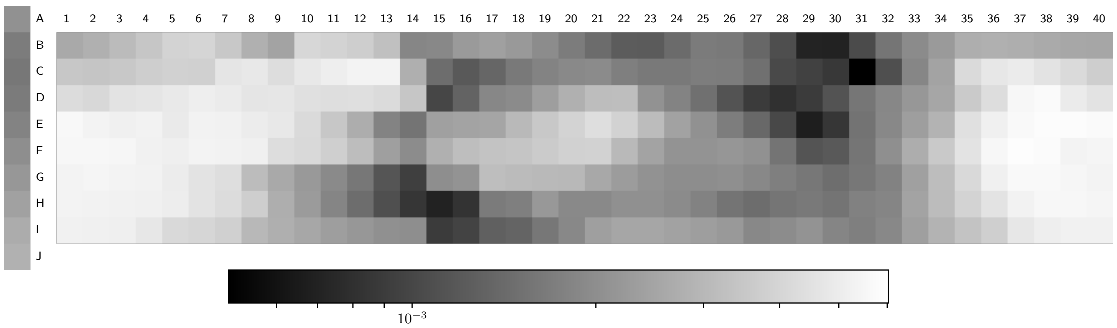


Fig. S9. The estimated salience-weighted (SW) prior over the 330 WCS chips. This prior was interpolated from the salience data of Gibson et al. (21).

We repeated our analysis exactly as described in the main text, but this time with the SW source. Our results show that in this case as well, naturally occurring color naming systems lie near the theoretical limit (Fig.S10A), and that IB achieves better scores than RKK+ (Table S6). Therefore, these results appear to be robust across the three reasonable source distribution we considered.

A comparison of Table S6 and Table 1 shows that the quantitative results with the SW source are similar to the results with the uniform source, and not as good as the results with the LI source. This can also be seen qualitatively by looking at Fig.S11 and Fig.S10B, which were evaluated for the SW source. Note that the effect of the SW source on the performance of the model is not specific to the IB principle — both IB and RKK+ do not fit the data well when evaluated with the SW source compared to the LI source or even to the uniform source. One possible explanation is that the SW source is strongly biased towards warm (reds/yellows) colors and does not weigh achromatic colors (in particular black and white) properly. This can clearly be seen in Fig.S9, and in Gibson et al.'s salience data before our interpolation. Although Gibson et al. argue that warm colors are more useful for communication than cool colors, and in that sense the SW source make sense, it seems unlikely that dark/light colors would have the low communicative need assigned to them by the SW prior.

Table S6. Quantitative evaluation (SW source)

Source	Model	ε_l	gNID	NID	β_l
SW	IB	0.24 (±0.09)	0.40 (±0.14)	0.54 (±0.12)	1.05 (±0.02)
	RKK+	0.96 (±0.22)	0.65 (±0.08)	0.51 (±0.10)	

Averages over left-out languages ±1 SD. Lower values of ε_l, gNID and NID are better.

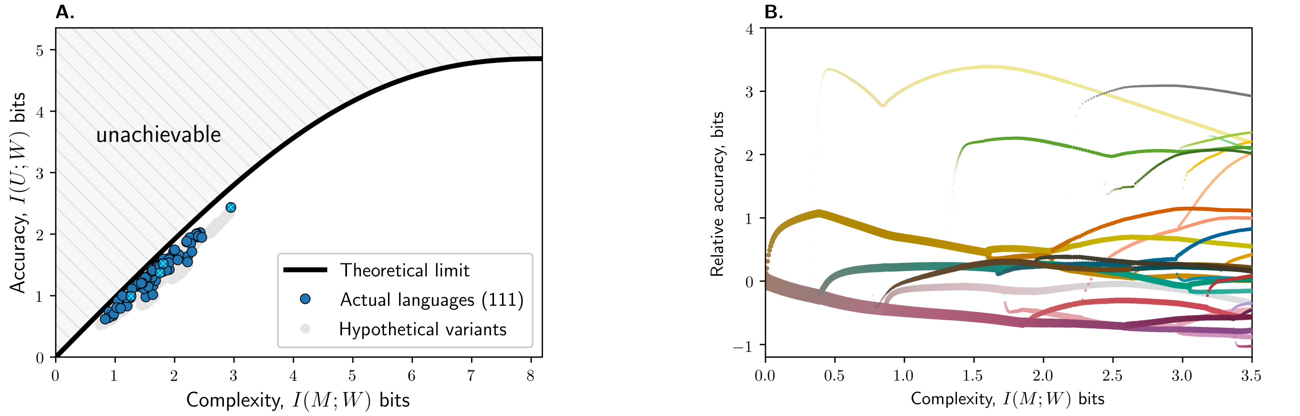


Fig. S10. SW source. Information plane (**A**) and bifurcation diagram (**B**) evaluated for the SW source. For more details see captions of Fig.3 and Fig.5 in the main text.

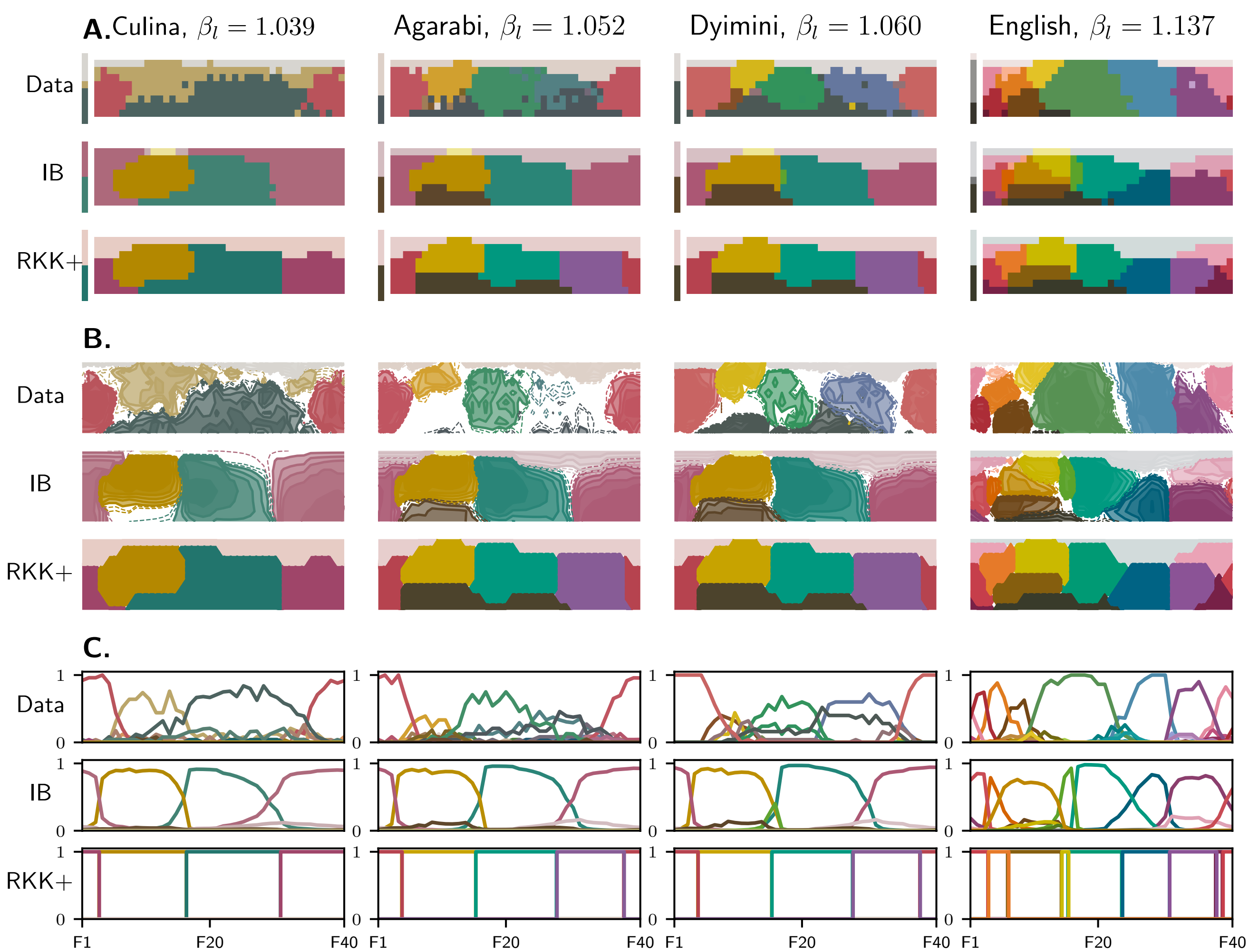


Fig. S11. SW source. Mode maps (**A**), contour plots (**B**) and naming probabilities along row F of the WCS palette (**C**), for the color naming distributions (data) and for the IB and RKK+ models. These plots are similar to Fig.4 in main text, where the only difference is that they were evaluated with respect to the SW source.

8. Hypothetical color naming systems

Is it a trivial result that naturally occurring color naming systems lie near the IB curve? Perhaps any 'reasonable-seeming' color naming system would lie near the curve, whether or not it is similar to naming systems found in the world's languages. Randomly generated color naming systems will typically lie close to the origin in the information plane. Such systems are non-informative and are thus not useful for color categorization. Therefore, in order to show that it is not trivial that naturally occurring color naming systems lie near the IB curve (and far from the origin), we considered two types of hypothetical color naming systems that maintain some informative structure about color space.

8.1. Rotation analysis. Following (20), we constructed a control set of 39 hypothetical variants for each language which were obtained by rotating its color naming distribution in the hue dimension across the columns of the WCS palette. Examples of a few hypothetical variants of Culina are shown in Fig.S12. $r = 0$ corresponds to the actual language, $r = 2$ corresponds to a shift of two columns to the right, and $r = -2$ corresponds to a shift of two columns to the left.

If languages are shaped by pressure for information-theoretic efficiency as defined by IB, we would expect that naturally occurring color naming systems would be more efficient than their hypothetical variants. To test this, for each rotated color naming system, $q_{l,r}$, we evaluated the deviation from optimality, or efficiency loss, in the same way we evaluated ε_l for the actual language, i.e. $\varepsilon_{l,r} = \min_\beta \frac{1}{\beta}(\mathcal{F}[q_{l,r}] - \mathcal{F}^*_\beta)$. We compared the efficiency of the language and the efficiency of its variants by considering $\varepsilon_{l,r} - \varepsilon_l$ (Δ efficiency loss) for IB with the full LI source. Fig.S13 shows that 93% of the languages are more efficient than all of their hypothetical variants. The remaining 7% are more efficient than most of their variants, and the preferred rotation is attained at a small $|r|$.

However, one could argue that these results are an outcome of the LI source, which was estimated with respect to the unrotated color naming systems. We therefore repeated this analysis with the uniform source. Fig.S14) shows that the results in this case are similar. This suggests that the actual languages are indeed more efficient than their hypothetical variants. The advantage of the actual languages can be explained by their alignment with the irregular structure of CIELAB space (20), which influences the accuracy of communication in the IB model. We also repeated this rotation analysis for colors that are represented in CIELUV space, and obtained similar results.

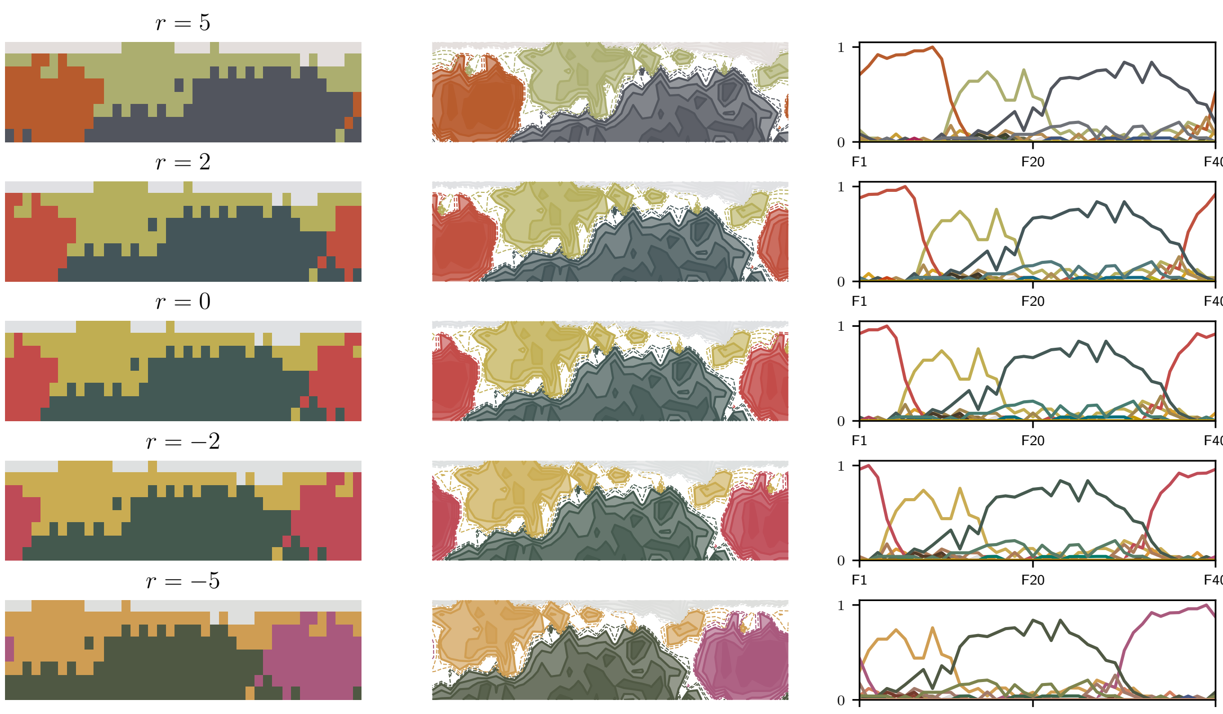


Fig. S12. Rotation example. Hypothetical variants for Culina obtained by rotating its color naming distribution in the hue dimension across the columns of the WCS palette. $r = 0$ corresponds to the actual language, $r = 2$ corresponds to a shift of two columns to the right, and $r = -2$ corresponds to a shift of two columns to the left. Colors correspond to the color centroid of each category, and columns correspond to mode maps (left), contour plots of the naming distribution (middle) and conditional probabilities along row F of the WCS palette (right).

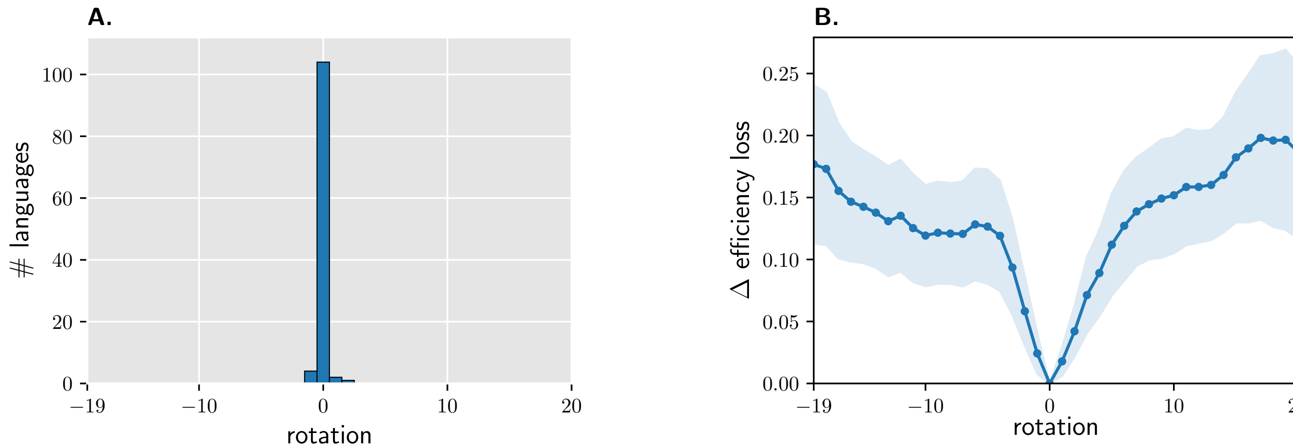


Fig. S13. Rotation analysis for the full LI source. A. Histogram of the most efficient rotation across languages. Rotation 0 corresponds to the actual language, and it is the most efficient for 93% of the languages in our data. **B.** Differences between the efficiency loss of the rotated language and the actual language, Δ efficiency loss $= \varepsilon_{l,r} - \varepsilon_l$. Lower values are better. Blue curve is the average across languages, and the colored region corresponds to ±1 SD across languages.

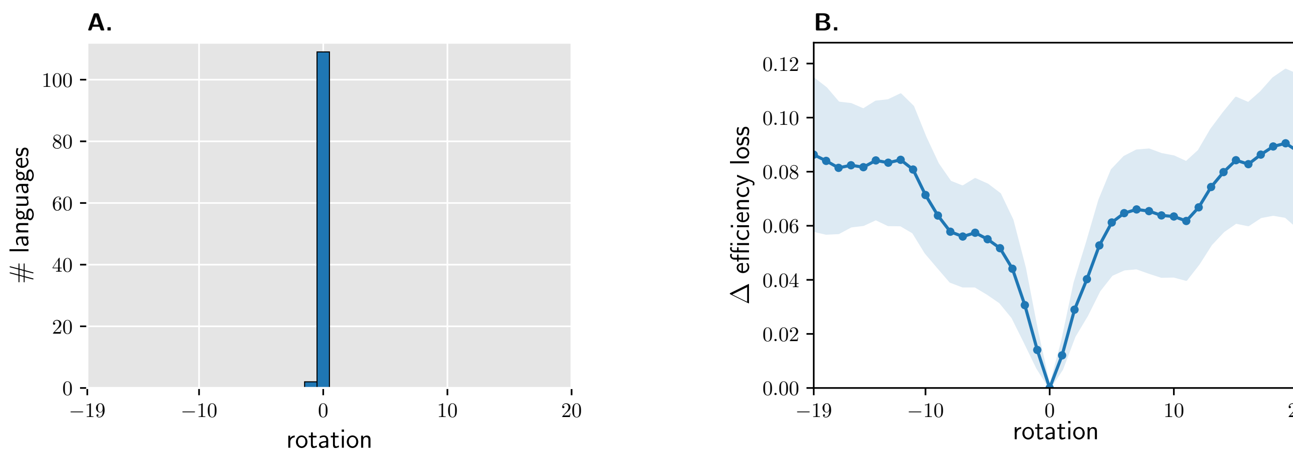


Fig. S14. Rotation analysis for the uniform source. A. Histogram of the most efficient rotation across languages. Rotation 0 corresponds to the actual language, and it is the most efficient for 98% of the languages in our data. **B.** Differences between the efficiency loss of the rotated language and the actual language, Δ efficiency loss $= \varepsilon_{l,r} - \varepsilon_l$. Lower values are better. Blue curve is the average across languages, and the colored region corresponds to ±1 SD across languages.

8.2. Structured control set based on random Gaussians. We considered another set of structured hypothetical systems in which the naming distribution is defined by random Gaussians over CIELAB space. We constructed a hypothetical system with K categories by (1) randomly selecting K chips c_w as representatives for categories $w = 1\dots, K$; (2) assigning to each category a random covariance matrix Σ_w; and (3) defining the color naming distribution by

$$q(w|m_c) \propto \exp\left(-\frac{1}{2}(c-c_w)^\top \Sigma_w^{-1}(c-c_w)\right) . \quad \text{[S25]}$$

Σ_w induces a random transformation of CIELAB space and its eigenvalues are exponentially distributed with mean $\sigma^2 = 64$, which matches the level of perceptual uncertainty we used for constructing the color meaning space. We generated these random matrices as follows: a 3×3 diagonal matrix D was generated by sampling $D_{ii} \sim \text{Exp}(\frac{1}{\sigma^2-1})+1$, and a 3×3 matrix A was generated by sampling uniformly $A_{ij} \in [0, 1]$. The singular value decomposition of $A^\top A$ was evaluated, i.e. $A^\top A = U\Lambda V^\top$. Finally, $\Sigma_w = UDV^\top$.

We constructed these hypothetical systems with $K = 3, \dots, 20$. For each K we sampled 100 systems, yielding a total of 1,800 hypothetical systems (see Fig.S15 for a few examples). We evaluated these systems with the IB model based on the full LI source ($\varepsilon_l = 0.33 \pm 0.1$, gNID $= 0.39 \pm 0.16$, NID $= 0.44 \pm 0.13$) and the uniform source ($\varepsilon_l = 0.36 \pm 0.08$, gNID $= 0.47 \pm 0.15$,NID $= 0.5 \pm 0.13$). In both cases, these hypothetical systems are less efficient on average than the actual languages we considered.

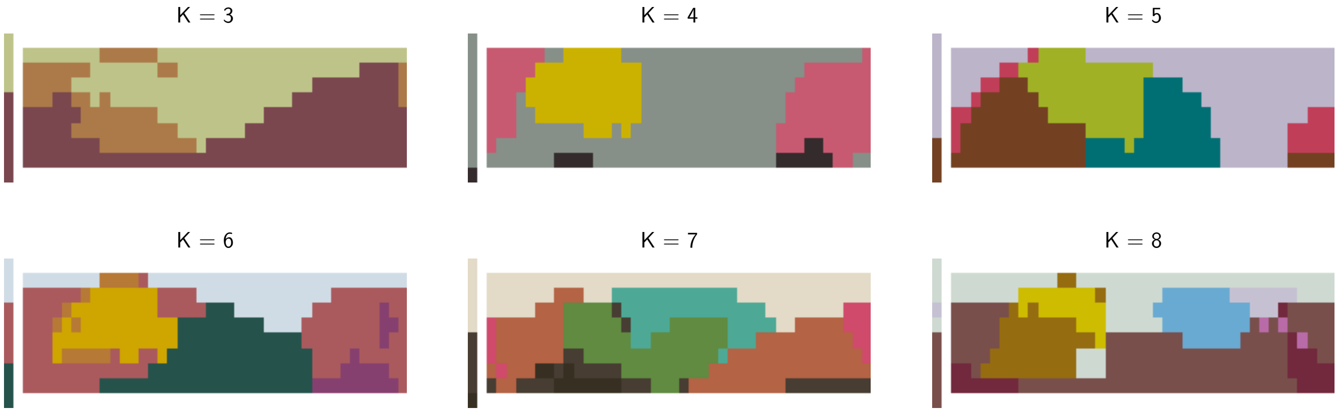


Fig. S15. Examples of hypothetical color naming systems based on K random Gaussians in CIELAB space.

9. Sensitivity analysis

In this section we test the sensitivity of our results to small errors in the structure of the meaning space, $\mathcal{M}$, that we assumed. To do so, we injected a small perturbation to each m_c and re-evaluated IB and RKK+ with the full LI source. We injected the perturbation by first drawing i.i.d. Gaussian variables $Z_{c,u} \sim \mathcal{N}(0, 0.01)$, and defining the perturbed model by $m'_c(u) \propto m_c(u)e^{Z_{c,u}}$. The results, summarized in table S7, are almost identical to the results without perturbation, which suggests that our analysis is robust to small amounts of noise in the perceptual model.

Table S7. Quantitative evaluation with the perturbed meaning space

Source	Model	ε_l	gNID	NID	β_l
	IB	0.18 (±0.07)	0.18 (±0.10)	0.31 (±0.07)	1.03 (±0.01)
LI	IB-C	0.18 (±0.07)	0.21 (±0.08)	0.30 (±0.08)	1.04 (±0.02)
	RKK+	0.70 (±0.23)	0.46 (±0.10)	0.31 (±0.10)	

Numbers correspond to averages over languages ±1 SD. Lower values are better for ε, gNID and NID.

10. Predictions for all languages

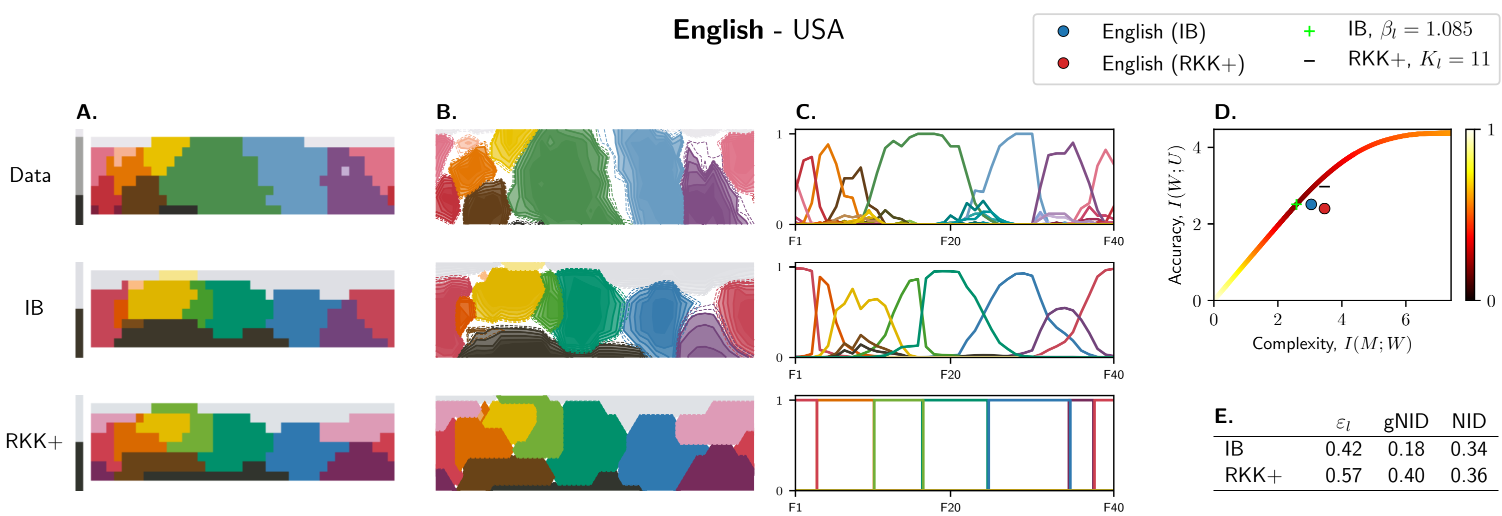
In this section the predictions of the IB model and RKK+ model for all 111 languages in our data are presented. These are based on the full LI source discussed in section 2.

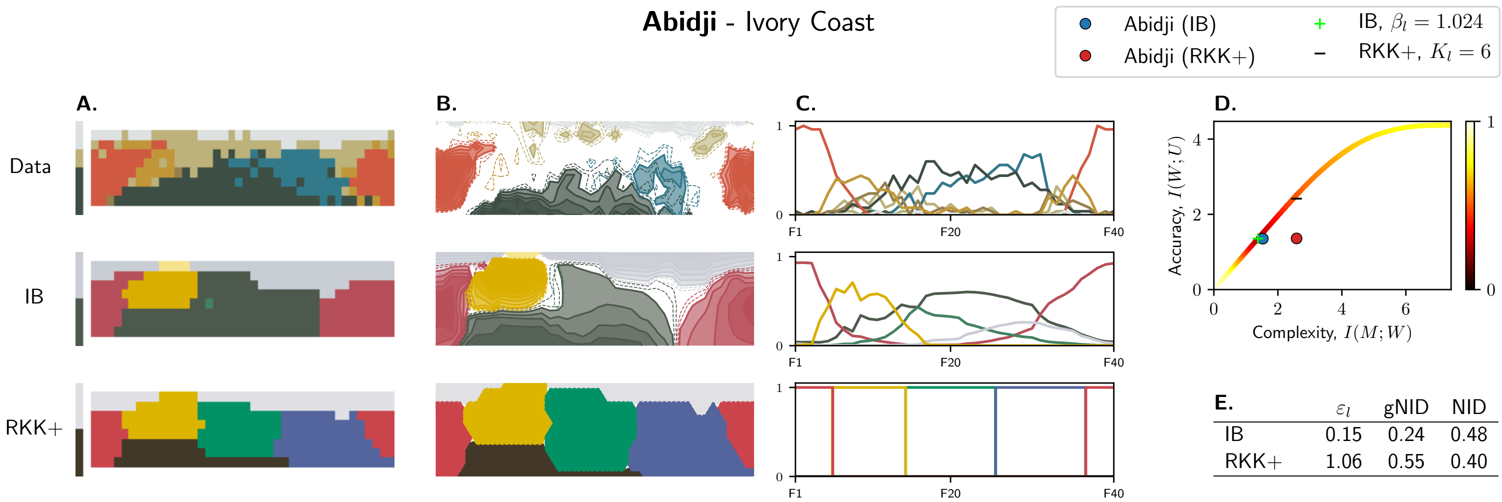
A-C. Similarity between color naming distributions of languages (data rows) and the corresponding optimal IB encoders at β_l (IB rows) and RKK+ encoders for K_l. Each color category is represented by the centroid color of the category. IB predicts soft partitions of color space while RKK+ predicts hard partitions. **A.** Mode maps: each chip is colored according to its modal category. **B.** Contours of the naming distribution. Solid lines correspond to level sets between 0.5 to 0.9; dashed lines correspond to level sets of 0.4 and 0.45. **C.** Naming probabilities along the hue dimension of row F in the WCS palette.

D. Information plane. Complexity-accuracy tradeoffs for each language according to IB (blue dots) and RKK+ (red dots) are compared with the theoretical optima of both principles (green cross and black line respectively). The IB curve is the same as in Fig.3, and also defines the theoretical limit for RKK+ (see section 4.6). Colors along the curve reflect gNID values between the language and the IB systems along the curve.

E. Quantitative evaluation of IB and RKK+. ε_l measures the extent to which language l deviates from the optimum predicted by the model. gNID measures the dissimilarity between the language's color naming distribution and the predicted encoder at β_l in IB or at K_l in RKK+. NID measures the dissimilarity between the model's and language's mode maps. Lower values of ε_l, gNID and NID are better.

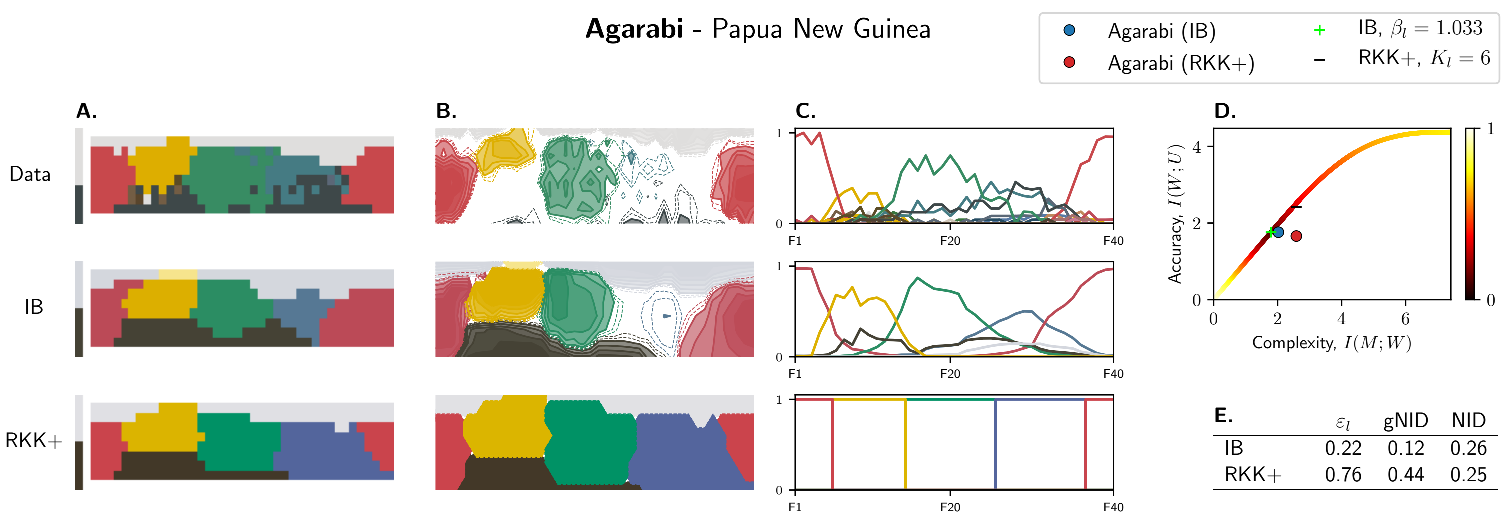
* WCS languages excluded from the estimation of the LI source and from our quantitative model evaluation due to insufficient data (see section 2). For some of these languages we could not evaluate the RKK+ encoders because some chips were not named by any major color term. In such cases the RKK+ prediction is not shown.
Excluded languages: Amuzgo, Camsa, Candoshi, Chayahuita, Chiquitano, Cree, Garífuna (Black Carib), Ifugao, Micmac, Nahuatl, Papago (O'odham), Slave, Tacana, Tarahumara (Central), Tarahumara (Western).





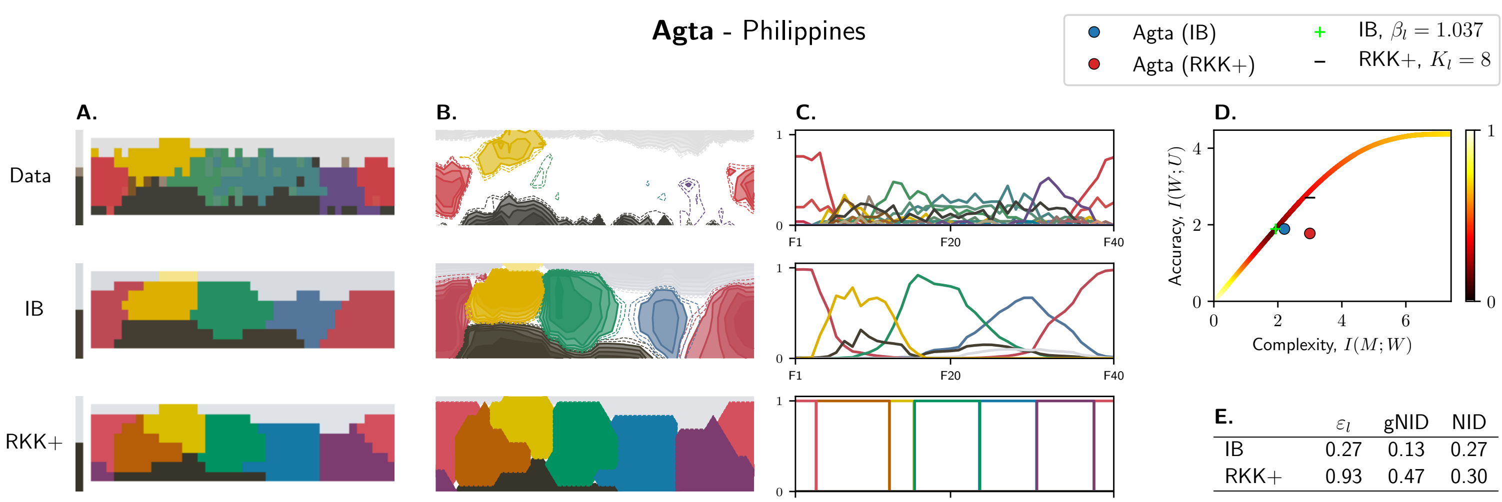
E.

	ε_l	gNID	NID
IB	0.15	0.24	0.48
RKK+	1.06	0.55	0.40



E.

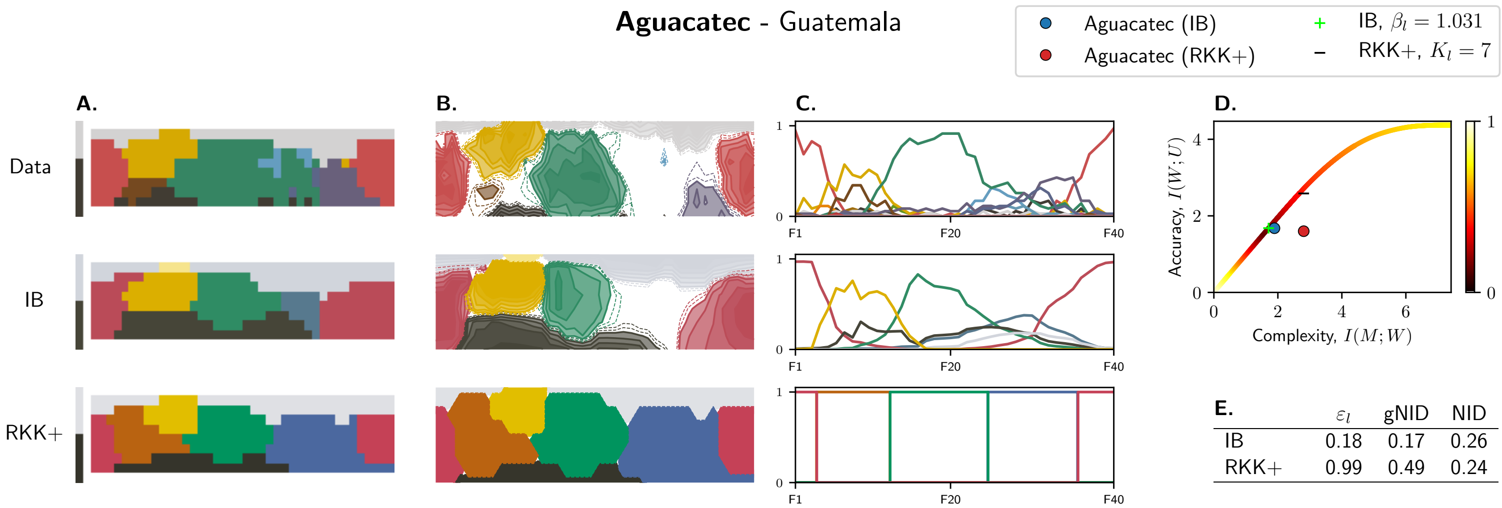
	ε_l	gNID	NID
IB	0.22	0.12	0.26
RKK+	0.76	0.44	0.25



E.

	ε_l	gNID	NID
IB	0.27	0.13	0.27
RKK+	0.93	0.47	0.30

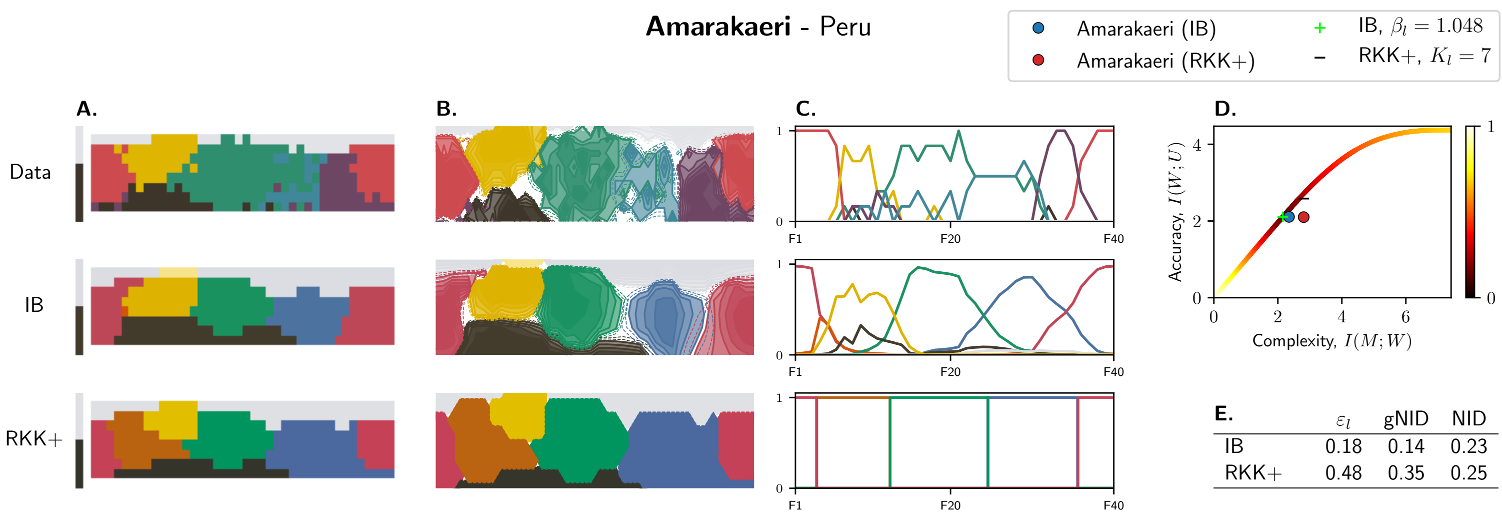
Aguacatec - Guatemala



E.

	ε_l	gNID	NID
IB	0.18	0.17	0.26
RKK+	0.99	0.49	0.24

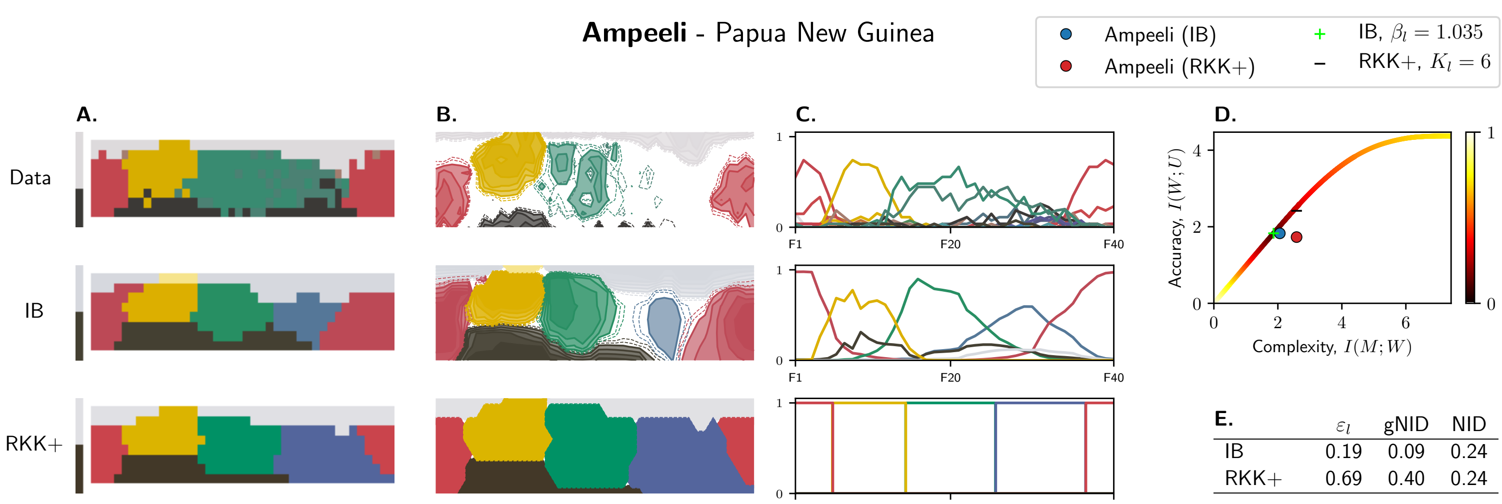
Amarakaeri - Peru



E.

	ε_l	gNID	NID
IB	0.18	0.14	0.23
RKK+	0.48	0.35	0.25

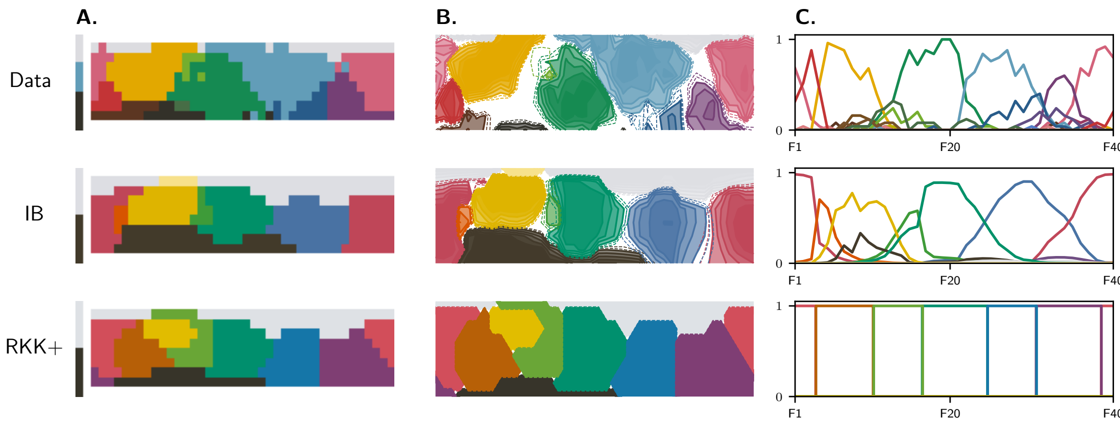
Ampeeli - Papua New Guinea

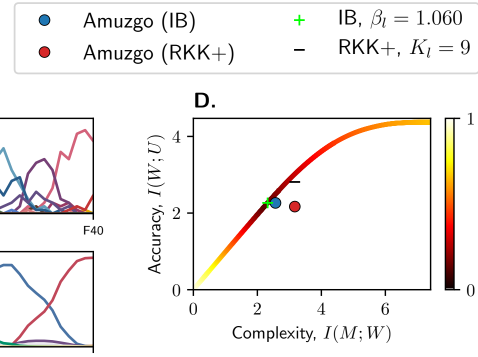


E.

	ε_l	gNID	NID
IB	0.19	0.09	0.24
RKK+	0.69	0.40	0.24

* **Amuzgo** - Mexico

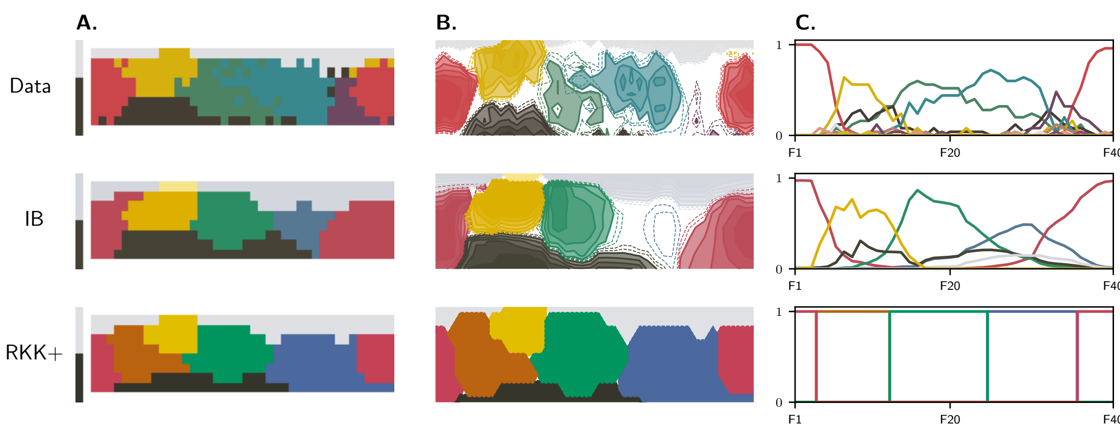


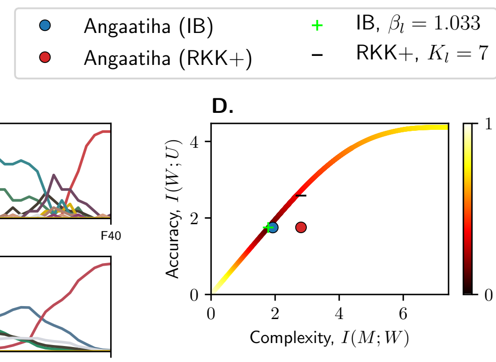


E.

	ε_l	gNID	NID
IB	0.23	0.13	0.35
RKK+	0.64	0.35	0.27

Angaatiha - Papua New Guinea

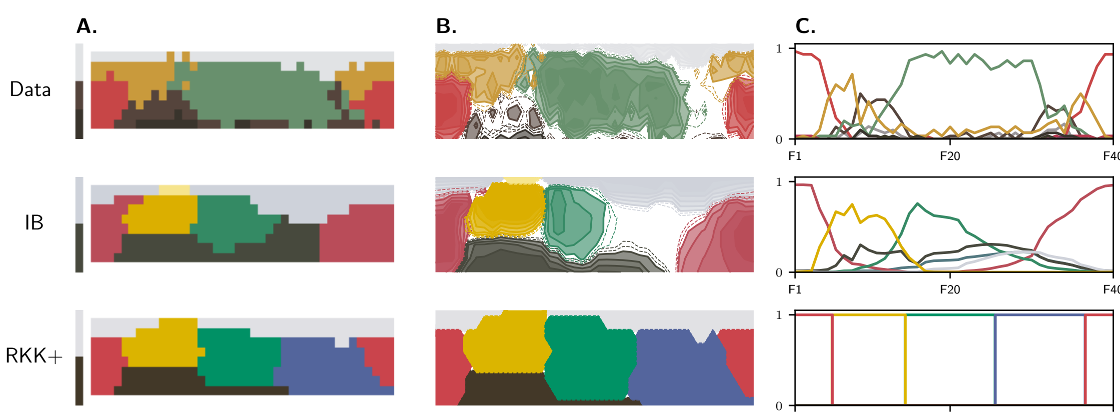


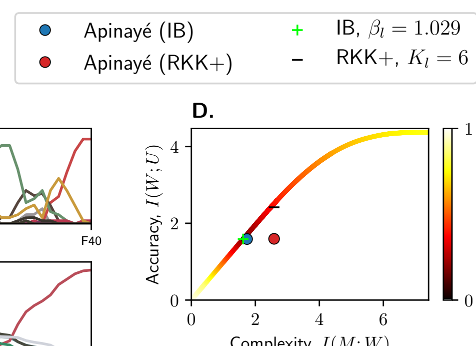


E.

	ε_l	gNID	NID
IB	0.14	0.11	0.22
RKK+	0.83	0.44	0.23

Apinayé - Brazil

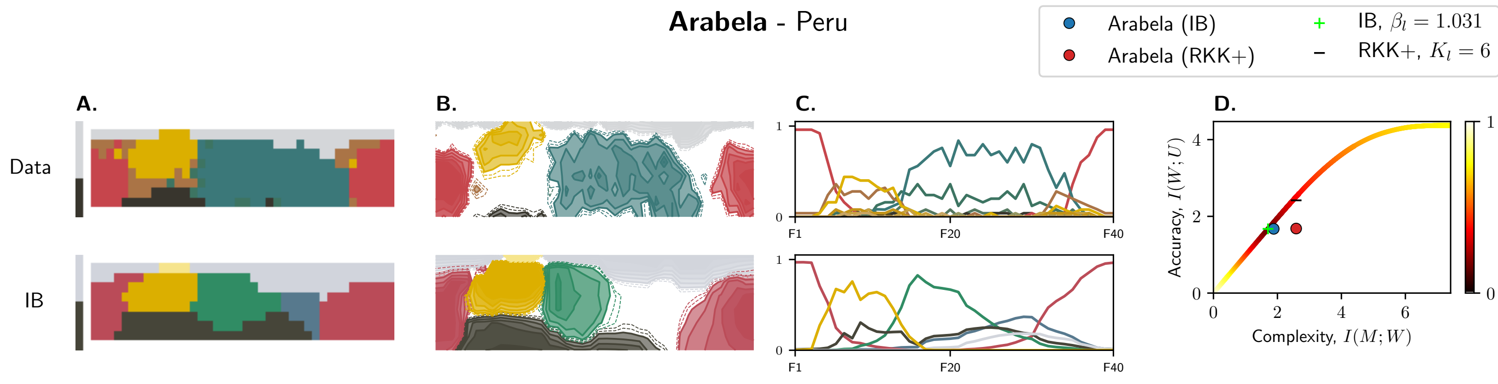




E.

	ε_l	gNID	NID
IB	0.12	0.21	0.39
RKK+	0.82	0.45	0.30

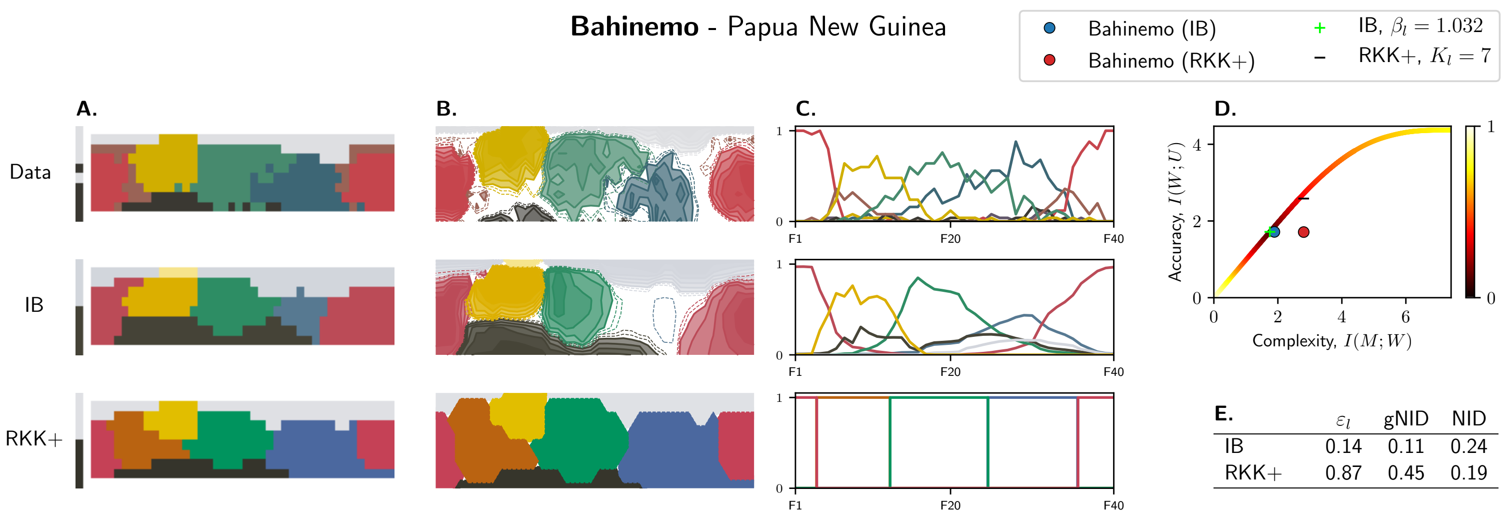
Arabela - Peru



E.

	ε_l	gNID	NID
IB	0.17	0.14	0.27
RKK+	0.73	0.42	0.23

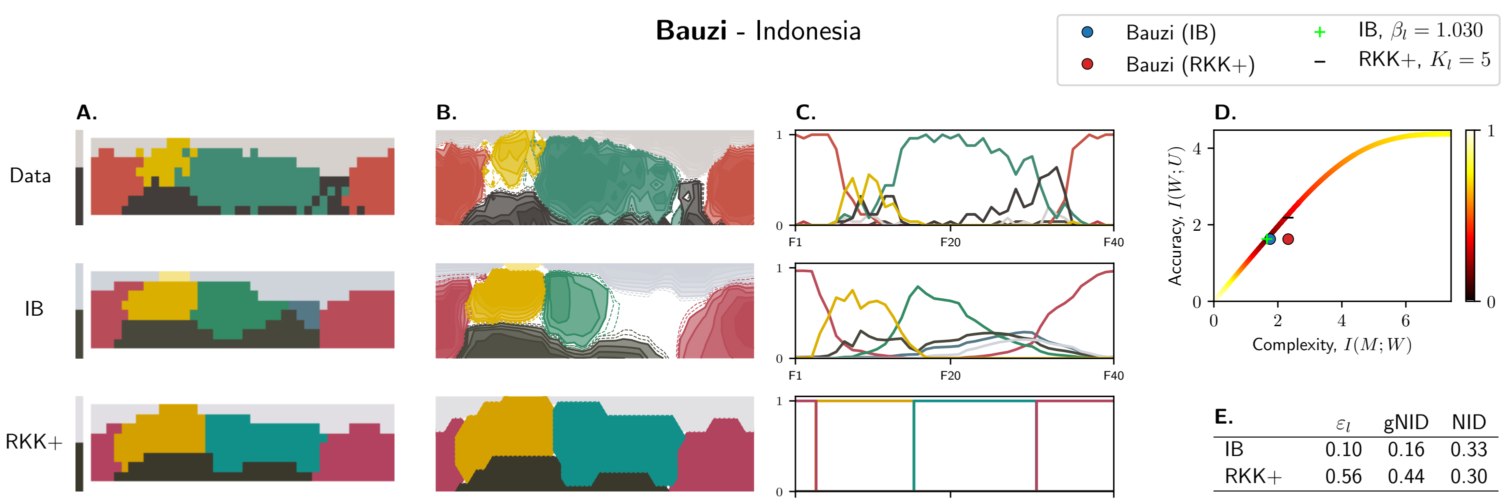
Bahinemo - Papua New Guinea



E.

	ε_l	gNID	NID
IB	0.14	0.11	0.24
RKK+	0.87	0.45	0.19

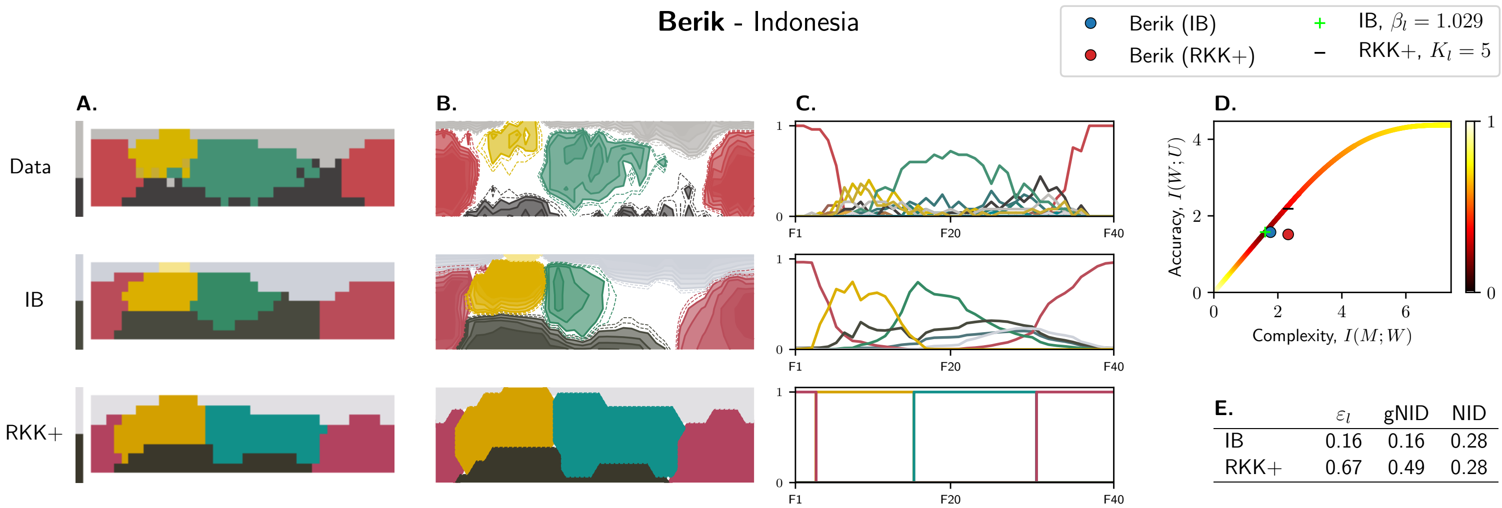
Bauzi - Indonesia



E.

	ε_l	gNID	NID
IB	0.10	0.16	0.33
RKK+	0.56	0.44	0.30

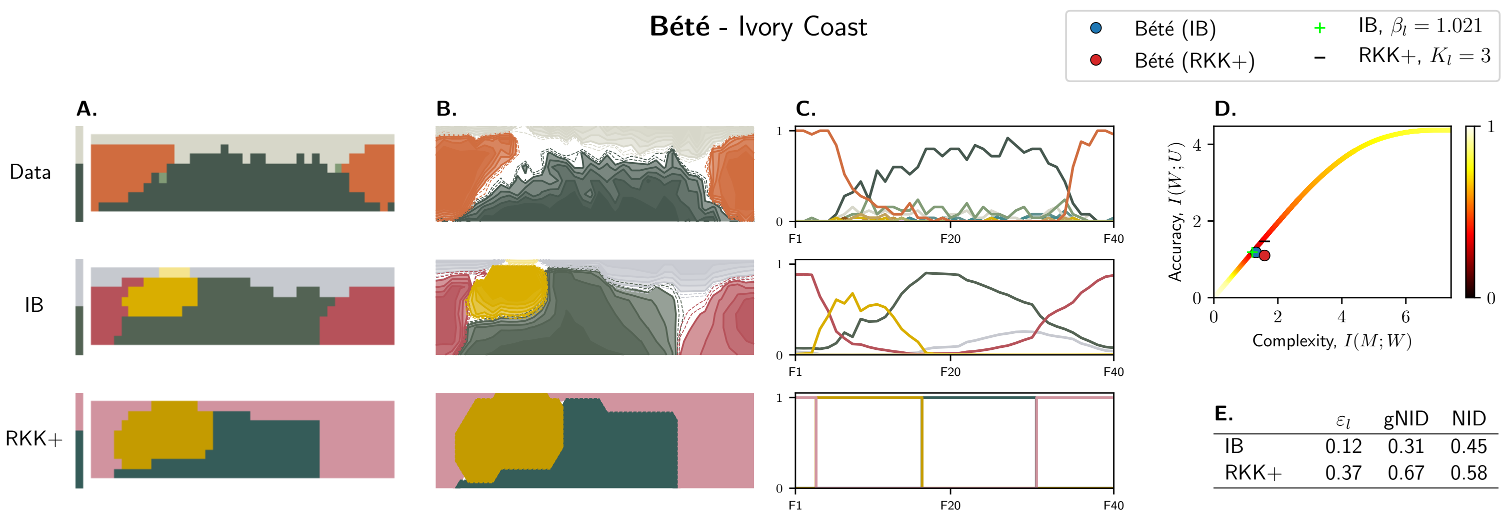
Berik - Indonesia



E.

	ε_l	gNID	NID
IB	0.16	0.16	0.28
RKK+	0.67	0.49	0.28

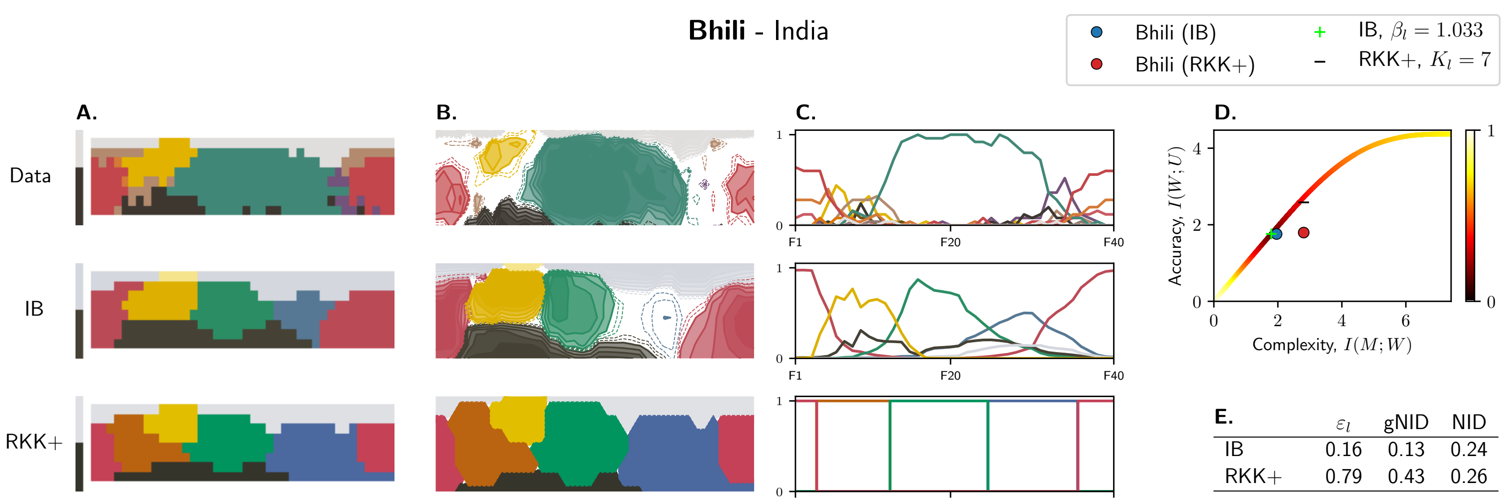
Bété - Ivory Coast



E.

	ε_l	gNID	NID
IB	0.12	0.31	0.45
RKK+	0.37	0.67	0.58

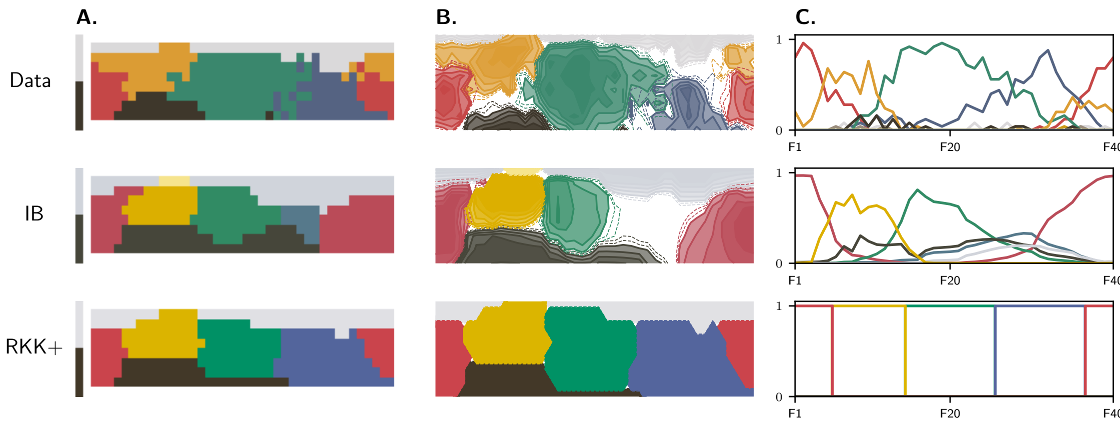
Bhili - India

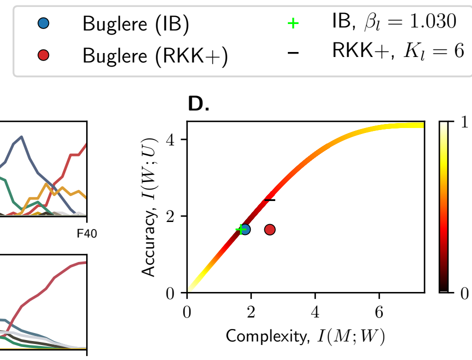


E.

	ε_l	gNID	NID
IB	0.16	0.13	0.24
RKK+	0.79	0.43	0.26

Buglere - Panama

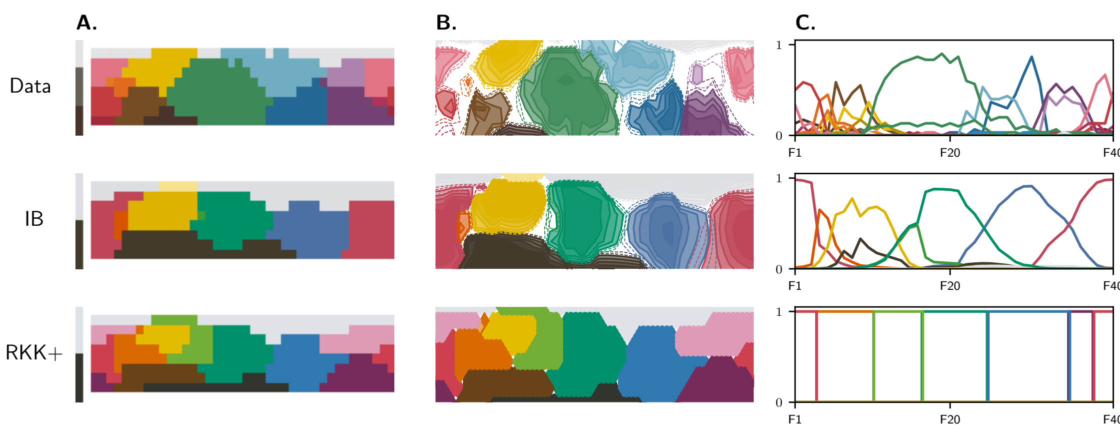


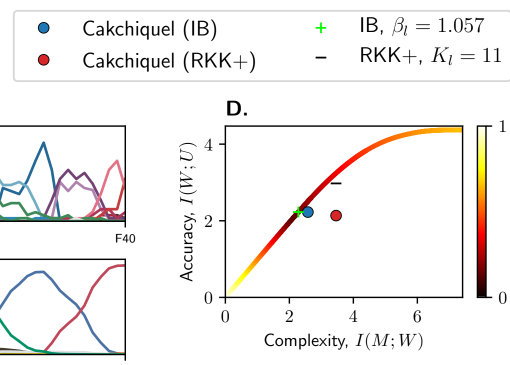


E.

	ε_l	gNID	NID
IB	0.13	0.14	0.28
RKK+	0.77	0.43	0.22

Cakchiquel - Guatemala

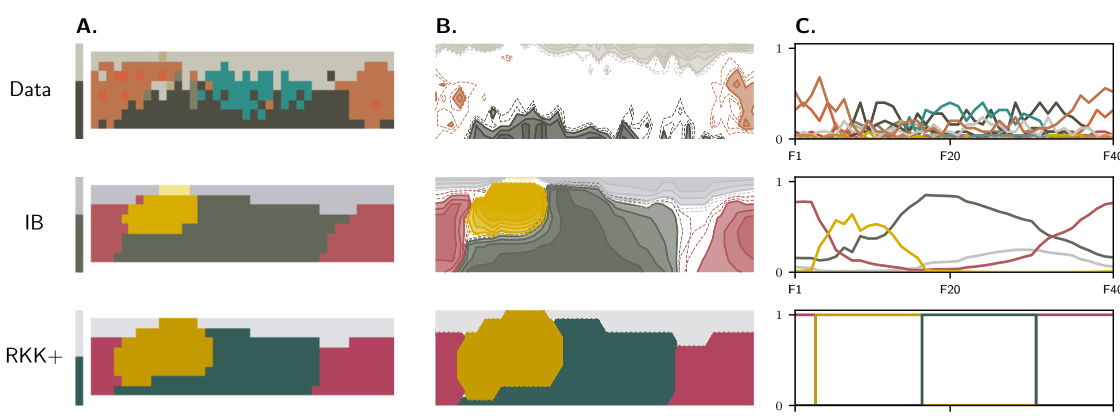


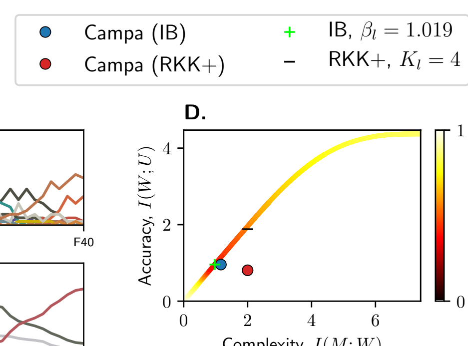


E.

	ε_l	gNID	NID
IB	0.28	0.15	0.35
RKK+	0.84	0.41	0.23

Campa - Peru

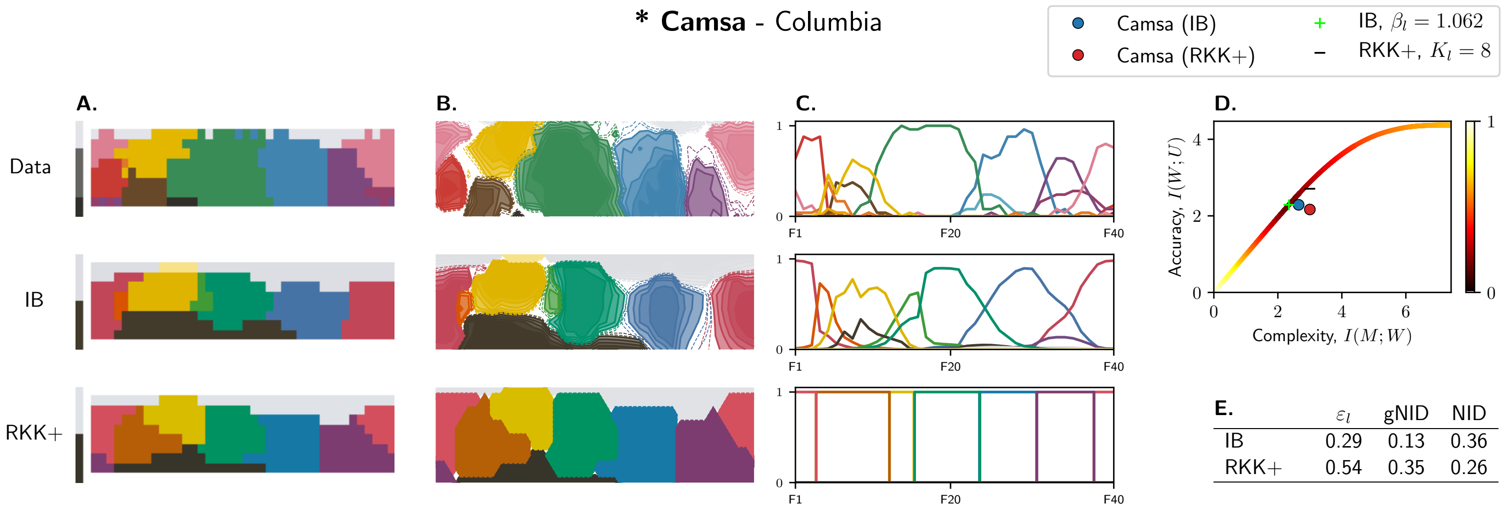




E.

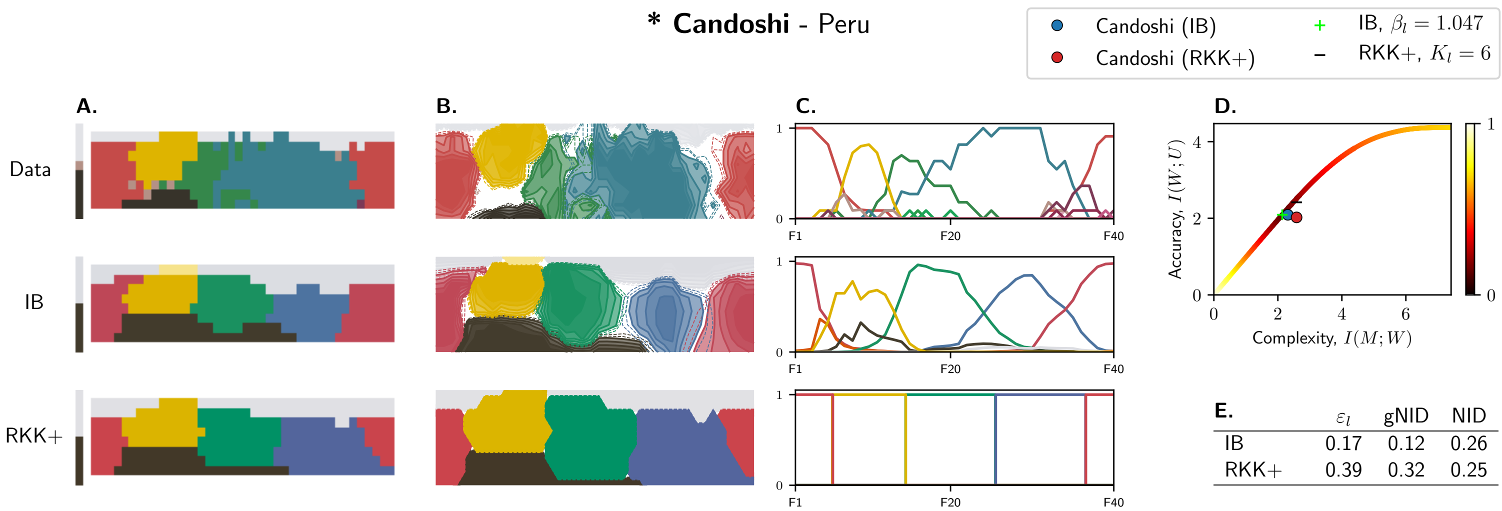
	ε_l	gNID	NID
IB	0.19	0.39	0.42
RKK+	1.07	0.73	0.47

* **Camsa** - Columbia



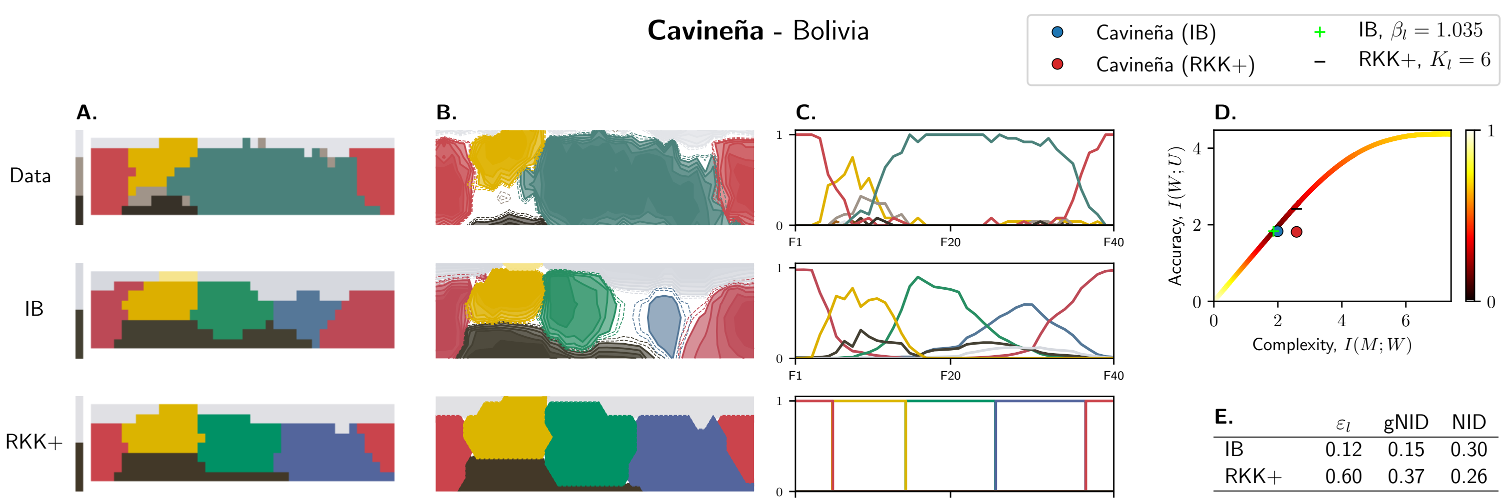
	ε_l	gNID	NID
IB	0.29	0.13	0.36
RKK+	0.54	0.35	0.26

* **Candoshi** - Peru



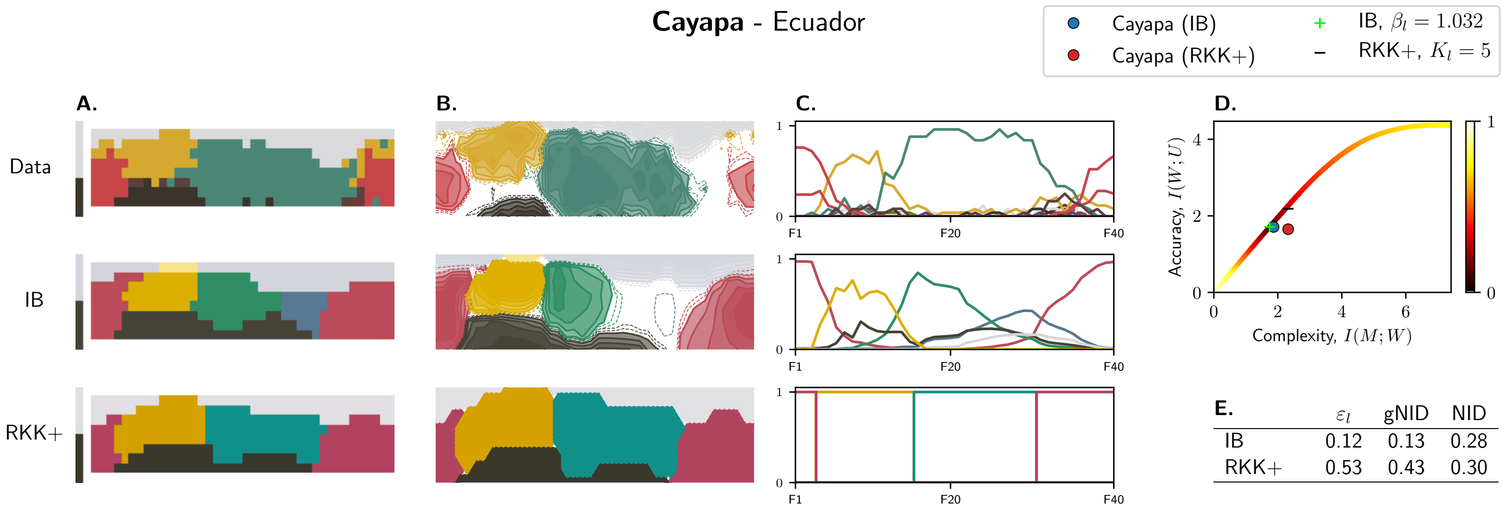
	ε_l	gNID	NID
IB	0.17	0.12	0.26
RKK+	0.39	0.32	0.25

Cavineña - Bolivia



	ε_l	gNID	NID
IB	0.12	0.15	0.30
RKK+	0.60	0.37	0.26

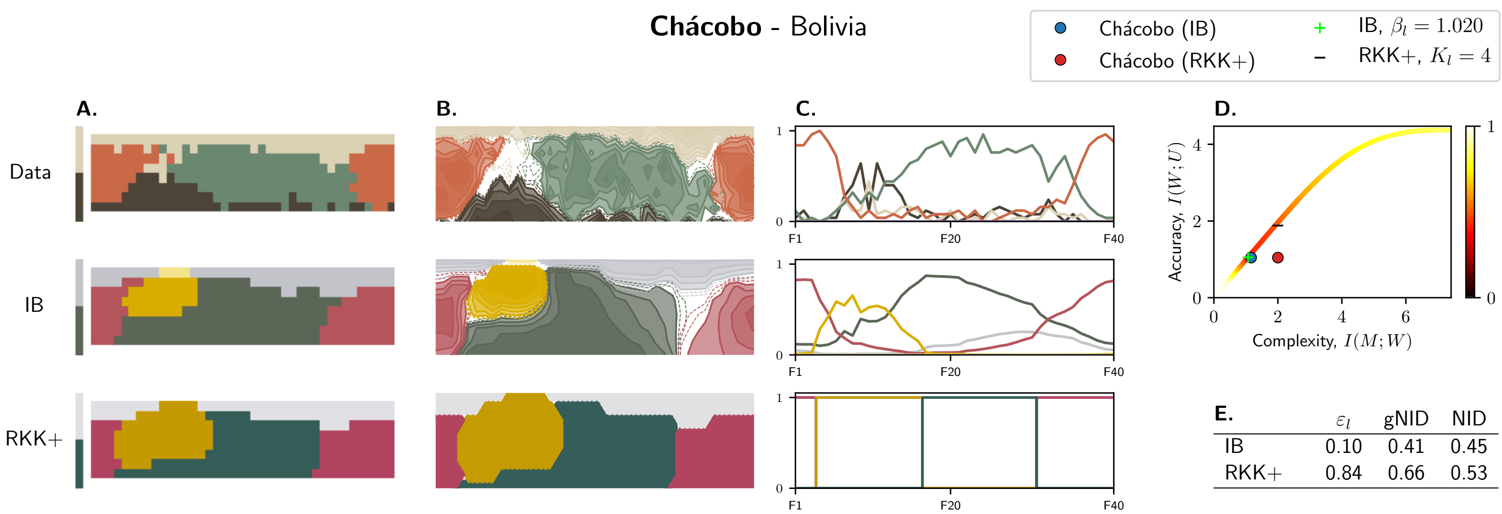
Cayapa - Ecuador



E.

	ε_l	gNID	NID
IB	0.12	0.13	0.28
RKK+	0.53	0.43	0.30

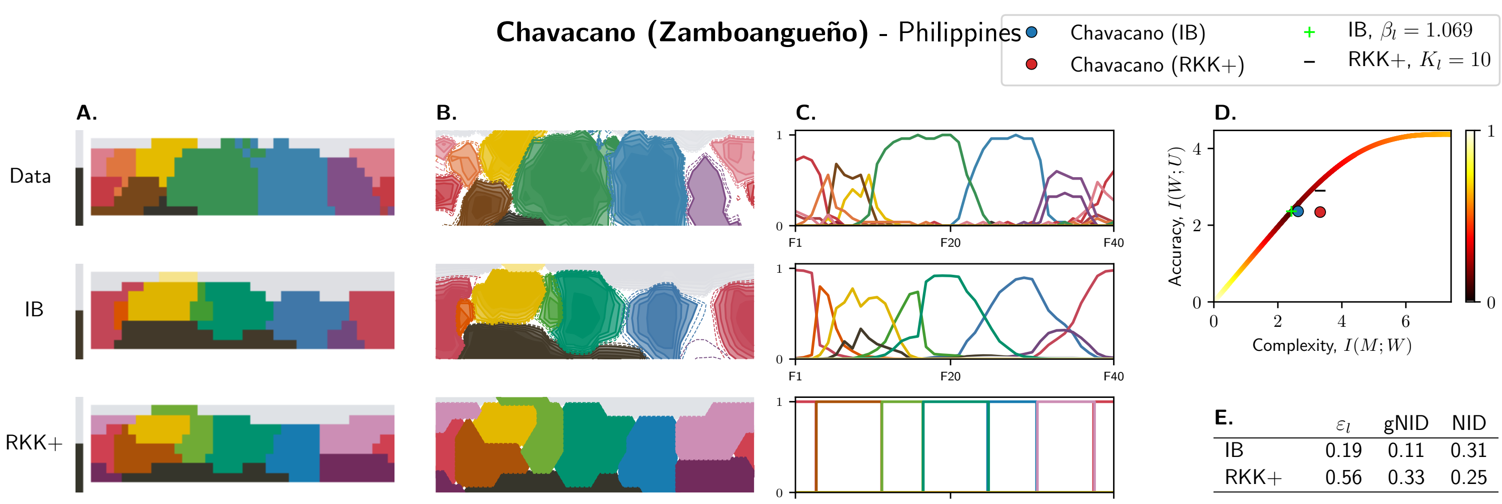
Chácobo - Bolivia



E.

	ε_l	gNID	NID
IB	0.10	0.41	0.45
RKK+	0.84	0.66	0.53

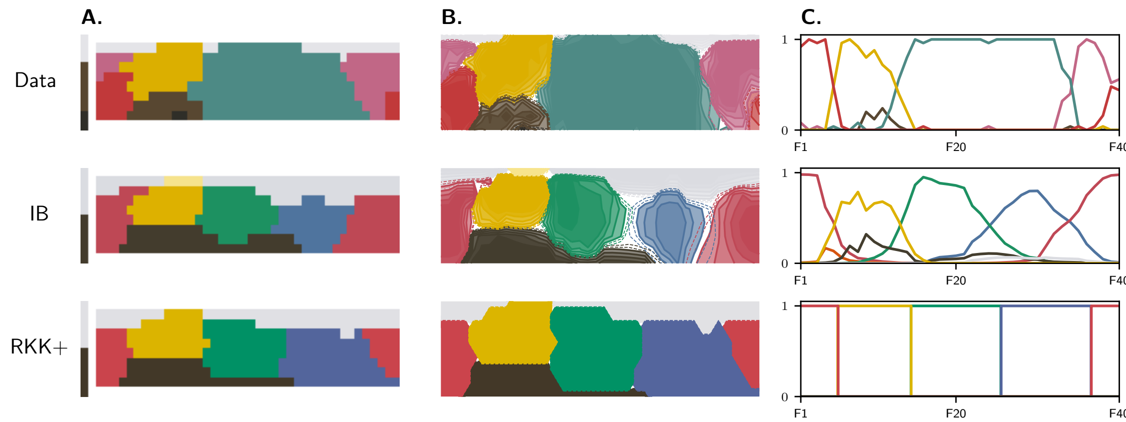
Chavacano (Zamboangueño) - Philippines

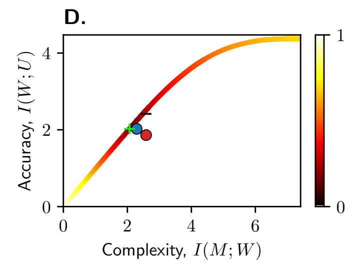


E.

	ε_l	gNID	NID
IB	0.19	0.11	0.31
RKK+	0.56	0.33	0.25

* **Chayahuita** - Peru

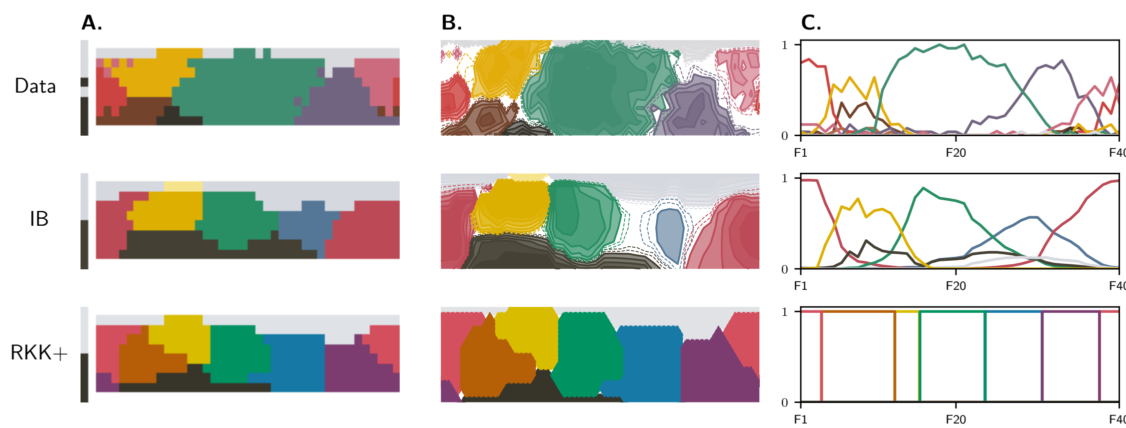


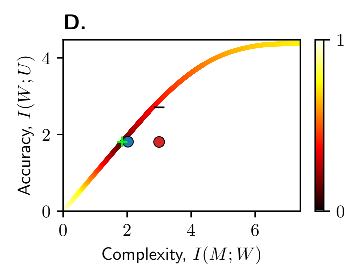


E.

	ε_l	gNID	NID
IB	0.21	0.20	0.34
RKK+	0.55	0.39	0.33

Chinantec - Mexico

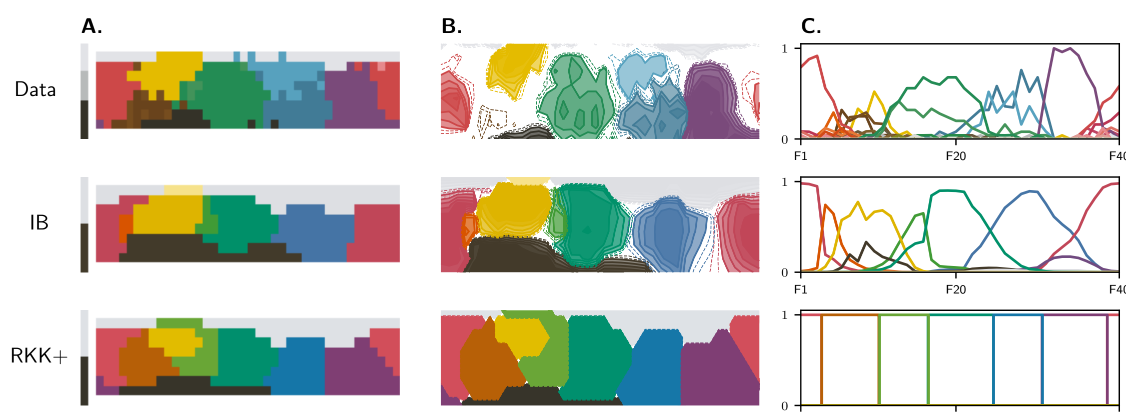


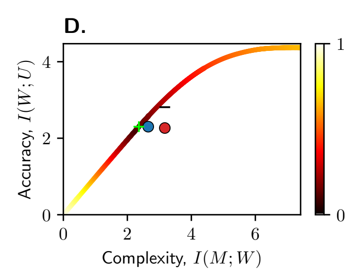


E.

	ε_l	gNID	NID
IB	0.17	0.14	0.34
RKK+	0.90	0.46	0.29

* **Chiquitano** - Bolivia

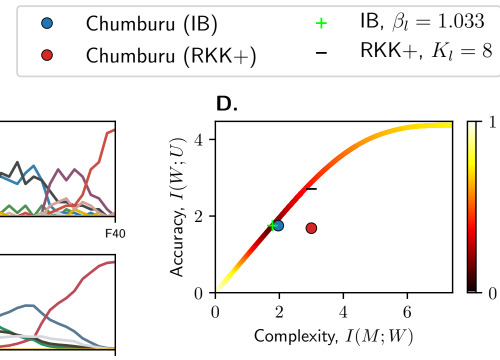


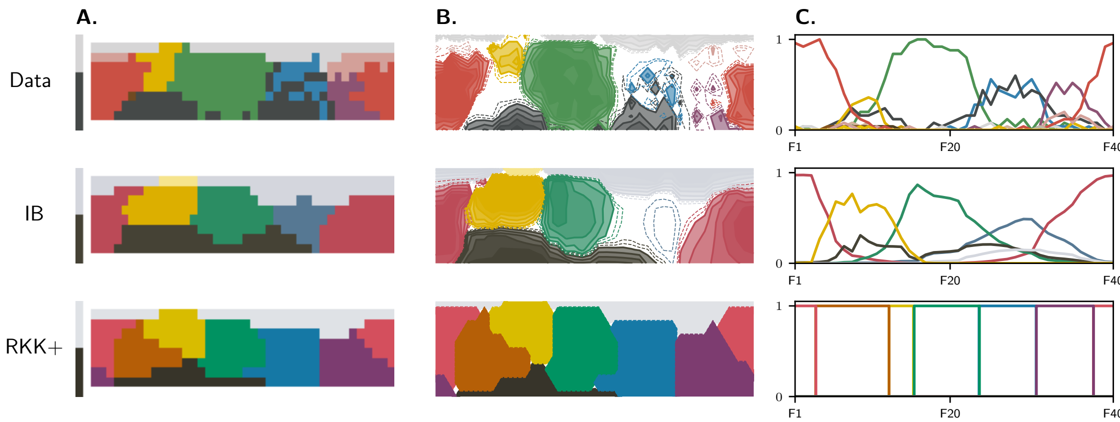


E.

	ε_l	gNID	NID
IB	0.28	0.11	0.27
RKK+	0.54	0.33	0.21

Chumburu - Ghana

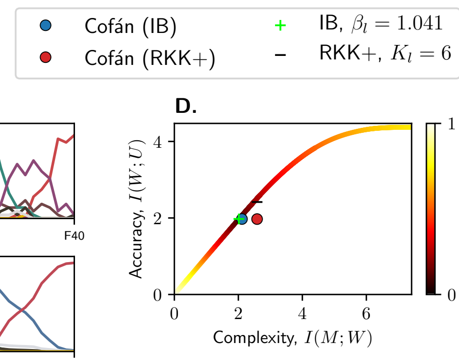


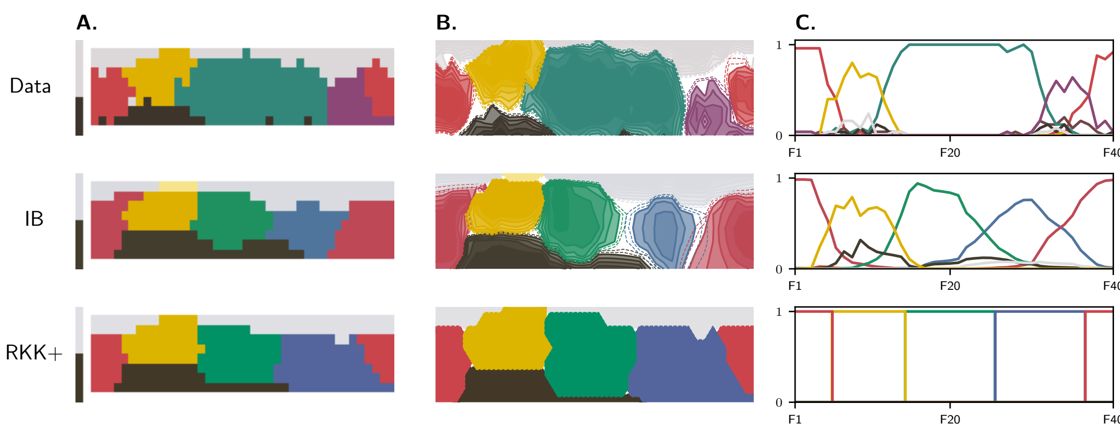


E.

	ε_l	gNID	NID
IB	0.18	0.15	0.27
RKK+	1.02	0.49	0.31

Cofán - Ecuador

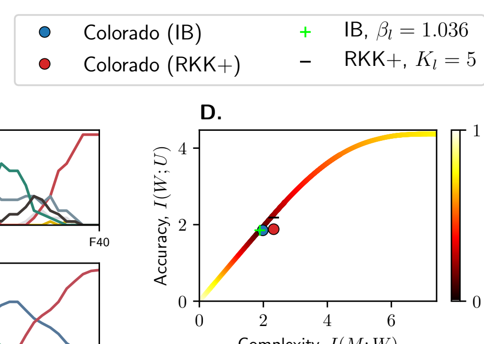


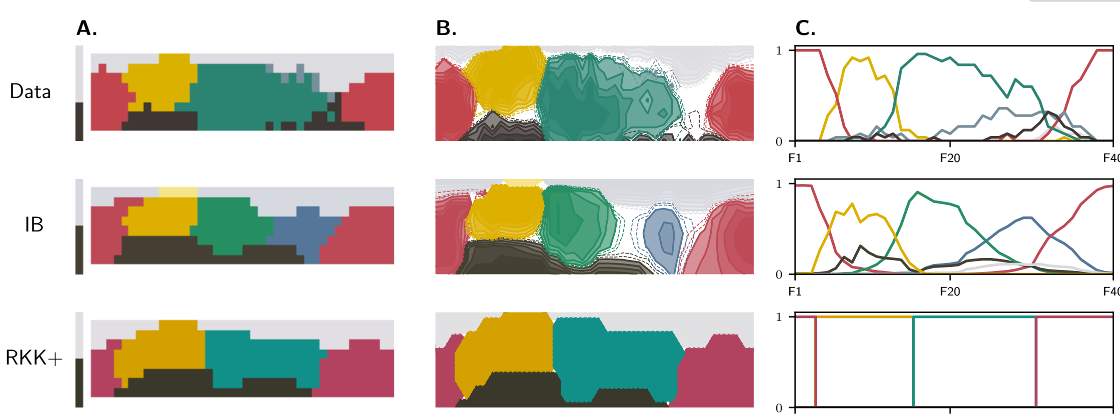


E.

	ε_l	gNID	NID
IB	0.10	0.10	0.25
RKK+	0.45	0.33	0.21

Colorado - Ecuador

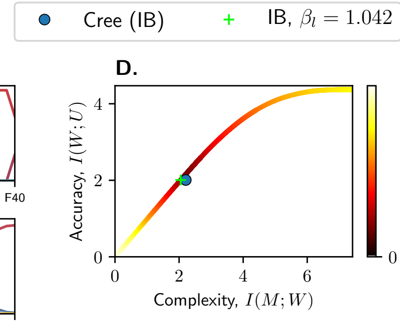


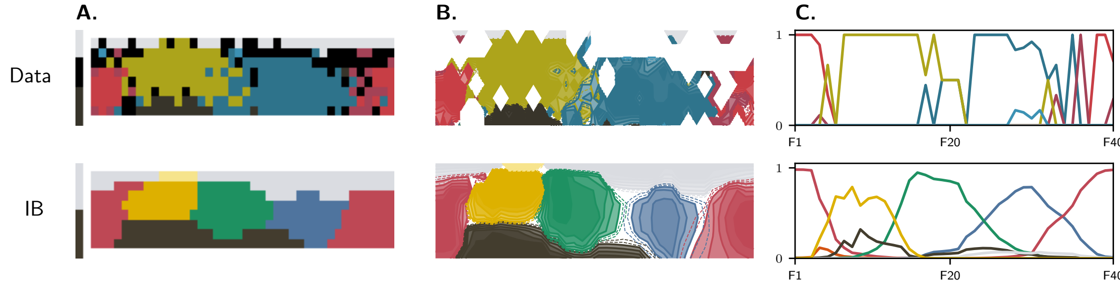


E.

	ε_l	gNID	NID
IB	0.09	0.11	0.27
RKK+	0.30	0.35	0.24

* **Cree** - Canada

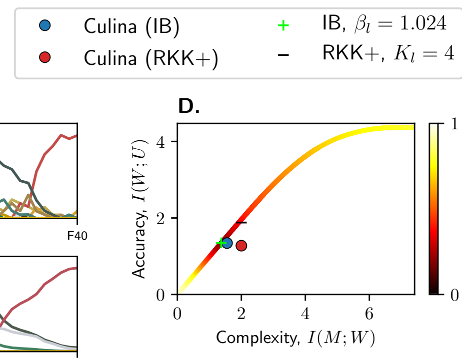


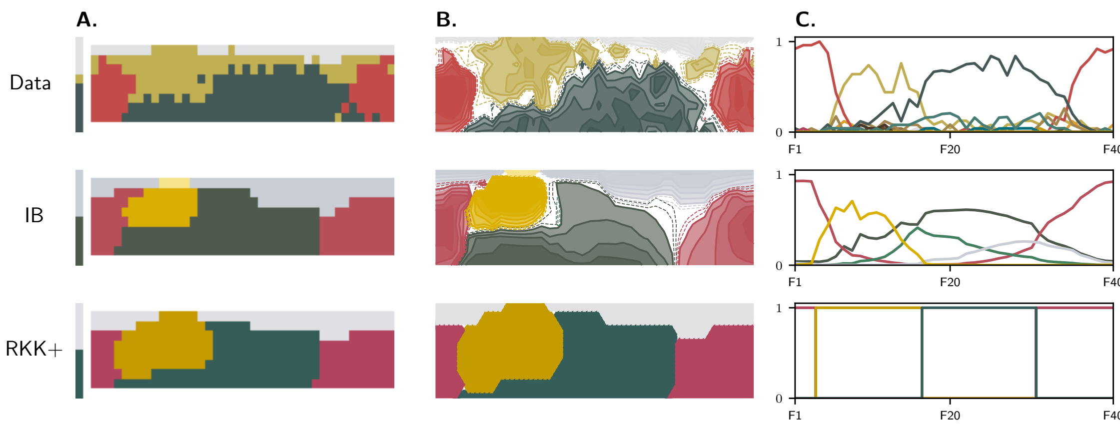


E.

	ε_l	gNID	NID
IB	0.15	0.22	0.37

Culina - Peru, Brazil

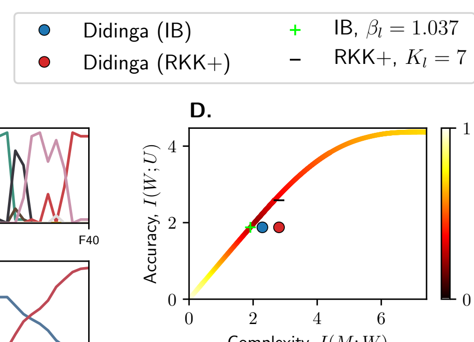


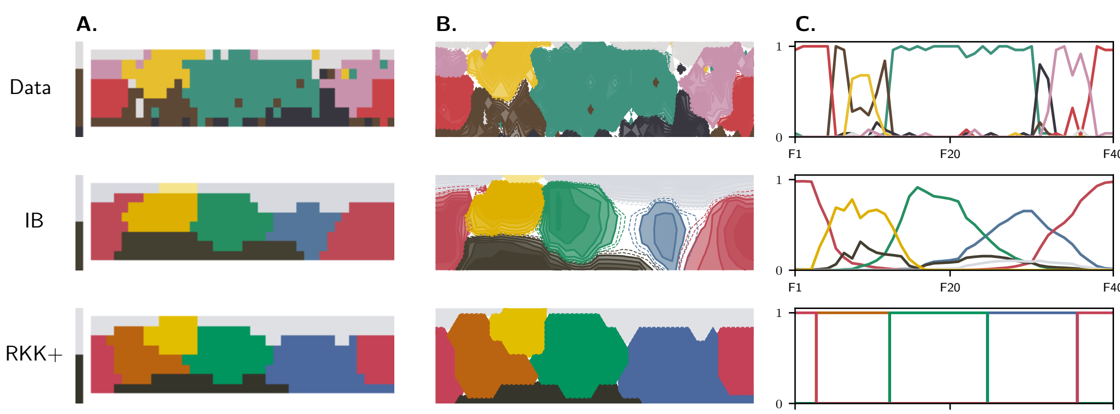


E.

	ε_l	gNID	NID
IB	0.18	0.19	0.31
RKK+	0.61	0.50	0.32

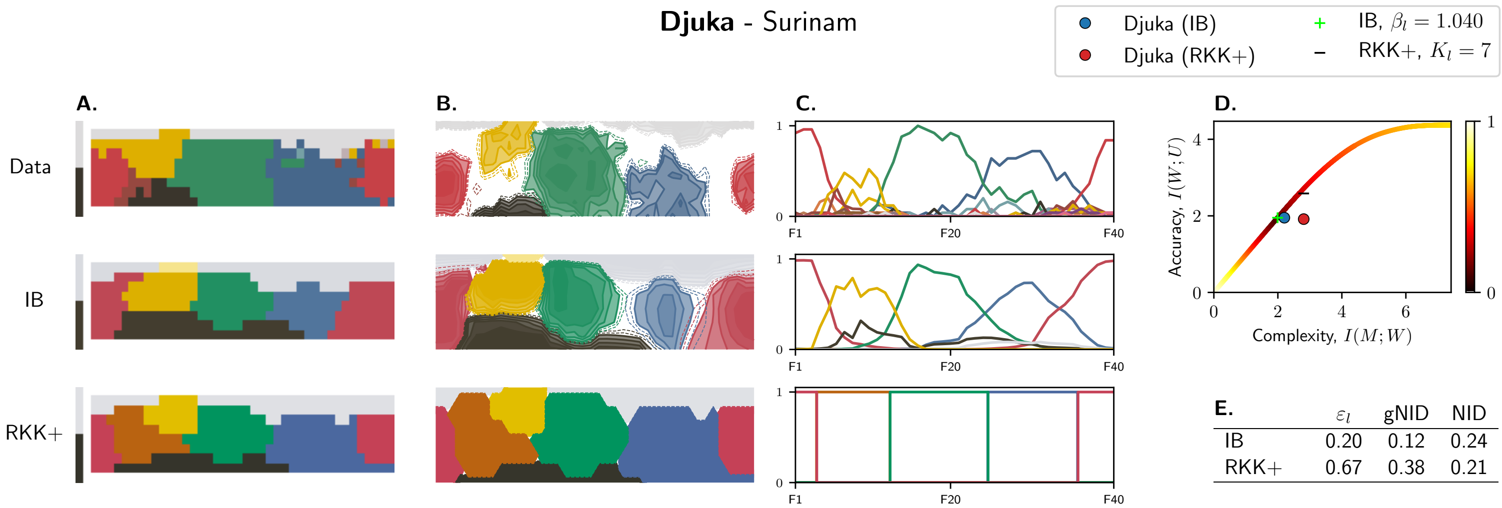
Didinga - Sudan



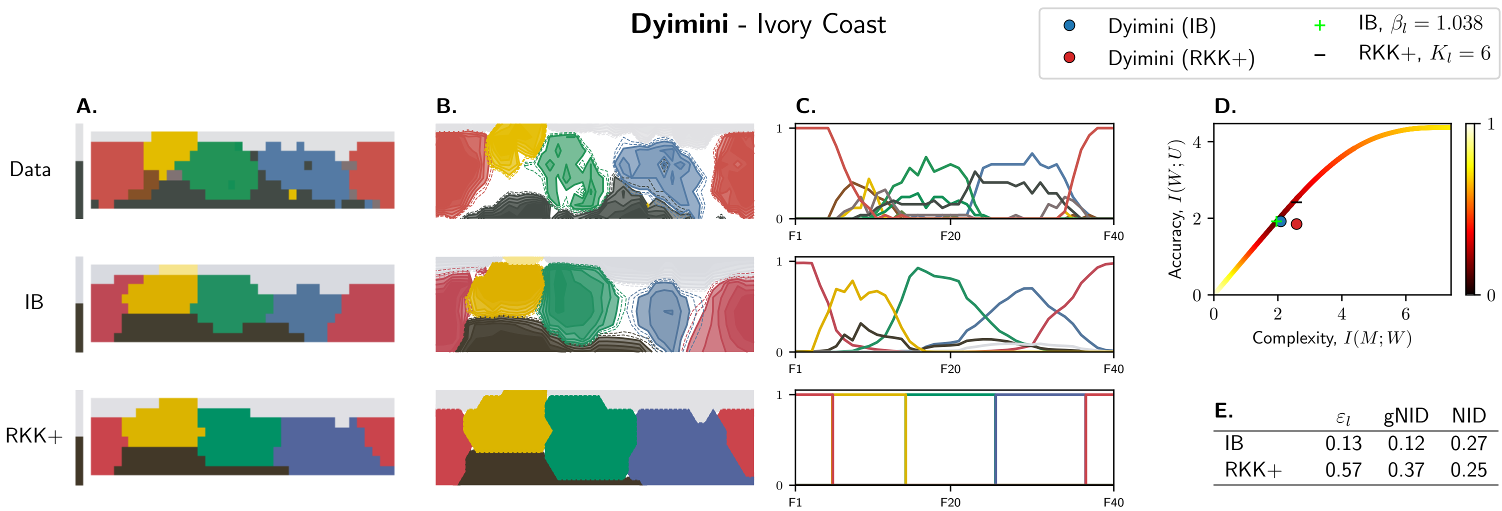


E.

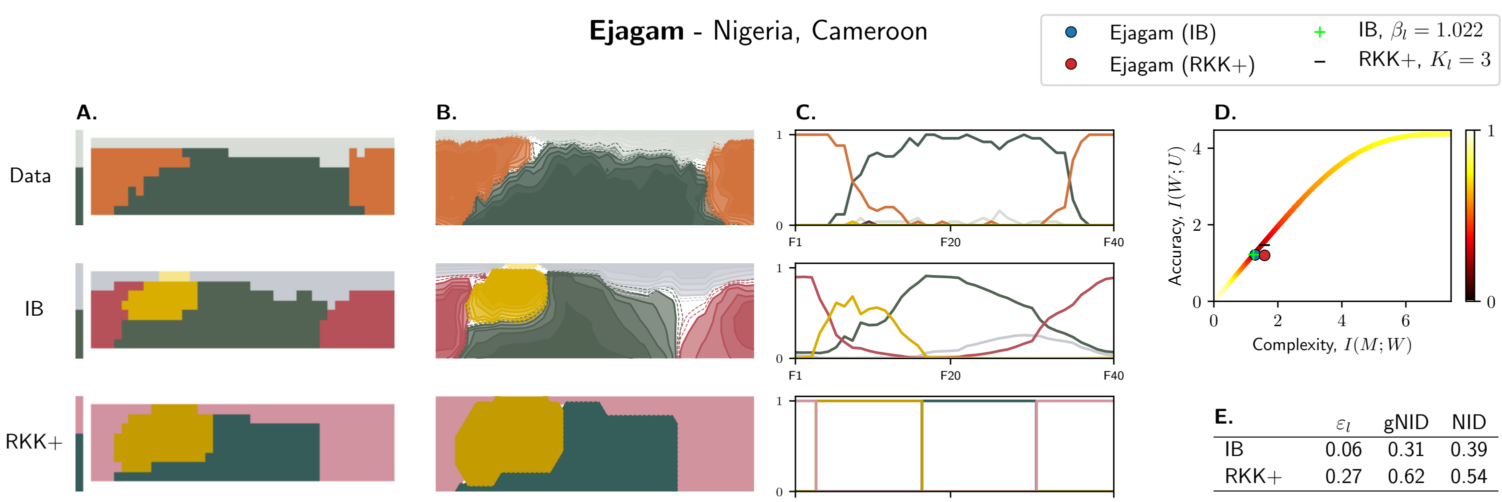
	ε_l	gNID	NID
IB	0.36	0.28	0.38
RKK+	0.71	0.45	0.37



	ε_l	gNID	NID
IB	0.20	0.12	0.24
RKK+	0.67	0.38	0.21

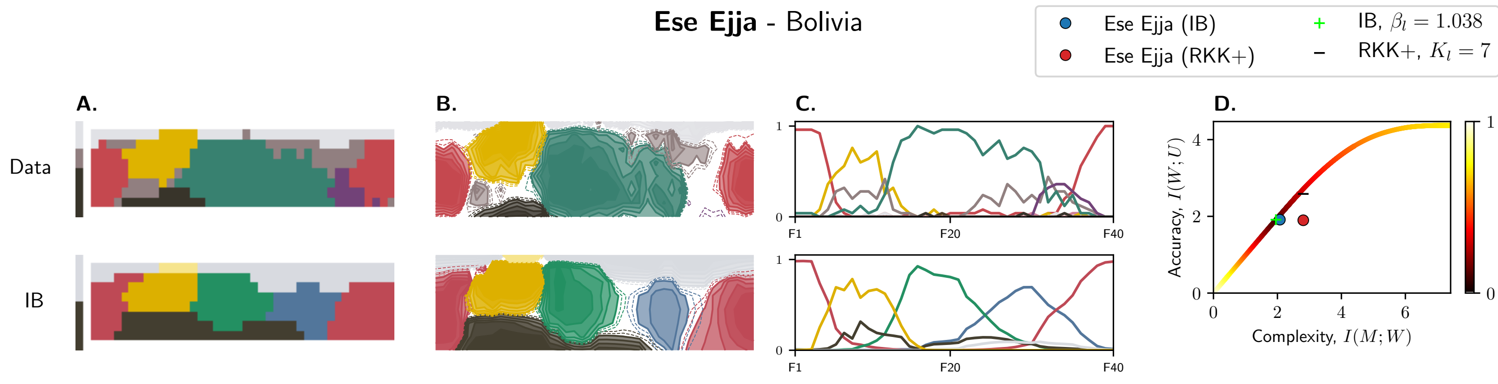


	ε_l	gNID	NID
IB	0.13	0.12	0.27
RKK+	0.57	0.37	0.25



	ε_l	gNID	NID
IB	0.06	0.31	0.39
RKK+	0.27	0.62	0.54

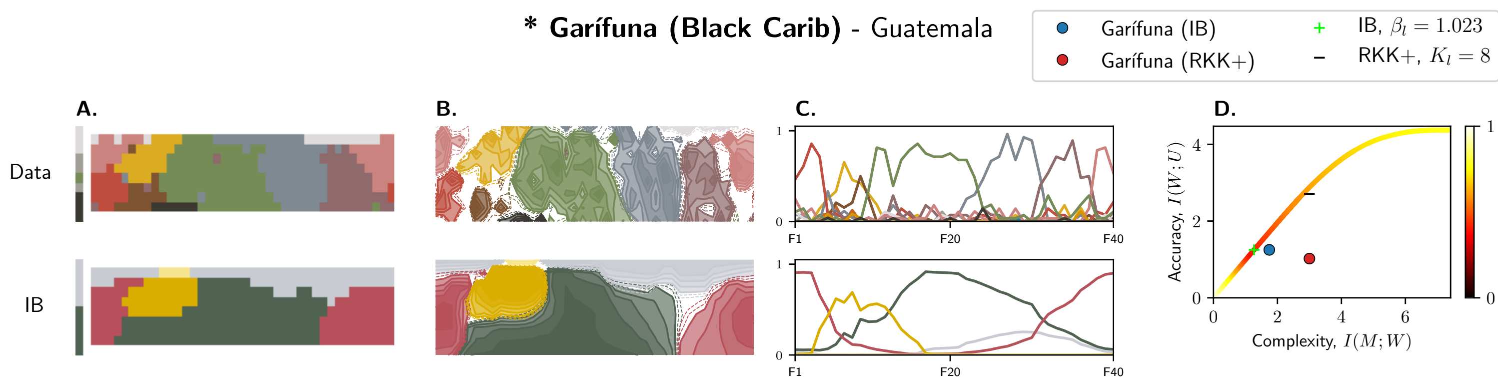
Ese Ejja - Bolivia



E.

	ε_l	gNID	NID
IB	0.12	0.08	0.24
RKK+	0.69	0.39	0.25

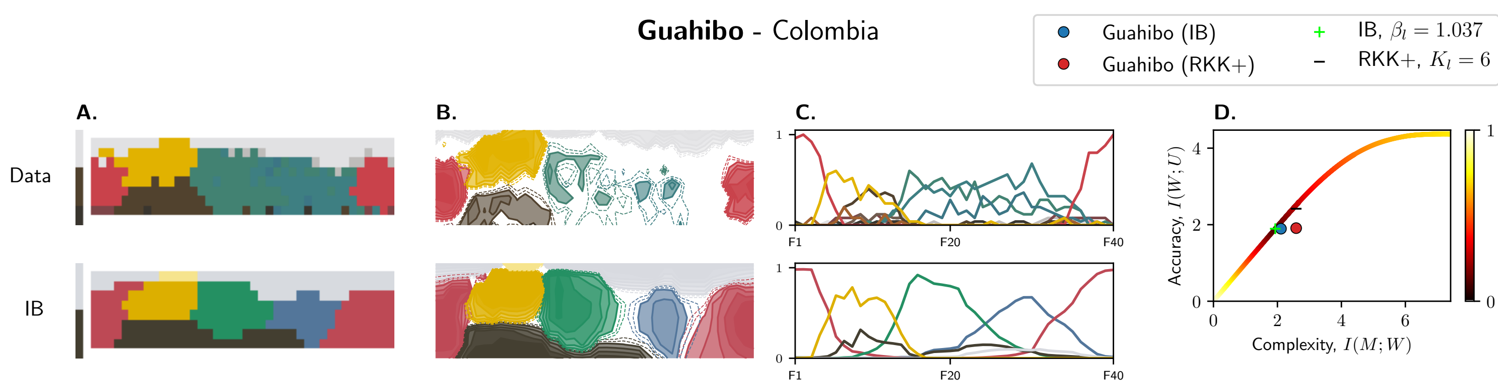
* **Garífuna (Black Carib)** - Guatemala



E.

	ε_l	gNID	NID
IB	0.47	0.41	0.54
RKK+	1.69	0.73	0.41

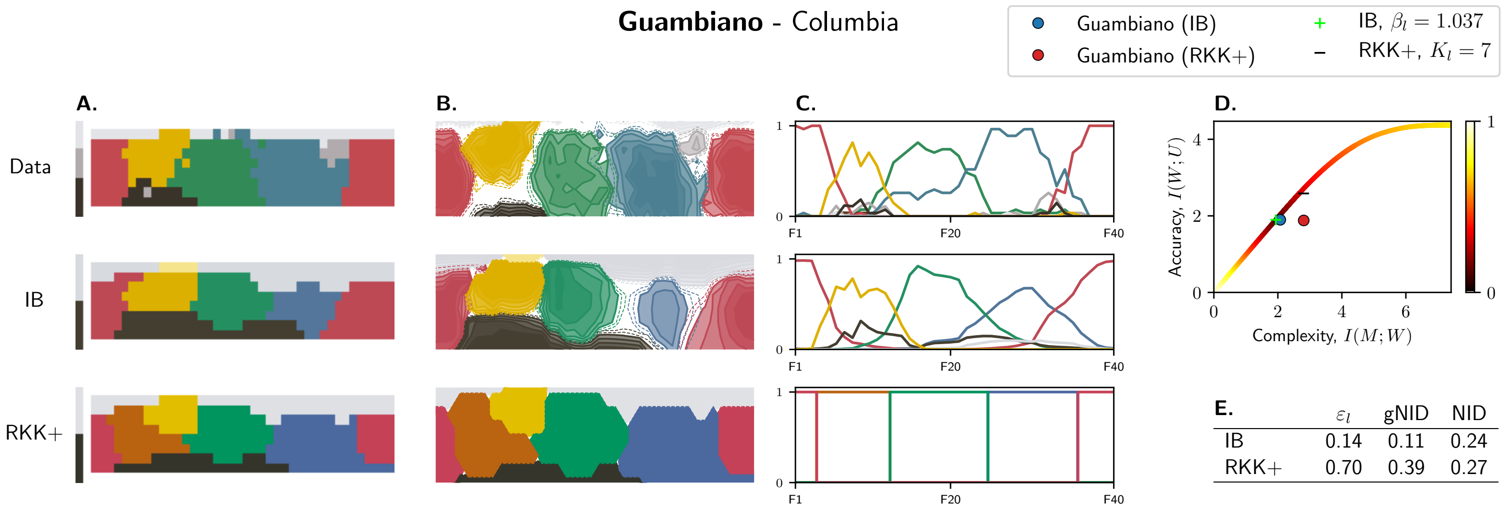
Guahibo - Colombia



E.

	ε_l	gNID	NID
IB	0.17	0.09	0.29
RKK+	0.51	0.32	0.19

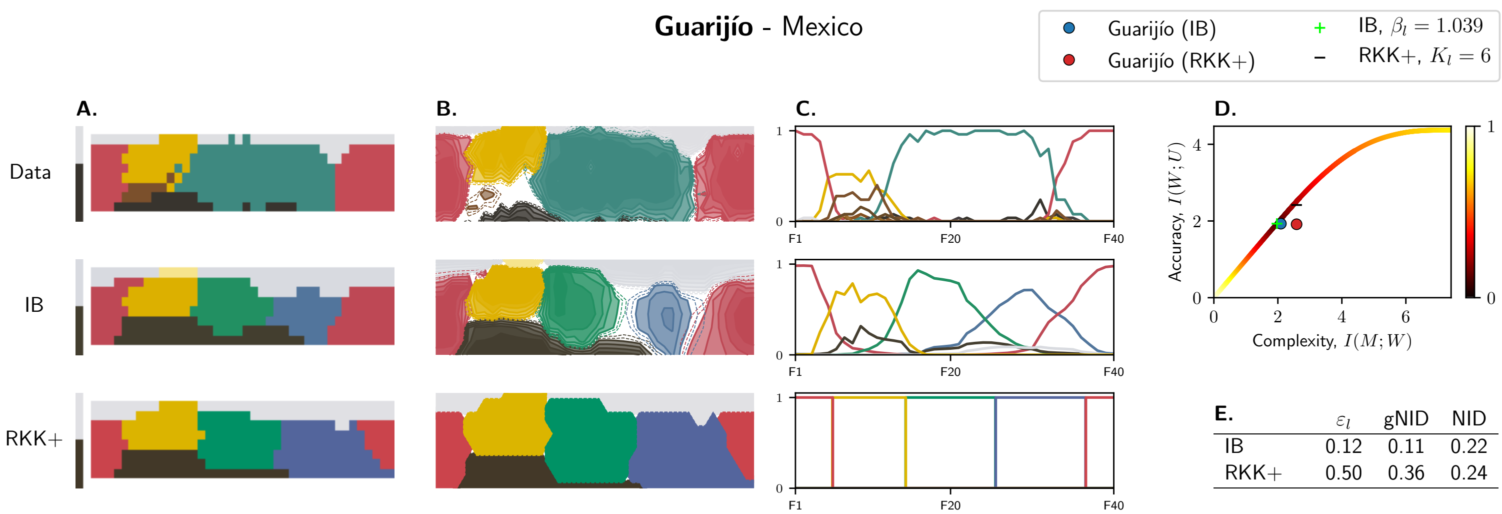
Guambiano - Columbia



E.

	ε_l	gNID	NID
IB	0.14	0.11	0.24
RKK+	0.70	0.39	0.27

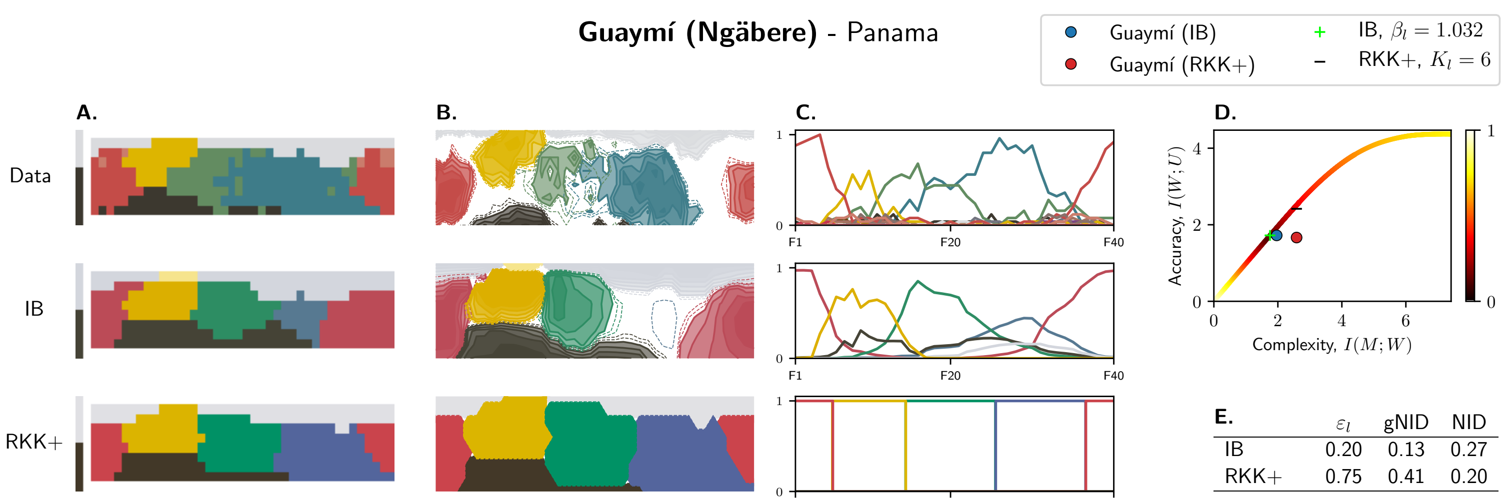
Guarijío - Mexico



E.

	ε_l	gNID	NID
IB	0.12	0.11	0.22
RKK+	0.50	0.36	0.24

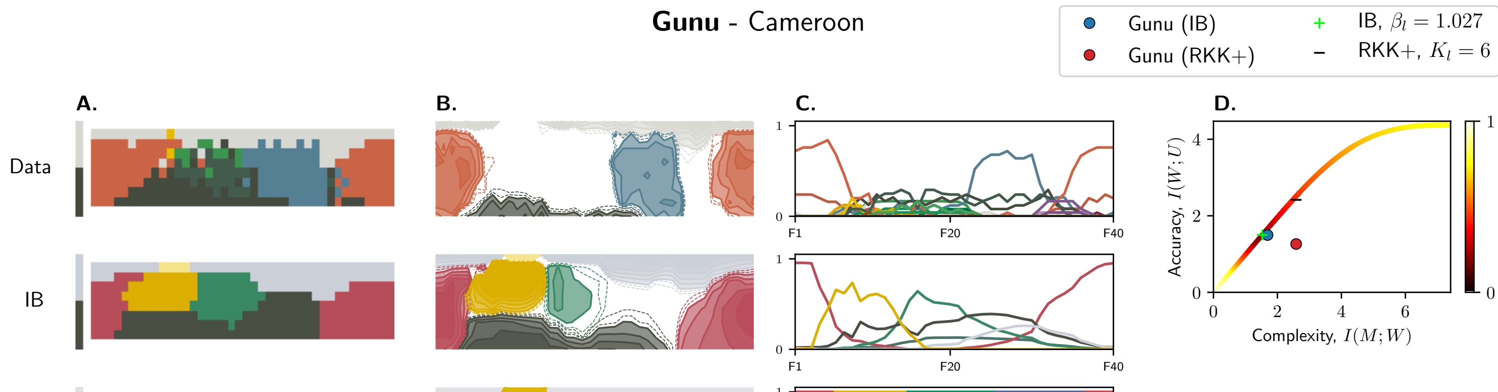
Guaymí (Ngäbere) - Panama



E.

	ε_l	gNID	NID
IB	0.20	0.13	0.27
RKK+	0.75	0.41	0.20

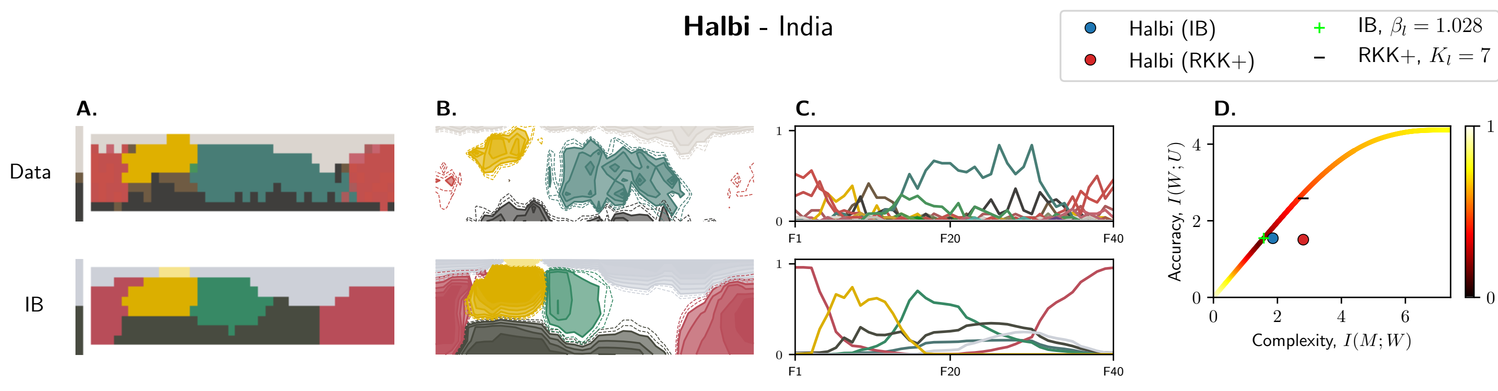
Gunu - Cameroon



E.

	ε_l	gNID	NID
IB	0.16	0.20	0.41
RKK+	1.15	0.58	0.41

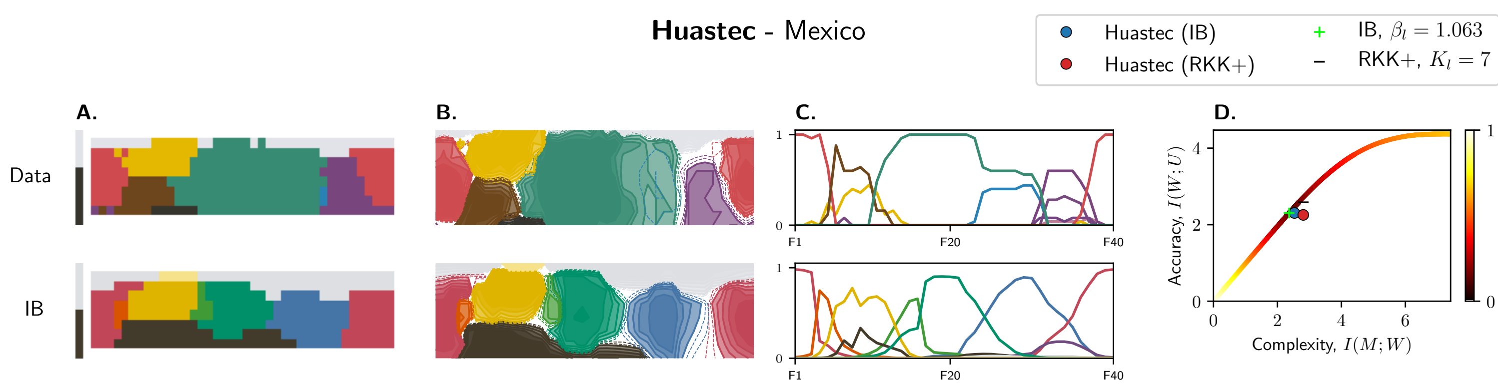
Halbi - India



E.

	ε_l	gNID	NID
IB	0.28	0.18	0.37
RKK+	1.08	0.51	0.28

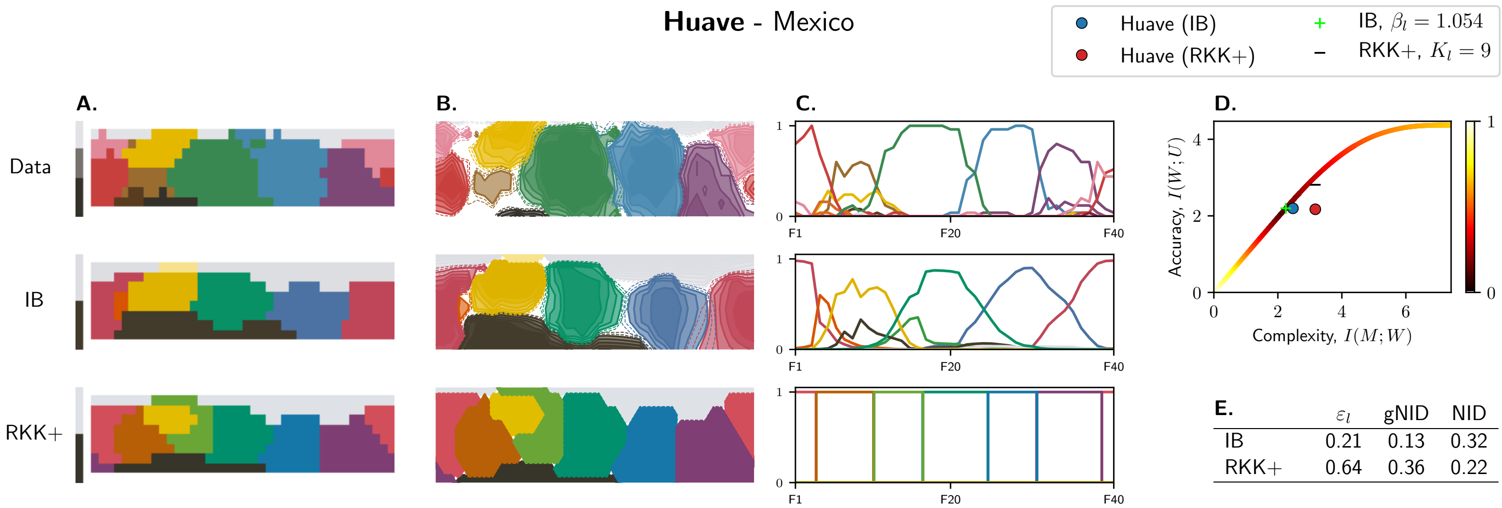
Huastec - Mexico



E.

	ε_l	gNID	NID
IB	0.16	0.15	0.28
RKK+	0.33	0.31	0.27

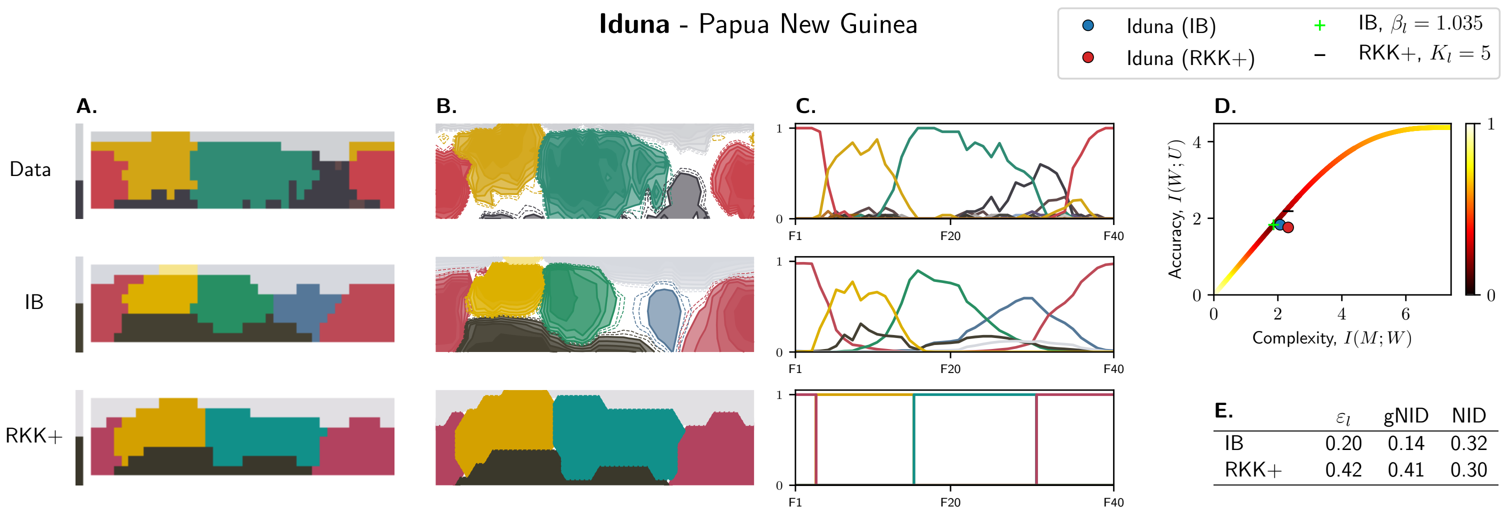
Huave - Mexico



E.

	ε_l	gNID	NID
IB	0.21	0.13	0.32
RKK+	0.64	0.36	0.22

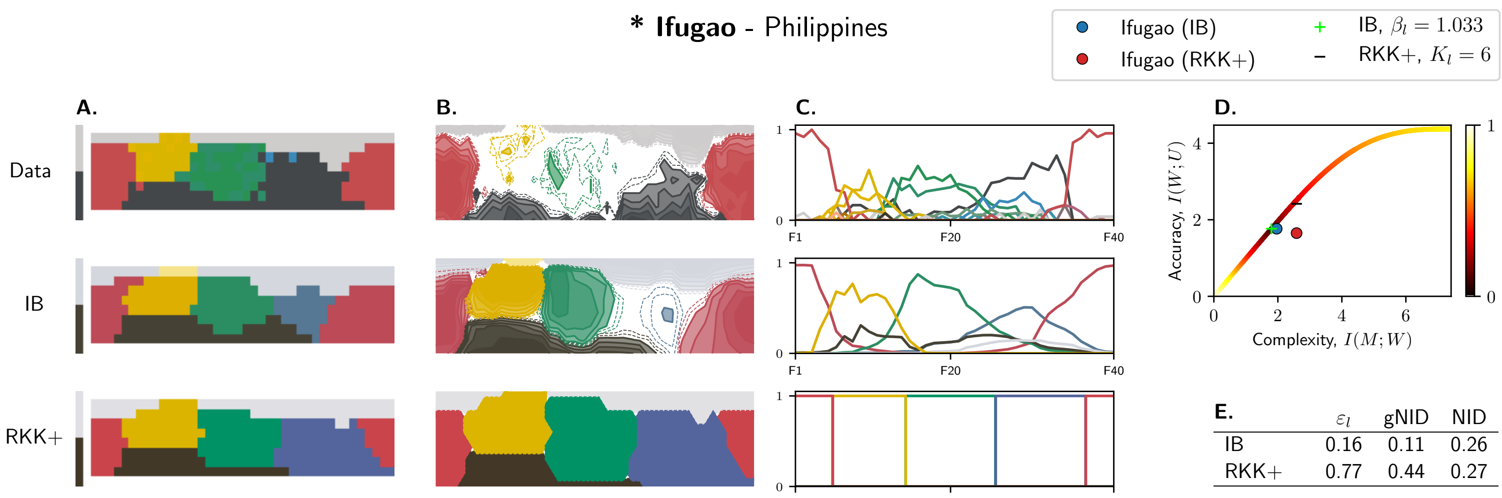
Iduna - Papua New Guinea



E.

	ε_l	gNID	NID
IB	0.20	0.14	0.32
RKK+	0.42	0.41	0.30

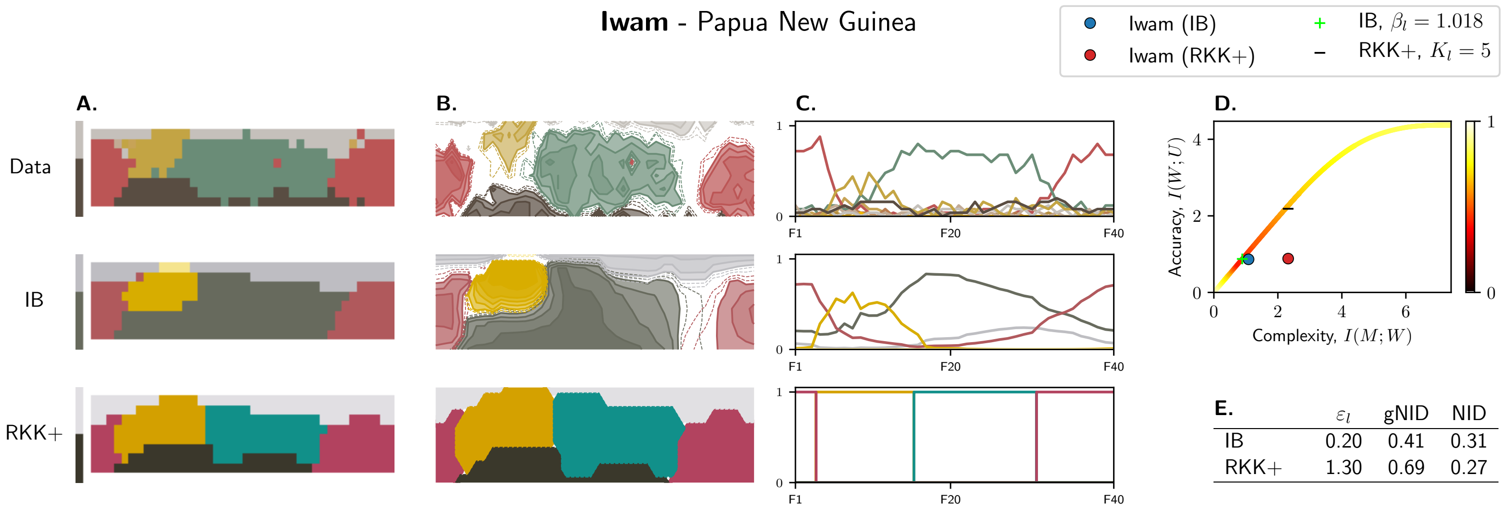
* **Ifugao** - Philippines



E.

	ε_l	gNID	NID
IB	0.16	0.11	0.26
RKK+	0.77	0.44	0.27

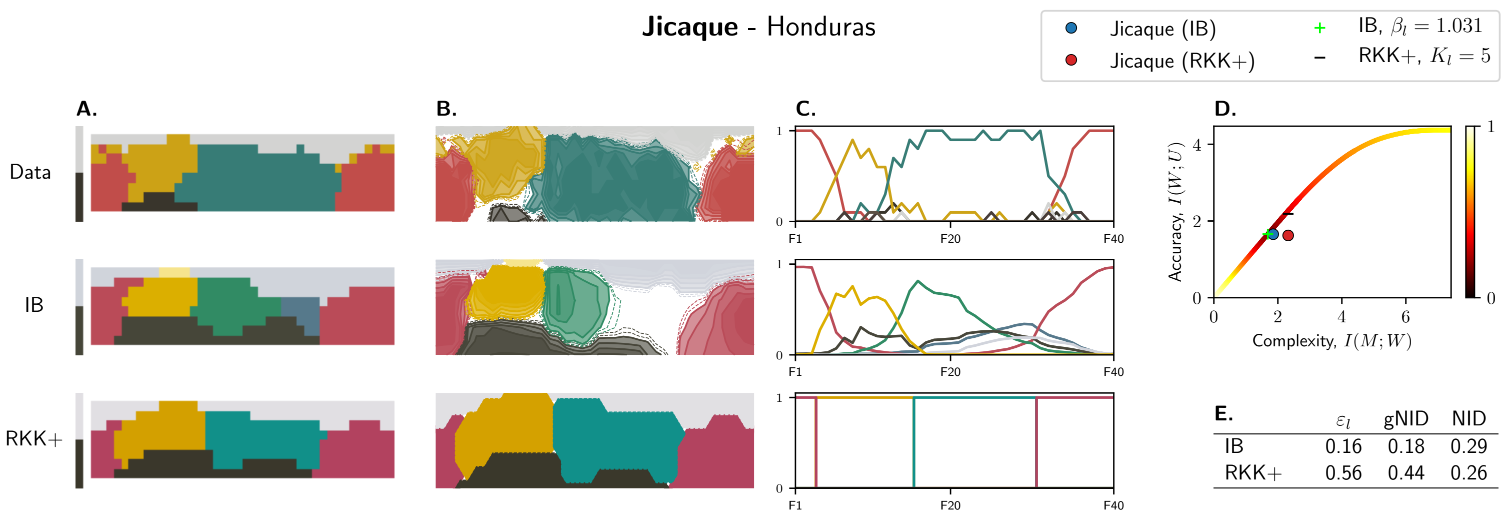
Iwam - Papua New Guinea



E.

	ε_l	gNID	NID
IB	0.20	0.41	0.31
RKK+	1.30	0.69	0.27

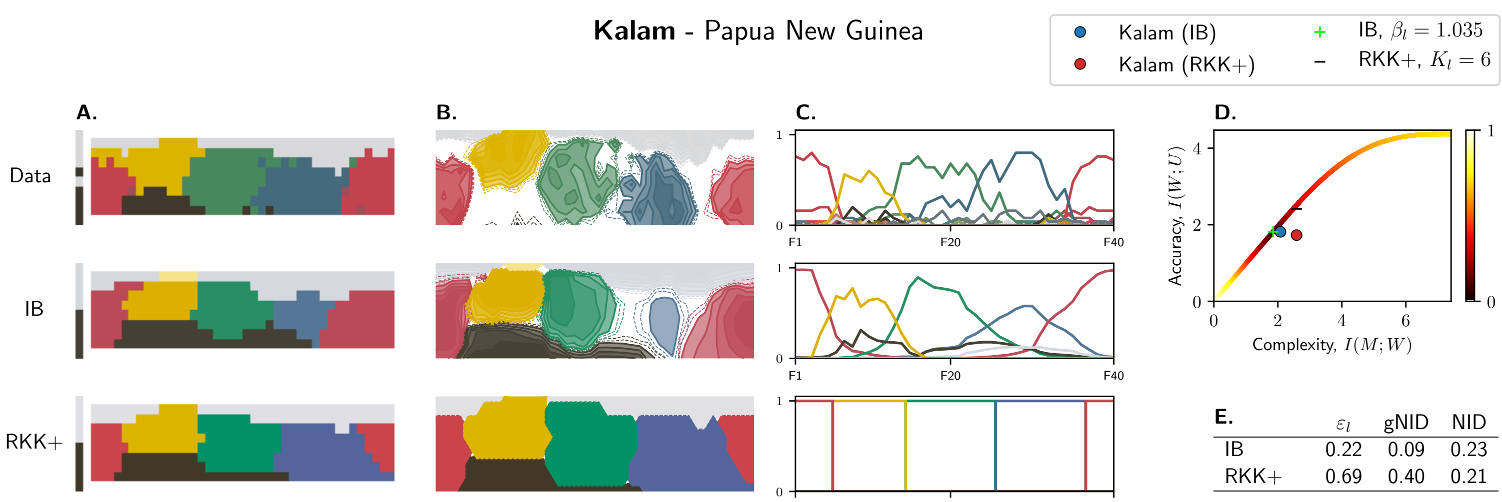
Jicaque - Honduras



E.

	ε_l	gNID	NID
IB	0.16	0.18	0.29
RKK+	0.56	0.44	0.26

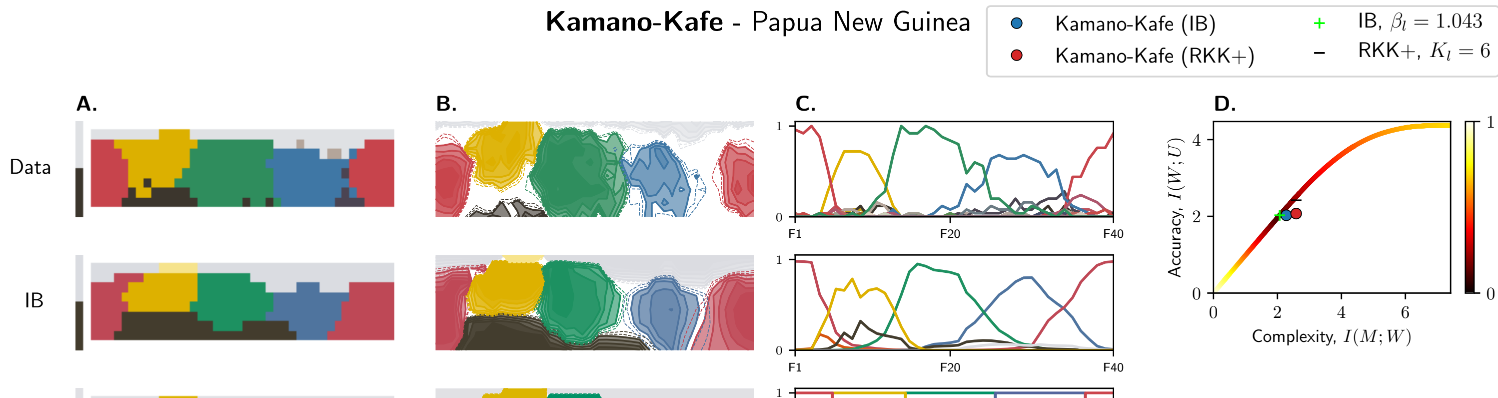
Kalam - Papua New Guinea



E.

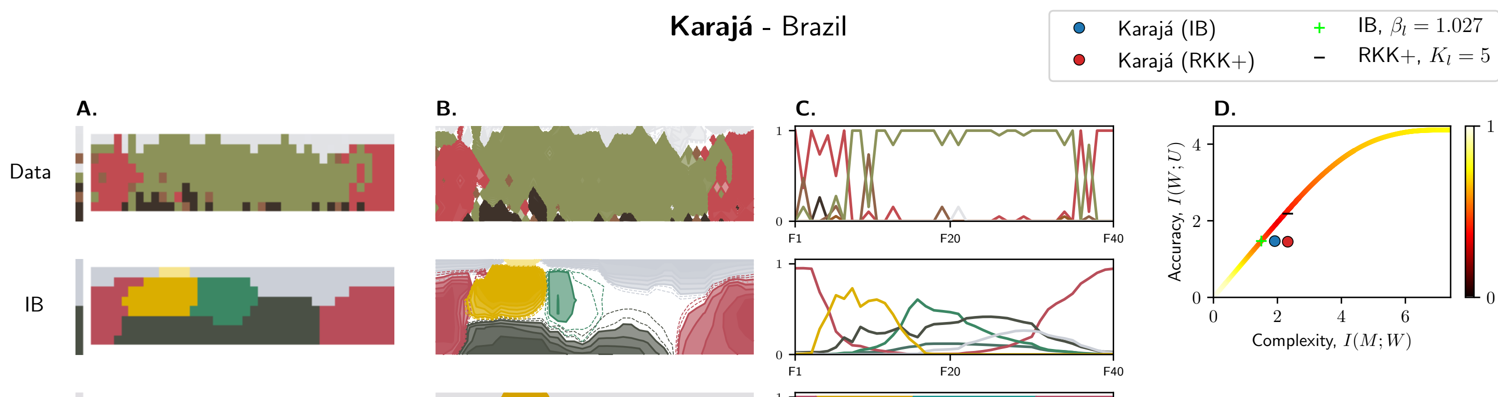
	ε_l	gNID	NID
IB	0.22	0.09	0.23
RKK+	0.69	0.40	0.21

Kamano-Kafe - Papua New Guinea



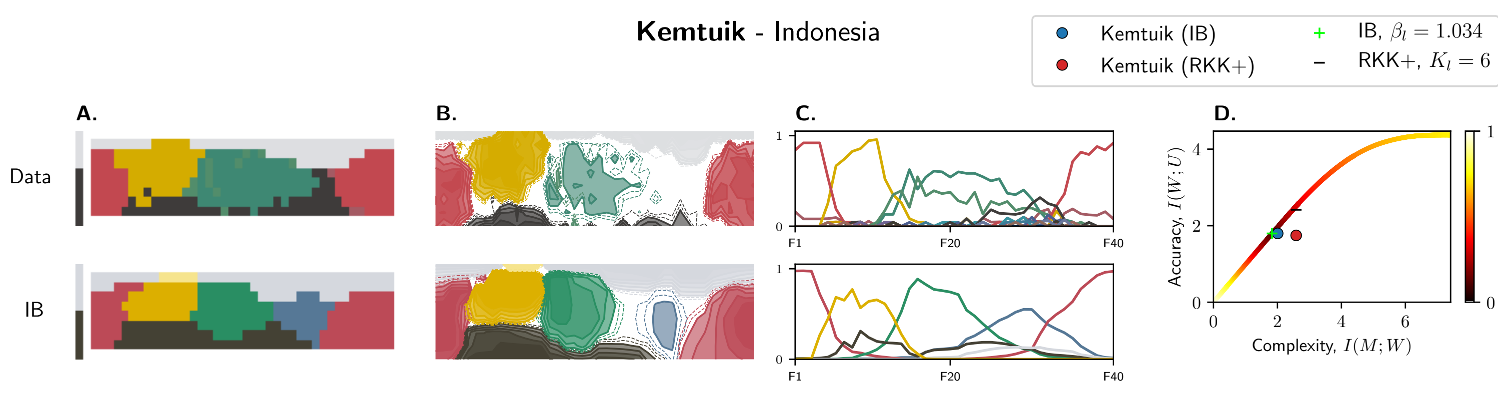
	ε_l	gNID	NID
IB	0.20	0.10	0.18
RKK+	0.34	0.28	0.17

Karajá - Brazil



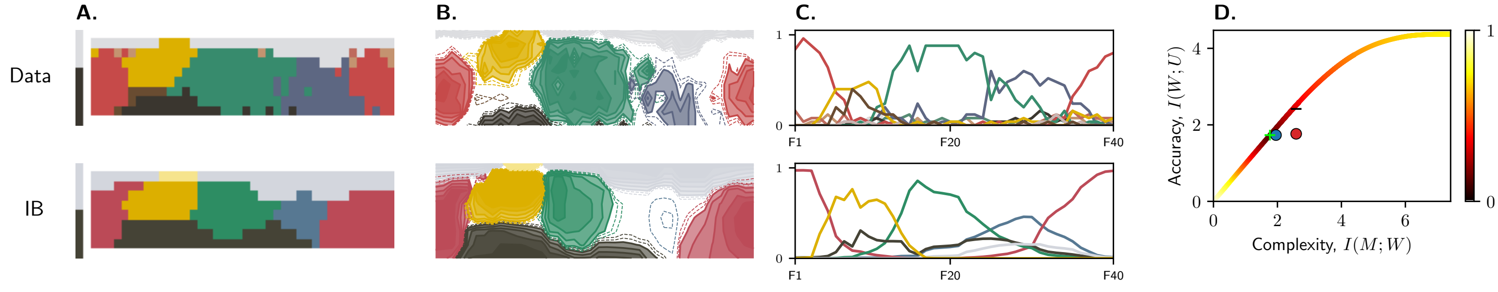
	ε_l	gNID	NID
IB	0.41	0.51	0.50
RKK+	0.73	0.53	0.49

Kemtuik - Indonesia



	ε_l	gNID	NID
IB	0.17	0.12	0.26
RKK+	0.67	0.40	0.28

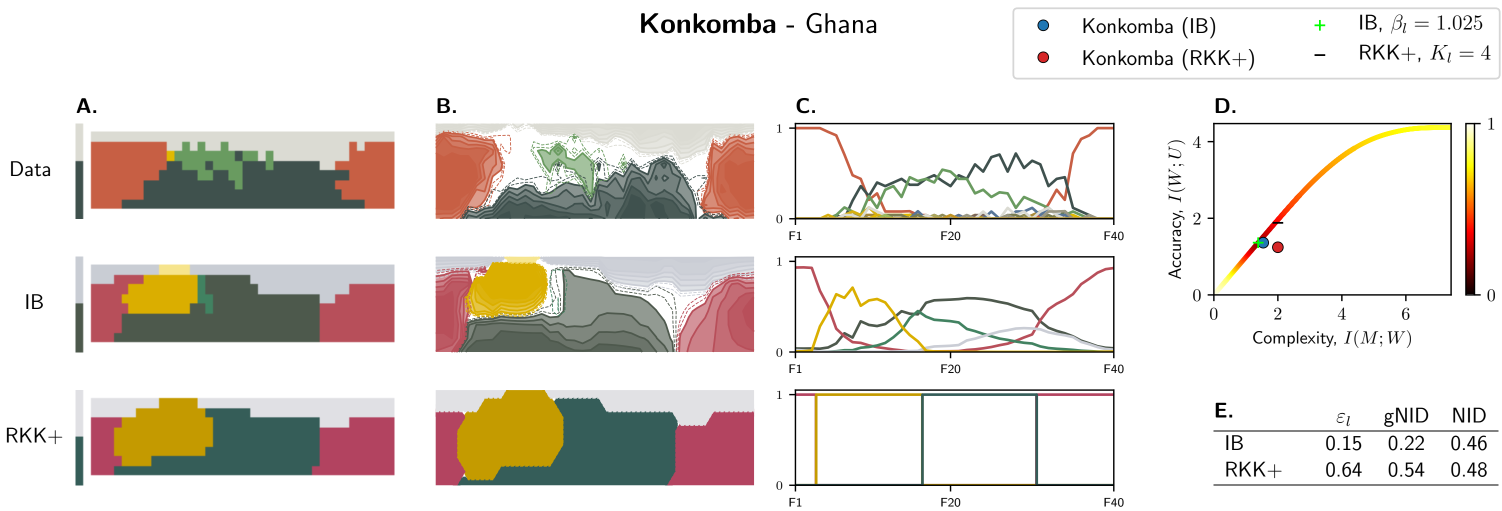
Kokni (Kokoni) - India



E.

	ε_l	gNID	NID
IB	0.18	0.11	0.23
RKK+	0.65	0.39	0.19

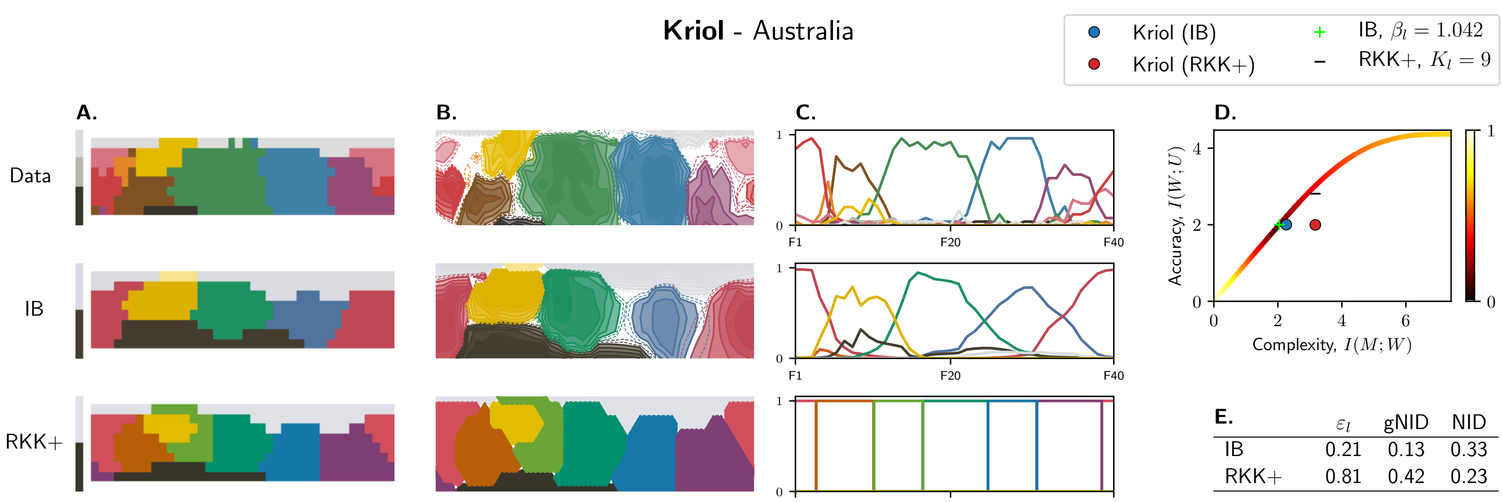
Konkomba - Ghana



E.

	ε_l	gNID	NID
IB	0.15	0.22	0.46
RKK+	0.64	0.54	0.48

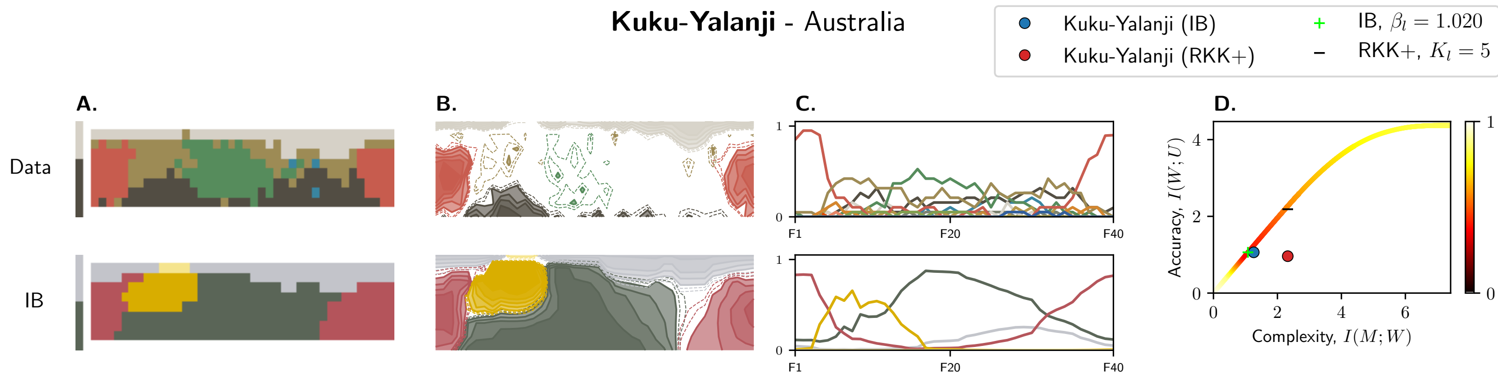
Kriol - Australia



E.

	ε_l	gNID	NID
IB	0.21	0.13	0.33
RKK+	0.81	0.42	0.23

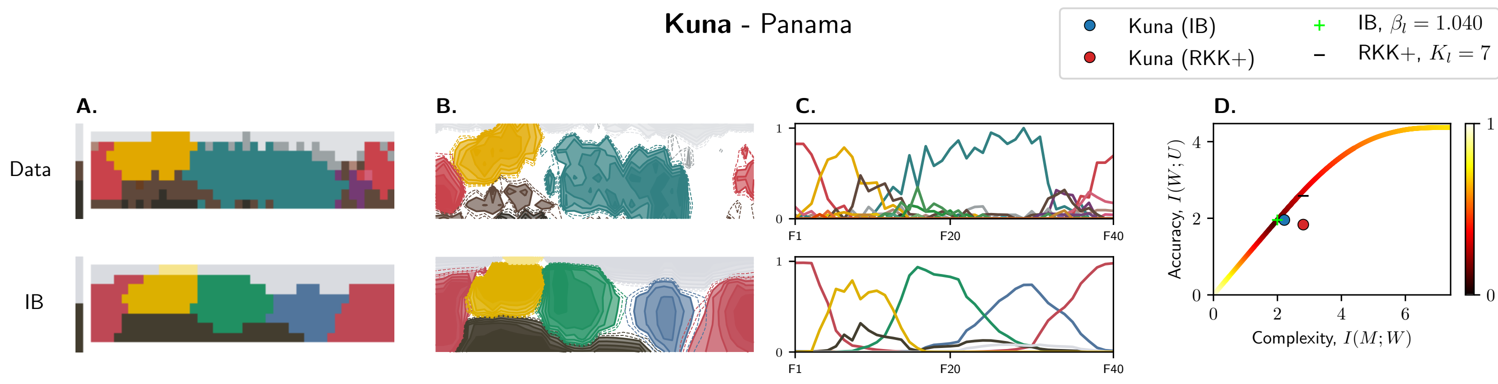
Kuku-Yalanji - Australia



E.

	ε_l	gNID	NID
IB	0.18	0.37	0.40
RKK+	1.22	0.67	0.37

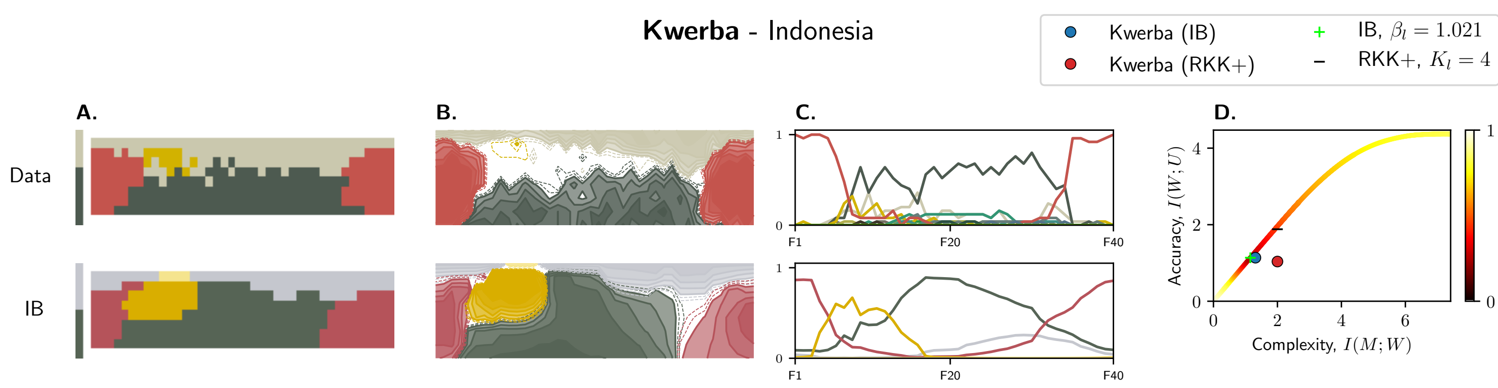
Kuna - Panama



E.

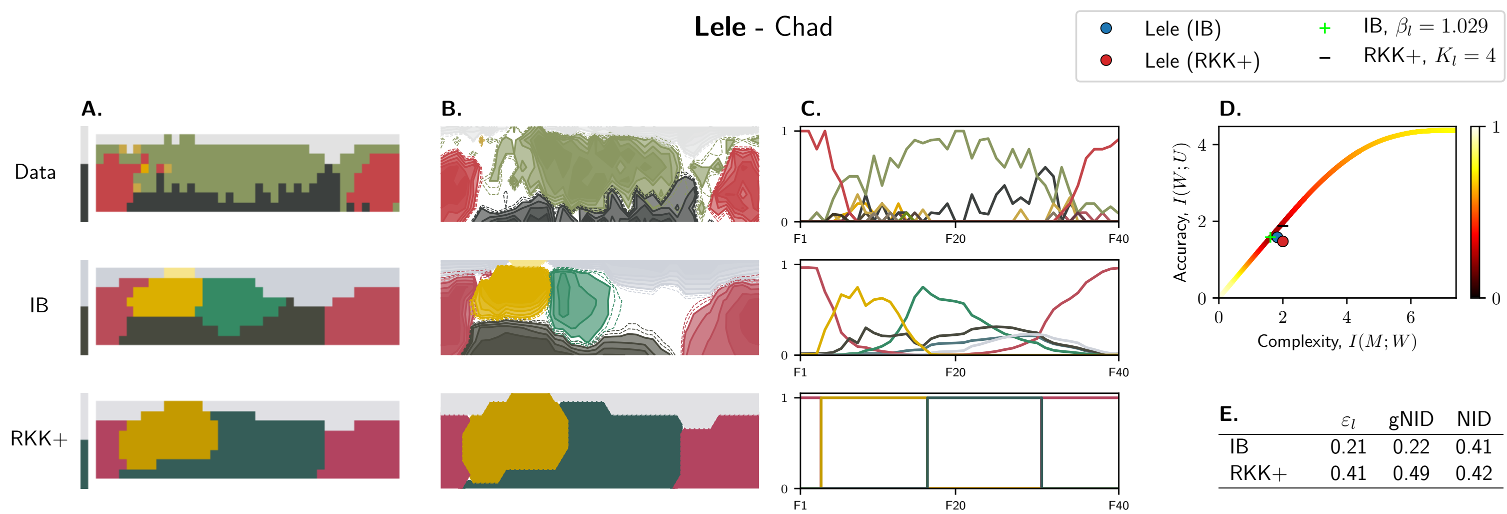
	ε_l	gNID	NID
IB	0.21	0.12	0.28
RKK+	0.75	0.42	0.27

Kwerba - Indonesia

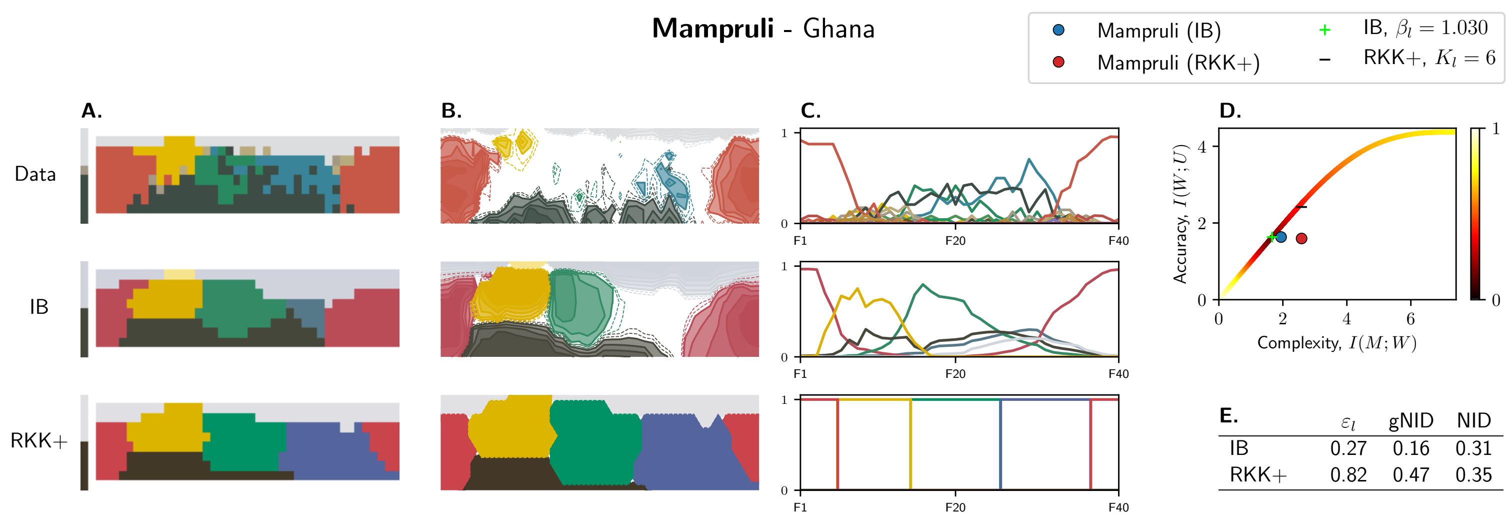


E.

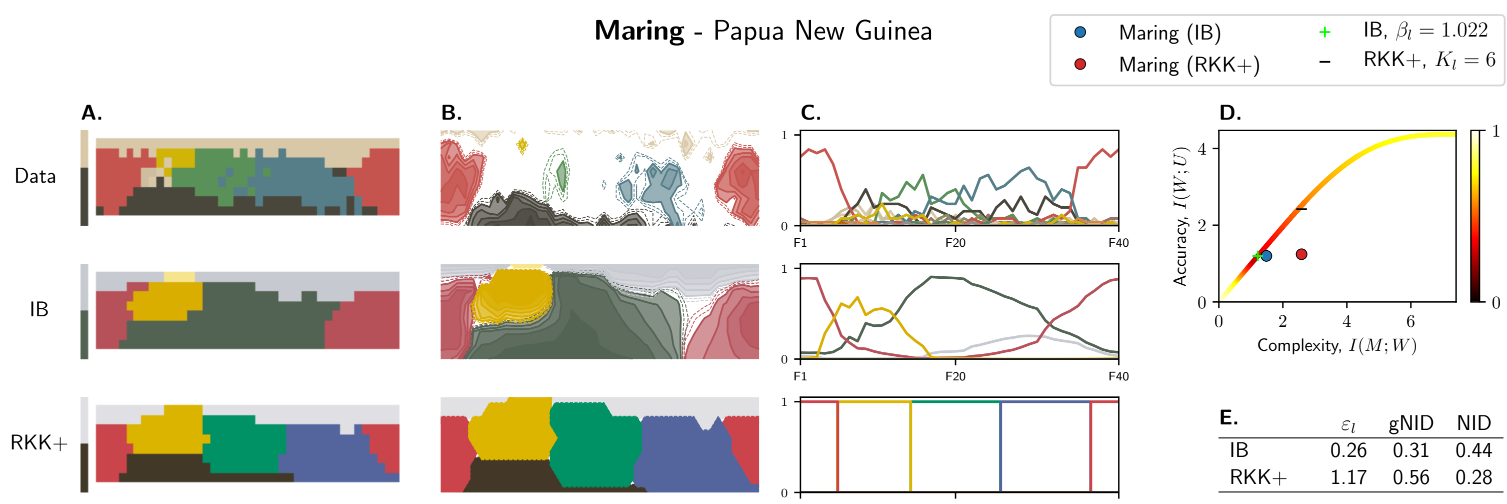
	ε_l	gNID	NID
IB	0.15	0.25	0.34
RKK+	0.84	0.61	0.44



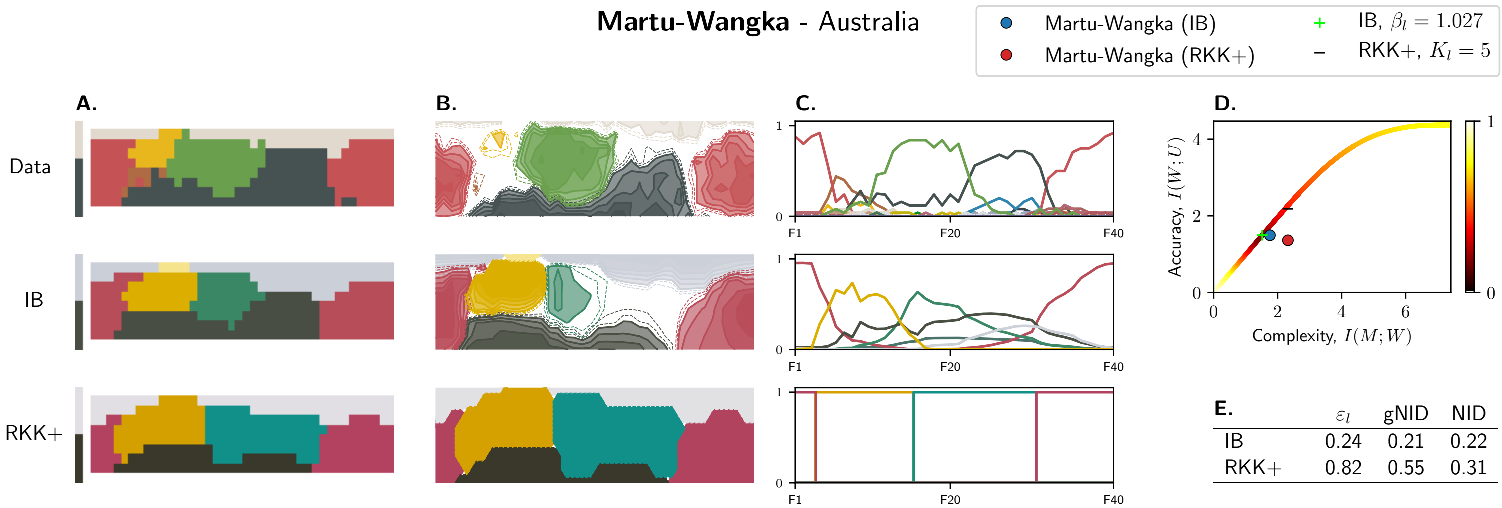
	ε_l	gNID	NID
IB	0.21	0.22	0.41
RKK+	0.41	0.49	0.42



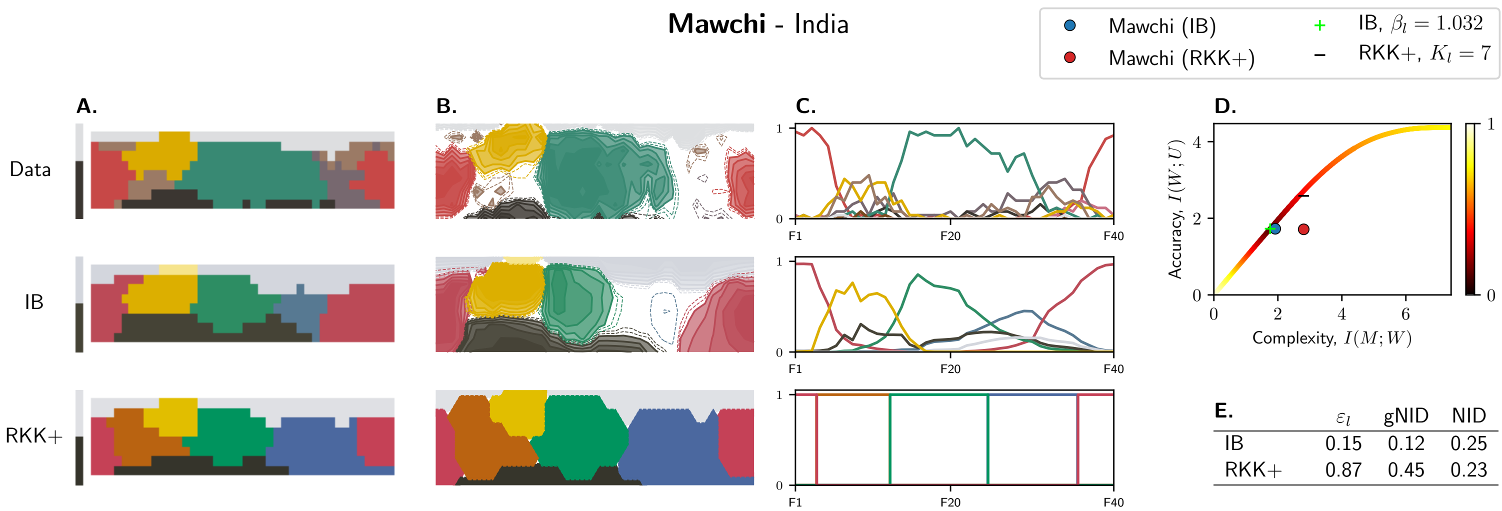
	ε_l	gNID	NID
IB	0.27	0.16	0.31
RKK+	0.82	0.47	0.35



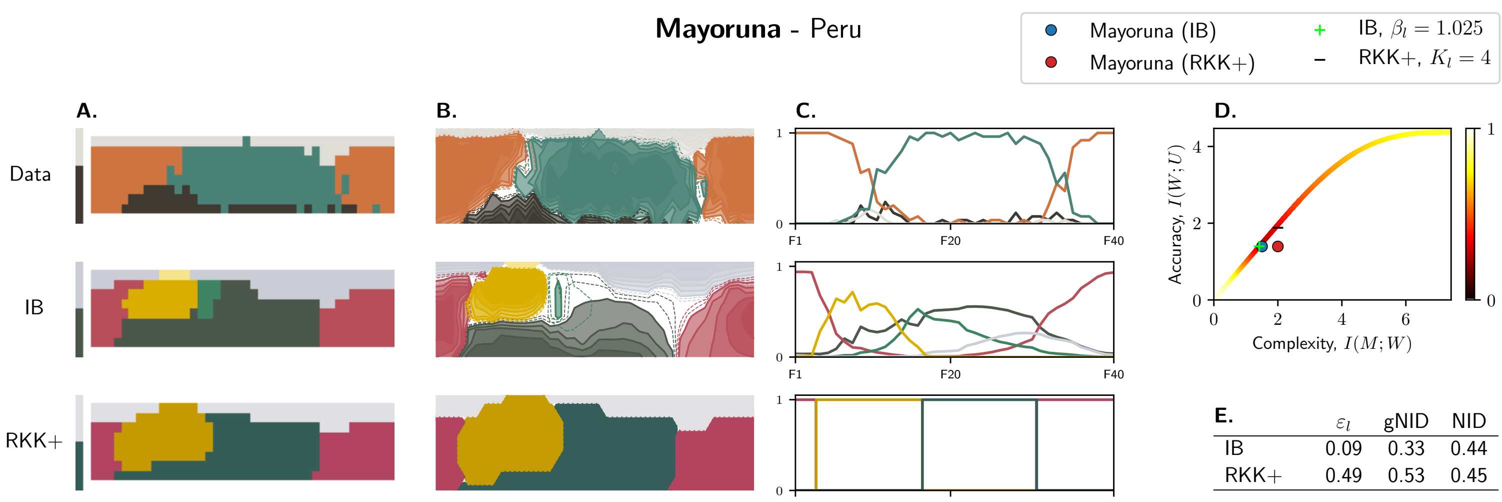
	ε_l	gNID	NID
IB	0.26	0.31	0.44
RKK+	1.17	0.56	0.28



	ε_l	gNID	NID
IB	0.24	0.21	0.22
RKK+	0.82	0.55	0.31

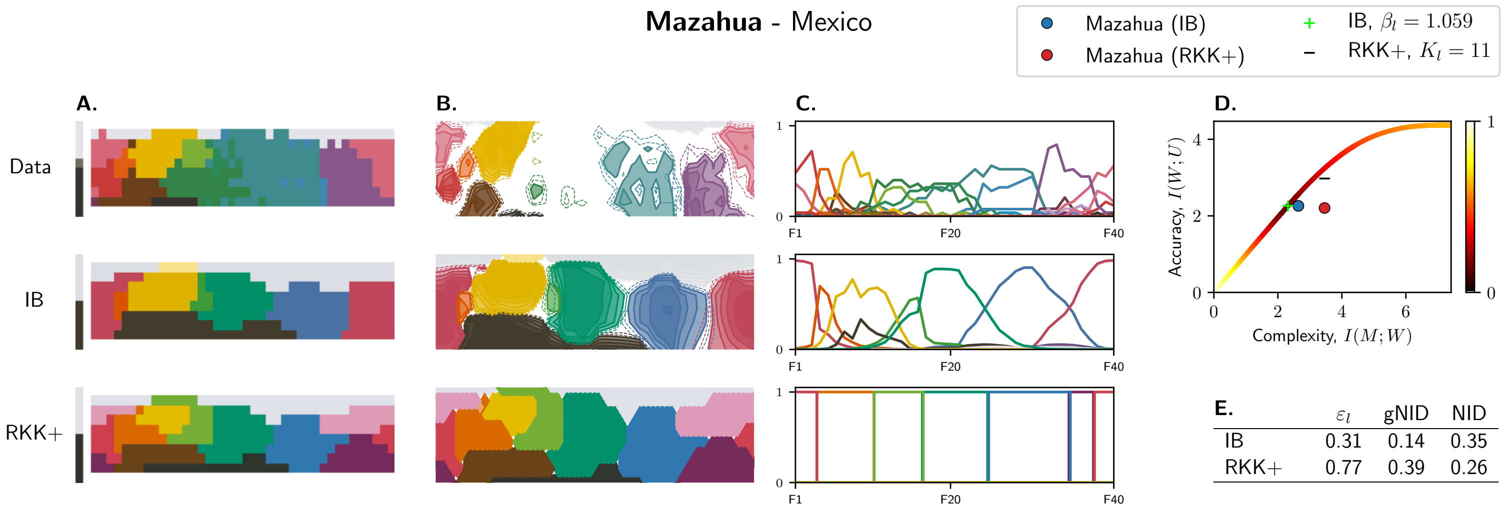


	ε_l	gNID	NID
IB	0.15	0.12	0.25
RKK+	0.87	0.45	0.23



	ε_l	gNID	NID
IB	0.09	0.33	0.44
RKK+	0.49	0.53	0.45

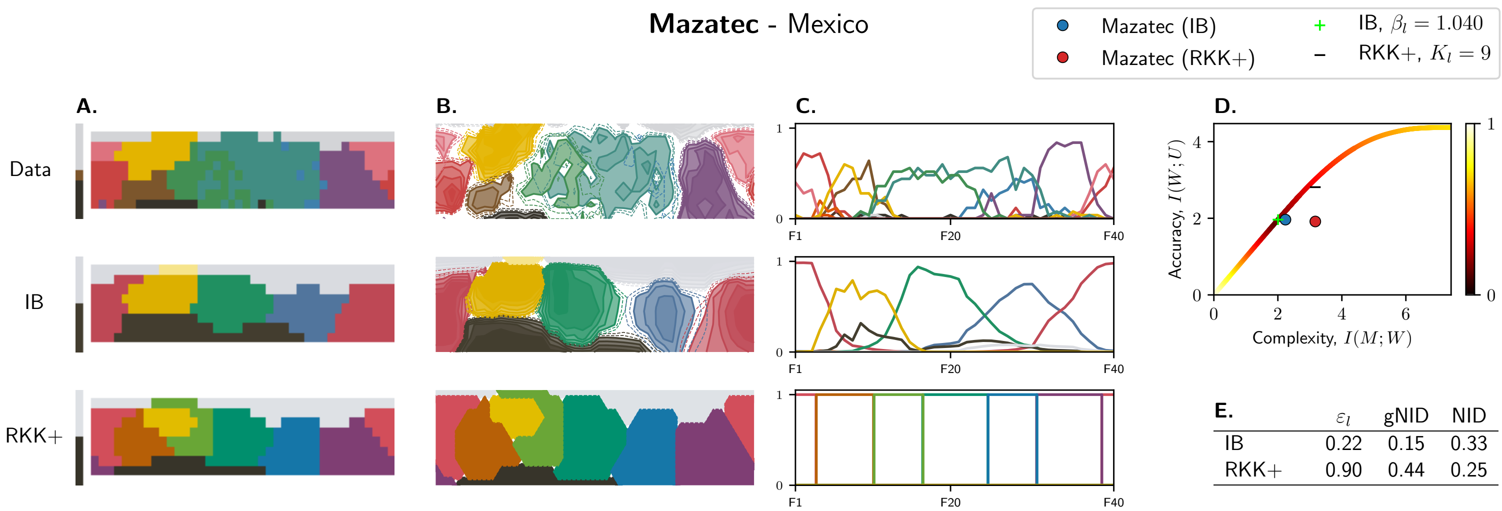
Mazahua - Mexico



E.

	ε_l	gNID	NID
IB	0.31	0.14	0.35
RKK+	0.77	0.39	0.26

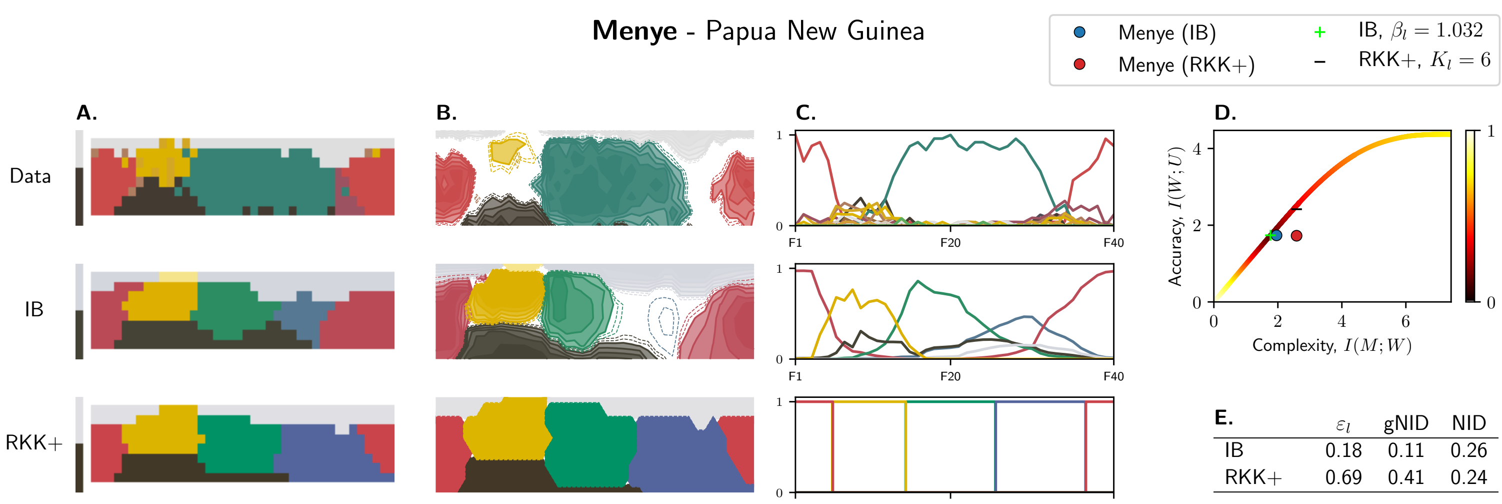
Mazatec - Mexico



E.

	ε_l	gNID	NID
IB	0.22	0.15	0.33
RKK+	0.90	0.44	0.25

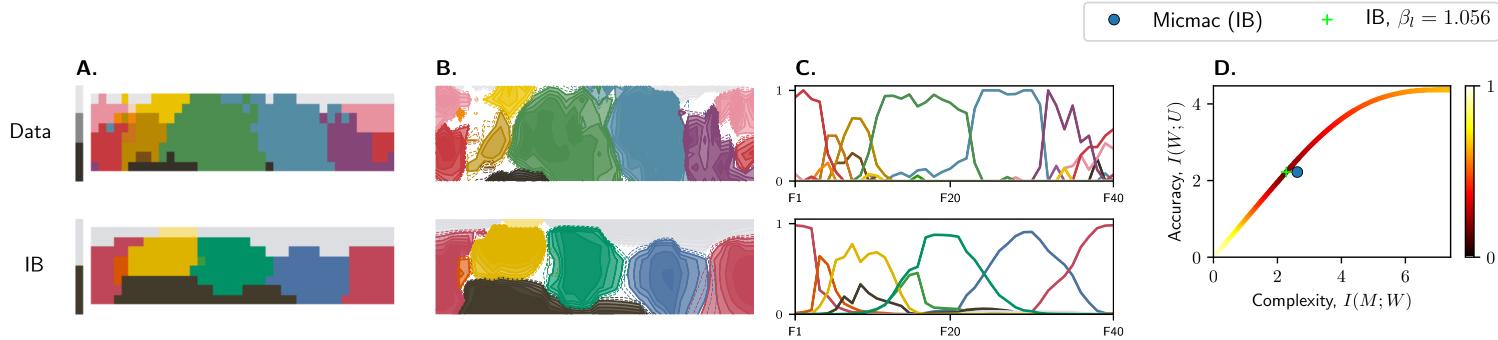
Menye - Papua New Guinea



E.

	ε_l	gNID	NID
IB	0.18	0.11	0.26
RKK+	0.69	0.41	0.24

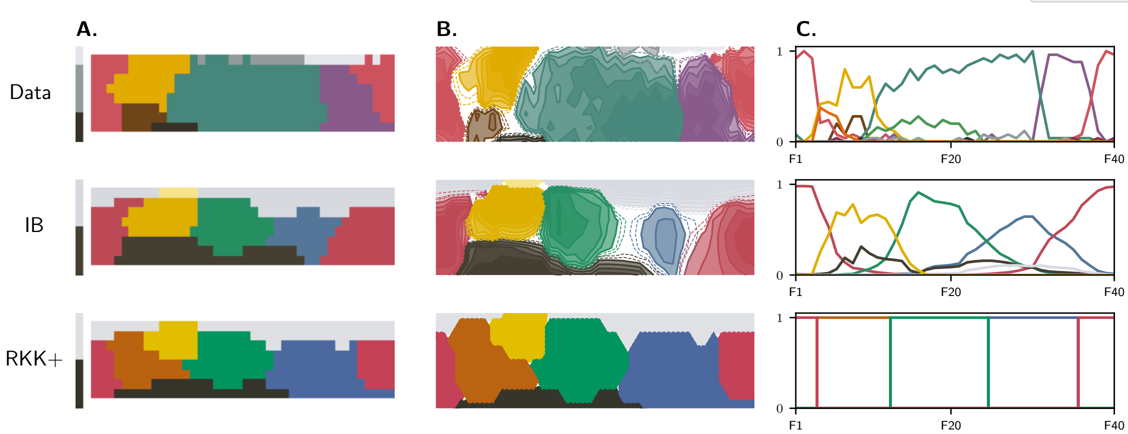
* **Micmac** - Canada

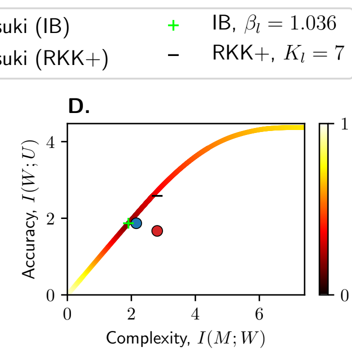


E.

	ε_l	gNID	NID
IB	0.33	0.16	0.37

Mikasuki - USA

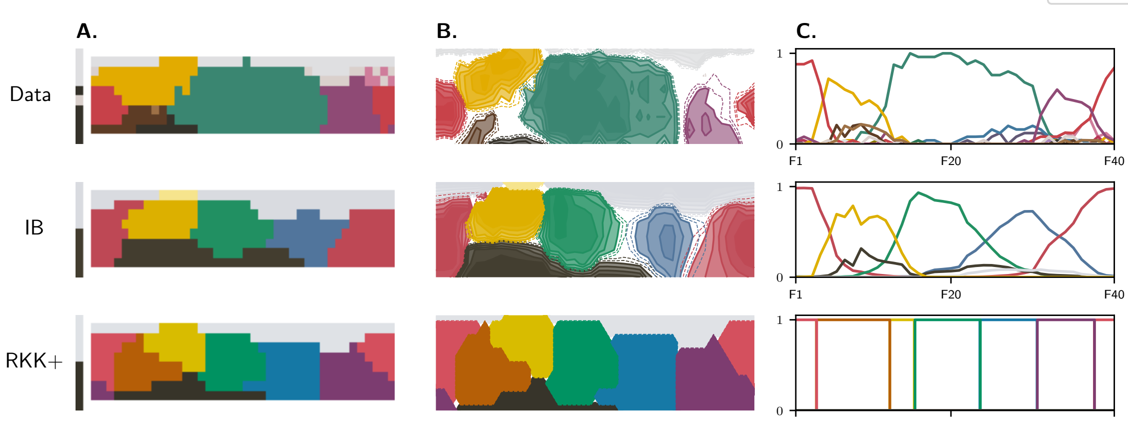


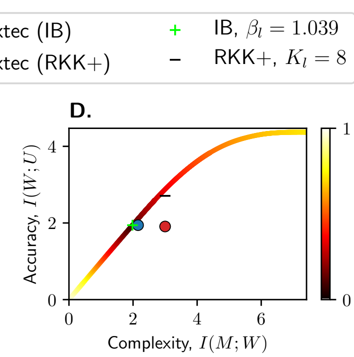


E.

	ε_l	gNID	NID
IB	0.24	0.18	0.39
RKK+	0.91	0.49	0.37

Mixtec - Mexico



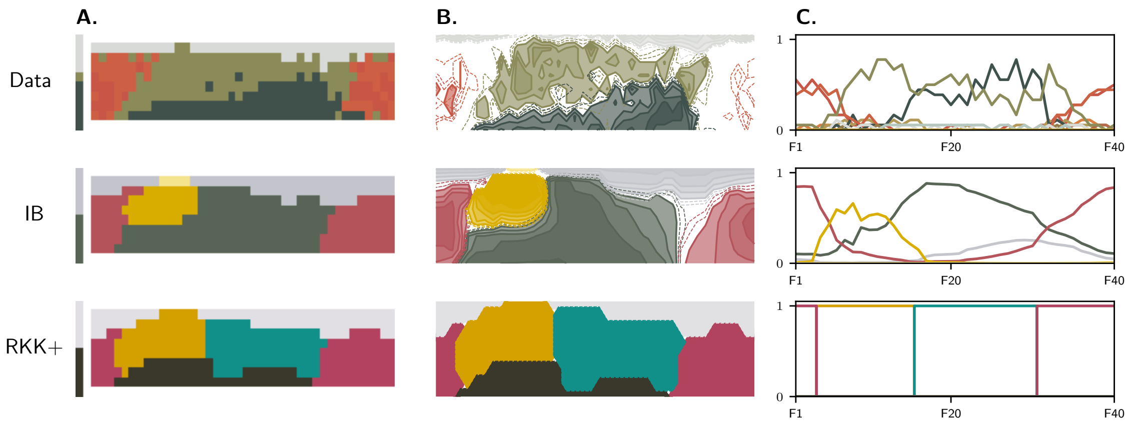


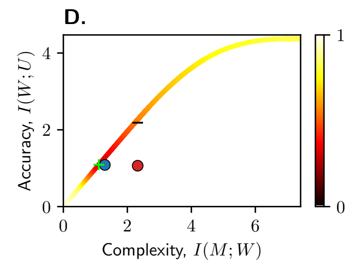
E.

	ε_l	gNID	NID
IB	0.17	0.10	0.28
RKK+	0.80	0.41	0.27

Mundu - Sudan

- Mundu (IB)
- Mundu (RKK+)
- IB, $\beta_l = 1.020$
- RKK+, $K_l = 5$



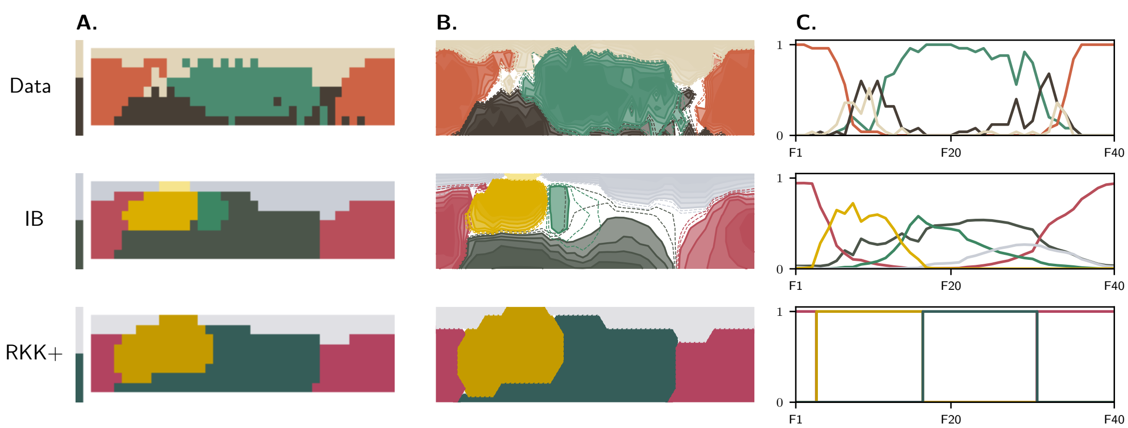


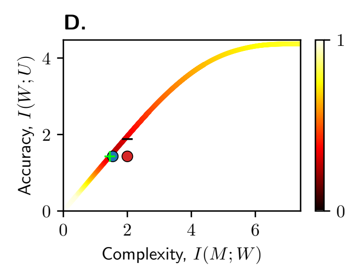
E.

	ε_l	gNID	NID
IB	0.19	0.30	0.45
RKK+	1.12	0.64	0.45

Múra-Pirahá - Brazil

- Múra-Pirahá (IB)
- Múra-Pirahá (RKK+)
- IB, $\beta_l = 1.026$
- RKK+, $K_l = 4$



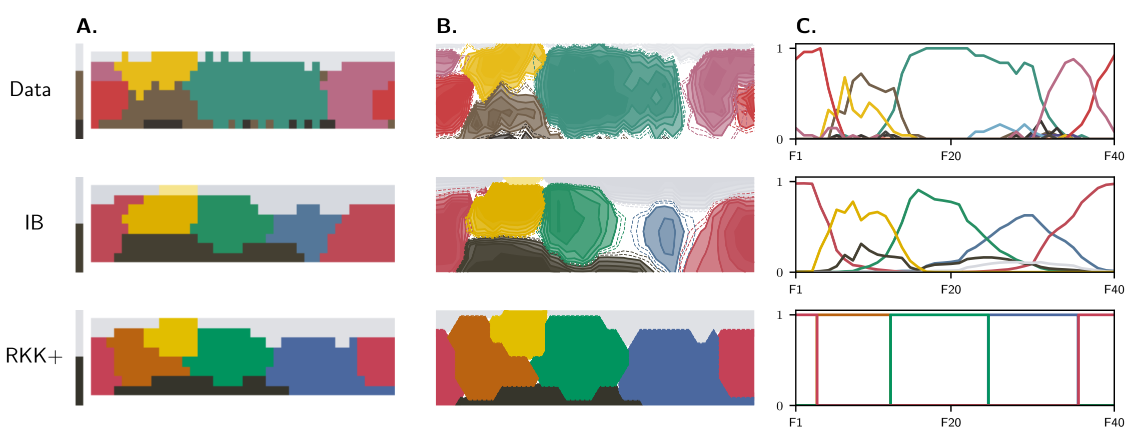


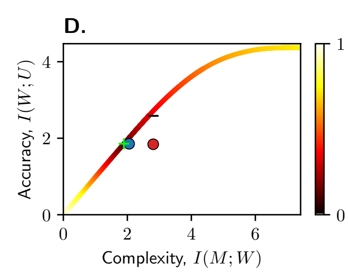
E.

	ε_l	gNID	NID
IB	0.09	0.31	0.45
RKK+	0.45	0.55	0.50

Murle - Sudan

- Murle (IB)
- Murle (RKK+)
- IB, $\beta_l = 1.036$
- RKK+, $K_l = 7$

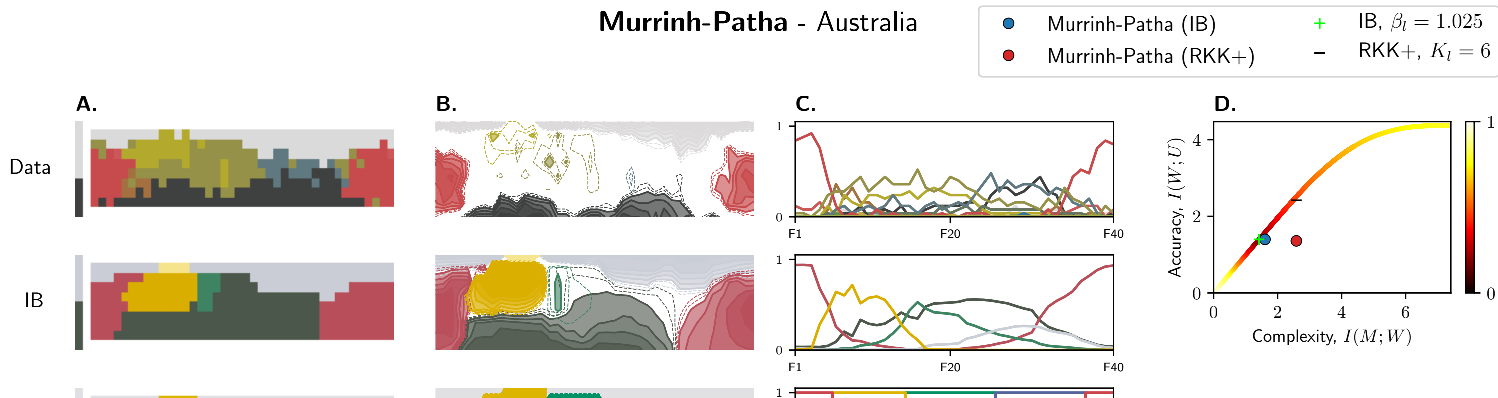




E.

	ε_l	gNID	NID
IB	0.16	0.13	0.30
RKK+	0.74	0.44	0.30

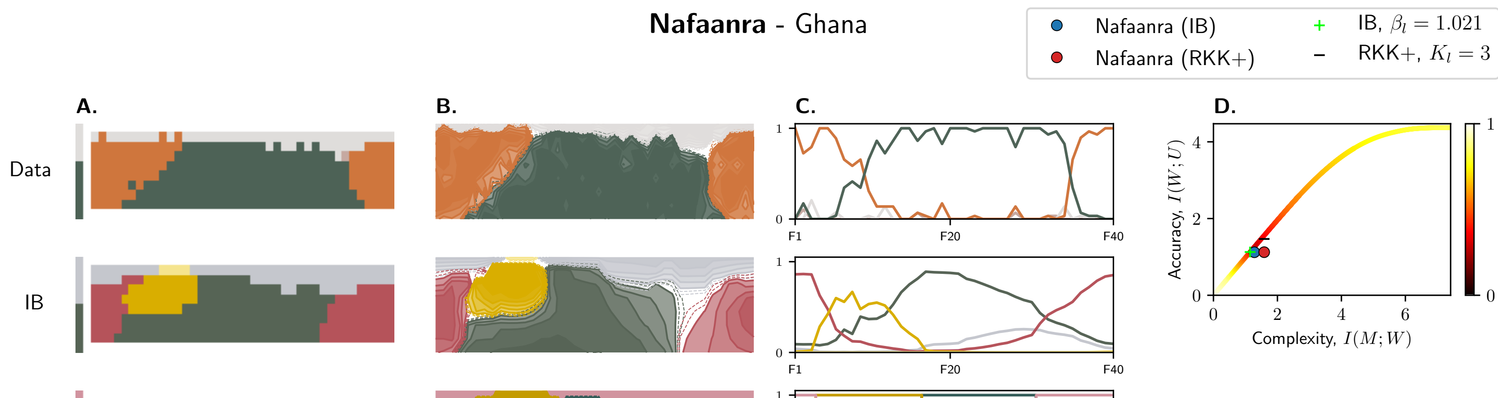
Murrinh-Patha - Australia



E.

	ε_l	gNID	NID
IB	0.18	0.20	0.38
RKK+	1.06	0.53	0.30

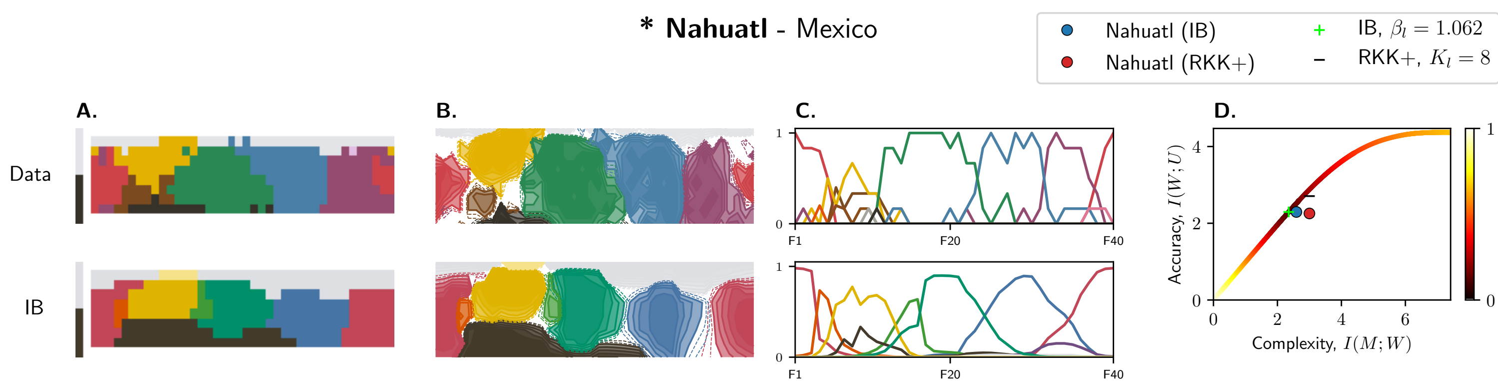
Nafaanra - Ghana



E.

	ε_l	gNID	NID
IB	0.14	0.41	0.42
RKK+	0.34	0.61	0.52

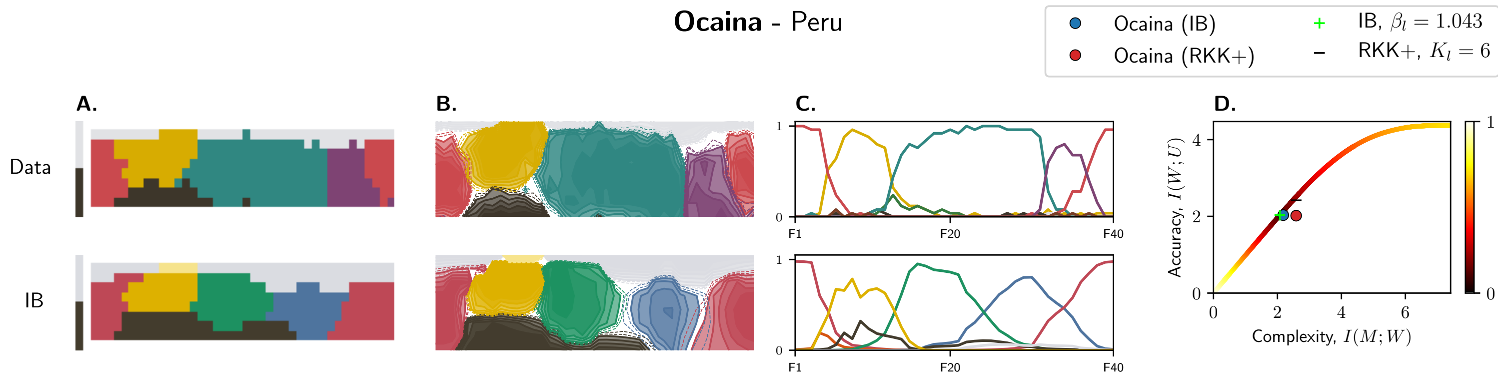
* **Nahuatl** - Mexico



E.

	ε_l	gNID	NID
IB	0.23	0.11	0.25
RKK+	0.45	0.32	0.20

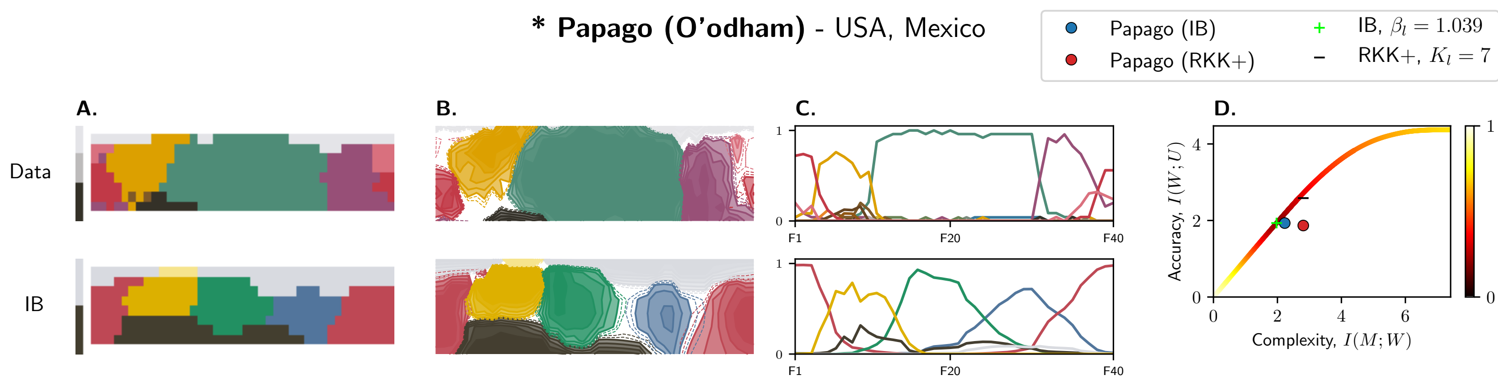
Ocaina - Peru



E.

	ε_l	gNID	NID
IB	0.10	0.11	0.24
RKK+	0.40	0.31	0.20

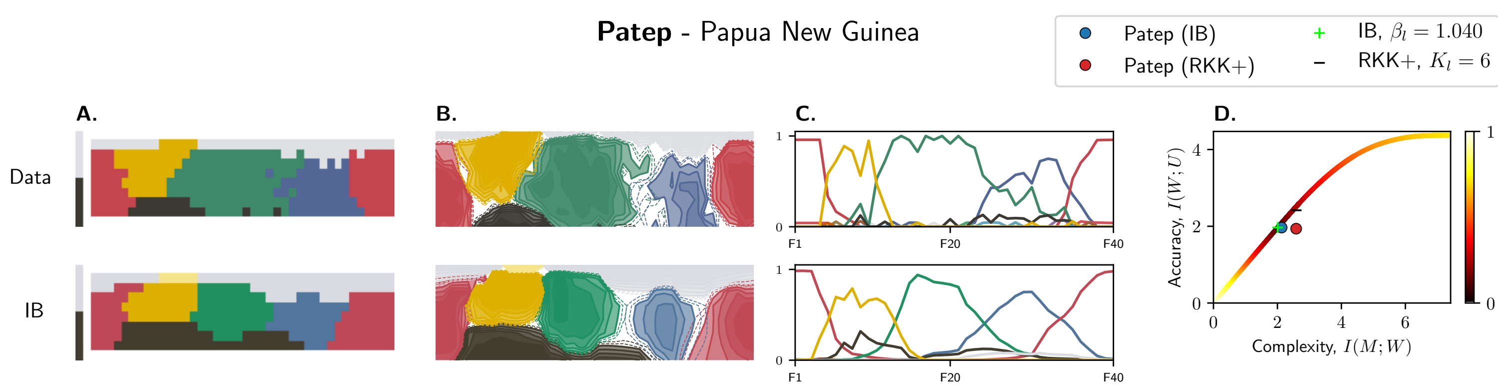
* **Papago (O'odham)** - USA, Mexico



E.

	ε_l	gNID	NID
IB	0.25	0.19	0.38
RKK+	0.72	0.41	0.32

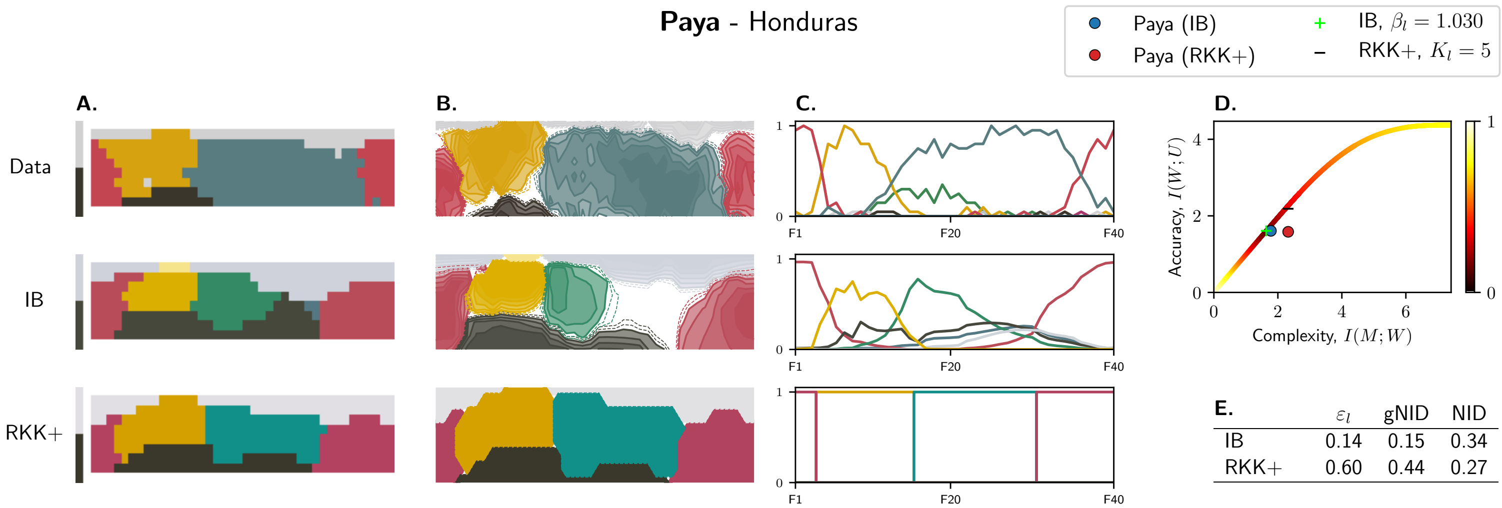
Patep - Papua New Guinea



E.

	ε_l	gNID	NID
IB	0.12	0.11	0.22
RKK+	0.48	0.33	0.20

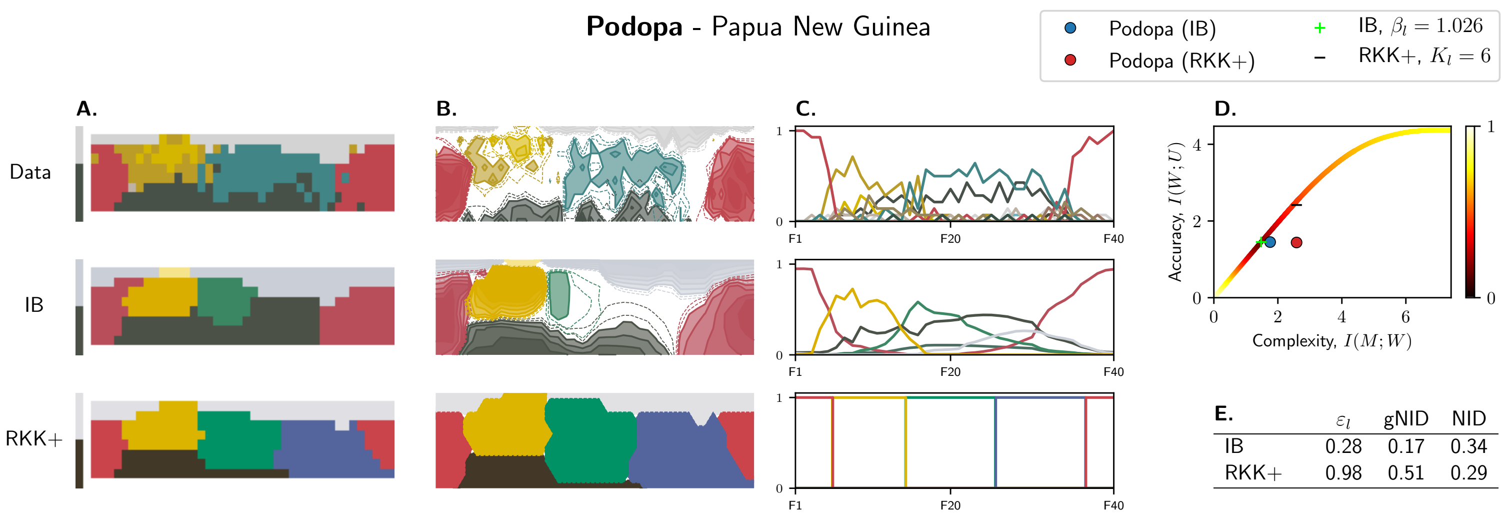
Paya - Honduras



E.

	ε_l	gNID	NID
IB	0.14	0.15	0.34
RKK+	0.60	0.44	0.27

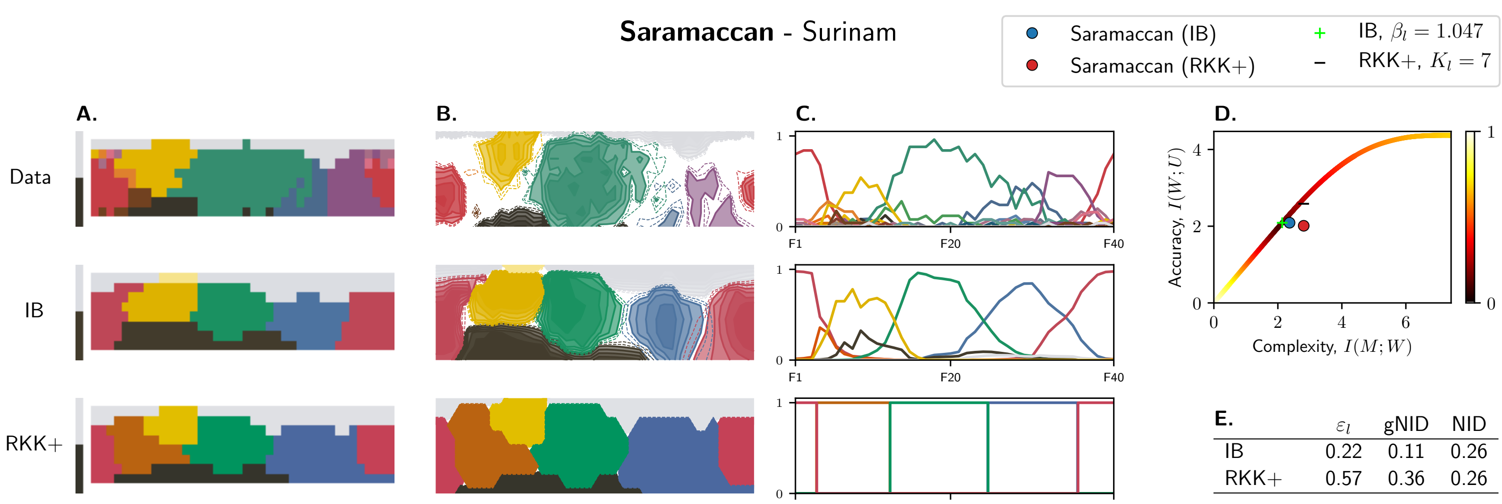
Podopa - Papua New Guinea



E.

	ε_l	gNID	NID
IB	0.28	0.17	0.34
RKK+	0.98	0.51	0.29

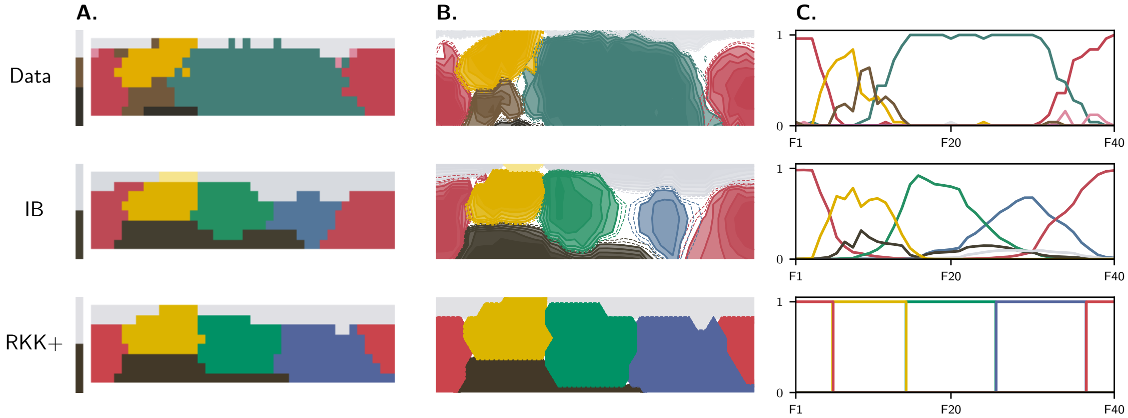
Saramaccan - Surinam

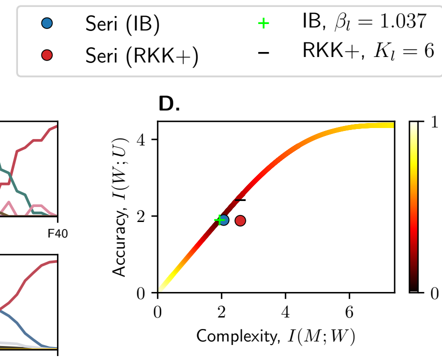


E.

	ε_l	gNID	NID
IB	0.22	0.11	0.26
RKK+	0.57	0.36	0.26

Seri - Mexico

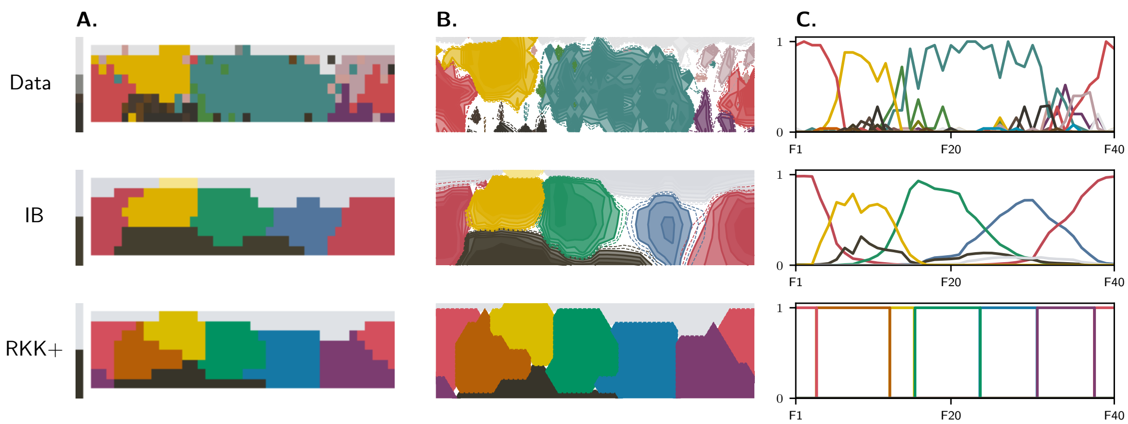


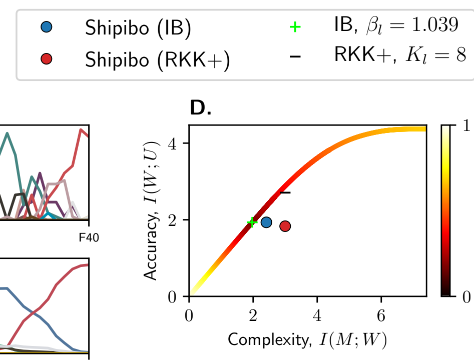


E.

	ε_l	gNID	NID
IB	0.13	0.12	0.28
RKK+	0.54	0.36	0.26

Shipibo - Peru

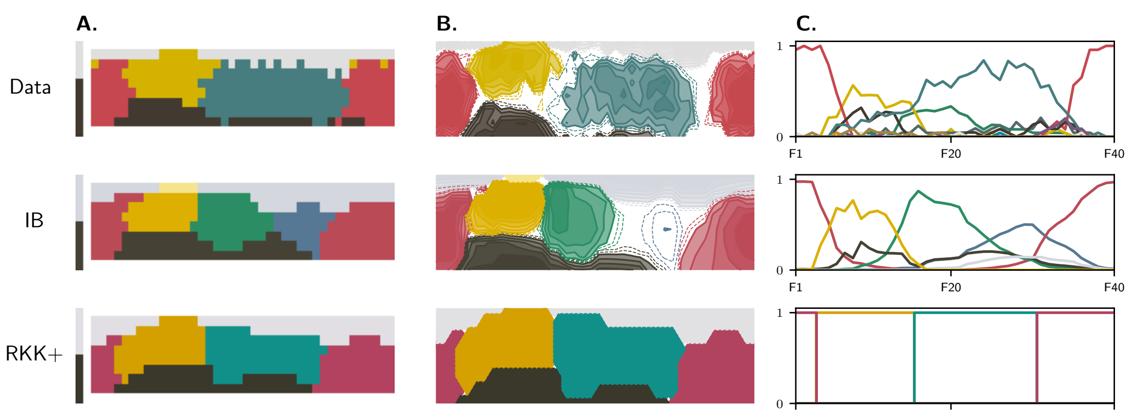


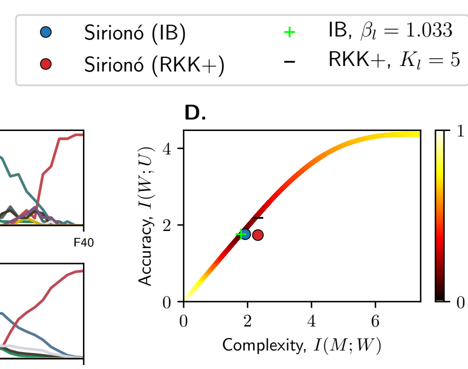


E.

	ε_l	gNID	NID
IB	0.43	0.17	0.33
RKK+	0.88	0.46	0.33

Sirionó - Bolivia

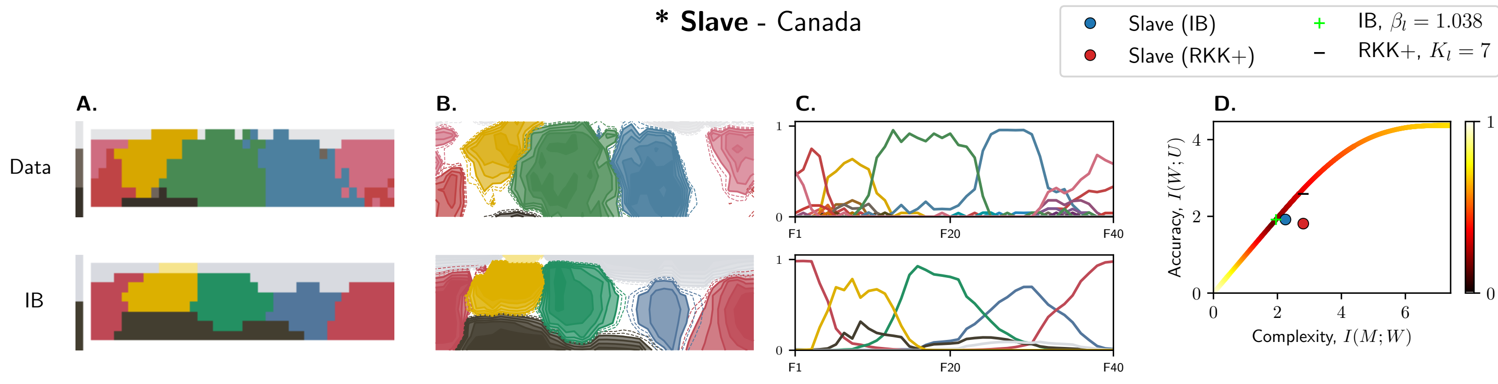




E.

	ε_l	gNID	NID
IB	0.13	0.11	0.25
RKK+	0.45	0.40	0.26

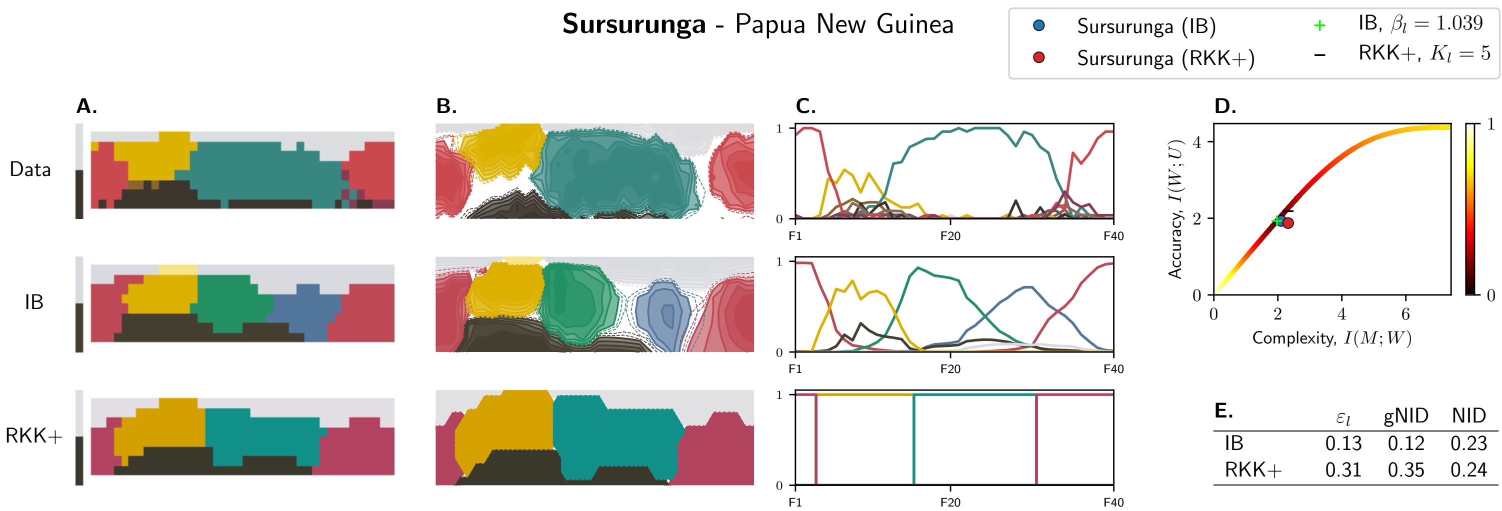
* **Slave** - Canada



E.

	ε_l	gNID	NID
IB	0.28	0.16	0.34
RKK+	0.77	0.42	0.25

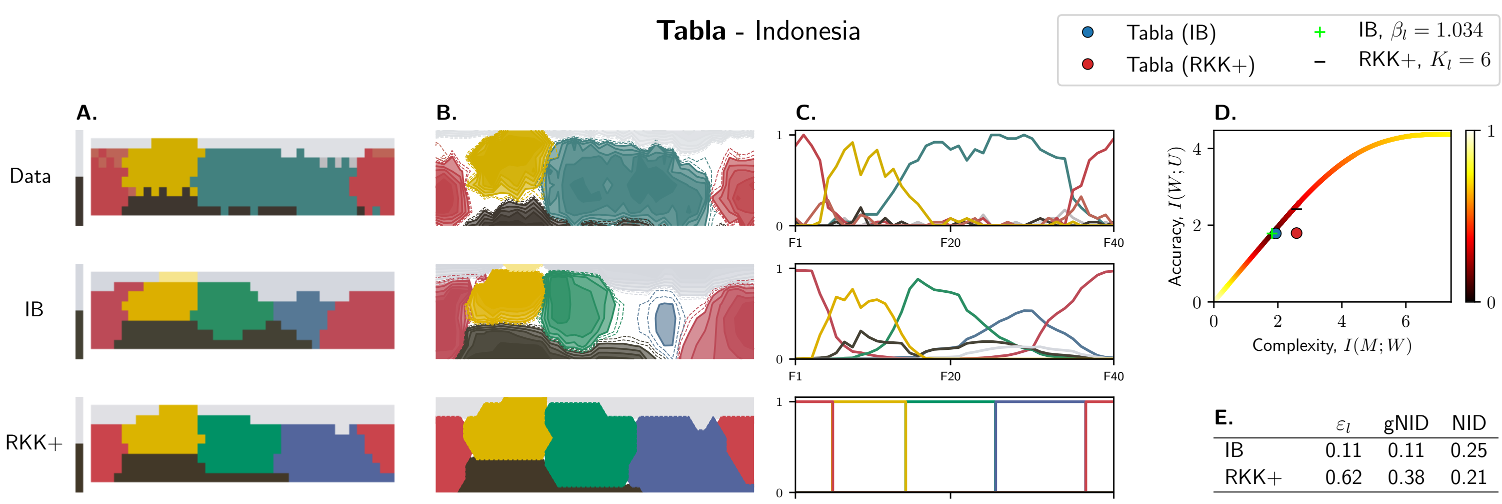
Sursurunga - Papua New Guinea



E.

	ε_l	gNID	NID
IB	0.13	0.12	0.23
RKK+	0.31	0.35	0.24

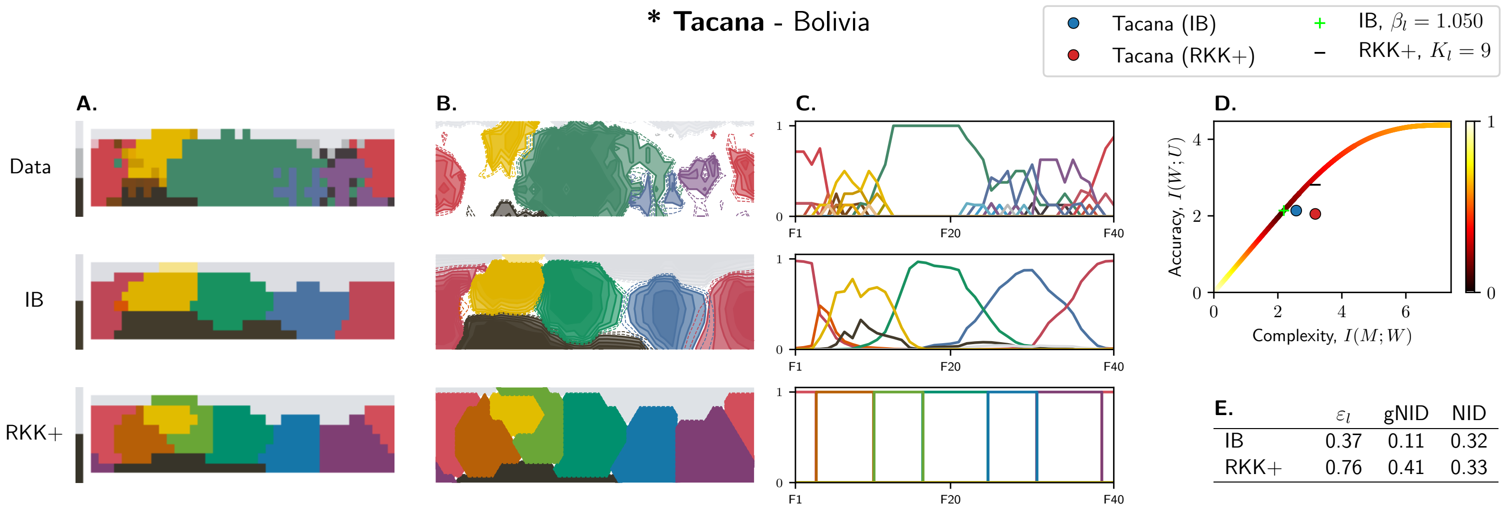
Tabla - Indonesia



E.

	ε_l	gNID	NID
IB	0.11	0.11	0.25
RKK+	0.62	0.38	0.21

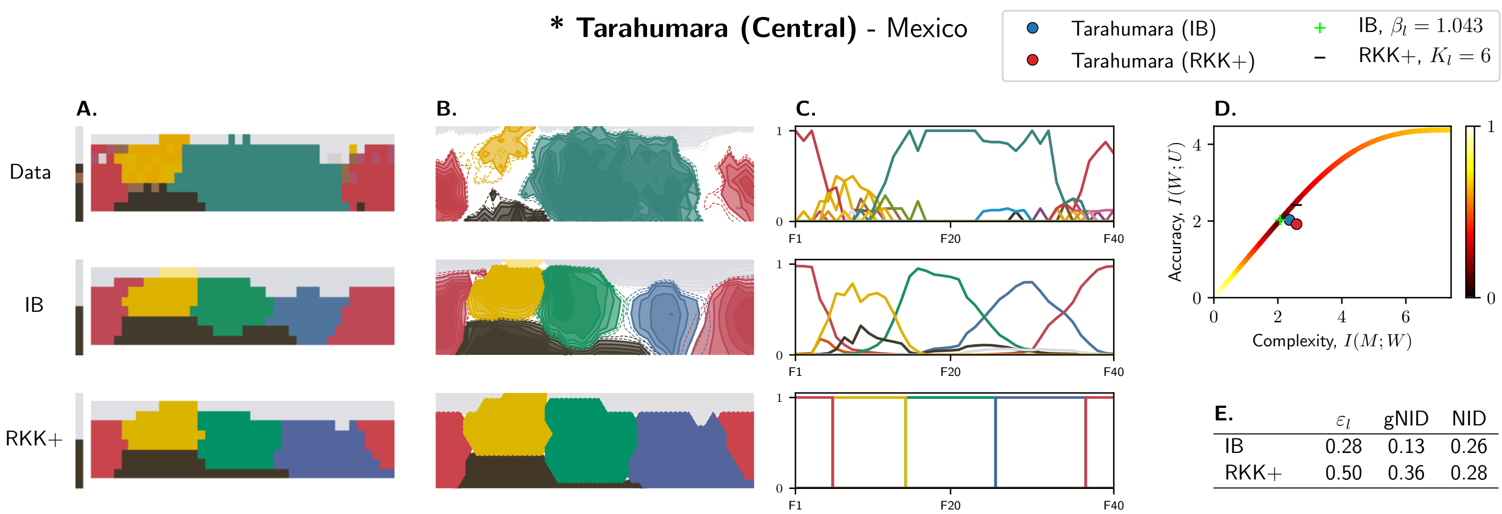
* **Tacana** - Bolivia



E.

	ε_l	gNID	NID
IB	0.37	0.11	0.32
RKK+	0.76	0.41	0.33

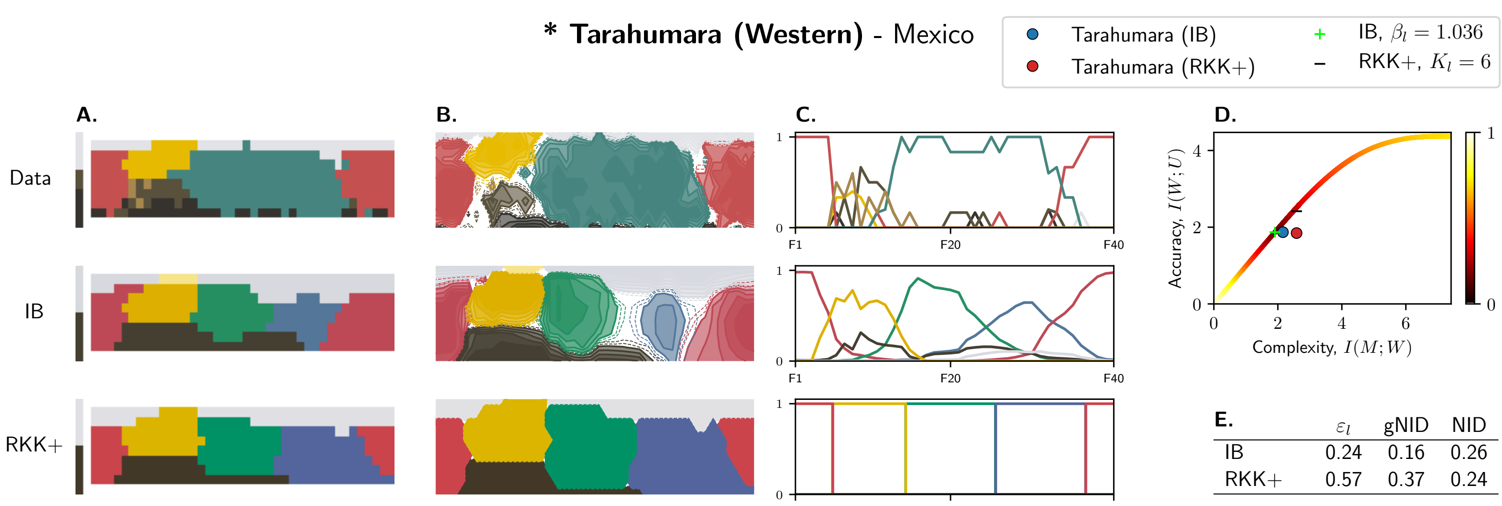
* **Tarahumara (Central)** - Mexico



E.

	ε_l	gNID	NID
IB	0.28	0.13	0.26
RKK+	0.50	0.36	0.28

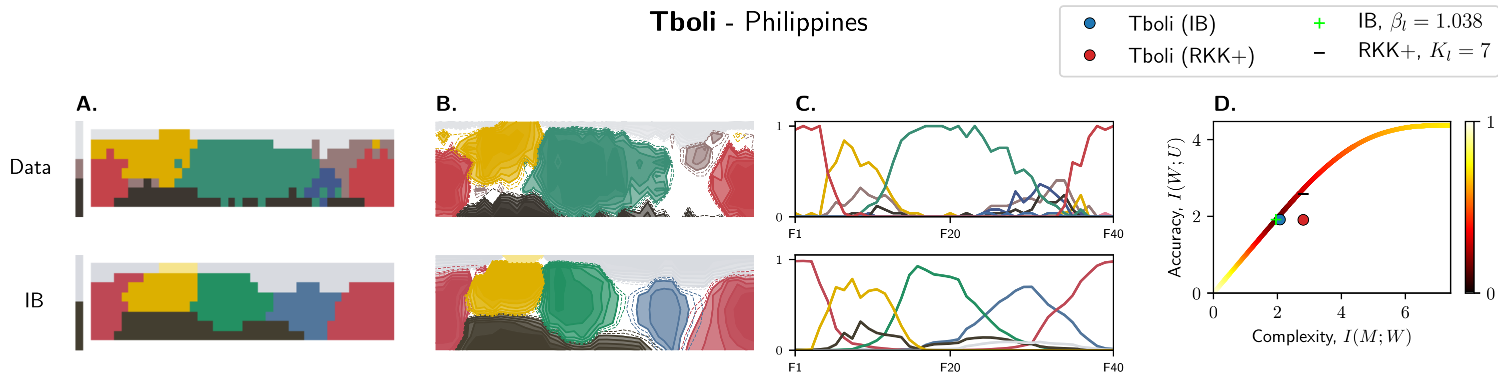
* **Tarahumara (Western)** - Mexico



E.

	ε_l	gNID	NID
IB	0.24	0.16	0.26
RKK+	0.57	0.37	0.24

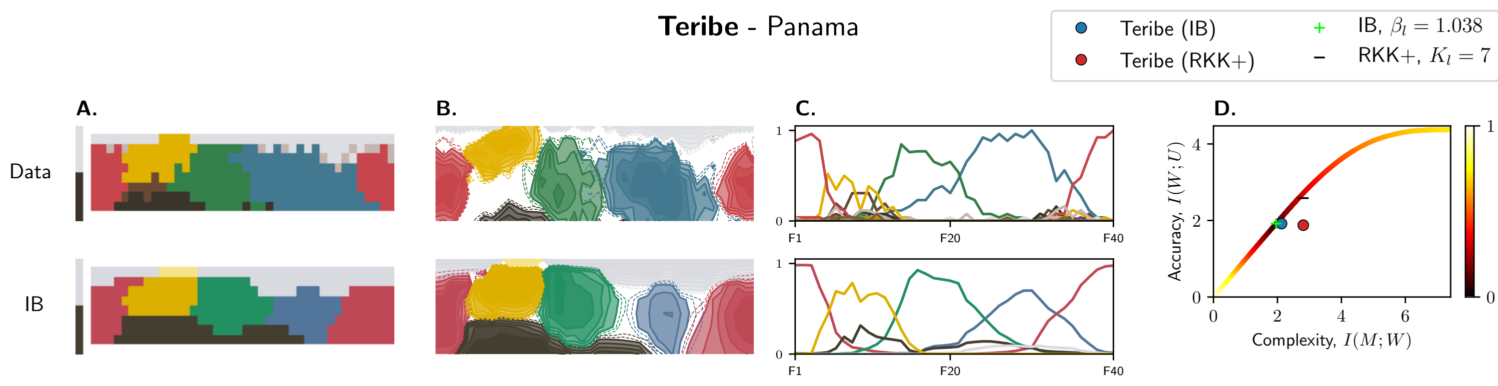
Tboli - Philippines



E.

	ε_l	gNID	NID
IB	0.12	0.12	0.24
RKK+	0.68	0.38	0.29

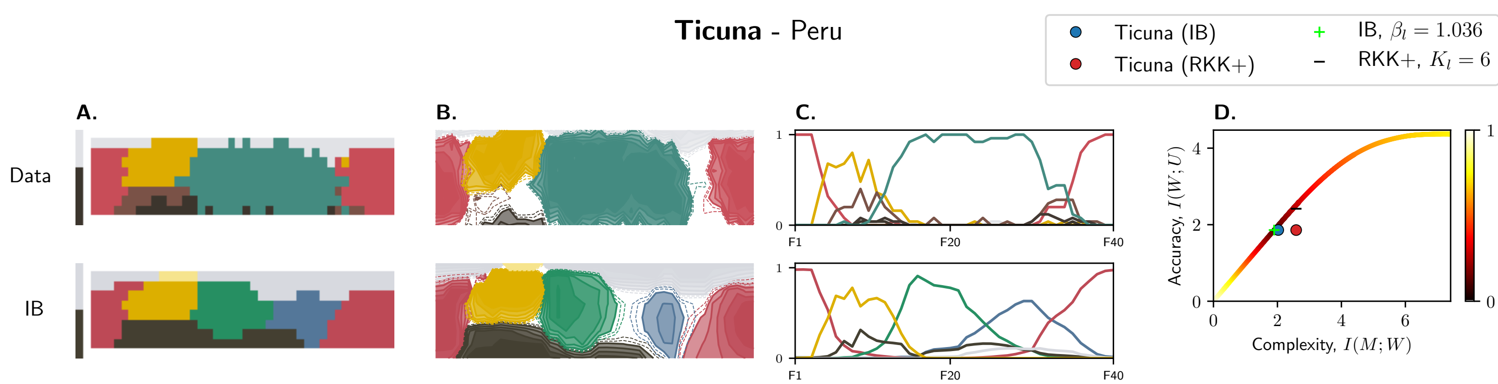
Teribe - Panama



E.

	ε_l	gNID	NID
IB	0.16	0.09	0.23
RKK+	0.71	0.38	0.21

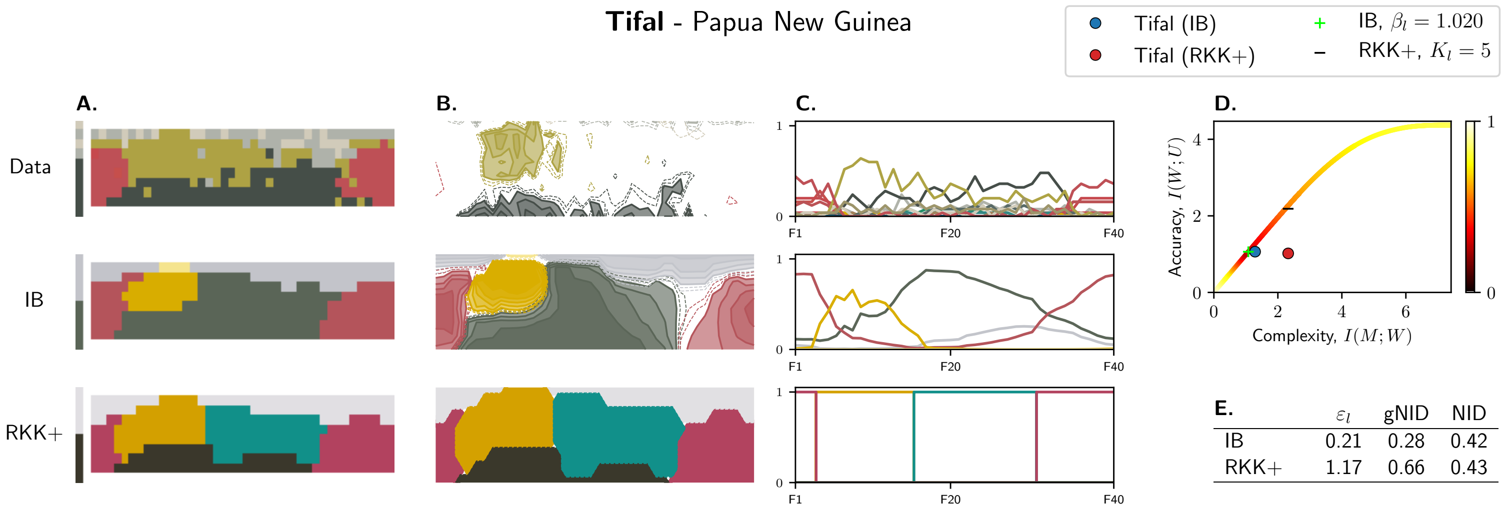
Ticuna - Peru



E.

	ε_l	gNID	NID
IB	0.12	0.12	0.25
RKK+	0.56	0.37	0.25

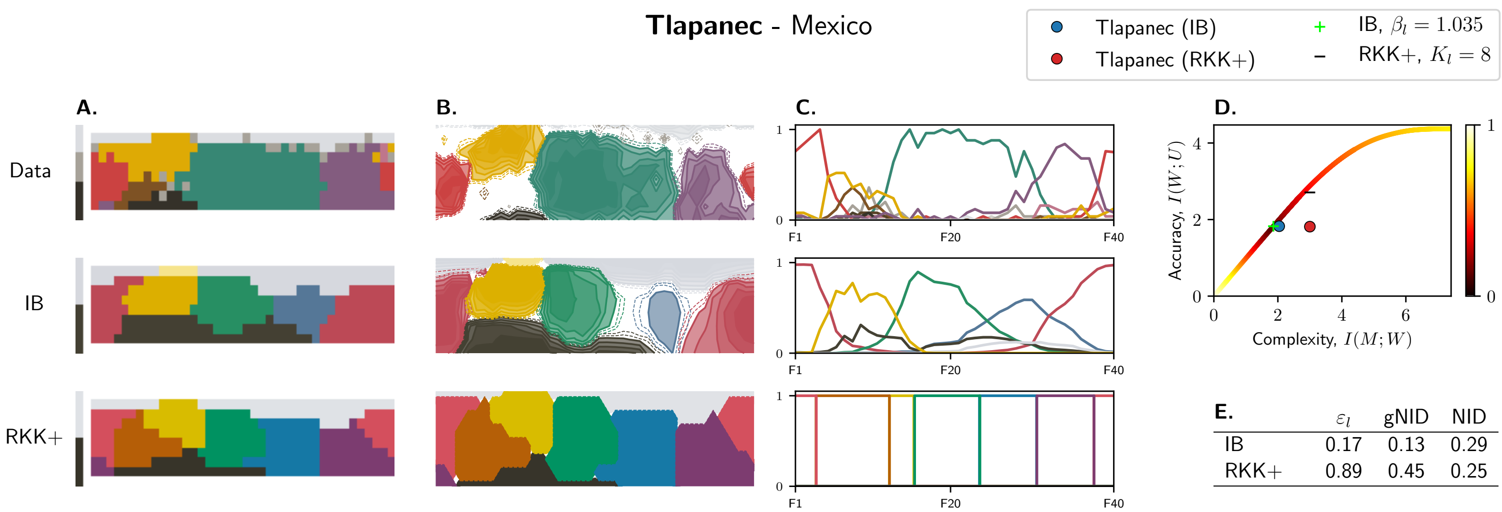
Tifal - Papua New Guinea



E.

	ε_l	gNID	NID
IB	0.21	0.28	0.42
RKK+	1.17	0.66	0.43

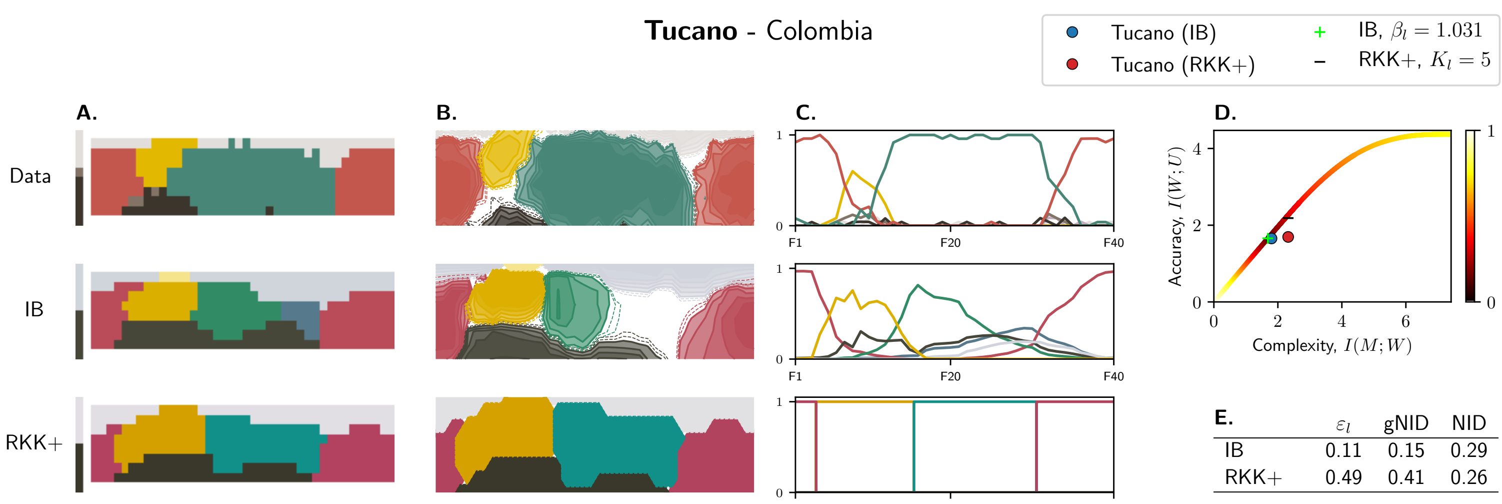
Tlapanec - Mexico



E.

	ε_l	gNID	NID
IB	0.17	0.13	0.29
RKK+	0.89	0.45	0.25

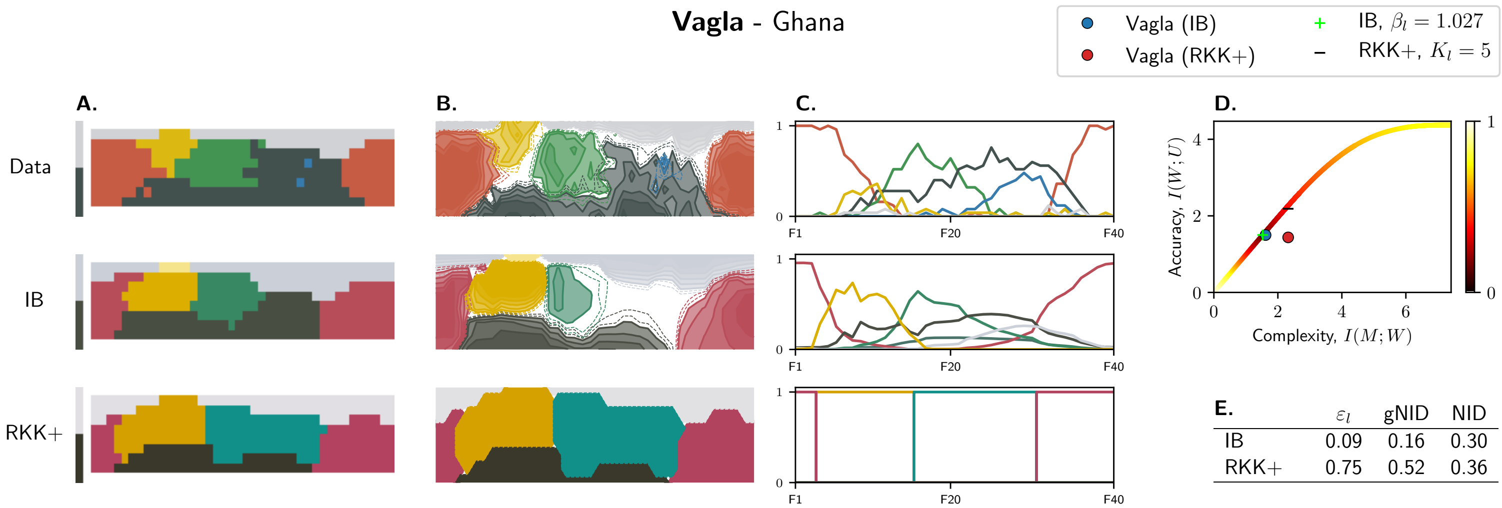
Tucano - Colombia



E.

	ε_l	gNID	NID
IB	0.11	0.15	0.29
RKK+	0.49	0.41	0.26

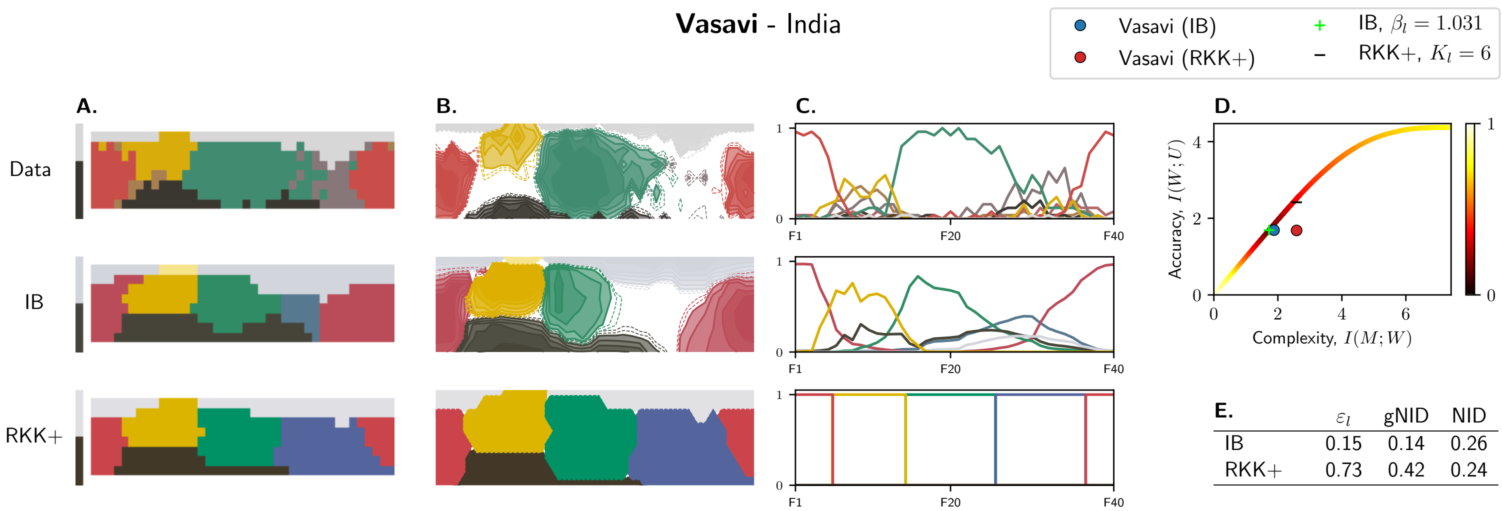
Vagla - Ghana



E.

	ε_l	gNID	NID
IB	0.09	0.16	0.30
RKK+	0.75	0.52	0.36

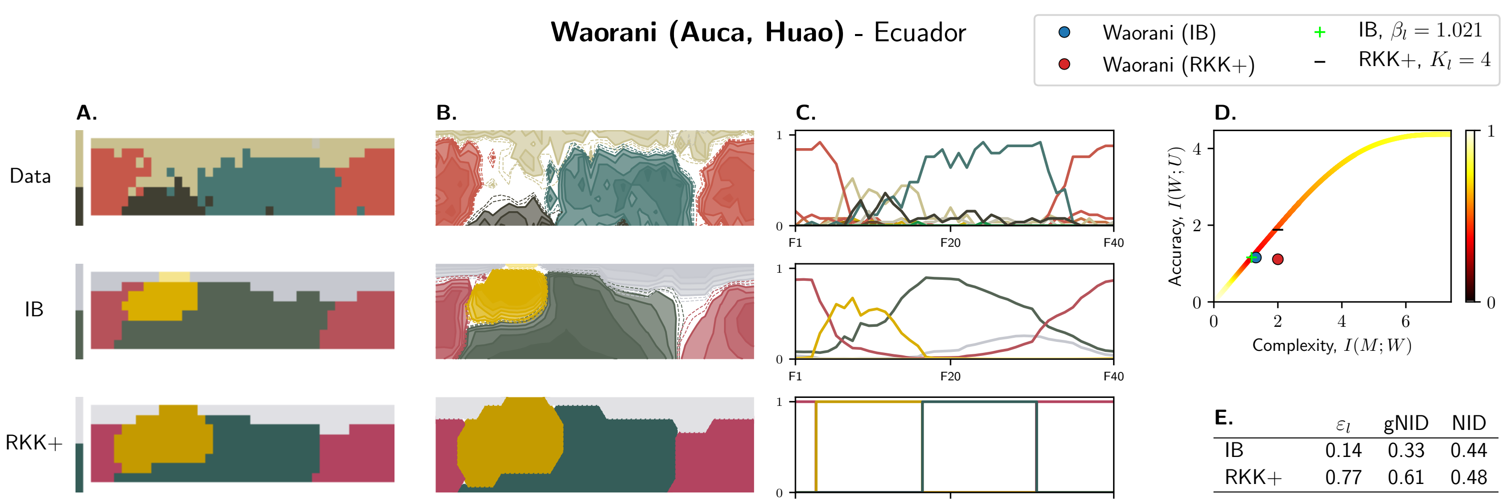
Vasavi - India



E.

	ε_l	gNID	NID
IB	0.15	0.14	0.26
RKK+	0.73	0.42	0.24

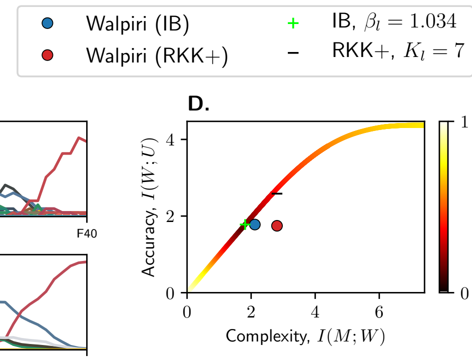
Waorani (Auca, Huao) - Ecuador

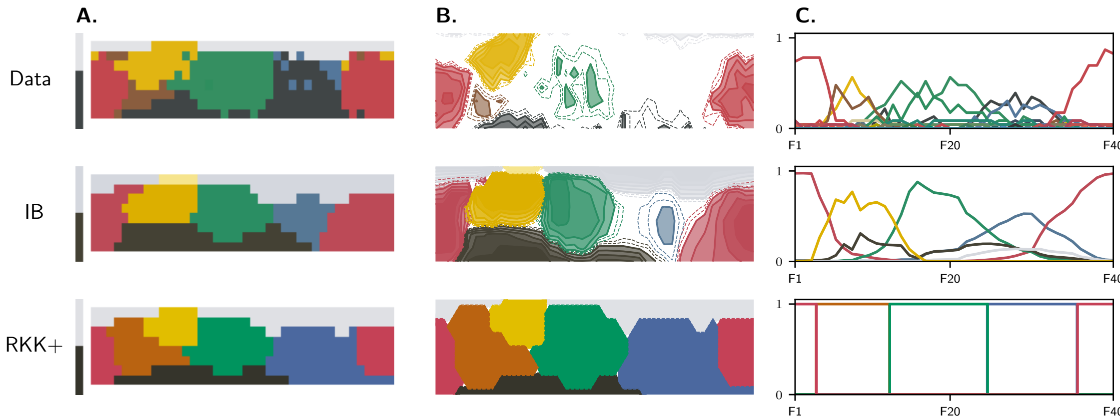


E.

	ε_l	gNID	NID
IB	0.14	0.33	0.44
RKK+	0.77	0.61	0.48

Walpiri - Australia

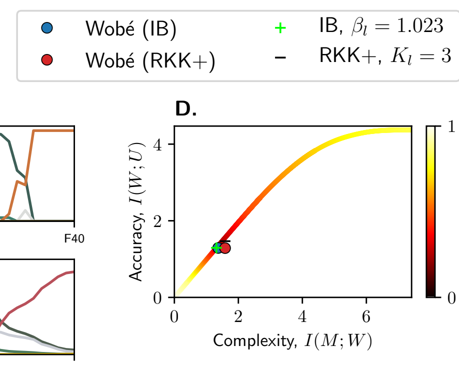


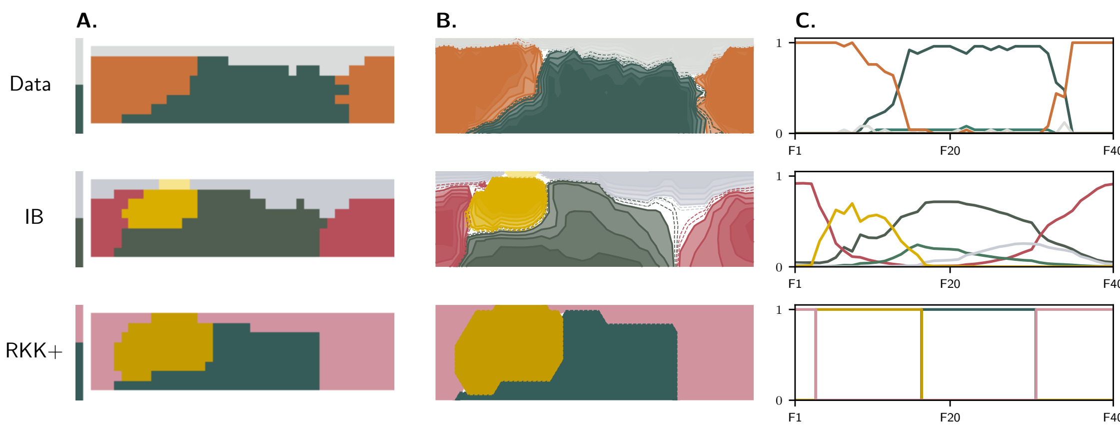


E.

	ε_l	gNID	NID
IB	0.29	0.14	0.24
RKK+	0.84	0.44	0.26

Wobé - Ivory Coast

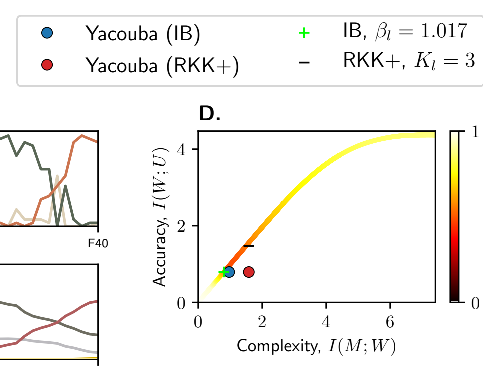


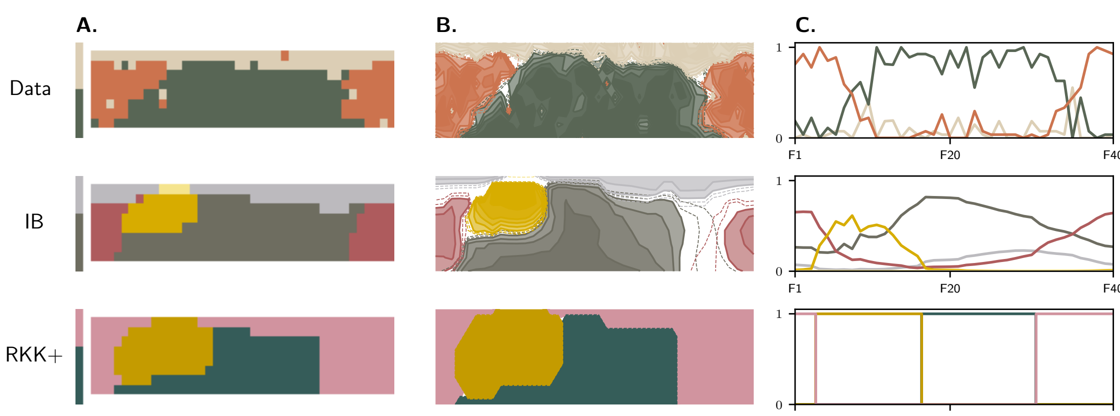


E.

	ε_l	gNID	NID
IB	0.06	0.35	0.40
RKK+	0.18	0.55	0.48

Yacouba - Ivory Coast

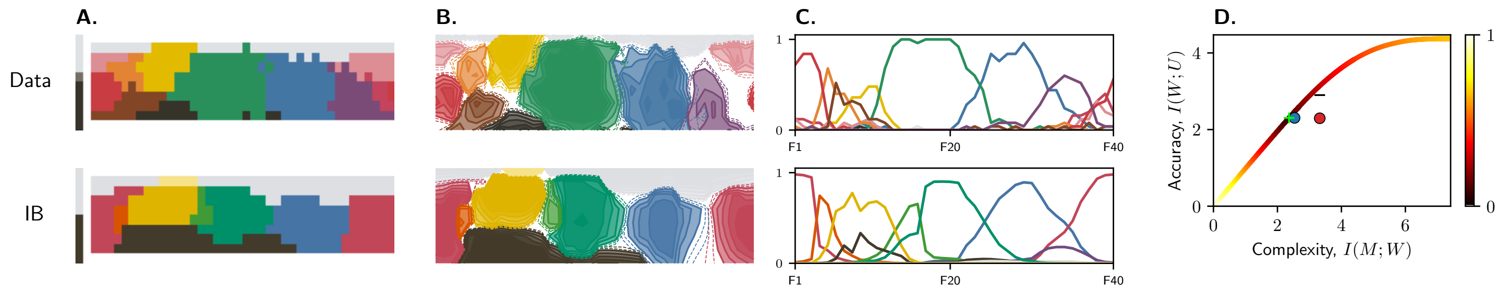




E.

	ε_l	gNID	NID
IB	0.16	0.61	0.43
RKK+	0.68	0.73	0.60

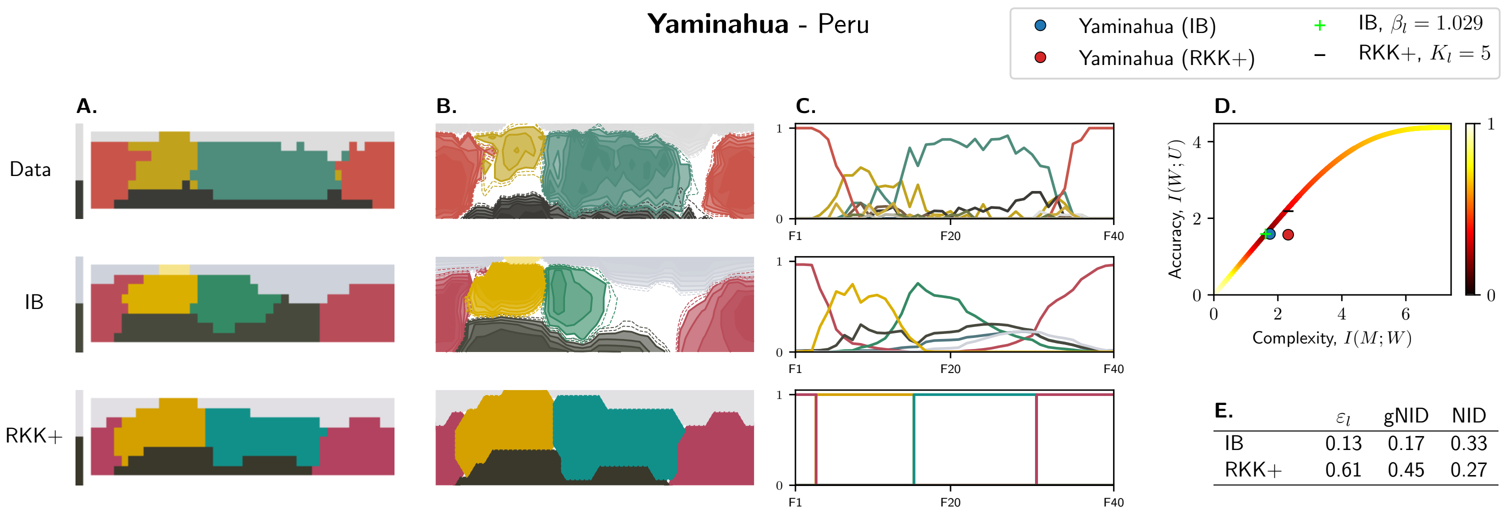
Yakan - Philippines



E.

	ε_l	gNID	NID
IB	0.16	0.10	0.27
RKK+	0.60	0.33	0.20

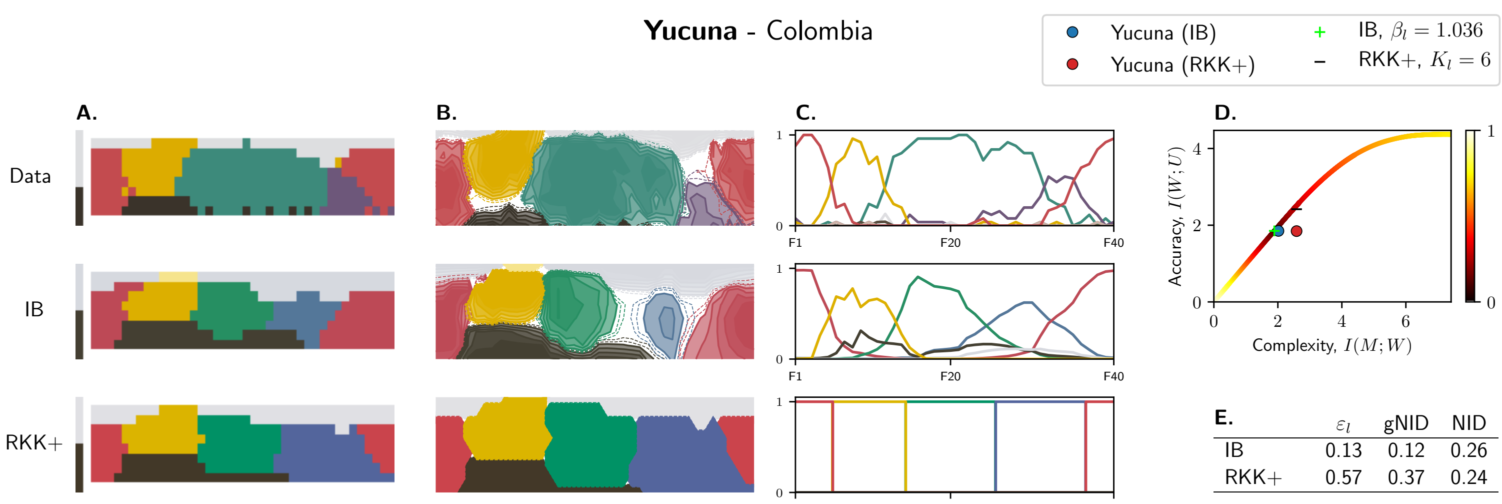
Yaminahua - Peru



E.

	ε_l	gNID	NID
IB	0.13	0.17	0.33
RKK+	0.61	0.45	0.27

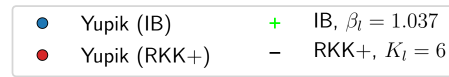
Yucuna - Colombia

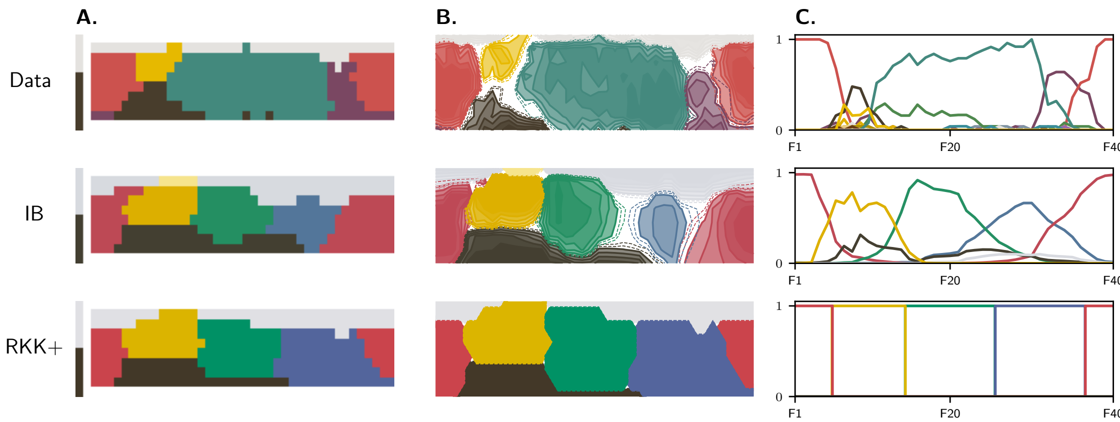


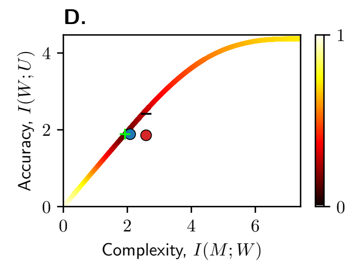
E.

	ε_l	gNID	NID
IB	0.13	0.12	0.26
RKK+	0.57	0.37	0.24

Yupik - USA



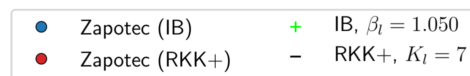


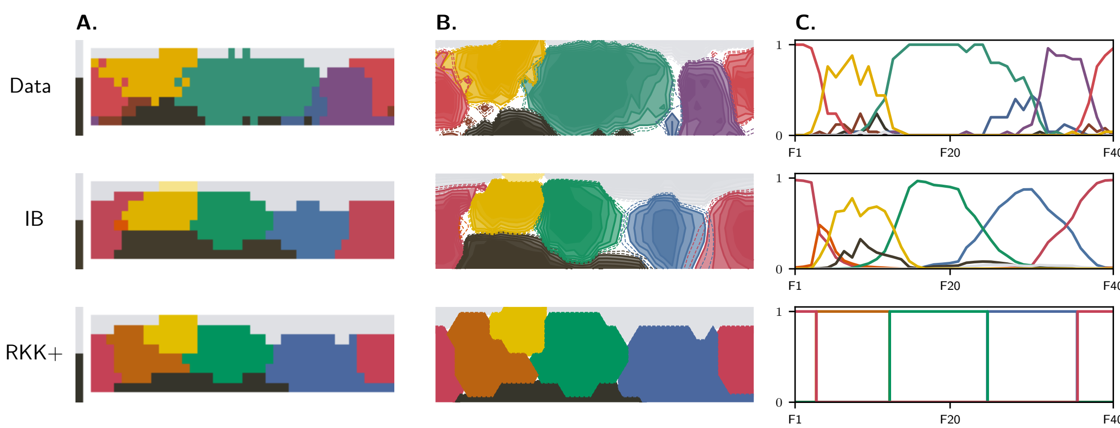


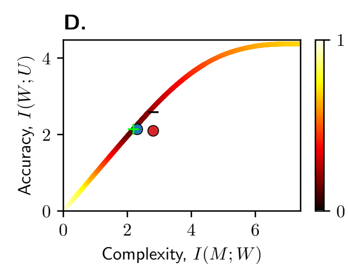
E.

	ε_l	gNID	NID
IB	0.16	0.12	0.26
RKK+	0.56	0.37	0.25

Zapotec - Mexico







E.

	ε_l	gNID	NID
IB	0.12	0.11	0.28
RKK+	0.49	0.33	0.28

References

1. Tishby N, Pereira FC, Bialek W (1999) The Information Bottleneck method in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*.
2. Regier T, Kemp C, Kay P (2015) Word meanings across languages support efficient communication in *The Handbook of Language Emergence*, eds. MacWhinney B, O'Grady W. (Wiley-Blackwell, Hoboken, NJ), pp. 237–263.
3. Gilad-Bachrach R, Navot A, Tishby N (2003) An information theoretic tradeoff between complexity and accuracy in *Proceedings of the 16th Annual Conference on Learning Theory*.
4. Harremoës P, Tishby N (2007) The Information Bottleneck revisited or how to choose a good distortion measure in *IEEE International Symposium on Information Theory*. pp. 566–571.
5. Pereira F, Tishby N, Lee L (1993) Distributional clustering of English words in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. pp. 183–190.
6. Shamir O, Sabato S, Tishby N (2010) Learning and generalization with the Information Bottleneck. *Theoretical Computer Science* 411(29-30):2696–2711.
7. Rose K (1998) Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE* 86(11):2210–2239.
8. Elidan G, Friedman N (2005) Learning hidden variable networks: The Information Bottleneck approach. *Journal of Machine Learning Research* 6:81–127.
9. Slonim N, Tishby N (1999) Agglomerative Information Bottleneck in *Advances in Neural Information Processing Systems (NIPS)*. pp. 617–623.
10. Blahut R (1972) Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory* 18(4):460–473.
11. Arimoto S (1972) An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory* 18(1):14–20.
12. Bernardo JM (2005) Reference analysis in *Bayesian Thinking Modeling and Computation*, Handbook of Statistics, eds. Dey D, Rao C. (Elsevier) Vol. 25, pp. 17 – 90.
13. Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *JMLR* 11:2837–2854.
14. Abbott JT, Griffiths TL, Regier T (2016) Focal colors across languages are representative members of color categories. *Proceedings of the National Academy of Sciences* 113(40):11178–11183.
15. Roberson D, Davies I, Corbett G, Vandervyver M (2005) Free-sorting of colors across cultures: Are there universal grounds for grouping? *Journal of Cognition and Culture* 5(3):349–386.
16. Roberson D, Davies I, Davidoff J (2000) Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General* 129(3):369–398.
17. Roberson D, Davidoff J, Davies IR, Shapiro LR (2005) Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology* 50(4):378 – 411.
18. Cibelli E, Xu Y, Austerweil JL, Griffiths TL, Regier T (2016) The Sapir-Whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PLOS ONE* 11(7):1–28.
19. Mokrzycki W, Tatol M (2012) Colour difference ΔE - a survey. *Machine Graphic and Vision* 8.
20. Regier T, Kay P, Khetarpal N (2007) Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences* 104(4):1436–1441.
21. Gibson E, et al. (2017) Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences* 114(40):10785–10790.
22. Lindsey DT, Brown AM (2014) The color lexicon of American English. *Journal of Vision* 14(2):17.