Relation between the number of peaks and the number of reciprocal sign epistatic interactions

Raimundo Saona[1], Fyodor A. Kondrashov[1], Ksenia A. Khudiakova[1*]

Abstract

Empirical assays of fitness landscapes suggest that they may be rugged, that is having multiple fitness peaks. Such fitness landscapes, those that have multiple peaks, necessarily have special local structures, called reciprocal sign epistasis (Poelwijk et al. (2011)). Here, we investigate the quantitative relationship between the number of fitness peaks and the number of reciprocal sign epistatic interactions. Previously it has been shown (Poelwijk et al. (2011)) that pairwise reciprocal sign epistasis is the necessary but not sufficient condition for the existence of multiple peaks. Applying discrete Morse theory, which to our knowledge has never been used in this context, we extend this result by giving the minimal number of reciprocal sign epistatic interactions required to create a given number of peaks.

Keywords

Fitness landscapes; Multiple peaks; Morse theory; Reciprocal sign epistasis

Competing Interests

Declarations of interest: none.

Mathematics subject classification

92D15

1 Introduction

The fitness landscape is the relationship between genotypes and their fitness. Availability of high throughput methods and next generation sequencing started to experimentally characterize aspects of different fitness landscapes. Due to the enormity of the underlying genotype space (Wright (1932); Maynard Smith (1970)), the experimental approaches are limited to assaying fitness of: (a) closely

[1]Institute of Science and Technology Austria, 1 Am Campus, Klosterneuburg, 3400, Austria
*corresponding author; kseniia.khudiakova@ist.ac.at

related genotypes (Sarkisyan et al. (2016); Melamed et al. (2013); Romero and Arnold (2009); de Visser and Krug (2014)); or (b), very restricted genotype spaces such as the interaction of a small number of protein sites (Wittmann, Yue and Arnold (2021); Kuo et al. (2020); Pokusaeva et al. (2019)). Nevertheless, the number of assayed genotypes in a single landscape is becoming larger in recent studies (Russ et al. (2020); Bryant et al. (2021)) and it appears that the experimental characterization of a sufficiently large fitness landscape with multiple fitness peaks may be attainable within the next decade. Therefore, there is a need for development of computational methods (Wittmann, Yue and Arnold (2021); Alley et al. (2019); Bryant et al. (2021); Russ et al. (2020); Biswas et al. (2021)) and theory (Zhou and McCandlish (2020)) that can improve the description of experimental fitness landscape datasets, such as obtaining an estimate of the number of isolated peaks. Here, we use Morse theory to calculate the minimal number of reciprocal epistatic interactions for a given number of peaks on a landscape.

Epistasis is the interaction of allele states of the genotype, which shapes the fitness landscape. When the impact of allele states on fitness is independent of each other, there is no epistasis and the resulting fitness landscape is smooth and has a single peak. Epistasis can lead to a more rugged fitness landscape and decrease the number of paths of high fitness between genotypes. Epistasis that makes the impact of an allele state on fitness stronger or weaker is called *magnitude epistasis*. On the other hand, epistasis that causes the contribution of an allele state on fitness to change its sign (e.g., a beneficial mutation becomes deleterious) is called *sign epistasis* (Weinreich, Watson and Chao (2005)). When the two allele states at different loci change the sign of their respective contribution to fitness then this interaction is called reciprocal sign epistasis. In a simple example of this principle, in a two loci two allele model, there are four genotypes, 00, 01, 01 and 11. The following landscape is shaped by sign epistasis when genotypes 00, 01, 10 and 11 have fitnesses of 1,-1,1 and 1, respectively. Reciprocal sign epistasis is present when the fitnessses of 00, 01, 10 and 11 genotypes are 1,-1,-1 and 1, respectively.

Of course, the effect of an allele state can depend on more than just one other locus, or site, in the genome. When allele states in different loci impact each other then the epistasis is higher-order. Higher-order epistasis if found frequently in the characterized fitness landscapes (Weinreich et al. (2013)), and it is clear that it has important evolutionary consequences (Kondrashov and Kondrashov (2001); Canale et al. (2018); Fragata et al. (2019); de Visser and Krug (2014)). However, models that allow studying such epistasis are at an early stage of their development (Crona, Krug and Srivastava (2021); Crona, Greene and Barlow (2013); Crona (2020)).

The evolutionary consequences of epistasis may be especially important when it leads to multiple local peaks. In that case, a population can get stuck on a suboptimal peak, decreasing the ability of evolution to find an optimal solution.

Using a combinatorial argument, Poelwijk et al. (2011) showed the following qualitative property: reciprocal sign epistasis is necessary for the existence of multiple peaks. Using Morse theory, we approach a more quantitative description of this relationship. This work might be the first formal use

of Morse theory to study fitness landscapes.

2 Outline of the method

Morse theory studies the properties of some discrete structures (such as graphs) and special functions defined on them. In particular, the strong Morse inequality relates the number of critical points with the Betti numbers of the underlying structure. In our case, we approximate the genotypes space structure as a graph where vertices are binary sequences (genotypes) and edges connect those genotypes within one-mutation distances. To apply Morse theory to our question of interest, we also consider edges between those vertices that are separated by reciprocal sign epistasis. The Morse function assigns a number to the vertices and all of their edges. On the vertices, the Morse function value is the corresponding genotype's fitness, while on edges the value is tailored for applicability of Morse theory.

Because we model genotypes as binary sequences the sequence space is a hypercube. Also, we consider only fitness landscape with no strictly neutral mutation, i.e. all direct neighbours of a vertex have slightly different fitness values.

The strong Morse inequality allows us to quantify the necessary condition for the existence of multiple peaks.

Theorem 1 (Strong Morse inequality)**.** *Consider m_i the number of critical points of the order i, and b_i the ith Betti number of the graph. Then, for all $i \geq 1$,*

$$\sum_{n=0}^{i}(-1)^{i-n}m_n \geq \sum_{n=0}^{i}(-1)^{i-n}b_n\,.$$

In particular, we will make use of the case $i = 1$, i.e

$$m_1 - m_0 \geq b_1 - b_0\,.$$

The main insight is, we define the graph and the Morse function in such a way that m_0 are peaks, m_1 are reciprocal sign epistatic interactions, and $b_0 = 1$. Finally, by definition of Betti numbers, $b_1 \geq 0$. Using the Theorem 1 for $i = 1$, we show the following result.

Theorem 2 (Quantification of epistatic interactions)**.**

$$\#\ \textit{reciprocal sign epistatic interactions}\ \geq \#\ \textit{peaks}\ - 1\,.$$

3 Formal proof

The combinatorial argument used in Poelwijk et al. (2011) shows that between any two peaks there must be a path connecting them. The minimum fitness along this path is part of a reciprocal sign epistatic interaction. Our result is intuitively explained by induction over the number of peaks as follows. The base case is when there are only two peaks, which is already explored in Poelwijk et al. (2011). Now, consider a fitness landscape and introduce the third peak. This new peak must be connected to all previous peaks through some reciprocal sign epistatic interaction, and the question is if a new peak introduces another such interaction.

The paths connecting the new peak to the old ones may use already existing epistatic interactions, but we show that introducing this new peak must imply the creation of a new reciprocal sign epistatic interaction. To make this last step in the proof formal, we use discrete Morse theory.

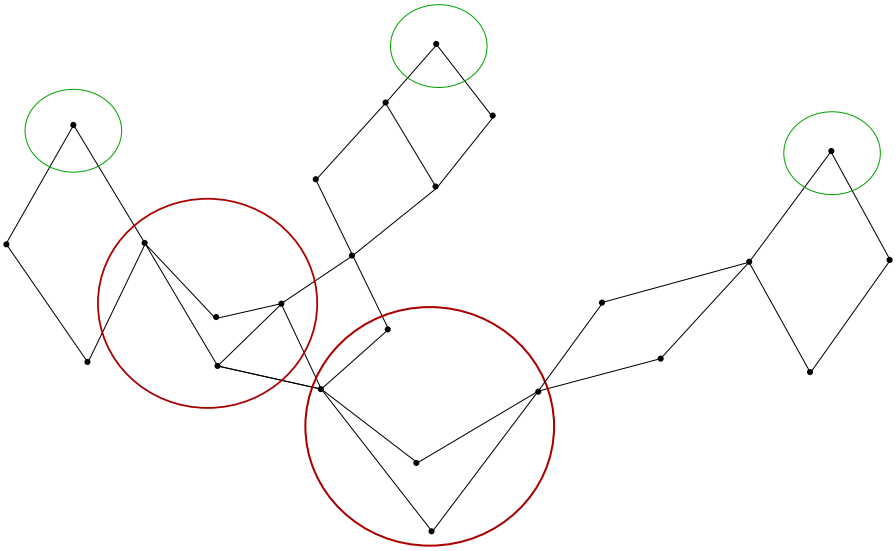

Figure 1: Proof by drawing

We introduce all the necessary concepts before explaining the proof step by step.

3.1 Necessary definitions (gentle version)

In this section we introduce the two terms used in Theorem 1: *critical points* and *betti numbers*. Since we work only on (undirected) graphs, the general definitions are simplified to the following.

Definition 3.1 (Betti numbers)**.** *Let $G = (V, E)$ be a graph. The zero-th Betti number (b_0) is the number of connected components in G. The first Betti number (b_1) equals $|E| + b_0 - |V|$, usually called cyclomatic number.*

Remark 3.1 (Betti numbers in connected graphs)**.** *Let $G = (V, E)$ be a connected graph. Then, $b_0 = 1$ and $b_1 = |E| + 1 - |V|$. Since G is connected, $|E| \geq |V| - 1$, therefore $b_1 \geq 0$.*

Definition 3.2 (Critical)**.** *Let $G = (V, E)$ be a graph and $f : V \cup E \to \mathbb{R}$ a function. We say that a vertex $v \in V$ is critical if, for all edges e containing v we have that $f(e) > f(v)$.*

We say that an edge $e = \{u, v\} \in E$ is critical if $f(e) > \max\{f(u), f(v)\}$.

We denote m_0 the number of critical vertices and m_1 the number of critical edges.

3.2 Necessary definitions (option 2)

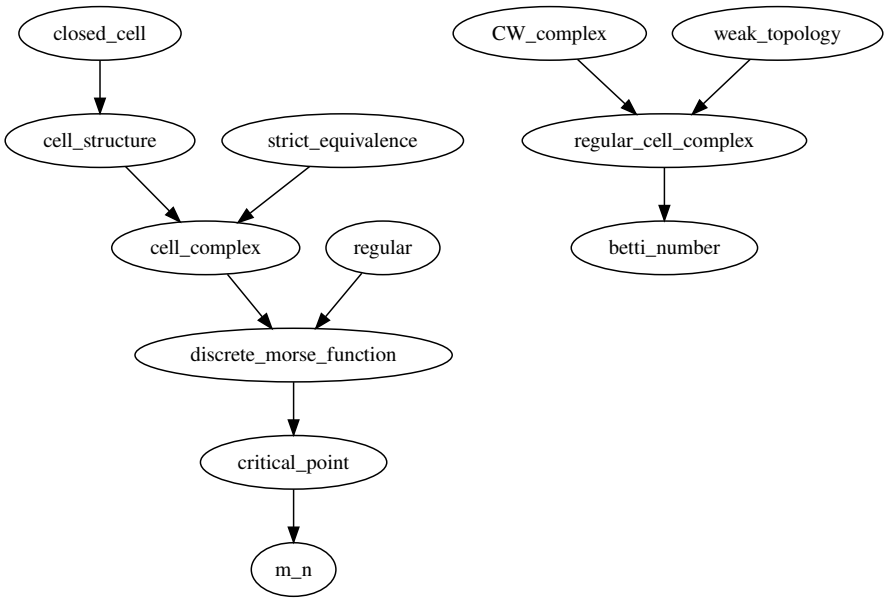

Figure 2: Definitions dependency

The following standard definitions were taken from Forman (1998); Lundell (1969) and are a formal repetition of the simplified version we presented before.

Definition 3.3 (Closed cell)**.** *Consider some space M and the set $E_n := [0, 1]^n \subseteq \mathbb{n}$. If there is an homeomorphism $\varphi : E_n \to \sigma_n \subset M$, we say that σ_n is a n-dimensional closed cell and φ a*

characteristic map for the cell σ.

Let τ and σ be two closed cells. We write $\tau > \sigma$ if $\sigma \in \partial\tau$.

Definition 3.4 (Cell structure)**.** *Let M be a set. A cell structure on M is a pair (M, Φ), where $\Phi = \{\varphi\}_{\varphi\in\Phi}$ is a collection of maps of closed cells into X satisfying the following conditions.*

1. *Injective interior. For all $\varphi \in \Phi$, if $E_n = \mathrm{dom}(\varphi)$, then φ is injective in $E_n \setminus \partial E_n$, where ∂ is the boundary operator.*
2. *Partition of M. The set $\{\varphi(E_n \setminus \partial E_n) : \varphi \in \Phi, E_n = \mathrm{dom}(\varphi)\}$ partition M.*
3. *Boundaries are lower dimensional cells. For all $\varphi \in \Phi$, if $E_n = \mathrm{dom}(\varphi)$, then $\varphi(\partial E_n) \in \{\psi(E_k \setminus \partial E_k) : \psi \in \Phi, E_k = \mathrm{dom}(\varphi), k \leq n-1\}$.*

Definition 3.5 (Strict equivalence)**.** *We say that two cell structures (M, Φ) and (M, Φ') are strictly equivalent if there is a one-to-one correspondence between Φ and Φ' such that a characteristic function with domain E_n corresponds to a characteristic function with domain E_n, and corresponding functions differ only by a reparametrization of their domain.*

Definition 3.6 (Cell complex)**.** *Let M be a set. A cell complex on M is an equivalence class of cell structures (M, Φ) under the equivalence relation of strict equivalence. A cell complex on M will be denoted by a pair (M, K), where $K = K_\Phi$ for some representative cell structure (X, Φ). The set K is called the set of (closed) cells of (X, K).*

Moreover, we will abuse the notation and write $K = \{\sigma : \sigma \in \Phi\}$, the set of all cells.

In particular, every graph $G = (V, E)$ can be seen as a cell complex where all the edges are closed cells of dimension one, and all vertices are closed cells of dimension zero.

Definition 3.7 (Regular)**.** *Let (X, S) be a cell complex. A cell σ_n, face of τ_{n+1} whose characteristic map is ψ, is called regular if*

1. *$\psi : \psi^{-1}(\sigma_n) \to \sigma_n$ is an homeomorphism.*
2. *$\overline{\psi^{-1}(\sigma_n)} = [0, 1]^n$.*

We say that a cell complex is regular if all cells that are faces of other cells are regular.

In particular, every graph is a regular cell complex.

Definition 3.8 (Discrete Morse)**.** *Let (M, K) be a cell complex. Consider a function $f : K \to \mathbb{R}$ which assigns a value to each cell. We say that f is a discrete Morse function if it satisfies the following conditions. For all $\sigma_n \in K$,*

1. *If σ_n is an irregular face of τ_{n+1}, then $f(\tau) > f(\sigma)$. Moreover,*

$$|\{\tau_{n+1} > \sigma_n : f(\tau_{n+1}) \leq f(\sigma_n)\}| \leq 1\,.$$

2. *If ν_{n-1} is an irregular face of σ_n, then $f(\sigma) > f(\nu)$. Moreover,*

$$|\{\nu_{n-1} < \sigma_n : f(\sigma_n) \leq f(\nu_{n-1})\}| \leq 1\,.$$

3. *f is injective.*

Note that the previous point is not formally necessary, but we describe it here for convenience.

Definition 3.9 (Critical)**.** *Let (M, K) be a cell complex and $f : K \to \mathbb{R}$ a discrete Morse function. We say that $\sigma = \sigma_n \in K$ is critical if*

$$|\{\tau_{n+1} > \sigma_n : f(\tau_{n+1}) \leq f(\sigma_n)\}| = 0$$
$$|\{\nu_{n-1} < \sigma_n : f(\sigma_n) \leq f(\nu_{n-1})\}| = 0\,.$$

Moreover, we say that σ_n is a critical cell of index n.

Example 1 (Regularity and minima)**.** *If (M, K) is a regular cell complex, then for all discrete Morse functions we have that its minimum must occur on a vertex, which must be a critical point of index 0. This follows from the following observation: if $n \geq 1$, then every σ_n, n-dimensional cell, has at least two $(n-1)$-dimensional faces.*

Definition 3.10 (m_n)**.** *Let (M, K) be a cell complex and $f : K \to \mathbb{R}$ be a discrete Morse function. We denote m_n for the number of critical points of index n of f.*

Consider a cell complex (M, K). To define Betti numbers and then state Morse inequalities, we need to equip M with a topology. We can do so using the weak topology with respect to K, which is defined as follows.

Definition 3.11 (Weak topology w.r. K)**.** *Let (M, K) be a cell complex. Consider the following procedure.*

1. *Give each cell $\sigma \in K$ the quotient topology with respect to its characteristic function φ.*
2. *Give M the weak topology with respect to the subsets $\sigma \in K$, i.e., a set $A \subset M$ is closed if and only if for all $\sigma \in K$ we have that $A \cap \sigma$ is closed in σ.*

The resulting topology on M is called the weak topology with respect to K.

Let us now define special cell complexes which also have a topology.

Definition 3.12 (CW complex)**.** *Let (M, K) be a cell complex, where M is also a M Hausdorff space. We say that (M, K) is a CW complex if the following conditions hold.*

1. *M has the weak topology with respect to K.*
2. *K is finite.*

Naturally, we can always go from a finite regular cell complex to a CW complex by giving to M the weak topology with respect to K. With this, we can consider regular cell complexes as topological spaces and so define its Betti numbers as follows.

Definition 3.13 (Betti number)**.** *Let (M, K) be a regular cell complex and F a field. Define the n-th Betti number with coefficients in F as*

$$b_n := \dim H_n(M, F)\,.$$

Example 2 (b_0)**.** *The first Betti number b_0 represents the number of connected components in M. In particular, if M is connected, then $b_0 = 1$.*

3.3 Proof

Proof of Theorem 2. Our proof consists in the following steps:

1. Definition of a regular cell complex.
2. Connectedness of the cell complex.
3. Definition of a discrete Morse function.
4. Application of Morse inequality.

Definition of a regular cell complex. Consider a graph $G = (V, E)$. Let $V := \{0, 1\}^d$. Let set of edges $E := E_1 \cup E_2$ by defined in two steps: E_1 and E_2 has edges involving only sequences at humming distance one and two respectively. The set E_1 contains only edges that connects a sequence with its fittest mutation, if it exists. Formally,

$$E_1 := \{\{u, v\} : d(u, v) = 1, W(u) < W(v) = \max\{W(v') : d(u, v') = 1\}\}\,.$$

In the other hand, E_2 contains edges that connect the two highest points of a reciprocal sign epistasis. Formally,

$$E_2 := \{\{u, v\} : d(u, v) = 2, \forall y \in V \quad d(u, y) = 1 \wedge d(v, y) = 1 \Rightarrow W(y) < W(u) \wedge W(y) < W(v)\}\,.$$

We can easily see that the graph G is a finite regular cell complex by considering $K := V \cup E$, i.e. the cells are both vertices and edges.

Connectedness of the cell complex. We now prove that G is connected, and therefore $b_0 = 1$. First note that any vertex is connected to a peak. Indeed, from any vertex, by following the path of

fittest mutations, we can go to a peak by edges in E_1. Therefore, we only need to prove that all peaks are connected.

By contradiction, assume that there are $K_1, \dots, K_r$ connected components of G. In each component there might be multiple peaks. Let us come back to the usual sequence graph $G_S = (\{0,1\}^d, E_{d_1})$, where E_{d_1} contains all edges connecting sequences at humming distance one. Consider the path P_1 that connects two peaks in different components and has the highest minimum value, i.e.,

$$P_1 \in \underset{P \text{ path in } G_S}{\operatorname{argmax}} \{\min\{W(v) : v \in P\} : \exists i \neq j, \exists v_1 \in K_i, v_2 \in K_j \text{ peaks } st\, v_1 \overset{P}{\longleftrightarrow} v_2\}\,.$$

Without loss of generality, assume that P_1 connects $v_1 \in K_1$ and $v_2 \in K_2$. Denote v_m by the vertex in P_1 that achieves the minimum fitness and divide in the following way: $P_1 = P_1^1 v_m P_1^2$. Note that we can assume that all vertices in P_1^1 are in K_1, i.e. $V(P_1^1) \subseteq K_1$. Indeed, if it was not the case, consider $v' \in P_1^1 \cap K_1^c$. Since $v' \notin K_1$, by following the fittest mutation, it is connected to a peak v_2' which is not in K_1. Consider a new path P_1' that goes from v_1 to v' and then to v_2'. Note that the minimum fitness value in P_1' is higher than the one in P_1 and P_1' also connects two different connected components, which is a contradiction. Therefore, $V(P_1^1) \subseteq K_1$. Similarly, we get that $V(P_1^2) \subseteq K_2$.

Denote $u_m^1 \in K_1$ the vertex in $K_1 \cap P_1^1$ closest to v_m, similarly denote u_m^2 the vertex in $K_2 \cap P_1^2$ closest to v_m. First notice that $\{u_m^1, u_m^2\} \in E_2$, i.e. there a reciprocal sign epistasis between vertices with high fitness. Indeed, if this were not the case, we could connect them through another mutation that does not involve v_m and create a path P_1' with a higher minimum value, which is a contradiction.

Since $u_m^1 \in K_1$, we can follow the fittest mutation path until a peak $u_1 \in K_1$ and similarly for u_m^2 to a peak $u_2 \in K_2$. Consider the path $Q_1 = Q_1^1 v_m Q_1^2$, where $u_1 \overset{Q_1^1}{\longleftrightarrow} u_m^1$ and $u_2 \overset{Q_1^2}{\longleftrightarrow} u_m^2$. Note that, by definition of E_1, we have that $Q_1^1, Q_1^2 \subseteq E_1$. Then, the vertex $u_1 \in K_1$ and $u_2 \in K_2$ are connected by $Q_1^1 Q_1^2$, using the edge $\{u_m^1, u_m^2\} \in E_2$ to fill the gap. But this is a contradiction because K_1 and K_2 were two different connected components. Therefore, G is connected.

Definition of a discrete Morse function. Consider the function $f : K \to \mathbb{R}$ given by the following.

- For all $v \in V$,
$$f(v) = -W(v)\,.$$
- For all $e = u, v \in E_1$,
$$f(e) = \frac{f(u) + f(v)}{2}\,.$$
- For all $e = u, v \in E_2$,
$$f(e) = C\,,$$
where $C > \max\{-W(v) : v \in V\}$.

Notice that, since the fitness landscape has no strictly neutral mutations, we can perturb f to get an injective function where the relationship between adjacent cells is preserved. Then, f is a discrete Morse function.

Application of Morse inequality. By Theorem 1, we have that

$$m_1 - m_0 \geq b_1 - b_0 \,.$$

Since M is connected, we have that $b_0 = 1$. By definition of Betti numbers, and since M is connected, $b_1 \geq 0$ (see Remark 3.1). The number of critical vertices is m_0 and the number of critical edges is m_1. By construction, the only critical vertices are peaks and the only critical edges are those in E_2, i.e. edges that represent reciprocal sign epistasis. Therefore,

$$\#\text{ reciprocal sign epistatic interactions } \geq \#\text{ peaks } - 1\,.$$

□

4 Discussion

We have shown that the multipeaked fitness landscape necessarily has no fewer pairwise reciprocal sign epistatic interactions than the number of fitness peaks minus one. This extends the result of Poelwijk et al. (2011) stating that the reciprocal sign epistasis is a necessary condition for multiple peaks. Additionally, our study showcases the application of discrete Morse theory to fitness landscapes.

As discussed in Poelwijk et al. (2011), reciprocal sign epistasis is not a sufficient condition for multiple peaks. Similarly, we do not show how to estimate the number of peaks from the number of epistatic interactions.

A sufficient condition for multiple peaks in terms of local interactions was given in a later work (Crona, Greene and Barlow (2013)): reciprocal sign epistasis leads to multiple peaks if there is no sign epistasis in any other pair of loci.

The complication of deducing the global properties of fitness landscapes from the local properties of epistasis between specific sites arises due to the multidimensionality of the fitness landscape: local peaks formed by a pairwise epistatic interaction can be bypassed through a different dimension. Therefore, the condition formulated in terms of the pairwise epistatic interaction cannot be sufficient. One needs to know the full fitness landscape: to deduce that the fitness landscape has multiple peaks, one has to know that there is no sign epistasis in any other pairwise interaction (Crona, Greene and Barlow (2013)).

For a quantitative result converse to ours, we anticipate that higher-order epistatic interactions have to be considered, which leads to the requirement of full information about the fitness landscape. We expect that this result can be obtained with a suitable definition of the higher-order epistasis. Such a result could be useful, for example, to study the empirical fitness landscapes if the number of mutations under consideration is small enough to make an almost complete description of the landscape feasible.

5 Acknowledgements

We are grateful to Herbert Edelsbrunner and Jeferson Zapata for helpful discussions. This work was supported by the ERC Consolidator (771209—CharFL) and the FWF Austrian Science Fund (I5127-B) grants to FAK.

6 References

Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M. and Church, G.M., 2019. Unified rational protein engineering with sequence-based deep representation learning. 16(12), pp.1315–1322. Available from: https://doi.org/10.1038/s41592-019-0598-1.

Biswas, S., Khimulya, G., Alley, E.C., Esvelt, K.M. and Church, G.M., 2021. Low-n protein engineering with data-efficient deep learning. 18(4), pp.389–396. Available from: https://doi.org/10.1038/s41592-021-01100-y.

Bryant, D.H., Bashir, A., Sinai, S., Jain, N.K., Ogden, P.J., Riley, P.F., Church, G.M., Colwell, L.J. and Kelsic, E.D., 2021. Deep diversification of an AAV capsid protein by machine learning. 39(6), pp.691–696. Available from: https://doi.org/10.1038/s41587-020-00793-4.

Canale, A.S., Cote-Hammarlof, P.A., Flynn, J.M. and Bolon, D.N., 2018. Evolutionary mechanisms studied through protein fitness landscapes. 48, pp.141–148. Available from: https://doi.org/10.1016/j.sbi.2018.01.001.

Crona, K., 2020. Rank orders and signed interactions in evolutionary biology. *elife*, 9, p.e51004. Available from: https://doi.org/10.7554/eLife.51004.

Crona, K., Greene, D. and Barlow, M., 2013. The peaks and geometry of fitness landscapes. *Journal of theoretical biology*, 317, pp.1–10. Available from: https://doi.org/https://doi.org/10.1016/j.jtbi.2012.09.028.

Crona, K., Krug, J. and Srivastava, M., 2021. Geometry of fitness landscapes: Peaks, shapes and universal positive epistasis. 2105.08469.

Forman, R., 1998. Morse theory for cell complexes. *Advances in mathematics*, 134(1), pp.90 – 145. Available from: https://doi.org/https://doi.org/10.1006/aima.1997.1650.

Fragata, I., Blanckaert, A., Dias Louro, M.A., Liberles, D.A. and Bank, C., 2019. Evolution in the light of fitness landscape theory. *Trends in ecology evolution*, 34(1), pp.69–82. Available from: https://doi.org/https://doi.org/10.1016/j.tree.2018.10.009.

Kondrashov, F.A. and Kondrashov, A.S., 2001. Multidimensional epistasis and the disadvantage of sex. *Proceedings of the national academy of sciences*, 98(21), pp.12089–12092. https://www.pnas.org/content/98/21/12089.full.pdf, Available from: https://doi.org/10.1073/pnas.211214298.

Kuo, S.T., Jahn, R.L., Cheng, Y.J., Chen, Y.L., Lee, Y.J., Hollfelder, F., Wen, J.D. and Chou, H.H.D., 2020. Global fitness landscapes of the shine-dalgarno sequence. 30(5), pp.711–723. Available from: https://doi.org/10.1101/gr.260182.119.

Lundell, A.T., 1969. *The topology of cw complexes*. New York: Van Nostrand Reinhold.

Maynard Smith, J., 1970. Natural selection and the concept of a protein space. 225(5232), pp.563–564. Available from: https://doi.org/10.1038/225563a0.

Melamed, D., Young, D.L., Gamble, C.E., Miller, C.R. and Fields, S., 2013. Deep mutational scanning of an RRM domain of the saccharomyces cerevisiae poly(a)-binding protein. 19(11), pp.1537–1551. Available from: https://doi.org/10.1261/rna.040709.113.

Poelwijk, F.J., Tănase-Nicola, S., Kiviet, D.J. and Tans, S.J., 2011. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of theoretical biology*, 272(1), pp.141 – 144. Available from: https://doi.org/https://doi.org/10.1016/j.jtbi.2010.12.015.

Pokusaeva, V.O., Usmanova, D.R., Putintseva, E.V., Espinar, L., Sarkisyan, K.S., Mishin, A.S., Bogatyreva, N.S., Ivankov, D.N., Akopyan, A.V., Avvakumov, S.Y., Povolotskaya, I.S., Filion, G.J., Carey, L.B. and Kondrashov, F.A., 2019. An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *Plos genetics*, 15(4), pp.1–30. Available from: https://doi.org/10.1371/journal.pgen.1008079.

Romero, P.A. and Arnold, F.H., 2009. Exploring protein fitness landscapes by directed evolution. 10(12), pp.866–876. Available from: https://doi.org/10.1038/nrm2805.

Russ, W.P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M. and Ranganathan, R., 2020. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502), pp.440–445. https://www.science.org/doi/pdf/10.1126/science.aba3304, Available from: https://doi.org/10.1126/science.aba3304.

Sarkisyan, K.S., Bolotin, D.A., Meer, M.V., Usmanova, D.R., Mishin, A.S., Sharonov, G.V., Ivankov, D.N., Bozhanova, N.G., Baranov, M.S., Soylemez, O., Bogatyreva, N.S., Vlasov, P.K., Egorov, E.S.,

Logacheva, M.D., Kondrashov, A.S., Chudakov, D.M., Putintseva, E.V., Mamedov, I.Z., Tawfik, D.S., Lukyanov, K.A. and Kondrashov, F.A., 2016. Local fitness landscape of the green fluorescent protein. 533(7603), pp.397–401. Available from: https://doi.org/10.1038/nature17995.

Visser, J.A.G. de and Krug, J., 2014. Empirical fitness landscapes and the predictability of evolution. 15(7), pp.480–490. Available from: https://doi.org/10.1038/nrg3744.

Weinreich, D.M., Lan, Y., Wylie, C.S. and Heckendorn, R.B., 2013. Should evolutionary geneticists worry about higher-order epistasis? *Current opinion in genetics development*, 23(6), pp.700–707. Genetics of system biology. Available from: https://doi.org/https://doi.org/10.1016/j.gde.2013.10.007.

Weinreich, D.M., Watson, R.A. and Chao, L., 2005. PERSPECTIVE:SIGN EPISTASIS AND GENETIC CONSTRAINT ON EVOLUTIONARY TRAJECTORIES. 59(6), p.1165. Available from: https://doi.org/10.1554/04-272.

Wittmann, B.J., Yue, Y. and Arnold, F.H., 2021. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell systems*. Available from: https://doi.org/https://doi.org/10.1016/j.cels.2021.07.008.

Wright, S., 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the xi international congress of genetics*, 8, pp.209–222.

Zhou, J. and McCandlish, D.M., 2020. Minimum epistasis interpolation for sequence-function relationships. 11(1). Available from: https://doi.org/10.1038/s41467-020-15512-5.