

An analysis of NK landscapes: Interaction structure, statistical properties and expected number of local optima

Jeffrey Buzas and Jeffrey Dinitz

Abstract—Simulated landscapes have been used for decades to evaluate search strategies whose goal is to find the landscape location with maximum fitness. Understanding properties of landscapes is important for understanding search difficulty. This paper presents a novel and transparent characterization of NK landscapes and derives an analytic expression representing the expected number of local optima.

We prove that NK landscapes can be represented by parametric linear interaction models where model coefficients have meaningful interpretations. We derive the statistical properties of the model coefficients, providing insight into how the NK algorithm parses importance to main effects and interactions. An important insight derived from the linear model representation is that the rank of the linear model defined by the NK algorithm is correlated with the number of local optima, a strong determinant of landscape complexity and search difficulty. We show that the maximal rank for an NK landscape is achieved through epistatic interactions that form partially balanced incomplete block designs. Finally, an analytic expression representing the expected number of local optima on the landscape is derived, providing a way to quickly compute the expected number of local optima for very large landscapes.

Index Terms—Balanced Incomplete Block Design, Orthant Probability, Walsh function

I. INTRODUCTION

NK LANDSCAPES, defined over two decades ago, continue to play an important role in the evaluation of evolutionary search algorithms, and understanding properties of landscapes is important for understanding search difficulty. Applications of simulated landscapes include modeling the capacity of enzymes to catalyze reactions or ligands to bind to proteins, and the clinical effectiveness of medical treatments [12], [4].

NK landscapes are defined by a straightforward, tunable algorithm where N specifies the number of binary features or loci and K the degree of epistatic interactions among the loci [11]. NK landscapes are convenient because there are only two tunable parameters (N and K), and yet they provide a very rich set of landscapes. There is a large literature exploring properties of NK landscapes, though there are few analytic results, with [6] and [13] notable exceptions.

This paper provides a novel perspective on NK landscapes and generalizations of NK landscapes, proving that these landscapes can be characterized by parametric linear models

comprised of main effects and interaction effects where the model coefficients have meaningful interpretations. The NK algorithm induces a statistical distribution on the parametric model coefficients. We derive the distribution of model coefficients, showing how the NK algorithm, for $K \ll N$, automatically assigns the largest expected magnitude to main effects, with the expected magnitude of interaction effects typically decreasing with increasing order of interaction.

The linear model representation of the landscape suggests that model rank should provide a measure of landscape complexity. A simple method of assessing rank is provided, and we determine conditions on N and K sufficient for the existence of designs that achieve maximum rank. We then show that the expected number of local optima is proportional to an orthant probability, which can be calculated with reasonable speed and accuracy for very large landscapes. Interestingly and surprisingly, rank is both positively *and* negatively correlated with the number of local optima. For fixed N , it is well-known [5] that the number of local optima increases with K (i.e. a positive correlation with landscape rank). We show that when N and K are both fixed, there is a strong *negative* correlation between rank and number of local optima for classic NK landscapes. The statistical distribution of model effect coefficients provides an explanation for this counter intuitive phenomenon.

In related work, a proposal to use linear models with main effects and interactions to construct landscapes was explored in [16], [17] and [18]. These authors examined the effect of epistatic interactions on the properties of landscapes using metrics common in the experimental design literature, and they noted the equivalence between Walsh function decompositions and interaction models. An analysis of NK landscapes using Walsh functions was given in [8].

II. GENERALIZED NK LANDSCAPES AND INTERACTION MODELS

In this section we define the NK landscape and interaction models. The models are represented as linear models in matrix form, as this representation facilitates the study of model properties. For simplicity, the model is first discussed for K constant across loci. The model is then generalized to allow varying K across loci.

A. NK landscapes

A general landscape is defined by a triple (χ, d, f) where χ is a set of locations, d is a distance measure and $f : \chi \rightarrow \mathbb{R}$

J. Buzas and J. Dinitz are with the Department of Mathematics and Statistics, University of Vermont, Burlington, VT, 05401 USA e-mail: (see <http://www.cems.uvm.edu/jbuzas>).

Manuscript received February 19, 2013; revised ??.

is a ‘‘fitness’’ function. NK landscapes are a map from $\chi = (\{0, 1\})^N$ to \mathbb{R} where the fitness function f is built from N binary loci and epistatic interactions formed between each locus and K other loci where K can range from 0 to $N - 1$.

To describe NK landscapes more fully, for each $k = 0, 1, \dots, 2^N - 1$ let $\mathbf{x}_k \in (\{0, 1\})^N$ represent k written in base 2 and represented as a (binary) vector of length N . For $i = 1, \dots, N$, let \mathbf{w}_i denote a $2^{K+1} \times 1$ vector of independent random weights, where each component has mean μ/N and variance $\sigma^2/N > 0$ and otherwise arbitrary probability distribution. Let $\mathbf{f}_i(\mathbf{x})$ denote a function from $(\{0, 1\})^N \rightarrow E^{2^{K+1}}$ where $E^{2^{K+1}}$ denotes unit coordinate vectors in $\mathbb{R}^{2^{K+1}}$, i.e. vectors of the form $(0, 0, 1, 0, \dots, 0)$. The definition of the function $\mathbf{f}_i(\mathbf{x})$ depends on the epistatic interactions to the i th locus. The function is defined explicitly below.

For $k = 0, 1, \dots, 2^N - 1$, define $p_k = \sum_{i=1}^N \mathbf{f}_i(\mathbf{x}_k) \mathbf{w}_i$. Note that p_k is the landscape fitness at location \mathbf{x}_k and that $E[p_k] = \mu$ and $\text{Var}(p_k) = \sigma^2$ as each p_k is comprised of a sum of N independent weights. The NK landscape is defined by the fitness-location pairs (p_k, \mathbf{x}_k) coupled with Hamming distance.

B. Matrix representation of generalized NK landscapes

We begin by providing an explicit description for the construction and representation of NK and generalized NK landscapes as linear models in matrix form. For $i = 1, \dots, N$, let $V_i = \{i_1, i_2, \dots, i_{K_i+1}\}$ where $\{i_1, i_2, \dots, i_{K_i+1}\} \subset \{1, 2, \dots, N\}$ and $i_1 < i_2 < \dots < i_{K_i+1}$ with $i_j = i$ for some j . V_i denotes the i th interaction set, comprised of locus i and K_i loci that interact with locus i . Note that K isn’t restricted to be constant across loci, and there are no restrictions on the number of times locus i can appear in the interaction sets. This represents a generalization of NK landscapes, with the classic NK landscape occurring as the special case when $K_i \equiv K$ and each locus appears in exactly $K + 1$ interaction sets. See [1], [2] and [21] for similar generalizations. An additional generalization would be to not require locus i to be a member of V_i . The results in this article would still hold for this generalization.

With a slight abuse of notation, for each i write $\mathbf{f}_i(\mathbf{x}) = \mathbf{f}_i(x_{i_1}, \dots, x_{i_{K_i+1}})$ where $x_{i_1}, \dots, x_{i_{K_i+1}}$ are the elements of \mathbf{x} corresponding to V_i . Thinking of $\{x_{i_1}, \dots, x_{i_{K_i+1}}\}$ as a binary number, let $e_i(x_{i_1}, \dots, x_{i_{K_i+1}})$ be the decimal representation of this number. Define the $1 \times 2^{K_i+1}$ vector $\mathbf{f}_i(x_{i_1}, \dots, x_{i_{K_i+1}}) = (0, \dots, 0, 1, 0, \dots, 0)$ where the 1 occurs in column $e_i(x_{i_1}, \dots, x_{i_{K_i+1}}) + 1$. Alternatively, if $I_{2^{K_i+1}}$ is the identity matrix of size 2^{K_i+1} , then $\mathbf{f}_i(x_{i_1}, \dots, x_{i_{K_i+1}})$ is the $(e_i(x_{i_1}, \dots, x_{i_{K_i+1}}) + 1)$ th row of $I_{2^{K_i+1}}$.

To write the generalized NK model in matrix form, consider the $1 \times C$ vector $\mathbf{f}(\mathbf{x}) = \mathbf{f}_1(\mathbf{x}) \mid \mathbf{f}_2(\mathbf{x}) \mid \dots \mid \mathbf{f}_N(\mathbf{x})$ where $C = \sum_{i=1}^N 2^{K_i+1}$ and \mid denotes column concatenation. Define the $2^N \times C$ model matrix

$$F = \begin{pmatrix} \mathbf{f}(\mathbf{x}_0) \\ \mathbf{f}(\mathbf{x}_1) \\ \vdots \\ \mathbf{f}(\mathbf{x}_{2^N-1}) \end{pmatrix}. \quad (1)$$

With this definition, the generalized NK landscape is $\mathbf{p} = F\mathbf{w}$ where $\mathbf{w} = (\mathbf{w}_1^T \mid \dots \mid \mathbf{w}_N^T)^T$ denotes the $C \times 1$ vector of independent random weights.

Example: To illustrate the definition of \mathbf{f}_i and the matrix F , suppose $N = 3$, $K_i = 1$ for $i = 1, 2, 3$ and $V_1 = \{1, 2\}$, $V_2 = \{2, 3\}$ and $V_3 = \{1, 3\}$. A generic location on the landscape is represented by $\mathbf{x} = (x_1, x_2, x_3)$ and the 2^3 locations are $\{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_7\} = \{(0, 0, 0), (0, 0, 1), (0, 1, 0), \dots, (1, 1, 1)\}$. Then

$$\mathbf{f}_1(x_1, x_2) = \begin{cases} (1, 0, 0, 0) & \text{if } x_1 = 0, x_2 = 0; \\ (0, 1, 0, 0) & \text{if } x_1 = 0, x_2 = 1; \\ (0, 0, 1, 0) & \text{if } x_1 = 1, x_2 = 0; \\ (0, 0, 0, 1) & \text{if } x_1 = 1, x_2 = 1. \end{cases}$$

The functions $\mathbf{f}_2(x_2, x_3)$ and $\mathbf{f}_3(x_1, x_3)$ are defined similarly. The matrix F is then

$$F = \begin{pmatrix} \mathbf{f}(\mathbf{x}_0) \\ \mathbf{f}(\mathbf{x}_1) \\ \vdots \\ \mathbf{f}(\mathbf{x}_7) \end{pmatrix} = \begin{pmatrix} \mathbf{f}_1(0, 0) & \mathbf{f}_2(0, 0) & \mathbf{f}_3(0, 0) \\ \mathbf{f}_1(0, 0) & \mathbf{f}_2(0, 1) & \mathbf{f}_3(0, 1) \\ \mathbf{f}_1(0, 1) & \mathbf{f}_2(1, 0) & \mathbf{f}_3(0, 0) \\ \vdots & \vdots & \vdots \\ \mathbf{f}_1(1, 1) & \mathbf{f}_2(1, 1) & \mathbf{f}_3(1, 1) \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

To our knowledge, NK landscapes have never been formalized using the matrix representation given here. The rank of F is a measure of the richness of the landscape, as it gives the dimension of the domain for the vector \mathbf{p} . We will show that the rank of F is determined by N , $\{K_i\}_{i=1}^N$ and the structure of the interaction sets $\{V_i\}_{i=1}^N$, and that using rank as a measure of complexity provides refinement beyond using only N and $\{K_i\}_{i=1}^N$.

C. Interaction Model

Here we define the general form of parametric interaction models, and in the next section relate them to NK landscapes. Statisticians have long employed interaction models to study the effects of multiple ‘treatments’ on an outcome, see for example [14]. Interaction models are straightforward to define and model parameters have meaningful interpretations. In the evolutionary computing literature, these models seem to have received little attention, with the exception of [16],[17] and [18].

To define a general interaction model, for $l = 1, \dots, L$, let $U_l \subset \{1, 2, \dots, N\}$ and define $\tilde{x}_i = 2x_i - 1$ for $i = 1, \dots, N$ and where as before $x_i \in \{0, 1\}$. Mathematical properties of the interaction model and interpretation of model parameters

are most easily obtained using the transformed values $\tilde{x}_i \in \{-1, 1\}$.

The general form of an interaction model with L terms is

$$q(\tilde{\mathbf{x}}) = \sum_{l=1}^L \beta_{U_l} \prod_{j \in U_l} \tilde{x}_j$$

where the β 's are coefficients that can take any value in \mathbb{R} and where we conform with the convention that when $U_l = \emptyset$, $\prod_{j \in U_l} \tilde{x}_j \equiv 1$.

Example: Consider a model with $N = 3$ loci and $U_1 = \emptyset$, $U_2 = \{1\}$, $U_3 = \{2\}$, $U_4 = \{3\}$, $U_5 = \{1, 2\}$, $U_6 = \{1, 3\}$, $U_7 = \{2, 3\}$. The interaction model is $q(\tilde{\mathbf{x}}) = \beta_0 + \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \beta_3 \tilde{x}_3 + \beta_{12} \tilde{x}_1 \tilde{x}_2 + \beta_{13} \tilde{x}_1 \tilde{x}_3 + \beta_{23} \tilde{x}_2 \tilde{x}_3$. To simplify notation, we write β_1 instead of $\beta_{\{1\}}$ etc. In this example, β_1, β_2 and β_3 are main effects coefficients while β_{12}, β_{13} and β_{23} are two-loci interaction coefficients.

The intercept (β_0) and coefficients of the main effects and interaction terms have meaningful interpretations. The intercept coefficient represents the average of the fitness values across the entire landscape. For the general interaction model, it is not difficult to show that the main effect $2\beta_i$ represents the difference in fitness values when locus i is varied between $\tilde{x}_i = 1$ and $\tilde{x}_i = -1$, averaged over the values of the other $N - 1$ loci. $4\beta_{ij}$ represents the difference of differences between fitness values for $\tilde{x}_i = 1$ and $\tilde{x}_i = -1$ when $\tilde{x}_j = 1$ and $\tilde{x}_j = -1$, averaged over all other \tilde{x}_k for $k \neq i, j$. In general, higher order interaction coefficients are interpreted as average differences between lower order interactions. For example, $8\beta_{ijk}$ represents the average difference in the two factor interaction between loci i and j when $\tilde{x}_k = 1$ and $\tilde{x}_k = -1$.

III. GENERALIZED NK LANDSCAPES AS INTERACTION MODELS

Here we show that generalized NK landscapes can be expressed as linear interaction models, and that the interactions that are included in the model are completely determined by the interaction sets V_i , $i = 1, \dots, N$. We also derive the statistical properties of the interaction model coefficients and show how to construct classic NK landscapes that maximize the number of interaction terms.

Let 2^{V_i} denote the power set for V_i , i.e. the set V_i and all its subsets, including the empty set. Let $T = \bigcup_{i=1}^N 2^{V_i}$, and consider the interaction model

$$p(\tilde{\mathbf{x}}) = \sum_{U \in T} \beta_U \prod_{j \in U} \tilde{x}_j. \quad (2)$$

Evaluated at the 2^N transformed values $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{2^N}$, the model can be written in matrix notation as $\mathbf{p} = \tilde{F}\beta$ where \tilde{F} is an appropriately defined $2^N \times L$ matrix, L is the number of elements in T and β is an $L \times 1$ vector of coefficients. The random vector β of main effects and interactions has distributional properties dependent on the probability distribution of the vector of weights \mathbf{w} and the structure of the interaction sets V_i . The distributional properties of β are studied in section III-A.

Note that the model defined in (2) always contains an intercept and all main effects terms.

Example: Consider again the example in Section II-B, where $N = 3$, $K_i = 1$ for $i = 1, 2, 3$ and $V_1 = \{1, 2\}$, $V_2 = \{2, 3\}$ and $V_3 = \{1, 3\}$. Then $T = (\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\})$ and

$$\tilde{F} = \begin{pmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_{12} \\ \beta_{23} \\ \beta_{13} \end{pmatrix}.$$

The first column of \tilde{F} corresponds to the intercept β_0 , columns two through four are for the main effects, and columns five through seven are for the two loci interactions. Column five, for example, corresponds to the interaction between loci 1 and 2, and is obtained by taking the product of the elements of columns two and three, which correspond to the main effects for loci 1 and 2.

The following proposition establishes that a generalized NK landscape defined by interaction sets V_1, \dots, V_N is equivalent to the interaction model given by (2), thereby clearly establishing the nonzero interaction effects induced by the NK algorithm. The proof is given in the appendix.

Proposition 1. *Let \mathbf{F} denote the model matrix for the generalized NK landscape model defined in (1), and $\tilde{\mathbf{F}}$ the model matrix for the interaction model defined in (2). Then $\mathcal{C}(\mathbf{F}) = \mathcal{C}(\tilde{\mathbf{F}})$ where $\mathcal{C}(\cdot)$ denotes column space.*

Equation (2) shows that the NK algorithm constructs an interaction model in an interesting way. Note that the NK algorithm dictates that the interaction model contain all sub-interactions contained in higher order interactions. For example, if an NK landscape has a fourth order interaction between loci $\{1, 2, 3, 4\}$, then it also has the four third-order and six second-order interactions defined by all subsets of size three and two formed from the four interacting loci.

However, knowing which interactions have non-zero coefficients is not sufficient to fully understand the structure of NK landscapes. A complete understanding requires knowing the statistical properties of the main effect and interaction coefficients.

A. Induced properties of interaction model coefficients

While we have established that generalized NK landscapes can be represented by interaction models and shown how the epistatic interaction sets define the interaction terms included in the model, additional insight into landscape properties and a complete understanding of the representation requires knowledge of the distributional properties of the interaction model coefficients. This is addressed in the next proposition, with proof given in the appendix.

Proposition 2. *Consider a generalized NK landscape defined by interaction sets V_1, V_2, \dots, V_N and given by $\mathbf{p} =$*

$Fw = \tilde{F}\beta$ with weight vector \mathbf{w} where $E[w_i] = \mu/N$ and $\text{Var}[w_i] = \sigma^2/N$. For $U \in T$, let β_U denote the coefficient of the interaction term corresponding to U , and $I(\cdot)$ the indicator function, i.e.

$$I(U \in 2^{V_i}) = \begin{cases} 0 & \text{if } U \notin 2^{V_i}; \\ 1 & \text{if } U \in 2^{V_i}. \end{cases}$$

Then

$$E[\beta_\emptyset] = \mu$$

$$E[\beta_U] = 0, \quad \text{for } U \neq \emptyset$$

$$\text{Cov}[\beta_U, \beta_{U^*}] = 0, \quad \text{for } U \neq U^*$$

and

$$\text{Var}[\beta_U] = \frac{\sigma^2}{N} \sum_{i=1}^N 2^{-(K_i+1)} I(U \in 2^{V_i}).$$

Some remarks are in order. First, if the weights \mathbf{w} are normally distributed, then the interaction model coefficients, which are linear functions of \mathbf{w} , are independent and normally distributed with the indicated means and variances. For other distributions on \mathbf{w} , the exact distribution of the coefficients is an often intractable convolution problem. However, as both N and K increase, the central limit theorem can be invoked to show that the coefficients will be approximately normally distributed (contrast with approximate normality of the fitnesses, which only requires N large). Regardless of the distribution of \mathbf{w} , the coefficients are uncorrelated.

For \mathbf{w} normally distributed, $E[|\beta_U|] = \sqrt{\text{Var}[\beta_U]2/\pi}$, i.e. the expected magnitude of β_U is proportional to its standard deviation. More generally, the variance represents the expected magnitude of the square of the coefficient ($\text{Var}[\beta_U] = E[\beta_U^2]$ for $U \neq \emptyset$). With this perspective, it is interesting to note how the NK algorithm assigns importance (magnitude) to the interaction terms. The expected squared magnitude of β_U depends on the frequency with which $U \in 2^{V_i}$. Then for the classic NK model where K is constant and $K \ll N$, main effects have the largest expected magnitude, second order interactions would typically have larger expected magnitude than third order interactions and so on. On the other hand, when $K = N - 1$, all coefficients have the same expected magnitude because each power set 2^{V_i} contains main effects and interactions of all orders.

When the coefficients are normally distributed, independence of model coefficients means that, for example, knowing the magnitude of a two loci interaction provides no information on the magnitudes of the corresponding main effects of the loci—the NK algorithm assigns magnitudes of the coefficients completely independently.

The interaction model representation of a landscape with binary loci is equivalent to the representation given by Walsh functions [18]. While NK landscapes have been studied from the perspective of Walsh functions [8], the above results provide a transparent and explicit analysis of the Walsh coefficients for generalized NK landscapes, showing exactly which Walsh coefficients are nonzero and the statistical properties of these coefficients. For example, note that Theorems 4 and 5 in [8] follow immediately from the representation given in

equation (2).

An interesting property of generalized NK landscapes gleaned from Propositions 1 and 2 is that two landscapes defined by different interaction sets can lead to landscapes with the same set of main effects and interactions, i.e. identical matrices \tilde{F} , but the variances of the coefficients of the landscapes can be different, suggesting that the properties of the resultant landscapes may also be different. Consider, for example, the two landscapes defined by the following interaction sets: $V_{1A} = \{1, 2, 3, 4\}$, $V_{2A} = \{2, 3\}$, $V_{3A} = \{1, 3\}$, $V_{4A} = \{1, 3, 4\}$, $V_{5A} = \{2, 5\}$ and $V_{1B} = \{1, 4\}$, $V_{2B} = \{1, 2, 3, 4\}$, $V_{3B} = \{3\}$, $V_{4B} = \{4\}$, $V_{5B} = \{2, 5\}$. Clearly $T_A = \bigcup_{i=1}^N 2^{V_{iA}} = \bigcup_{i=1}^N 2^{V_{iA}} = T_B$ so that the landscapes have identical column space, but by Proposition 2, the variances of the terms are not all identical. Consider, for example, the main effect for locus 1: From Proposition 2, $\text{Var}[\beta_{1A}] = 7\sigma^2/16N > \text{Var}[\beta_{1B}] = 5\sigma^2/16N$.

B. Rank of NK landscapes and maximal rank models

Here we show that the rank of the model matrix for an NK landscape is determined by the interaction sets $\{V_i\}_{i=1}^N$ that define the landscape. We derive the maximum achievable rank and describe how interaction sets can be constructed to maximize rank.

Proposition 3. $\text{rank}(\tilde{F}) = \mathcal{M}(T)$ where $\mathcal{M}(\cdot)$ denotes counting measure and where $T = \bigcup_{i=1}^N 2^{V_i}$.

Proof. The result follows immediately from noting that the complete set of linear and interaction terms of all orders, $\{1, x_1, x_2, \dots, x_N, x_1*x_2, \dots, x_{N-1}*x_N, x_1*x_2*x_3, \dots, x_1*x_2*\dots*x_N\}$ comprise a Hadamard matrix, a linearly independent set of (orthogonal) column vectors. \square

It follows from the proof of Proposition 3 that $\text{rank}(\tilde{F}) \equiv [\text{number of model interactions} - (N+1)]$, i.e. model rank is a one-to-one function of the total number of interactions in an NK landscape.

The next results give an upper bound on the rank of an NK landscape, and provide conditions under which maximal rank designs exist for classic NK landscapes.

Proposition 4. For the classic NK landscape (K constant across loci), $\max\{\text{rank}(\tilde{F})\} \leq \min\{2^N, N2^{K+1} + 1 - N(K+1)\}$ where the max is over all possible interaction sets $\{V_i\}_{i=1}^N$ defining the landscape for fixed values of N and K .

Proof. The maximum rank occurs when there are no overlaps in interactions between the inputs to the loci. In this case each locus will contribute $\sum_{j=2}^{K+1} \binom{K+1}{j} = 2^{K+1} - (K+2)$ interactions of order two or higher. Summing over the N loci and including the intercept and main effects terms gives the result. \square

Proposition 4 begs the question as to when maximal rank classic landscapes exist. Equation (2) shows that all classic NK landscapes contain a main effect corresponding to each loci. Then for both N and K fixed, maximizing rank is equivalent to maximizing the number of interactions in the model. From the proof of Proposition 4, it is evident that a

necessary and sufficient condition is that each set V_i contain unique pairs of loci, as this ensures no redundant two factor interactions, and by extension no redundant higher order interactions. We now give explicit cases when the maximum rank given in Proposition 4 above can be achieved. The proof of the following proposition is given in the appendix.

Proposition 5. *There exist classic NK landscapes of maximal rank for the following values of N and K .*

- 1) $K = 2$ and all $N \geq 7$.
- 2) $K = 3$ and all $N \geq 13$.
- 3) $K = 4$ and all $N \geq 21$ except possibly $N = 22$.
- 4) $K = 5$ and all $N \geq 31$ except possibly $N = 32, 33, 34$.
- 5) $K = 6$ and all $N \geq 51$.
- 6) $K = 7$ and $N = 57, 64, 67, 69$ and all $N \geq 71$.
- 7) $K = 8$ and $N = 73, 89$ and all $N \geq 91$.
- 8) $K = 9$ and $N = 91$ and all $N \geq 111$.

It is not difficult to extend Proposition 5 to larger values of K , and the approach for doing so is contained in the proof of the result. It is worth noting, and not difficult to prove, that a necessary condition for existence of maximal rank landscapes is that $N \geq K^2 + K + 1$, because otherwise at least one pair of loci must occur together in more than one set V_i .

Proposition 5 established the existence of maximum rank NK landscapes. We now discuss the construction of these landscapes, which is easily achieved through the use of what we term NK difference sets. The combinatorial theory underlying these sets is detailed in the proof of Proposition 5 given in the appendix. Table I provides the needed difference sets for up to $K = 9$.

To construct a maximal rank classic NK landscape, recall that an NK landscape is completely determined by its interaction sets $\{V_i\}_{i=1}^N$. Interaction sets resulting in maximal rank designs are defined by incrementing the elements in the difference sets by one (modulo N) until N sets have been defined. This process is illustrated in the following example.

Example: Consider constructing an NK landscape achieving maximum rank when $N = 7$ and $K = 2$. From Table I, the NK difference set is $\{0, 1, 3\}$. Then define $V_1 = \{1, 2, 4\}$, $V_2 = \{2, 3, 5\}$, $V_3 = \{3, 4, 6\}$, $V_4 = \{4, 5, 7\}$, $V_5 = \{5, 6, 1\}$, $V_6 = \{6, 7, 2\}$, $V_7 = \{7, 1, 3\}$. Note that increments are modulo N , i.e. increments exceeding N “wrap around”.

TABLE I: NK difference sets used for constructing maximum rank designs. The second column gives values of N such that maximal NK landscapes exist.

K	N	NK Difference Set
2	$N \geq 7$	$\{0, 1, 3\}$
3	$N \geq 13$	$\{0, 1, 4, 6\}$
4	$N \geq 23$	$\{0, 2, 7, 8, 11\}$
5	$N \geq 35$	$\{0, 1, 4, 10, 12, 17\}$
6	$N \geq 51$	$\{0, 1, 4, 10, 18, 23, 25\}$
7	$N \geq 71$	$\{0, 4, 5, 17, 19, 25, 28, 35\}$
8	$N \geq 91$	$\{0, 2, 10, 24, 25, 29, 36, 42, 45\}$
9	$N \geq 111$	$\{0, 1, 6, 10, 23, 26, 34, 41, 53, 55\}$

The design constructed in the example is shown in Fig-

ure 1a. The i th column of the figure gives the elements in V_i . Notice that no pair of loci appears together more than once across columns. The landscape resulting from this design has rank $N2^{K+1} + 1 - N(K+1) = 36$, resulting from an intercept, 7 main effects, 21 two factor interactions and 7 three factor interactions.

Contrast with Figure 1b which shows the design resulting from choosing adjacent loci for the epistatic interactions. This design has rank 29, resulting from an intercept, 7 main effects, 14 two factor interactions and 7 three factor interactions.

IV. NUMBER OF LOCAL OPTIMA

The number of local optima is perhaps the strongest measure of landscape ruggedness and search difficulty [10], [19], [22]. For fixed N , it has been established empirically that the number of local optima increase with K [5]. The result is not surprising as increasing K increases the landscape model rank by increasing both the number and order of interactions defining the landscape. An unexplored question is the association between the number of local optima and landscape rank when N and K are both fixed.

We begin by providing an analytic expression for the expected number of local optima on NK and generalized NK landscapes.

Proposition 6. *Consider a generalized NK landscape defined by interaction sets V_1, V_2, \dots, V_N and weight vector \mathbf{w} where the elements of \mathbf{w} are independent normal random variables with $E[w_i] = \mu/N$ and $\text{Var}[w_i] = \sigma^2/N$. Then the expected number of local optima on the landscape is given by $2^N \Phi(\mathbf{0}; \Sigma)$ where*

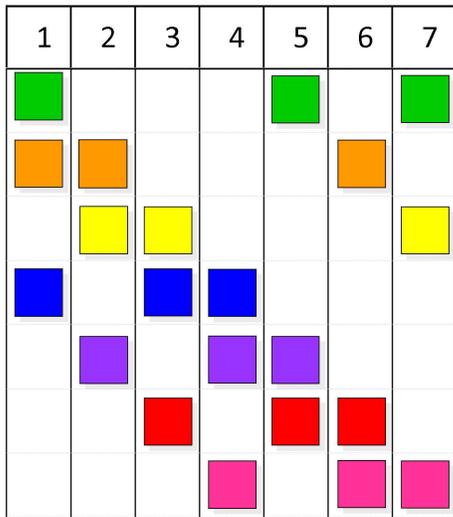
$$\Phi(\mathbf{0}; \Sigma) = \int_0^\infty \int_0^\infty \dots \int_0^\infty e^{-\frac{1}{2} \mathbf{z}^T \Sigma^{-1} \mathbf{z}} dz_1 dz_2 \dots dz_N$$

and where Σ is an $N \times N$ symmetric matrix with elements σ_{ij} given by

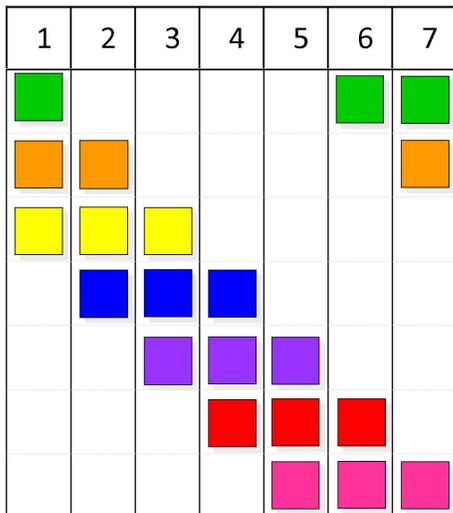
$$\sigma_{ij} = \begin{cases} \frac{2\sigma^2}{N} \sum_{k=1}^N I(i \in V_k) & \text{if } i = j; \\ \frac{\sigma^2}{N} \left(\sum_{k=1}^N I(i \in V_k) + \sum_{k=1}^N I(j \in V_k) \right. \\ \left. - \sum_{k=1}^N I(i \in V_k \text{ or } j \in V_k) \right) & \text{if } i \neq j. \end{cases}$$

Proof. Let p^* denote the fitness of a randomly selected location on the landscape, and let p_i^* denote the fitness one hamming distance away obtained by flipping the value of locus i . Let $\mathbf{z} = (p^* - p_1^*, p^* - p_2^*, \dots, p^* - p_N^*)$ denote the $N \times 1$ vector of fitness differences. Then the probability that the random location with fitness p^* is a local optima is given by the orthant probability $\Pr(\mathbf{z} > 0)$, and the expected number of local optima on the landscape is then $2^N \Pr(\mathbf{z} > 0)$. When \mathbf{w} is jointly normally distributed, it follows that \mathbf{z} is multivariate normal because it is a linear function of \mathbf{w} . The mean of \mathbf{z} is clearly zero, and the variance matrix is straightforward to derive and is given by Σ defined above. The derivation is omitted. The orthant probability $\Pr(\mathbf{z} > 0)$ is therefore given by $\Phi(\mathbf{0}; \Sigma)$. \square

Proposition 6 assumes \mathbf{w} is multivariate normal. The utility of this assumption is that the expected number of local optima



(a) Maximal Rank Design



(b) Adjacent Loci Design

Fig. 1: Interaction sets for maximal and adjacent loci designs.

then depends on the multivariate normal orthant probability $\Phi(\mathbf{0}; \Sigma)$, allowing us to take advantage of the extensive and decades long research on the numerical computation of multivariate normal probabilities (e.g. [7] and references therein). We are then able to estimate the number of local optima for very large landscapes without having to generate an actual landscape and check individually whether each location is a

local optimum. Note the normality assumption implies that the expected number of local optima is a function of only N and Σ .

For arbitrary probability distributions for the weights \mathbf{w} , computation of the orthant probability $\Pr(\mathbf{z} > \mathbf{0})$ is typically an intractable N dimensional integral. When the weights are non-normal and N, K are both large, the distribution of \mathbf{z} (see proof of Proposition 6) should be approximately normal by the central limit theorem, and normal orthant probabilities should then provide reasonable approximations to the expected number of local optima.

To examine the utility of the analytic expression representing the expected number of local optima for weights distributed other than normal, we performed a small simulation study. For $N = 15, 20$ and $K = 4$, we generated landscapes using interaction sets constructed from adjacent loci and where the random weights \mathbf{w} were generated from either the normal (to provide a baseline), uniform or chi-squared distribution on one degree of freedom (a highly skewed distribution). 100 landscapes were generated for the six combinations of N and weight distribution. For each landscape, we counted the number of local optima by brute force. Table II shows the means and standard deviations of the number of local optima and the expected number of local optima computed via the normal orthant probability. Of course when the weights are normally distributed, the empirical mean matches the expected number of local optima to within a small fraction of a standard deviation. The match was nearly as good for uniformly distributed weights. However, for the highly skewed chi-squared distribution, the expected number of local optima computed assuming normality well-overestimated the empirical mean. Simulations for other values of N and K produced similar results (not shown).

TABLE II: Empirical study of the number of local optima. Table entries are mean number of local optima with standard deviations in parentheses computed from 100 simulated landscapes with $K = 4$.

	Expected	Normal	Uniform	Chi-square
$N = 15$	121	121 (19)	123 (22)	92 (21)
$N = 20$	599	612 (141)	619 (141)	444 (100)

An empirical study was done to explore the relation between landscape rank and the number of local optima for both classic and generalized NK landscapes. For each combination of $N = 25, 50, 100$ and $K = 1, 2, \dots, 7$, we first generated interaction sets for 20 classic NK landscapes. For each value of N and K , maximal rank landscapes were generated when they were known to exist, and an adjacent loci interaction design was also generated. Additional classic NK landscapes were generated by randomly selecting an $N \times N$ latin square using the methods in [9] and then randomly selecting $K + 1$ rows from the latin square. The resulting N columns of $K + 1$ elements comprise the N interaction sets that define a classic NK landscape. For each landscape, the rank and the expected number of local optima were computed. Multivariate normal orthant probabilities representing the expected number of local

optima were computed using the "mvtnorm" package in the R computing environment, see [7].

As seen in Figure 2, there is a strong positive association between the expected number of local optima and landscape rank for fixed N and increasing K . To two decimal places, the correlations in Figure 2 for $N = 25, 50, 100$ are each 0.98 with respective 95% confidence intervals (0.97, 0.99); (0.97, 0.99); (0.97, 0.98). Figure 3 provides additional resolution with separate plots for $K = 3, 4, 5$ when $N = 50$, showing very strong *negative* correlation between rank and the number of local optima when N and K are both fixed. To two decimal places, the correlations for $K = 3, 4, 5$ are each -1.00 with the upper endpoints for 95% confidence intervals each ≤ -0.99 (computed on the non-log scale). Results for $N = 25$ and 100 were similar. Note that landscapes computed using adjacent loci to define interaction sets resulted in landscapes with significantly lower rank and larger expected number of local optima than randomly chosen landscapes.

The strong negative correlation seen in Figure 3 is perhaps surprising. Recall that landscape rank is equivalent to the number of terms in the interaction model representation, so that larger rank corresponds to more interaction terms. It is well known that landscapes with main effects but no interactions have only a single peak. It would seem that additional interaction terms would translate on average to more rugged landscapes, but this is not the case when N and K are both fixed.

This phenomenon is explained, at least partially, by noting that when Proposition 2 is applied to any maximal design for an NK landscape, it follows that main effects have variance $(K + 1)2^{-(K+1)}\sigma^2/N$ and interactions of all orders have variance $2^{-(K+1)}\sigma^2/N$ giving a ratio of $(K + 1)/1$, demonstrating that main effects can have considerably more influence than interactions in maximal designs. This observation would clearly extend to designs that are nearly maximal. Conversely, for an adjacent loci design, the variation is spread more equitably among main effects and interactions. For example, applying Proposition 2, the ratio of variances for a main effect and two factor interaction is $(K + 1)/K$ for the adjacent loci design.

Note also from Proposition 2 that for fixed N and K , the variance of the main effects and the *sum* of the variances of interaction coefficients is constant for all classic NK landscapes, i.e. these quantities do not vary regardless of how the interaction sets are defined. It follows that for fixed N and K , designs that increase the number of interaction terms must have a decreased average variance for the interaction terms. In other words, additional model complexity achieved through the addition of interaction terms is necessarily offset by a reduction in their expected magnitude.

Additional simulations were done to assess the relation between rank and number of local optima for *generalized* NK landscapes. To construct generalized NK landscapes, we kept the sizes of the interaction sets V_i constant for a given

landscape, but did not restrict the number of times that locus i could appear in the interaction sets V_j for $i \neq j$ (for classic NK landscapes, locus i appears in exactly K interaction sets V_j for $i \neq j$).

We randomly generated 20 generalized NK landscapes for each combination of $N = 25, 50, 100$ and $K = 1, 2, \dots, 7$. A strong positive association is again seen between rank and expected number of local optima as K varies with N fixed, see Figure 4. To two decimal places, the correlations and 95% confidence intervals in Figure 4 for $N = 25, 50, 100$ are respectively 0.99 (0.98, 0.99); 0.98 (0.98, 0.99); 0.98 (0.98, 0.99).

However, the additional resolution provided by Figure 5 shows a weak *positive* association between landscape rank and expected number of local optima. Much of the strength of the association is driven by the maximal design, and if the data point for the maximal design is left out, there is no statistically significant correlation between rank and expected number of local peaks for generalized NK landscapes for N and K both fixed. To two decimal places, the correlations for $K = 3, 4, 5$ in figure 5 are respectively (95% confidence intervals in parentheses) 0.57 (0.17, 0.81); 0.81 (0.57, 0.92); 0.71 (0.39, 0.88). An explanation for the correlation in Figure 5 is that for generalized NK landscapes the expected magnitude of main and interaction effects can vary, and it is possible to have expected magnitudes for a subset of interactions larger than those for a subset of main effects.

V. CONCLUSION

Representation of NK landscapes as linear models in matrix form provides a transparent connection to parametric linear interaction models, and provides a straightforward means for deriving the statistical properties of interaction model coefficients induced by the NK landscape algorithm. The interaction model representation coupled with distributional properties of model coefficients provides new insights into properties of NK landscapes. Expressing the expected number of local optima as a multivariate normal orthant probability provides additional insight into aspects of NK landscapes that affect the number of local optima, and also allows for quick computation of the expected number of local optima.

The gaps in the data points seen in the horizontal axes of Figure 4 and Figure 5 represent gaps in the ranks of NK landscape models and clearly illustrate that classic NK landscapes only represent a small subset of possible interaction designs, see also [8] for a similar observation. While our definition of generalized NK landscapes would fill in some of the gaps, the definition still requires that lower order interactions that are contained in higher order interactions appear in the model.

The representation of NK landscapes as interaction models immediately suggests a more general definition that allows for main effects and interactions of any order to appear together without restriction. Such landscapes could be constructed with any rank ranging from 1 to 2^N , and main effect and interaction

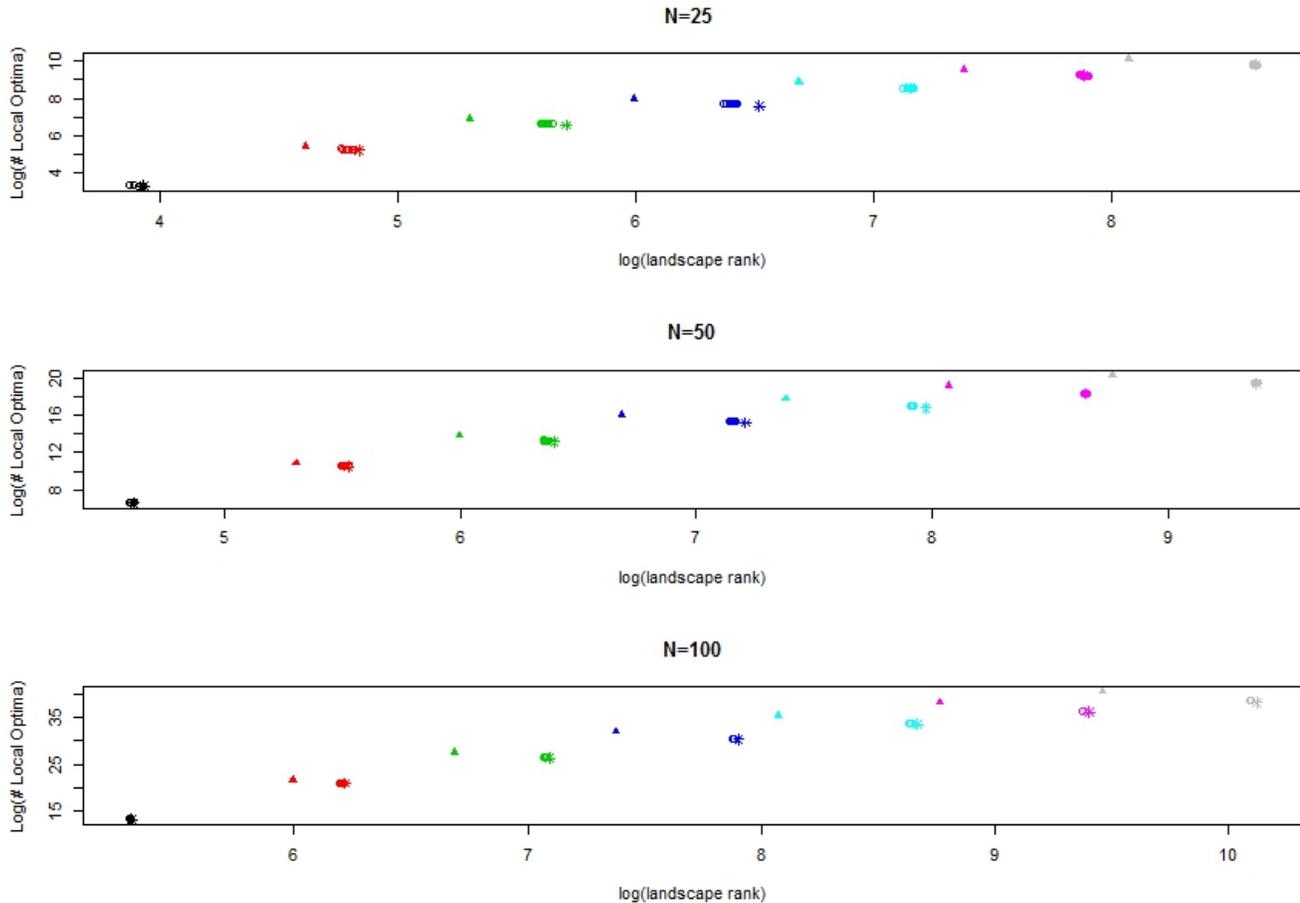


Fig. 2: Classic NK landscapes: Expected number of local optima versus landscape rank (on log-log scale) for $K=1$ to 7. $\star \equiv$ maximal rank design, $\blacktriangle \equiv$ adjacent loci design, $\circ \equiv$ random design. Colors indicate different values for K .

coefficients could be generated with arbitrary variances and correlation structures. In this vein, [16] defined a class of tunable landscapes allowing general interaction structures and studied the performance of GA on these landscapes.

Using arguments similar to those in Section IV, it is not difficult to show that for a general interaction model with normally distributed coefficients, the expected number of local optima is represented by a normal orthant probability that depends entirely on the variance/covariance matrix of the vector of fitness differences. However, it is not obvious how to select interactions and assign variances to coefficients that will result in landscapes with large numbers of local peaks, or more generally that are difficult to search. There are few results regarding the size of normal orthant probabilities as a function of properties of the variance/covariance matrix, though [20] provide potentially useful results in this context.

An analysis of properties of multiobjective NK landscapes, including the structure of Pareto optimal sets, is given in [23]. It seems possible to extend the methods for computing the expected number of local optima given in this article to find the probability that a landscape location is Pareto optimal in the multi-objective case.

Characterizing the search difficulty of landscapes is itself

difficult [15], and the number of local optima is an imperfect determinant of search difficulty [10], [15]. For example, the needle-in-a-haystack landscape, which can be represented as a parametric interaction model containing all possible interactions where the magnitudes of all main and interactions effects are equal, has equal fitness values at all locations with the exception of a single peak which can only be found by guessing. This illustrates that models with many interactions can exhibit a high degree of neutrality (changing location does not result in a change of fitness). While the maximal rank designs for fixed N and K had smaller expected number of local optima, they may possess other attributes that affect search difficulty.

An attractive feature of the NK algorithm is that a very rich set of landscapes can be generated with only two tuning parameters. Whether a procedure with comparable simplicity can be developed for constructing a more general class of interaction models that are difficult to search is an area for future research.

APPENDIX PROOF OF PROPOSITIONS 1, 2 AND 5

The following is a proof of Proposition 1.

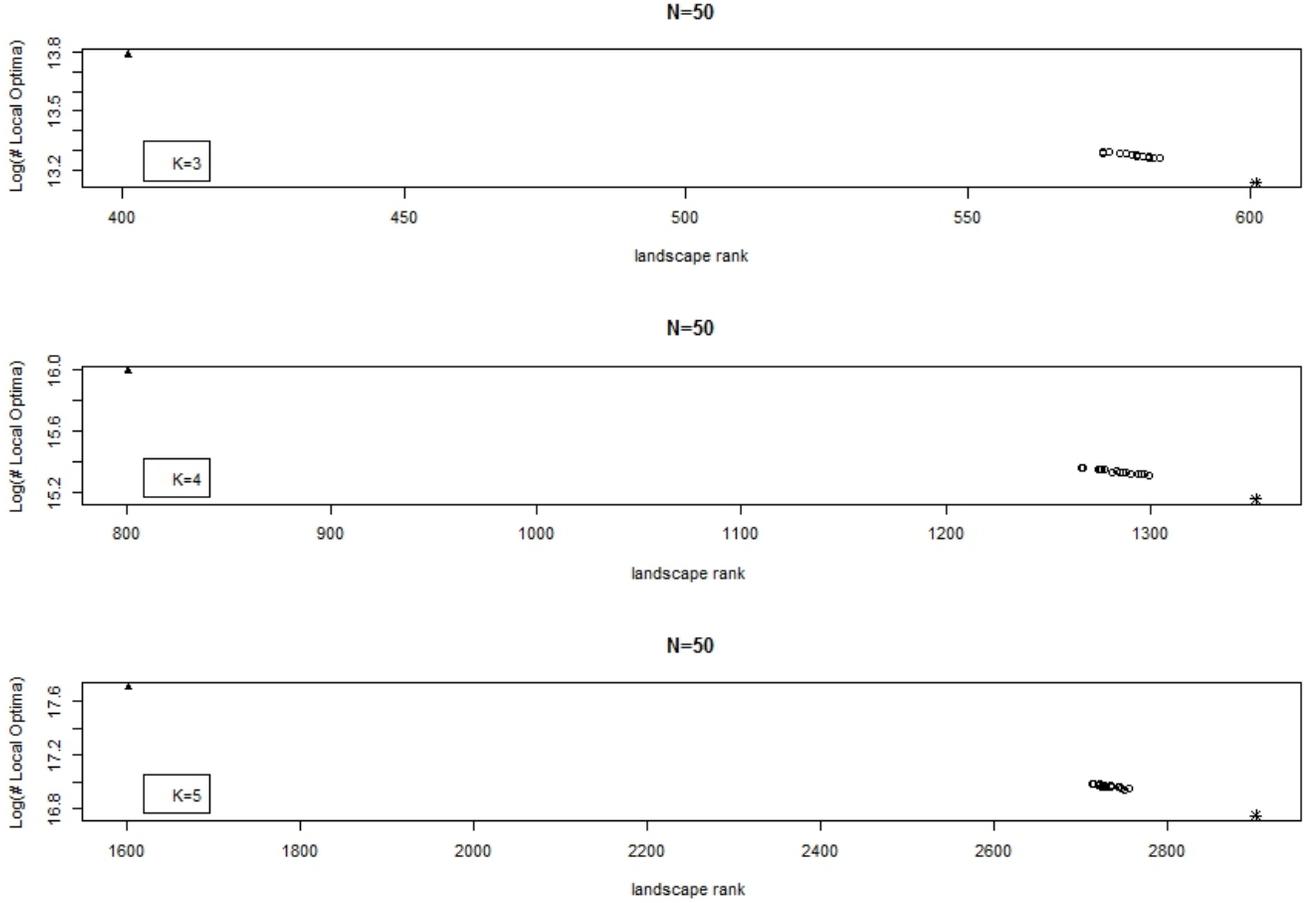


Fig. 3: Classic NK landscapes: Expected number of local optima (on log scale) versus landscape rank for $N = 50$, $K=3,4,5$. $\star \equiv$ maximal rank design, $\blacktriangle \equiv$ adjacent loci design, $\circ \equiv$ random design.

Proof. We first construct a matrix representation for the interaction model. For the i th locus, let $S_{i,1}(\bar{\mathbf{x}})$ denote the vector of covariates corresponding to the inputs for locus i . For example, if $V_1 = \{1, 4, 6, 9\}$, then $S_{1,1} = (x_1, x_4, x_6, x_9)$. Note that there are $K_i + 1$ elements in $S_{i,1}$. More generally, let $S_{i,m}(\bar{\mathbf{x}})$ denote the set of $\binom{K_i+1}{m}$ m -order interactions for the inputs of locus i . For $i = 1, \dots, N$, define the $1 \times 2^{K_i+1}$ vector $\tilde{\mathbf{f}}_i(\mathbf{x}) = (1, S_{i,1}, S_{i,2}, \dots, S_{i,(K_i+1)})$, and note that $\tilde{\mathbf{f}}_i(\cdot)$ can take 2^{K_i+1} different values.

Consider the $1 \times C$ vector $\tilde{\mathbf{f}}(\mathbf{x}) = \{\tilde{\mathbf{f}}_1(\mathbf{x}), \tilde{\mathbf{f}}_2(\mathbf{x}), \dots, \tilde{\mathbf{f}}_N(\mathbf{x})\}$ where $C = \sum_{i=1}^N 2^{K_i+1}$. Define the $2^N \times C$ model matrix

$$\tilde{F}^* = \begin{pmatrix} \tilde{\mathbf{f}}(\mathbf{x}_1) \\ \tilde{\mathbf{f}}(\mathbf{x}_2) \\ \vdots \\ \tilde{\mathbf{f}}(\mathbf{x}_{2^N}) \end{pmatrix}. \quad (3)$$

Note that $\tilde{F}^* \neq \tilde{F}$ but clearly $\mathcal{C}(\tilde{F}^*) = \mathcal{C}(\tilde{F})$ as \tilde{F}^* is comprised of all the columns in \tilde{F} with some columns repeated. \tilde{F}^* is an overparameterized version of \tilde{F} , and in the following we show that $\mathcal{C}(\tilde{F}^*) = \mathcal{C}(F)$.

The proof uses two basic results. First, suppose two matrices have the same column space. Choose l rows from the first

matrix and form a new matrix by expanding this matrix by row concatenating each of the l rows j times. The same operations are applied to the second matrix, that is the corresponding l rows from the second matrix are repeated j times. The expanded matrices are then easily seen to have the same column space. The second result is that if $\mathcal{C}(A_l) = \mathcal{C}(B_l)$ for $l = 1, \dots, N$ then $\mathcal{C}(A) = \mathcal{C}(B)$ where $A = A_1 | A_2 | \dots | A_N$ and $B = B_1 | B_2 | \dots | B_N$ and where $|$ denotes column concatenation.

Let $\mathbf{y}_i = (x_i, x_{i_1}, \dots, x_{i_{K_i}})$ and let $\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, \dots, \mathbf{y}_{i,2^{K_i+1}}$ be the binary ordering of \mathbf{y}_i . Define

$$\tilde{a}_i = \begin{pmatrix} \tilde{\mathbf{f}}_i(\mathbf{y}_{i,1}) \\ \tilde{\mathbf{f}}_i(\mathbf{y}_{i,2}) \\ \vdots \\ \tilde{\mathbf{f}}_i(\mathbf{y}_{i,2^{K_i+1}}) \end{pmatrix}, \quad a_i = \begin{pmatrix} \mathbf{f}_i(\mathbf{y}_{i,1}) \\ \mathbf{f}_i(\mathbf{y}_{i,2}) \\ \vdots \\ \mathbf{f}_i(\mathbf{y}_{i,2^{K_i+1}}) \end{pmatrix}.$$

Then $\mathcal{C}(a_i) = \mathcal{C}(\tilde{a}_i)$ as \tilde{a}_i is a $2^{K_i+1} \times 2^{K_i+1}$ full rank matrix, and a_i is the $2^{K_i+1} \times 2^{K_i+1}$ identity matrix. Next, let F_i and \tilde{F}_i^* denote columns $\sum_{l=1}^{i-1} 2^{K_l+1} + 1$ to $\sum_{l=1}^i 2^{K_l+1}$ of F and \tilde{F}^* respectively. Note F_i and \tilde{F}_i^* are obtained from a_i and \tilde{a}_i by repeating rows of these matrices. Then by the first result described at the start of the proof, $\mathcal{C}(F_i) = \mathcal{C}(\tilde{F}_i^*)$.

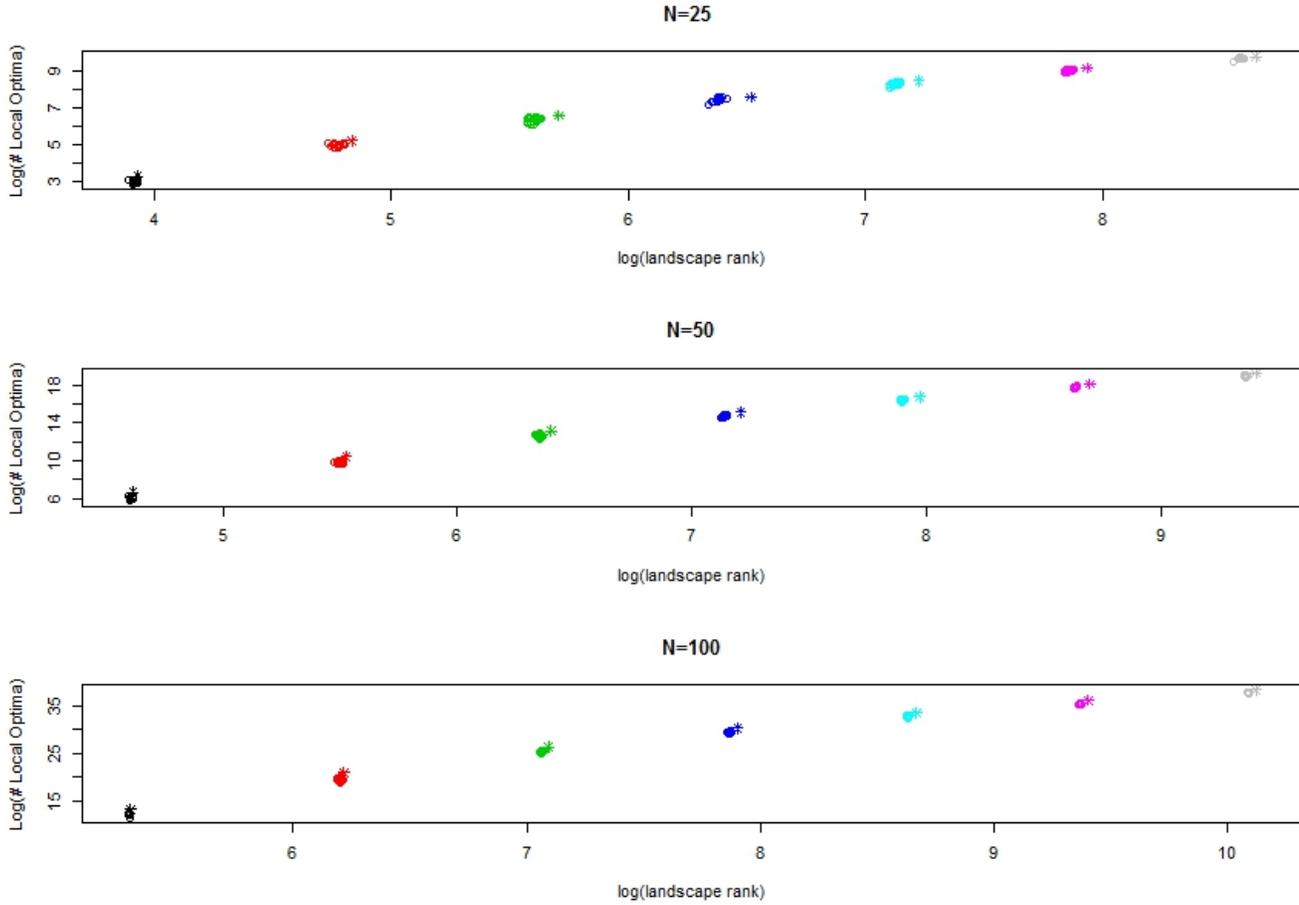


Fig. 4: Generalized NK landscapes: Expected number of local optima versus landscape rank (on log-log scale) for $K=1$ to 7. $\star \equiv$ maximal rank design, $\circ \equiv$ random design. Colors indicate different values for K .

Finally, the result follows from the fact that $\tilde{F}^* = \tilde{F}_1^* | \tilde{F}_2^* | \dots | \tilde{F}_N^*$ and $F = F_1 | F_2 | \dots | F_N$.

□

The following is the proof of Proposition 2.

Proof. Note that $2^N \beta_\emptyset = 1^T \tilde{F} \beta = 1^T F \mathbf{w}$ and the expectation of the RHS is easily seen to be $2^N \mu$ from which it follows that $E[\beta_\emptyset] = \mu$. Let $\tilde{\mathbf{F}}_{[k]}$ denote the k th column of \tilde{F} and $\beta_{U_{[k]}}$ the corresponding interaction model coefficient. Then for $k > 1$, $E[\beta_{U_{[k]}}] = E[\tilde{\mathbf{F}}_{[k]}^T \tilde{F} \beta] = E[\tilde{\mathbf{F}}_{[k]}^T F \mathbf{w}] = 0$ where the last equality follows from $1^T \tilde{\mathbf{F}}_{[k]} = 0$ whenever $k > 1$.

To obtain the variance/covariance results, note that $\tilde{F} \beta = F \mathbf{w}$ implies

$$\beta = 2^{-N} \tilde{F}^T F \mathbf{w}$$

where we use the fact that the columns of \tilde{F} are orthogonal, each with norm 2^N . Then

$$\text{Var}[\beta] = 2^{-2N} \tilde{F}^T F F^T \tilde{F} \times \text{Var}[\mathbf{w}] = \frac{\sigma^2 2^{-2N}}{N} \tilde{F}^T F F^T \tilde{F}.$$

To evaluate $\tilde{F}^T F$, let $\mathbf{f}_{i,j}$ represent the j th column of F_i where F_i represents the submatrix of F corresponding to the interaction set V_i (columns $\sum_{l=1}^{i-1} 2^{K_l+1} + 1$ to $\sum_{l=1}^i 2^{K_l+1}$

of F). Then it is not difficult to show that

$$\tilde{\mathbf{F}}_{[k]}^T \mathbf{f}_{i,j} = h(i, j, k) 2^{N-(K_i+1)} \quad (4)$$

where the function $h(i, j, k) \in \{-1, 1\}$ when $U_{[k]} \in 2^{V_i}$ and zero otherwise ($h(i, j, k)$ is defined explicitly below), and where $U_{[k]}$ denotes the interaction term corresponding to $\tilde{\mathbf{F}}_{[k]}$. Let \mathbf{g}_k^T denote the k th row of $\tilde{F}^T F$. Then it follows from the identity above that $\mathbf{g}_k^T \mathbf{g}_k = \sum_{i=1}^N 2^{K_i+1} (2^{N-(K_i+1)})^2 I(U_{[k]} \in 2^{V_k}) = \sum_{i=1}^N 2^{2N-(K_i+1)} I(U_{[k]} \in 2^{V_k})$, and then that

$$\begin{aligned} \text{Var}[\beta_{U_{[k]}}] &= \frac{\sigma^2}{N} \left(\frac{1}{2^N} \right)^2 \sum_{i=1}^N 2^{2N-(K_i+1)} I(U_{[k]} \in 2^{V_i}) \\ &= \frac{\sigma^2}{N} \sum_{i=1}^N 2^{-(K_i+1)} I(U_{[k]} \in 2^{V_i}). \end{aligned}$$

Finally, the covariance result will follow provided the rows of $\tilde{F}^T F$ are orthogonal. We first define the function $h(i, j, k)$ appearing in (4). For $j = 1, \dots, 2^{K_i+1}$, let $e_i^{-1}(j-1)$ denote the $1 \times 2^{K_i+1}$ vector representation of $j-1$ as a binary number. The elements of $e_i^{-1}(j-1)$, from left to right, correspond to values for $\{x_{i1}, x_{i2}, \dots, x_{iK_i+1}\}$, compare to

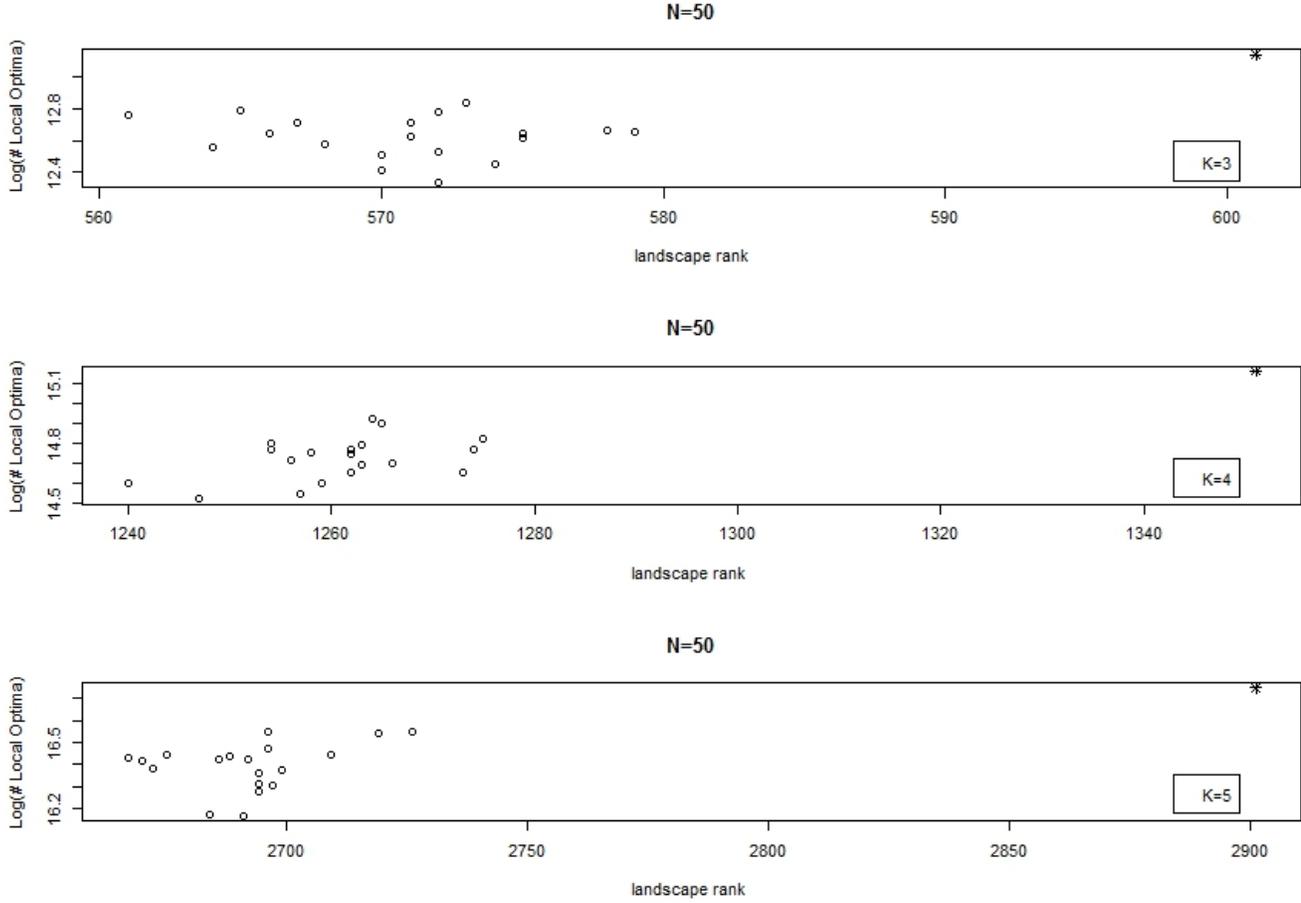


Fig. 5: Generalized NK landscapes: Expected number of local optima (on log scale) versus landscape rank for $N = 50$, $K=3,4,5$. $\star \equiv$ maximal rank design, $\circ \equiv$ random design.

the definition of $e_i(\cdot)$ in Section II-B. Define $h(i, j, k) = I(U_{[k]} \in 2^{V_i}) \prod_{U_{[k]}} 2(e_i^{-1}(j-1) - 1)$ where the right hand side represents the product of the values of the elements of $e_i^{-1}(j-1)$ corresponding to the loci in $U_{[k]}$. For $k=1$, we have $U_{[k]} = U_{[1]} = \emptyset$, and then define $h(i, j, 1) = 1$ for all i, j .

Example: Consider again the example in Section II-B, where $N = 3$, $K_i = 1$ for $i = 1, 2, 3$ and $V_1 = \{1, 2\}$, $V_2 = \{2, 3\}$ and $V_3 = \{1, 3\}$. The interaction model is $p = \beta_0 + \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \beta_3 \tilde{x}_3 + \beta_{12} \tilde{x}_1 \tilde{x}_2 + \beta_{13} \tilde{x}_1 \tilde{x}_3 + \beta_{23} \tilde{x}_2 \tilde{x}_3$.

For $i = 1$, $j = 3$, $e_1^{-1}(3-1) = (1, 0)$. Then for $k = 2$, $U_{[2]} = \{1\}$ and $h(1, 3, 2) = 1$. For $k = 3$, $U_{[3]} = \{2\}$ and $h(1, 3, 3) = -1$. For $k = 5$, $U_{[5]} = \{1, 2\}$ and $h(1, 3, 5) = 1 \times -1 = -1$. For $k = 6$, $U_{[6]} = \{1, 3\}$, $U_{[6]} \notin 2^{V_1}$ and $h(1, j, 6) = 0$ for $j = 1$ to 4.

The inner product of row k and l of $\tilde{F}^T F$ is

$$\sum_{i=1}^N \sum_{j=1}^{2^{K_i+1}} \tilde{\mathbf{F}}_{[k]}^T \mathbf{f}_{i,j} \tilde{\mathbf{F}}_{[l]}^T \mathbf{f}_{i,j} = \sum_{i=1}^N 2^{2N-2(K_i+1)} \sum_{j=1}^{2^{K_i+1}} h(i, j, k) h(i, j, l).$$

Note that $\sum_{j=1}^{2^{K_i+1}} h(i, j, k) h(i, j, l) = 0$ for all i follows from orthogonality of $\tilde{\mathbf{F}}_{[k]}$ and $\tilde{\mathbf{F}}_{[l]}$. Therefore the rows of $\tilde{F}^T F$ are orthogonal, and it follows that $\text{Cov}(\beta_U, \beta_{U^*}) = 0$ for $U \neq U^*$. \square

The following is a proof of Proposition 5. In proving Proposition 5, we also define and demonstrate the properties of the NK difference sets used to construct maximum rank designs.

We begin with the notion of a *packing design* or just *packing* for short. These are variants of well-known objects from combinatorial design theory, see [3]; for completeness we give the definition here.

A (N, κ) -*packing design*, consists of a set S on N elements (called "points") and a collection of subsets of S (called the "blocks") all of size $\kappa = K + 1$ with the property that each pair of points in S is contained in at most one of the blocks. To obtain an NK landscape of maximum rank, one can use the points and the blocks of a (N, κ) packing design to construct the N interaction sets defining the landscape. The next theorem gives this connection.

Proposition 7. *If there exists an (N, κ) packing design, then there exists a classic NK landscape with N loci and $K =$*

$\kappa - 1$ which achieves maximum rank.

Proof. As noted in the proof of Proposition 4, maximal rank designs will occur when each pair of loci occurs at most once. The existence of a $(N, \kappa = K + 1)$ packing design ensures that there are no redundant two factor interactions in the NK interaction set specification, and by extension no redundant interactions of order greater than two. Hence a NK landscape of maximal possible rank will result. \square

In order to construct NK landscapes via Proposition 7 we must construct (N, κ) packing designs. Constructing (N, κ) packing designs is made easier by employing so called difference methods from combinatorial design theory. Let \mathbb{Z}_n be the integers modulo n , so $\mathbb{Z}_n = \{0, 1, 2, \dots, n-1\}$ with addition modulo n . Define an (N, κ) difference set in \mathbb{Z}_N to be a subset $D = \{x_1, x_2, \dots, x_\kappa\} \subseteq \mathbb{Z}_N$ of size κ with the property that the list of all differences $x_i - x_j$ (where $x_i, x_j \in D$ and $i \neq j$) contains each nonzero element at most one time.

An example of a $(8, 3)$ difference set is $D = \{0, 1, 3\}$. Note that in \mathbb{Z}_8 , $1 - 0 = 1$, $0 - 1 = 7$, $3 - 0 = 3$, $0 - 3 = 5$, $3 - 1 = 2$, and $1 - 3 = 6$ and hence all of the differences are different as required.

We can use a (N, κ) difference set in \mathbb{Z}_N to construct an (N, κ) packing design and thus via Proposition 7 we will have constructed a classic NK landscape with N loci and $K = \kappa - 1$ which achieves maximum rank. The construction is given in the next proposition.

Proposition 8. *If there exists an (N, κ) -difference set in \mathbb{Z}_N , then there exists an (N, κ) packing design.*

Proof. Let $D = \{x_1, x_2, \dots, x_\kappa\}$ be an (N, κ) difference set in \mathbb{Z}_N . For $g \in \mathbb{Z}_N$ define $D+g = \{x_1+g, x_2+g, \dots, x_\kappa+g\}$. $D+g$ is called a translate of D . We claim the set of all translates of D (namely $\{D+g : g \in \mathbb{Z}_N\}$) form the blocks of an (N, κ) packing design. First note that there are exactly N of these translates (one for each element in \mathbb{Z}_N) and also note that there are N points (the elements of \mathbb{Z}_N). We must show that any pair of elements $a, b \in \mathbb{Z}_N$ that a and b are in at most one of the translates of D . Assume to the contrary that the pair $\{a, b\}$ occurs in two of the translates, say $\{a, b\} \subseteq D+p$ and $\{a, b\} \subseteq D+q$ where $p \neq q$ are both elements of \mathbb{Z}_N . Then $a = x_i + p$ and $b = x_j + p$ for some i and j and also $a = x_k + q$ and $b = x_m + q$ for some k and m . Hence $a - b = x_i - x_j = x_k - x_m$. From the difference property of D this implies that $i = k$ and $j = m$ which in turn says that $p = q$, a contradiction. Hence we have shown that the n translates of an (N, κ) difference set in \mathbb{Z}_N give the blocks of an (N, κ) packing design. \square

ACKNOWLEDGMENT

The authors would like to thank Stuart Kauffman, Jeffrey Horbar and Maggie Eppstein for discussions motivating this work.

REFERENCES

[1] L. Altenberg 1995. Genome growth and the evolution of the genotype-phenotype map. In *Evolution and Biocomputation: Computational Models of Evolution*, ed. Wolfgang Banzhaf and Frank H. Eeckman. Lecture Notes in Computer Science vol. 899. Springer-Verlag, pp. 205-259.

[2] L. Altenberg 1996. NK Fitness Landscapes. In *Handbook of Evolutionary Computation*, ed. T. Back, D. Fogel, and Z. Michalewicz. Institute of Physics Press.

[3] C. Colbourn and J. Dinitz, *Handbook of Combinatorial Designs*, 2nd ed. Chapman & Hall, USA, 2007.

[4] M. Eppstein, J. Horbar, J. Buzas, and S. Kauffman, "Searching the Clinical Fitness Landscape". *PLOS | one*, November, Vol. 7, Issue 11, e49901. 2012.

[5] A. Eremeev, and C.R. Reeves, On confidence Intervals for the Number of Local Optima. *EvoWorkshops LNCS 2611*, Springer-Verlag, Berlin, pp.224-235 (2003).

[6] S. Evans and D. Steinsaltz, Estimating Some Features of NK Fitness Landscapes *The Annals of Applied Probability*, Vol. 12, No. 4 (Nov., 2002), pp. 1299-1321.

[7] A. Genz and F. Bretz, *Computation of Multivariate Normal and t Probabilities*, Lecture Notes in Statistics 195, Springer-Verlag, New York 2009.

[8] R. Heckendorn and D. Whitley, (1997). A Walsh analysis of NK landscapes. In Bäck, T. editor, *Proceedings of the Seventh International Conference on Genetic Algorithms*, pages 41-48. Morgan Kaufmann.

[9] M. Jacobson and P. Matthews, Generating uniformly distributed random Latin squares, *J. Combinatorial Design* 4 (1996), 405-437.

[10] L. Kallel, B. Naudts, and C. Reeves, Properties of Fitness Functions and Search Landscapes. In *Theoretical Aspects of Evolutionary Computing*, Eds. L. Kallel, B. Naudts, and A. Rogers. Springer, New York (2001).

[11] S. Kauffman, *Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press (1993).

[12] S. Kauffman, E. Weinberger, The NK model of rugged fitness landscapes and its application to maturation of the immune response. *J Theor Biol.* 1989 Nov 21;141(2):211-45 (1989).

[13] H. Kaul and S. Jacobson, Global Optima Results for the Kauffman NK Model. *Mathematical Programming*, Volume 106, 2006, 319-338 (2006).

[14] D. Montgomery, *Design and Analysis of Experiments*, 7th edition. Wiley, New York (2008).

[15] B. Naudts and L. Kallel, "A comparison of predictive measures of problem difficulty in evolutionary algorithms," *IEEE Transactions on Evolutionary Computation*, vol.4, no.1, pp.1-15, Apr 2000.

[16] C. Reeves (2000) Experiments with Tuneable Fitness Landscapes. In M.Schoenauer et al. (Eds.) *Parallel Problem Solving from Nature VI*, LNCS 1917, Springer, Berlin, pp.139-148.

[17] C. Reeves and C. Wright, Epistasis in Genetic Algorithms: An Experimental Design Perspective ,*Proceedings of the 6th International Conference on Genetic Algorithms* ,217-224, 1995.

[18] C. Reeves and C. Wright An Experimental Design Perspective on Genetic Algorithms, *Foundations of Genetic Algorithms* , 7-22, 1995.

[19] C. Reidys and P. Stadler, *Combinatorial Landscapes*, *SIAM Review*, 44, 3-54 (2002).

[20] Y. Rinott and T. Santner, An Inequality for Multivariate Normal Probabilities with Application to a Design Problem. *Ann. Statist.* Volume 5, Number 6, 1228-1234 (1977).

[21] R. Smith and J. Smith, New Methods for Tunable, Random Landscapes, *Foundations of Genetic Algorithms* 6, pp.47-67, Morgan Kaufmann, (2001).

[22] S. Verel, G. Ochoa, and M. Tomassini, Local Optima Networks of NK Landscapes With Neutrality. *IEEE Transactions on Evolutionary Computation*, Volume: 15, Issue: 6, Page(s): 783 - 797 (2011).

[23] S. Verel, A. Liefvooghe, L. Jourdan, and C. Dhaenens, On the structure of multiobjective combinatorial search space: MNK -landscapes with correlated objectives, *European Journal of Operational Research*, in press (2013).

Jeffrey Buzas Jeff Buzas is Professor of Mathematics and Statistics and Director of the Statistics Program at the University of Vermont. Dr. Buzas received his Ph.D. in statistics at North Carolina State University in 1993. His recent interest in evolutionary computation was the result of collaborations with computer scientists and physicians on how to model the search for better medical practice and treatments as a search on fitness landscapes.

Jeffrey Dinitz Jeff Dinitz is Professor of Mathematics and Statistics at the University of Vermont and holds a secondary appointment in the Department of Computer Science. He was Chair of the Department of Mathematics and Statistics from 1998 - 2004 and Chair of the Computer Science Department from 2010 - 2012. Dr. Dinitz received his Ph.D. in mathematics at The Ohio State University in 1980 with a specialty in combinatorial designs. Since 1980, Dr. Dinitz has published extensively in this area and in combinatorics in general with over 90 publications in refereed research journals. Dr. Dinitz is the managing editor-in-chief of the Journal of Combinatorial Designs and is also one of two coeditors of the Handbook of Combinatorial Designs.