nature portfolio

Peer Review File

The simplicity of protein sequence-function relationships

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In this manuscript "The simplicity of protein sequence-function relationships", the authors focus on a critical question about protein fitness landscapes, that is, how important is higher order epistasis in determining the sequence function relationship for proteins. The answer to this question has been cofounded historically by the influence of nonspecific epistasis, which is a nonlinear mapping from the phenotype to the measurement scale and may manifest as spurious higher order epistasis if not accounted for.

The authors developed a method for inferring the coefficients of specific epistatic interactions that does not rely on the choice of a reference genotype (referece-free analysis, RFA). The authors curated a diverse datasets of near combinatorically complete protein fitness landscapes. The authors then applied the RFA method together with a nonlinear sigmoidal function to jointly model nonspecific and specific epistasis in the chosen datasets. The main conclusion is that most protein fitness landscapes can be largely captured by only additive and pairwise interactions when nonspecific epistasis is accounted for.

Overall, I think this is a very important topic and that the paper well written. I think the evidence for the lack of higher order interactions supported by the very marginal increase in prediction performance in these datasets is convincing. However, I have some major concerns about RFA model's purported superiority over the classical reference-based analysis (RBA) and its significance in drawing main conclusion of the paper, which I listed below.

First, it is unclear what the role of RFA model is in this paper. The RFA model is great for inferring epistatic coefficients that are independent of a reference genotype. Furthermore, unlike the Fourier framework, coefficients of the RFA directly correspond to interactions among alleles across sites, making it an intuitive method for interpreting specific epistasis. However, this is not how the RFA model is mostly used in the paper. The main conclusion that there is negligible higher-order interaction in the test datasets has little to do with these desirable properties of RFA. And the same conclusion can be drawn by just fitting standard RBA regression using least squares. This is because although RFA and RBA have different parametrizations of the fitness landscape, they are indeed modeling in the exact same space of functions. Specifically, for any RFA model, there is a corresponding RBA model (and vice versa) that gives identical predictions everywhere in the sequence space. So if the modeling fitting was done correctly (with the right choice of regularization), they should produce the exact same predictions.

Second, I think there are some unfair comparisons between the RBA and RFA model. My overall criticism is: the RBA model was not fitted using the the same amount of data with appropriate regularized least squares regression like the RFA model. And when the RBA model is fitted in the correct way, most differences the authors highlighted would go away. In the supplement to the bioRxiv preprint, the authors pointed out "A reference-based effect of order k is the sum or subtraction of 2k phenotypes", meaning the effect of a mutation is equal to the difference between the single mutant and the WT; a pairwise coefficient is the sum of WT, single mutants, and double mutants, with alternating signs. When formulated this way, of course the RBA model performs less well than the RFA model, simple because it uses less data. In contrast, the right way to infer effects in an RBA model is to fit a model to all genotypes in the data, such that for any genotype, its phenotype = sum of mutation effects + sum of pairwise interactions + sum of higher order interactions. Because regardless of if there is a chosen WT, the fitness of a genotype is decomposed in the same way as the RFA model (except with an intercept WT term), so coefficients of the model should be inferred using all the data, instead of just the lower order mutants. This unfairness is manifested in the following places.

In Figure 1d, the authors show RBA models have much higher variance in the estimation of epistatic effects than the RFA model when measurement noise is present. I think if the RBA model coefficients, if inferred using least squares by fitting the RBA model to all the data will produce very similar behavior. That is because the same mutation/epistatic interaction in the RBA model appears as many time as in the RFA model.

In Figure 1e, the Fourier and RBA model apparently have lower out of sample R^2. Again, I think if the RBA model was fitted correctly, the performance should be similar.

In Figure 2C, what is the performance of the RFA model in the same setting? Again, if both models were fitted using least squares on the same amount of data, the performance should be similar. Furthermore, it appears (at least at first glance for some readers) that the RBA model does not perform as well as the RFA mode, even though panel A and C are measuring different things. Juxtaposition of these two panels is misleading.
Despite this, the RFA model does have its role (albeit in more nuanced places) in the paper, for example, measuring the sparsity of the fitness landscape (Fig 4) and the variance decomposition. But I do not reckon that it is the main selling point of the paper. I think a viable route for the authors is to put the comparison between RBA and RFA in less conspicuous places and focus on the main conclusion of the paper, which is the simplicity of protein sequence function relationships.

Reviewer #2 (Remarks to the Author):

The nature and prevalence of epistasis, as well as how it varies across different orders, has long been a central concern of evolutionary genetics and biochemistry, but experimental results were scarce. Advances in high-throughput phenotype-genotype mapping, such as deep mutational scanning, have produced complete high-order combinatorial mutational datasets for a number of targets. This has stimulated further interest in the quantitative understanding of the underlying fitness landscapes.

In this work, the authors present a formalism for the description of a mutational fitness landscape which they call “reference-free analysis” (RFA). They claim this framework is superior to alternative approaches in inferential robustness and parameter interpretability. They analyze 20 datasets across 13 targets with RFA and make several major claims: that protein sequence-function relationships are “simple,” that genetic architecture is “sparse,” and that nonspecific epistasis is attributable primarily to measurement saturation.

A commentary on an initial draft of this manuscript by Dupic et al. [PMID 38352387], who also produced several of the datasets analyzed here, made several critiques of these claims, some of which we find compelling and others less so. Briefly, we agree that RFA is equivalent to certain formulations of reference-based analysis (RBA) and Hadamard-transformed landscapes, and that the arguments for sparsity and simplicity require more care, but we believe this will prove an interesting and insightful intervention with some minor reframing.

In general, we find that the interest of this work lies in the comparative analysis across multiple datasets with a unified framework rather than in the value of RFA itself. With this in mind, we feel several of the major claims are tantalizing but insufficiently proven, and some of the language could be made more precise. Despite the issues discussed below, we emphasize that this work should be of broad interest to evolutionary and protein biologists and recommend it for revision at Nature Communications.

Reviewed by Angelica Lam, Christian Macdonald, Rosa Sanchez, and Willow Coyote-Maestas

Major issues:

There is undue emphasis on the novelty and superiority of RFA given that it is exactly equivalent to reference-based (RBA) and Hadamard approaches. A term-by-term comparison of RFA with an expansion in Walsh basis (see, e.g., Weinrich, et al. [PMID 24290990]) makes this clear. Given this, RFA can only be superior in terms of parameter interpretation. The authors recognize this and argue that RFA produces a maximally efficient and clear encoding of genetic architecture, but we feel this is both inadequate and unnecessary. The discussion of the superiority of RFA-derived parameters is too short and not concrete enough, and the emphasis on proving the superiority of RFA distracts from what we found to be the genuinely interesting biological results.

The claim that measurement bounding is the major cause of nonspecific epistasis is insufficiently argued. We believe this is an important and major claim in the work, and that a deeper understanding of the sources of nonspecific epistasis will be of major use to the field. As written, however, this manuscript only demonstrates that use of a sigmoid link function improves model fit with RFA. This does not support the larger claim that "the primary cause of nonspecific epistasis is phenotype bounding."

The authors claim that protein sequence-structure relationships are simple (as in, first-order models are sufficient to predict most phenotypes). This is an interesting and important result, and generally in line with other findings that high-order epistasis seems to be rare. The authors clearly demonstrate this with their work. We believe the interpretation of these results requires more care, however. "Simplicity" may be misleading. Often evolutionary dynamics on a landscape and not the landscape itself is the subject of interest, and it is not obvious that "simple" sequence-structure relationships will also produce simple evolutionary outcomes. Further, given the exact equivalence of (a form of) RBA and Hadamard approaches with RFA, the argument for RFA based on this form of simplicity is effectively a belief that terms in a sequence-structure relationship should be of order one, or an argument for a form of parsimony that has intuitive appeal but should be made more explicit.

The authors argue that genetic architecture is sparse (as in, few terms of a model are sufficient to predict most phenotypes). This is another important and interesting result which deserves additional care. In particular, we wonder why only a third-order model was used for all datasets, and whether RFA is any more sparse than RBA.

The utility of this work for experimentalists could be enhanced through some further comparison with reference-based approaches. One issue we see is that in this framework the individual parameters may be intractable: for example, the zero-order epistatic coefficient e_0 is defined with the same information as a n-th order coefficient e_n, that is, the entire combinatorial genotype space. In many DMS settings, a particular sequence is actually privileged for one reason or another, and so experimentalists anchor their exploration of sequence space, even at high orders, at that point. RFA may not be able to guide experiments due to the data dependencies of the parameters. The authors do mention that multiple screens across multiple homologs may be one approach to estimate an RFA landscape from limited data, but it is not obvious if this would work. We feel some further discussion of the phenomenology of the recovered parameters would be useful to guide potential applications of this approach.

The claims about sparsity and simplicity of protein sequence-function relationships are based upon datasets of proteins with relatively small genotype space (with the exception of ParB and GB1), and are based upon phenotypes (e.g., binding, fluorescence) that are treated independently and measured on a minimum/maximum spectrum. While these simpler representations may be necessary for the isolated analysis of a particular dataset, they do not consider dependency between phenotypes (e.g., binding and transcription activity of transcription factors) that may be encoded by more widespread and high-order epistatic interactions. Moreover, the authors demonstrate that their model can explain most of the phenotypic variance of avGFP with just first- and second-order effects despite its dense

epistatic network, but the relevant sites that engage in epistasis are structurally close together. Proteins that experience allosteric effects over longer distances may also be encoded by more widespread and high-order epistatic interactions. Given that the analyzed datasets also have existing structures, it is important to consider why the model can explain phenotypic variance with sparse and lower-order effects in the context of the protein's structure.

Minor issues:

Figure 1a - legends for the models would be helpful here.

Figure 1c: In the context and in the figure description, they don't mention what some of the variables are. For easier understanding of the passage and interpretation of data, variable definitions such as f0 and b0 would be good.

The authors claim that because effect terms for RBA methods are computed as a chain of sums and subtractions of individual variants, measurement errors propagate and snowball with epistatic order. They demonstrate this in Fig 1D using phenotypic data that was simulated from a known genetic architecture with normally distributed noise. Additional information about how this data was simulated and why it is representative of observed phenotypic datasets would make this claim stronger.

The choice of states and sites in Fig 1e seems somewhat arbitrary. Some explanation of the choices would improve clarity.

Figure 3B: This is a little tricky to grasp and some sort of cartoon might help the reader.

The equivalent plots to 4a for other datasets should be included.

Figure 6B: the average genetic score value (-7.8) is mentioned in the text. It may be nice to include this value on the plot as well.

Using cross-validated, out-of-sample R^2 to compare RFA and RBA may not be an apples-to-apples comparison. The implicit "reference" used by RFA is the mean phenotype, making it inherently more robust to cross-validation statistical measures than reference-based analyses that rely on a WT sequence and may suffer from cross-validation sets that do not have WT, WT-adjacent sequences.

While the authors do provide evidence that a small number of reference free terms can determine function, they also show that genetic architecture of proteins cannot be captured by sparse sampling. This is an interesting result and worthy of more discussion.

The authors apply a framework for understanding genetic architecture where genetic variation is analogous to change in energy, and functional vs. nonfunctional phenotypes are described in terms of the relative occupancy of a functional vs. nonfunctional state. For clarity, emphasizing that this is an apparent or effective free energy would be helpful.

Suggested citations:
A publication by Buda, Miton and Tokuriki [PMID: 38129396] performed a similar comparative analysis of combinatorial landscapes and made somewhat different claims.
Recent work from David McCandlish [PMIDs 32286265, 36129941, 35428271] has investigated genotype-phenotype map structures
In Tonner, et al. [PMID 35733251] the power of a number of alternative genotype-phenotype functional forms is compared, including linear and spline models.

Reviewer #3 (Remarks to the Author):

In their manuscript Park et al. describe a formalism to derive the genetic architecture of a protein, while accounting for unspecific epistatic effects. Different than other common methods their "reference free" approach sets the global average across all sequences as the zero order term /ground state and employs a sigmoid link function that is simultaneously fitted together with the epistatic coefficients for the model. The sigmoid link function is meant to capture all unspecific epistasis/non-linearities inherent to the data as e.g. caused by the measurement process. Using their method, they re-analyzed 20 previously published datasets using their and complementary frameworks. They show that heir framework appears to be more robust and furthermore suggests that the genetic architecture of protein function is rather simple, consisting mostly of additive terms, a few pairwise epistatic terms and almost no higher-order terms. This result generalizes recent observation made on e.g. protein stability (1) and tRNA functon(2) to the large variety of protein functions measured in the 20 analyzed datasets.

There are few points though that presented themselves while reading the manuscript.

While an intuitive choice for a link function, no rational is given for choosing a sigmoid function. It would be interesting whether the authors have tested other sensible link functions with a similar result. We support the idea of unspecific epistasis e.g. non linearities introduced due to the involved measurement process and are intrigued by the authors' results. However, the function is used as a catch all term without validating or exploring this idea further by using a set of measured experimental values quantifying the relationship between measured and latent phenotype. Such measurements could deliver the upper and lower bounds the authors use as parameters for the link function. Furthermore, in some cases the relation between assayed phenotype and function already follows a non-linear relationship e.g in cases were the impact of mutations on ddG binding itself is assayed, how would that be impacted by the unspecific link function?

Furthermore, while the reference free approach is suggested by the authors to have clear advantages in terms of robustness to noise, completeness of genotypes and number of epistatic orders studied (Fig.1), a recent publication by Dupic, Phillips and Desai (3) suggest that this is partly due to the way the epistatic effects were recalculated using the comparative methods.
We would be interested in the authors response to that and would like to know how their recalculation compares to the original methods used in the studies they selected as datasets. Especially, since the authors themselves state that the epistatic coefficients derived via the different methods can transformed into each other by a simple linear remapping: "Although the coefficients of any of the four formalisms can be converted into those of the others using a linear re-mapping …"
This would imply that the different results only differ in scale, but describe the same underlying genotype to phenotype mapping. As also noted in the
paper cited above, this would imply that the difference lies in the interpretation of the coefficients.

Using their method the authors find that the genetic architecture for almost all of the studied datastes is dominated by first and second order epistatic terms, if judged by the differences in out of sample R^2 (Fig2, Extended data Fig.1). However, this does not necessarily mean that the remaining higher order terms are not important/ crucial for the function. This could especially be true in cases where proteins are studied on a molecular level, were most of the variance migth arise from mutant effects on protein folding /abundance rather than on e.g. the (indireclty) measured funcional phenotype (e.g. binding).
As stated by the authors, the choice of reference vs. reference free should be contextual to the data and the biological question. In cases were a molecular phenotype is directly targeted and the question is whether two mutations improve synergistically over the wildtype an approach using the wildtype as actual physically realized state should be advantageous over the use of a global hypothetical average.

In summary, the present manusscript of Park et al., is a valuable contribution to the analysis of

genotype phenotype maps and while it is not necessarily superior to other methods it is important addition.

1. Faure AJ et al., The genetic architecture of protein stability. bioRxiv (2023); https://doi.org/10.1101/2023.10.27.564339
2. Domingo J, Diss G and Lehner B, Pairwise and higher-order genetic interactions during the evolution of a tRNA. Nature 558:117-121 (2018)
3. Dupic T, Phillips AM and Desai MM. Protein sequence landscapes are not so simple: on reference-free versus reference-based inference. bioRxiv (2024); https://doi.org/10.1101/2024.01.29.577800

We are grateful for the reviewers' thoughtful comments on our manuscript, "The simplicity of protein sequence-function relationships" (NCOMMS-24-07858). We have performed new analyses, changed numerous figures, and modified the text extensively to address their comments. Our responses are detailed below, with the reviewers' comments pasted in blue.

All three reviewers concurred on the validity and significance of our central finding: protein sequence-function relationships are simple when described from a global perspective, in that they can be effectively described by first- and second-order determinants and a simple form of nonlinearity. The reviewers' principal criticisms pertained to the novelty and significance of our new method of reference-free analysis (RFA). The most significant revisions that we made therefore involve clarifying the differences between RFA and existing formalisms and the advantages of RFA. An additional major area of commentary and revision was our treatment of nonspecific epistasis.

Response to Reviewer 1

> In this manuscript "The simplicity of protein sequence-function relationships", the authors focus on a critical question about protein fitness landscapes, that is, how important is higher order epistasis in determining the sequence function relationship for proteins. The answer to this question has been cofounded historically by the influence of nonspecific epistasis, which is a nonlinear mapping from the phenotype to the measurement scale and may manifest as spurious higher order epistasis if not accounted for.
>
> The authors developed a method for inferring the coefficients of specific epistatic interactions that does not rely on the choice of a reference genotype (referece-free analysis, RFA). The authors curated a diverse datasets of near combinatorically complete protein fitness landscapes. The authors then applied the RFA method together with a nonlinear sigmoidal function to jointly model nonspecific and specific epistasis in the chosen datasets. The main conclusion is that most protein fitness landscapes can be largely captured by only additive and pairwise interactions when nonspecific epistasis is accounted for.

1. > Overall, I think this is a very important topic and that the paper well written. I think the evidence for the lack of higher order interactions supported by the very marginal increase in prediction performance in these datasets is convincing.

   We appreciate that the reviewer finds our central conclusion convincing and significant.

2. > However, I have some major concerns about RFA model's purported superiority over the classical reference-based analysis (RBA) and its significance in drawing main conclusion of the paper, which I listed below.
   >
   > First, it is unclear what the role of RFA model is in this paper. The RFA model is great for inferring epistatic coefficients that are independent of a reference genotype. Furthermore, unlike the Fourier framework, coefficients of the RFA directly correspond to interactions among alleles across sites, making it an intuitive method for interpreting specific epistasis. However, this is not how the RFA model is mostly used in the paper. The main conclusion that there is negligible higher-order interaction in the test datasets has little to do with these desirable properties of RFA. And the same conclusion can be drawn by just fitting standard RBA regression using least squares. This is because although RFA and RBA have different parametrizations of the fitness landscape, they are indeed modeling in the exact same

space of functions. Specifically, for any RFA model, there is a corresponding RBA model (and vice versa) that gives identical predictions everywhere in the sequence space. So if the modeling fitting was done correctly (with the right choice of regularization), they should produce the exact same predictions.

R1 and R2 both commented that RFA and RBA-regression are identical or equivalent in some ways and that the significance of our development of RFA was therefore unclear or overstated. In our initial submission, we addressed RBA-regression to a limited extent, and we did so primarily in supplemental material. The reviewers' comments indicate to us that we need to address this issue more thoroughly and to do so in the main manuscript. We therefore performed several new analyses, modified main and supplemental figures, and rewrote the text to prominently address the differences between RFA and RBA-regression and the serious shortcomings of the latter. We believe that these changes establish clearly that RFA and RBA-regression are not equivalent or identical, that RFA represents a significant advance over RBA-regression, and that regression should not be used to estimate RBA models.

The major revisions can be found in Figs. 1c and 2c-e, Supplementary Fig. 1, and in the text at 233-272. The take-home message is that estimating RBA models using least-squares regression causes two distinct forms of strong bias, which lead to incorrect and anomalous conclusions; RFA does not suffer from these problems. We briefly summarize the argument here:

- The procedure for analyzing genetic architecture by RBA-regression as used in the literature is as follows: first use regression to fit a series of truncated RBA models to all the data to estimate the fraction of variance explained at each order, and then fit the complete model to estimate the specific terms of the RBA architecture.
- The first part of the RBA-regression procedure leads to a dramatically oversimplified estimate of the RBA genetic architecture, with greater variance attributed to low-order determinants and less to high-order terms than is true under the RBA formalism (Fig. 2). This bias of RBA-regression has already been established in the literature (Otwinowski & Plotkin, PMID 24843135), although it seems to have been insufficiently appreciated – perhaps because no alternative method has yet been available. The bias occurs because using regression to estimate any model finds the set of model coefficients that minimize the sum of squared error (SSE) across all genotypes; however, the true terms of truncated RBA models do not minimize the SSE across all genotypes. Rather, the RBA formalism is structured so that low-order terms exactly fit low-order mutants and do not attempt to predict higher-order mutants. When regression is used to fit a truncated RBA model, it finds the terms that minimize the SSE across all genotypes, which forces low-order terms to fit phenotypic variation that is produced by higher-order interactions that are excluded from the model. This allows the low-order model to explain more variance than it actually does under the true RBA architecture. This continues with each truncated model, systematically overestimating the variance explained at every order until the complete model is reached.
- The second bias arises in the last phase of the procedure, when the complete model is used to estimate the model terms. Here the estimated coefficients are subject to extreme propagation of measurement error, leading to estimates that deviate wildly from the true terms, with the magnitude of terms especially inflated at high orders. This occurs because estimation by regression using the complete model is identical to the exact

computation of RBA. We also show that RBA truncated models when fit by regression also overestimate the magnitude of the highest orders of terms that are included in the model (Supp. Fig. 1).

- When combined, these two forms of bias lead RBA-regression to the anomalous conclusion that high-order epistatic interactions explain little or no phenotypic variance but are somehow "widespread and significant" nevertheless (Dupic et al. *bioRxiv* 2024.01.29.577800).

- RFA does not suffer from these problems. RFA's terms at each order are defined as means over genotypes, so the true terms do minimize the SSE. RFA's terms and variance partition can therefore both be accurately estimated by regression. Moreover, each order of epistatic interaction produces patterns of variation that cannot be explained by lower orders, so using truncated models does not bias the variance partition (Fig. 1, Supp. Fig 1).

- In addition to these issues related to accuracy, there is a more fundamental sense in which RFA and RBA-regression are different: they model genetic architecture in different ways. The causal decompositions arrived at by RFA and RBA are fundamentally different, because the two approaches (whether estimated by regression or exact computation) parameterize sequence space differently, with model terms that correspond to different kinds of causal factors, which combine in different ways to yield the phenotype, and which consequently take on very different values. This deep difference is manifest whether or not the formalisms are estimated exactly or using regression (see Suppl. Fig. 5 of our paper for the different parameterizations of genetic architecture between RFA and RBA-regression).

- It is true that using regression to estimate RBA or RFA – or any other arbitrary model with the same degrees of freedom – will yield the same predicted phenotypes. But the purpose of analyzing genetic architecture is not to predict phenotypes (and certainly not to predict them using deliberately truncated models). Rather the purpose is to understand genetic architecture – to characterize the sequence features that cause variation in a protein's phenotype – and the RBA and RFA decompositions are very different, as described above.

- A corollary of the fact that any regression-coupled model will predict the same phenotypes is that they will also yield the same variance partition. This is a property of regression, not of the model formalisms. RFA is distinct from RBA in that the regression-based partition is an accurate estimate of the variance actually explained by the model. One could use RBA-regression to obtain the RFA variance partition, but without knowing the RFA formalism, the inferred partition would makes no sense; further, it would be misleading, because it would misrepresent the true variance partition under the causal decomposition being used.

RFA and RBA estimated by regression are therefore not equivalent in theory or in practice. The RFA model is structured so that its coefficients and variance partition can be estimated by regression. RBA is structured in such a way that neither the coefficients nor the variance partition can be accurately estimated by regression. The RBA formalism is appropriate for asking certain kinds of questions – those that pertain to the effect of mutations in the local neighborhood of a designated wild-type sequence -- but it should not be estimated by regression.

3. In the comment above, the reviewer suggested that, because regression predicts identical phenotypes irrespective of the model being used, we might have reached the same conclusions if we had used RBA-regression. This is an important issue, and we appreciate the impetus to clarify it.

   If we had used RBA-regression instead of RFA, we would have been led to numerous conclusions that are different from those that we drew, and several of these would have been erroneous. Specifically:

   - We would have falsely concluded that the RBA architecture is simple – when in fact the architecture of most proteins using the RBA formalism is quite complex, as demonstrated by our Fig. 4c and by numerous prior studies (e.g., Weinreich et al. PMID 16601193, Poelwijk et al. PMID 31527666, Phillips et al. PMID 34491198). The complexity of RBA architecture arises from the way in which the sequence-phenotype relationship is modeled relative to a wild-type reference, where the states themselves have no effects or interactions – which means that their effects must enter the modeled as higher-order interactions when those states are lost – as well as from the propagation of error. The architecture of proteins actually *is* complex from the RBA perspective. Had we used RBA-regression, we would have mischaracterized the RBA decomposition of sequence-phenotype relationships.
   - If we had used RBA-regression, we would have falsely concluded that the RBA variance partition does not depend on the reference protein chosen. In fact, the architecture of the proteins we analyzed does depend strongly on the particular sequence designated as wild-type protein, as shown in our Figs. 2a and Suppl. Fig S6. The dependence on wild-type is a feature of the RBA formalism, and it would be erroneous to conclude otherwise.
   - If we had used regression to estimate the RBA variance partition, we would have inferred a variance partition identical to that of RFA. But we would not understand its meaning or significance, because we would lack the RFA formalism under which that variance partition is correct or makes sense.
   - If we had fit model terms to the data using RBA-regression, we would have concluded that many high-order interactions are large, in contrast to our finding using RFA that the vast majority are very small (Supplementary Fig. 5). These apparently large high-order terms would also have undermined the conclusion of an apparently simple architecture suggested by the variance partition, and it would have (and should have!) called into question the method we were using. This is in contrast to what we actually found using RFA: the variance partition and inferred terms provide a consistent picture of a simple genetic architecture dominated by first- and second-order terms across the 20 datasets.
   - Our analysis of the sparsity of genetic architecture uses the inferred terms of the RFA model (Figs. 6 and 7). We found that we need a very small number of RFA terms to account for the vast majority of functional variation in virtually all proteins analyzed. If we had inferred terms by RBA-regression, the low-order terms would be poor predictors of the phenotypes, and we would need many high-order terms to reach high predictive accuracy. This would be the opposite of the sparse architecture that we found with RFA.
   - The RFA terms that we inferred are the basis of our findings with respect to the three protein datasets we explore as quantitative case studies (Fig. 8). If we had used RBA-regression to estimate the terms, we would have been led to radically different conclusions from those we reached in each case, which showed how the specific

quantitative effects of a few key amino acid states and pairs determine each protein's function and specificity.

We believe that these points establish that the findings in our manuscript are different from what we would have concluded if we had used RBA-regression and had never developed RFA. We appreciate the reviewer pushing us to make this clear in the manuscript, and we hope these changes are sufficient for this purpose.

4. Second, I think there are some unfair comparisons between the RBA and RFA model. My overall criticism is: the RBA model was not fitted using the same amount of data with appropriate regularized least squares regression like the RFA model. And when the RBA model is fitted in the correct way, most differences the authors highlighted would go away. In the supplement to the bioRxiv preprint, the authors pointed out "A reference-based effect of order k is the sum or subtraction of 2k phenotypes", meaning the effect of a mutation is equal to the difference between the single mutant and the WT; a pairwise coefficient is the sum of WT, single mutants, and double mutants, with alternating signs. When formulated this way, of course the RBA model performs less well than the RFA model, simple because it uses less data. In contrast, the right way to infer effects in an RBA model is to fit a model to all genotypes in the data, such that for any genotype, its phenotype = sum of mutation effects + sum of pairwise interactions + sum of higher order interactions. Because regardless of if there is a chosen WT, the fitness of a genotype is decomposed in the same way as the RFA model (except with an intercept WT term), so coefficients of the model should be inferred using all the data, instead of just the lower order mutants.

   We agree with the reviewer that RFA and RBA when computed exactly use different amounts of data to estimate coefficients, and this is a key cause of the greater susceptibility of RBA to noise. But this difference is not an artifact of an implementation choice. Rather, it is a feature of the formalisms themselves -- the different ways in which their coefficients are defined and the causes of phenotypic variation are decomposed into specific factors. RFA terms are defined as averages over large numbers of genotypes, whereas RBA terms are defined as sums/differences over large chains of individual genotypes. For the reasons described in items 2 and 3 above, the right way to implement the RBA model is not using regression, because this leads to bias and error in the estimates of model terms and the variance explained per order. In revising the manuscript, we have tried to make a clear distinction between the issues that arise from the RBA formalism itself and those that arise from issues in implementation.

5. R1 commented on Fig. 1d (which is now Fig. 2b): This unfairness is manifested in the following places. In Figure 1d, the authors show RBA models have much higher variance in the estimation of epistatic effects than the RFA model when measurement noise is present. I think if the RBA model coefficients, if inferred using least squares by fitting the RBA model to all the data will produce very similar behavior. That is because the same mutation/epistatic interaction in the RBA model appears as many time as in the RFA model.

   We appreciate the suggestion to infer model terms using RBA-regression for comparison to the performance of RFA in the face of measurement noise. We agree and have performed a new analysis and added new text to address this suggestion. This analysis shows that using RBA-regression to fit model terms to the data yields large error, especially at high orders (Fig. 2e). This result arises because RBA-regression with the complete model is equivalent to not using regression at all, so the inferred terms are exactly as unstable to noise as those

that are directly computed. We also show that errors occur if we estimate terms using RBA-regression with truncated models, because in this case low-order terms absorb variation caused by the unmodeled high-order effects that are excluded from the model (Fig. 2c and Supplementary Fig. 1).

6. R1 commented about Figure 1e (now 3b): In Figure 1e, the Fourier and BA model apparently have lower out of sample R^2. Again, I think if the RBA model was fitted correctly, the performance should be similar.

   We agree with the reviewer's prediction that RBA-regression is likely to yield similar R^2 as RFA, but the purpose of this figure is to compare RFA to Fourier and background-averaged analysis as part of the discussion of the differences among the latter three frameworks. The comparison of RFA to RBA and RBA-regression is made earlier in the paper. Having established earlier that RBA-regression would yield an erroneous estimate of the variance explained by the RBA formalism at each order, we do not subsequently use RBA-regression, in order to avoid mischaracterizing the genetic architecture. We explain this in the main text now at lines 258-272.

7. R1 commented on Fig. 2C and its relationship to Fig. 2A (now 4C and 4A). In Figure 2C, what is the performance of the RFA model in the same setting? Again, if both models were fitted using least squares on the same amount of data, the performance should be similar. Furthermore, it appears (at least at first glance for some readers) that the RBA model does not perform as well as the RFA mode, even though panel A and C are measuring different things. Juxtaposition of these two panels is misleading.

   We agree that these panels were positioned next to each other in a way that invited a direct visual comparison, which is not appropriate because the panels measure different things on their *y*-axes and involve analyzing different amounts of data. (Reviewer 2 made a similar comment, noting that this is not an apples-to-apples comparison, because the implicit "reference" used by RFA is the mean phenotype, which makes it inherently more robust to cross-validation.)

   This is a challenging issue: we do not want to invite an uncritical direct comparison, but a true apples-to-apples comparison is impossible. We cannot evaluate RBA by cross-validation (as in Fig. 4A) – estimating a model from a random subset of data and then characterizing predictive accuracy to the unsampled variants – because RBA at each order is constructed using only mutants at that order and below. Exact computation of RBA at any order therefore cannot be performed on a random sample of genotype space, because it requires all genotypes at the specified order. (We do not use RBA-regression as a substitute because, as we have shown, this approach mischaracterizes the model terms and the variance explained.) Conversely, we cannot fit RFA to only low-order mutants and predict high-order mutants (as in Fig. 4C), because RFA at every order is constructed to estimate its terms from genotypes across sequence space as a whole; moreover, there is no concept of "low-order" mutants in RFA because there is no wild-type sequence relative to which mutations can be defined.

   Our goal is therefore to be maximally clear about each analysis and to avoid direct comparison between them. We therefore modified the text and figures to precisely state what we are doing and to avoid inviting direct comparison between the out-of-sample R^2 for RFA and the low-to-high-order R^2 that we use to describe RBA. First, we moved Fig. 4c so that it is not directly juxtaposed against Fig. 4a. Second, we modified the y-axis labels and

the legend to indicate the different things being represented in these two plots. Finally, we rewrote the relevant text so that the metrics used and the claims based upon them are clear and precise, and so that we do not imply a claim that requires a direct comparison of the different metrics. Specifically, for RFA (lines 377-384): “We then used cross-validation to estimate the fraction of phenotypic variance explained at each order as the out-of-sample R^2 when the models are fit to a random subset of data and then used to predict the phenotypes of unsampled variants…. These analyses show that the genetic architecture of proteins is simple: knowing just the additive effects and pairwise interactions, coupled with a simple model of nonspecific epistasis, is sufficient to accurately predict and explain phenotypes across the entire ensemble of sequences. Higher-order interactions are not completely absent, but they are weak and limited to a very small fraction of genotypes.” For RBA (lines 413-420): “We also examined the 20 datasets using RBA. We exactly computed the first-, second-, and third-order RBA models, using the sigmoid link function with parameters that maximize predictive accuracy for all genotypes. We then used each fitted model to predict the phenotypes of the higher-order mutants, which were not used to compute the model. The median R^2 across datasets is <0.2 for all three model orders; the vast majority of phenotypic variation is thus left to be explained by higher-order epistasis (Fig. 4c). The RBA formalism therefore leads to a complex and idiosyncratic description of the genetic architecture of these proteins.”

We believe that we are now accurately describing the apples as apples and the oranges as oranges, and drawing appropriate conclusions from the data in each case. We appreciate being pushed to clarify this part of the analysis.

8. Despite this, the RFA model does have its role (albeit in more nuanced places) in the paper, for example, measuring the sparsity of the fitness landscape (Fig 4) and the variance decomposition. But I do not reckon that it is the main selling point of the paper. I think a viable route for the authors is to put the comparison between RBA and RFA in less conspicuous places and focus on the main conclusion of the paper, which is the simplicity of protein sequence function relationships.

   We appreciate this suggestion. However, we believe that for our inference about the simplicity and sparsity of proteins’ genetic architecture to be valid, development of RFA was necessary, for the reasons that we argued in items 2 and 3 above. Moreover, the reviewers’ comments indicate to us how important it is to make clear the differences between RFA and RBA-regression, rather than doing so in a less prominent place, as we did before. More generally, we have found in presenting this work that it is important for us to explain our method and how it differs from existing methods early in our argument so that the reader can understand why it leads to an accurate and concise decomposition of genetic architecture and why our conclusions differ from many prior studies.

Responses to Reviewer 2

The nature and prevalence of epistasis, as well as how it varies across different orders, has long been a central concern of evolutionary genetics and biochemistry, but experimental results were scarce. Advances in high-throughput phenotype-genotype mapping, such as deep mutational scanning, have produced complete high-order combinatorial mutational datasets for a number of targets. This has stimulated further interest in the quantitative understanding of the underlying fitness landscapes.

In this work, the authors present a formalism for the description of a mutational fitness

landscape which they call “reference-free analysis” (RFA). They claim this framework is superior to alternative approaches in inferential robustness and parameter interpretability. They analyze 20 datasets across 13 targets with RFA and make several major claims: that protein sequence-function relationships are “simple,” that genetic architecture is “sparse,” and that nonspecific epistasis is attributable primarily to measurement saturation.

Major issues

9. A commentary on an initial draft of this manuscript by Dupic et al. [PMID 38352387], who also produced several of the datasets analyzed here, made several critiques of these claims, some of which we find compelling and others less so. Briefly, we agree that RFA is equivalent to certain formulations of reference-based analysis (RBA) and Hadamard-transformed landscapes, and that the arguments for sparsity and simplicity require more care, but we believe this will prove an interesting and insightful intervention with some minor reframing.

   In general, we find that the interest of this work lies in the comparative analysis across multiple datasets with a unified framework rather than in the value of RFA itself. With this in mind, we feel several of the major claims are tantalizing but insufficiently proven, and some of the language could be made more precise. Despite the issues discussed below, we emphasize that this work should be of broad interest to evolutionary and protein biologists and recommend it for revision at Nature Communications.

   With respect to RFA and RBA-regression being equivalent, we understand that this idea was raised by Dupic et al. RFA and RBA-regression are not equivalent, however: they decompose genetic architecture in different ways, and RBA-regression leads to biased and anomalous conclusions – problems that do not occur with RFA. We have performed new analyses and made extensive changes to the text and figures to explain these issues clearly. Please see items 2 and 3 above for details. With respect to the idea that RFA and “Hadamard” approaches are equivalent, please see item 10 below.

10. There is undue emphasis on the novelty and superiority of RFA given that it is exactly equivalent to reference-based (RBA) and Hadamard approaches. A term-by-term comparison of RFA with an expansion in Walsh basis (see, e.g., Weinrich, et al. [PMID 24290990]) makes this clear. Given this, RFA can only be superior in terms of parameter interpretation. The authors recognize this and argue that RFA produces a maximally efficient and clear encoding of genetic architecture, but we feel this is both inadequate and unnecessary. The discussion of the superiority of RFA-derived parameters is too short and not concrete enough, and the emphasis on proving the superiority of RFA distracts from what we found to be the genuinely interesting biological results.

    We appreciate the impetus to make our presentation on these points more concrete and complete. With respect to the comment that RFA is exactly equivalent to so-called Hadamard approaches and that this is revealed using a term-by-term comparison, we understand that this idea is presented in the preprint of Dupic et al., but it is incorrect. Although RFA, Fourier analysis (FA), and background averaging (BA) all involve some kind of centering, the formalisms impose very different decompositions on the sequence-function relationship, with terms that have distinct meanings, are computed differently, combine in very different ways with respect to the phenotype, and therefore perform differently in the face of noise and partial sampling. The claim of equivalence may arise in part because Dupic et al. define RFA using an incorrect equation (their Eq. 1 is not RFA, which is correctly

defined in our Fig. 1b). Dupic et al.'s Eq. 1 is in fact FA in the special case where the number of states $q = 2$. Incorrectly defining RFA using an equation for FA could lead to the erroneous belief that FA and RFA are identical.

We agree that a term-by-term comparison is useful for understanding the differences and similarities among these formalisms. When $q = 2$, the terms of the three models are related in a simple fashion – by scaling factors that are unique to each order – but this relationship does not hold when $q > 2$. In the latter case, the terms of the formalisms can be interconverted only through a complicated set of equations. For example, with four states, the first-order RFA effect of state A at site i is related to the three Fourier coefficients w_i, y_i, and k_i as $e_i(A) = w_i - y_i - k_i$; each of the other first-order RFA terms involves a different set of signs. At second order, the RFA pairwise effect $e_{ij}(A, A) = w_i w_j - w_i y_j - w_i k_j - y_i w_j + y_i y_j + y_i k_j - k_i w_j + k_i y_j + k_i k_j$. Each third-order RFA interaction is a uniquely signed sum across 27 FA coefficients, and so on. When there are 20 states, the relationship becomes exponentially more complex: each first-order RFA term is a uniquely signed sum across 19 Fourier first-order coefficients, each second-order RFA term is a signed sum across 361 second-order FA coefficients, each third-order RFA terms is a signed sum across 6859 FA coefficients, etc. The interconversions between BA and RFA (and between BA and FA) are similarly complex, but in that case the signs are different and each term in the equation is weighted by its order. We have included these comparisons in the paper at Fig. 3a, Supplementary Text 1.2 and 1.3, and in the text at lines 277-297. We have emphasized in the text that these complex encodings in FA and RFA make it difficult to understand from the terms of the model how phenotype arises from genotype.

Another key difference among the three frameworks is in the way in which the predicted phenotype is a function of the model coefficients, and this has important practical implications. In RFA, the phenotype of any genotype is the sum of only those coefficients for the states found in the genotype (Fig. 3a). In FA and BA, each phenotype is a signed and weighted function over every coefficient in the entire model. As a result, RFA performs slightly better in the face of measurement noise and missing genotypes than FA and notable better than BA (Fig. 3b) when q is large. Moreover, although FA and BA were originated many years ago, they have almost never been applied to multi-state datasets because of their formal complexity. The development of RFA therefore greatly facilitated our analysis of multi-state datasets, most of which had never been systematically analyzed by genetic models. These issues are addressed in Figs. 3a and Fig 3b and in the text at lines 299-313.

We appreciate the reviewers' point that our discussion of the differences among these formalisms and the impacts of those differences was too short and insufficiently concrete. We hope that these modifications address this suggestion.

11. With respect to the comment about undue emphasis on the novelty of RFA, we agree that the most important part of the paper is the insight that protein architecture in these DMS datasets is simple and sparse. However, we could not have reached these conclusions if we had not developed RFA, and we believe it is important that the reader understand our method and how it differs from other approaches so that they can understand why we reach different conclusions from most prior studies and why these conclusions are valid.

    We have already made this case with respect to RBA in item 3 above. If we had used FA or BA, we would have thought that the genetic architecture is more complex than it is under RFA, because FA and BA are more sensitive to noise and partial sampling when the number of states is large, so low-order terms in those formalisms would explain less variance than

RFA does. Most significantly, the architecture would also have appeared to be considerably less sparse than we found it to be using RFA (under which a relatively small number of states and pairs explain 90% of variance), because in the 20-state spaces the effects of each amino acid is coded in FA and BA as 19 coefficients and each pair of amino acids as 361 terms. Further, we could not have made sense of the resulting architecture as the effect of a small number of particular residues and their pairwise interactions, because that is not what the terms of these architectures represent. (One might say that we could have analyzed the datasets using FA or BA and then "remapped" those terms to obtain the RFA architecture using the equations we present in the supplement, but that presupposes that we had developed the RFA formalism already – and in that case one might as well just use RFA directly.) Moreover, it would have been impossible for us to use FA and BA for our study, because there has been no accessible implementation to date for large sequence spaces.

We therefore believe that our biological conclusions about protein architecture depend on the fact that we developed RFA, and that the differences between RFA and the other formalisms need to be clearly explained. Our goal is not to say that RFA is better in all settings than other methods but to argue that RFA is useful for understanding global genetic architecture of proteins because it provides a maximally efficient and transparent encoding of the relationship between protein sequence and function, and it can be accurately estimated from noisy and partially sampled datasets. We have tried to make this case clearly by making revisions throughout the manuscript, particularly by better explaining the goals of interpreting global genetic architecture (lines 106-128) and more clearly explicating the ways in which RFA can achieve these particular goals compared to the other approaches in each ensuing section.

12. The claim that measurement bounding is the major cause of nonspecific epistasis is insufficiently argued. We believe this is an important and major claim in the work, and that a deeper understanding of the sources of nonspecific epistasis will be of major use to the field. As written, however, this manuscript only demonstrates that use of a sigmoid link function improves model fit with RFA. This does not support the larger claim that "the primary cause of nonspecific epistasis is phenotype bounding."

    We appreciate the reviewer's critical attention to this issue, because it plays an important role in our manuscript. We have worked to strengthen and clarify this argument by modifying the claim, performing new analysis, and rewriting the text to make logic of the argument clearer. Specifically, we now claim 1) phenotype bounding is a major cause of nonspecific epistasis, and 2) although the causes of nonspecific nonlinearity are likely to be complex and to vary among datasets, the simple sigmoid link function captures its most salient features. The evidence for this claim is as follows:

    (i) The sigmoid link function explicitly models the effect of phenotype bounding and the nonlinearity it imposes on observed genetic effect sizes.

    (ii) Using this model improves the fit to virtually all datasets (Fig. 5a).

    (iii) Using the model simplifies the inferred specific genetic architecture in virtually all datasets (Fig. 5b).

    (iv) The improvement in fit is directly proportional to the fraction of genotypes with phenotypes outside the dynamic range, and the correlation in very strong (Fig. 5d).

(v) Simulations show that when phenotypes are generated with phenotype bounds, the link function effectively captures this phenomenon and allows the RFA variance partition to be accurately inferred (Fig. 3c); and

(vi) Simulations also show that when phenotypes are generated without phenotype bounds under realistic forms of specific epistasis, the link function does not impose nonspecific epistasis and the correct RFA variance partition is inferred (Supplementary Fig. 2).

We believe that taken together this evidence provides strong support for the revised claim. The major changes in the text for this comment are at lines 317-358, 445-448 and 623-632 and in the figures cited above.

13. The authors claim that protein sequence-structure relationships are simple (as in, first-order models are sufficient to predict most phenotypes). This is an interesting and important result, and generally in line with other findings that high-order epistasis seems to be rare. The authors clearly demonstrate this with their work. We believe the interpretation of these results requires more care, however. “Simplicity” may be misleading. Often evolutionary dynamics on a landscape and not the landscape itself is the subject of interest, and it is not obvious that “simple” sequence-structure relationships will also produce simple evolutionary outcomes.

We appreciate the reviewers’ raising this point. We agree that simplicity of genetic architecture does not necessarily entail simplicity of evolutionary dynamics. Indeed, in a separate paper we showed that landscapes generated by only first- and second-order coefficients can engender complex evolutionary dynamics (Metzger et al. *eLife* 10.7554/eLife.88737.2). In the current manuscript, the point that we are making is that it is the genetic architecture that is simple, not the evolutionary process. To address the reviewer’s comment, we have explicitly defined a simple genetic architecture as one in which epistasis above order two is negligible (lines 407-411), and in the discussion we point out that the pairwise epistasis we observed could be sufficient to substantially affect evolutionary processes (lines 601-604).

14. Further, given the exact equivalence of (a form of) RBA and Hadamard approaches with RFA, the argument for RFA based on this form of simplicity is effectively a belief that terms in a sequence-structure relationship should be of order one, or an argument for a form of parsimony that has intuitive appeal but should be made more explicit.

We agree that we should be clear about the core criteria and values that underlie our evaluation of the various formalisms. With respect to the first part of this comment, we would like to reiterate that RFA is not equivalent to any of the other formalisms. With respect to the second part, we do not believe a priori that sequence-structure relationship should be of order one; in fact, our findings indicate that order two is quite important in the genetic architecture of many proteins. What we do believe is that a method for describing genetic architecture should provide the most concise possible description of causality in the system that explains the data, rather than unnecessarily invoking idiosyncrasy. This priority, which is consistent with the fundamentals of modern statistical and scientific inference, plays a key role in our evaluation of RFA vs. RBA and of the effect of incorporating nonspecific epistasis. We also believe that a method should be accurately estimable from real-world data, which distinguishes RFA from RBA (either by exact computation or regression). And we believe that transparency and interpretability, as well as robustness to noise, are also key values, which distinguish RFA from FA and BA.

We agree with the reviewer that these considerations ought to be made explicit, and we have modified the text to say this (see esp. lines 106-133, as well as 163-167).

15. The authors argue that genetic architecture is sparse (as in, few terms of a model are sufficient to predict most phenotypes). This is another important and interesting result which deserves additional care. In particular, we wonder why only a third-order model was used for all datasets, and whether RFA is any more sparse than RBA.

    We used third-order and lower RFA models because we already showed that these models explain virtually all of the phenotypic variance. Using higher-order models would lead to the same conclusion – that most variance is explained by the same small set of low-order terms – but with considerably more computation caused by introducing a huge number of additional terms (each of which affects fewer genotypes) and which we already know explain very little variance.

    We did not analyze sparsity under the RBA formalism because RBA does not – and is not intended to – provide a simple or sparse decomposition. Our data show that under RBA fourth- and higher-order terms explain most genetic variance in most of the 20 datasets (Fig. 4c); high-order terms influence progressively smaller numbers of genotypes, so predicting phenotypes across most of sequence space will by necessity require many more terms than RFA does. Consistent with this, Poelwijk et al. (PMC6746860) found that the avGFP dataset is not sparse under RBA but sparse under BA, and we have also shown that it is sparse under RFA. We did not perform RBA analysis for our 20 datasets, because at this point in the manuscript, our purpose is to characterize the sparsity of genetic architecture under the RFA formalism, not to compare it to RBA.

16. The utility of this work for experimentalists could be enhanced through some further comparison with reference-based approaches. One issue we see is that in this framework the individual parameters may be intractable: for example, the zero-order epistatic coefficient e_0 is defined with the same information as a n-th order coefficient e_n, that is, the entire combinatorial genotype space. In many DMS settings, a particular sequence is actually privileged for one reason or another, and so experimentalists anchor their exploration of sequence space, even at high orders, at that point. RFA may not be able to guide experiments due to the data dependencies of the parameters. The authors do mention that multiple screens across multiple homologs may be one approach to estimate an RFA landscape from limited data, but it is not obvious if this would work. We feel some further discussion of the phenomenology of the recovered parameters would be useful to guide potential applications of this approach.

    We appreciate the suggestion to directly address the experimental implications of our findings and model. We agree that RFA models cannot be estimated from DMS designs in which a wild-type sequence is used as the starting point for single-site or pairwise mutagenesis. But these kinds of datasets cannot be used to infer global genetic architecture under any formalism, even RBA or RBA-regression. Assessing global genetic architecture requires an experimental design that measures genotypes far from any individual reference sequence, and the need for this kind of data is not unique to RFA.

    The need for data from genotypes far from wild-type does not mean that RFA models are intractable. Our results indicate that estimating global genetic architecture is more tractable using RFA than previously thought, because it does not require near-complete combinatorial

sampling. Because of RFA's structure, its terms are particularly tractable for estimation using partially sampled datasets and truncated models, an advantage over other existing formalisms. Despite this encouraging news, however, characterizing even the sparse RFA architecture in large sequence spaces may often outstrip random sampling from such combinatorial libraries. Our findings suggest several options, including complete combinatorial assessment when sequence space is not too large, random sampling from combinatorial space as long as dynamic range is adequate, and low-order combinatorial assessment across several starting points when these conditions are not met. To address this comment, we discuss this issue more extensively than before in the discussion, and we note that further research will be necessary to illuminate the effectiveness of the last strategy in particular (lines 643-659).

17. The claims about sparsity and simplicity of protein sequence-function relationships are based upon datasets of proteins with relatively small genotype space (with the exception of ParB and GB1), and are based upon phenotypes (e.g., binding, fluorescence) that are treated independently and measured on a minimum/maximum spectrum. While these simpler representations may be necessary for the isolated analysis of a particular dataset, they do not consider dependency between phenotypes (e.g., binding and transcription activity of transcription factors) that may be encoded by more widespread and high-order epistatic interactions.

    We agree that these are important limitations. The "multidimensional" nature of allosteric phenotypes could indeed cause them to have more complex architectures, as could the fact that allostery involves differences in occupancy between multiple folded conformations. We have modified the discussion to acknowledge this point and its potential significance. We also agree about the small size of the sequence space examined in these datasets and have noted in this in the discussion. Please see lines 595-601 and 606-621.

18. The authors demonstrate that their model can explain most of the phenotypic variance of avGFP with just first- and second-order effects despite its dense epistatic network, but the relevant sites that engage in epistasis are structurally close together. Proteins that experience allosteric effects over longer distances may also be encoded by more widespread and high-order epistatic interactions. Given that the analyzed datasets also have existing structures, it is important to consider why the model can explain phenotypic variance with sparse and lower-order effects in the context of the protein's structure.

    As we understand this comment, R2 are considering whether the simple epistatic architecture that we observe might be an artefact of biased spatial distribution of mutated sites in the available datasets. The mutated sites in the proteins we analyzed are structurally clustered in some datasets but are dispersed in others, so the simplicity is unlikely to be an artefact of spatially biased site sampling. We agree that allosteric regulation as a phenotype is not addressed by these datasets, and the long-distance conformational dependencies that these phenotypes involve could involve more epistasis. We have modified the discussion at lines 5888-604 and 606-621 to acknowledge these points.

    The reviewers may also be asking for a structural explanation for why the pattern of epistatic interactions in the datasets is so simple. We also desire such an explanation. As we note in the discussion at lines 614-618, a possible explanation is that the datasets we examined all assayed functions carried out by a single protein fold, in which couplings may be largely static; when different conformations are present, these may impose different couplings, and functions that involve a shift between conformations would be expected to produce more

high-order epistasis. A less speculative structural explanation would require us to contrast datasets that manifest complex genetic architectures to those with simple architectures, but this is currently impossible, because none of the available datasets are complex. We modified the discussion to make this point and acknowledge the provisional nature of our physical understanding this point (lines 618-621).

Minor Issues

R2 made several helpful suggestions labeled as minor issues. We address them all below.

19. Figure 1a - legends for the models would be helpful here. We have described the toy examples in detail in the legend.

20. Figure 1c: In the context and in the figure description, they don't mention what some of the variables are. For easier understanding of the passage and interpretation of data, variable definitions such as f0 and b0 would be good. We have described the notation in detail in the legend (Fig. 3a in the updated manuscript).

21. The authors claim that because effect terms for RBA methods are computed as a chain of sums and subtractions of individual variants, measurement errors propagate and snowball with epistatic order. They demonstrate this in Fig 1D using phenotypic data that was simulated from a known genetic architecture with normally distributed noise. Additional information about how this data was simulated and why it is representative of observed phenotypic datasets would make this claim stronger. We have provided detail on the simulation condition in the legend for Fig. 1c in the updated manuscript, which is used to evaluate both RFA (in 1c) and RBA (in 2b). The purpose of these analyses is to evaluate the performance of the formalisms on noisy data. The amount of noise we simulated is typical of experimental datasets.

22. The choice of states and sites in Fig 1e seems somewhat arbitrary. Some explanation of the choices would improve clarity. In this figure, genetic architectures with an increasing number of states were evaluated to show that the performance of FA and BA relative to RFA drops with increasing number of states. The number of sites was adjusted to keep the total number of genotypes roughly constant.

23. Figure 3B: This is a little tricky to grasp and some sort of cartoon might help the reader. The equivalent plots to 4a for other datasets should be included. Point 6. We did not include an explanatory figure because of space constraint, but we explain the result in more detail in the corresponding results section.

24. Figure 6B: the average genetic score value (-7.8) is mentioned in the text. It may be nice to include this value on the plot as well. We have indicated the intercept with a vertical line and an indicator in the figure.

25. Using cross-validated, out-of-sample R^2 to compare RFA and RBA may not be an apples-to-apples comparison. The implicit “reference” used by RFA is the mean phenotype, making it inherently more robust to cross-validation statistical measures than reference-based analyses that rely on a WT sequence and may suffer from cross-validation sets that do not have WT, WT-adjacent sequences.

We agree that the figures do not present an apples-to-apples comparison and have addressed this comment in item 7 above.

26. While the authors do provide evidence that a small number of reference free terms can determine function, they also show that genetic architecture of proteins cannot be captured by sparse sampling. This is an interesting result and worthy of more discussion.

    We appreciate the reviewer's interest in this initially puzzling finding. We have included more detail at lines 491-502 explaining why the simple genetic architectures cannot be efficiently inferred by random sampling. Specifically, we show that a principal cause is the limited dynamic range of measurement, because genotypes outside the range provide little quantitative information on genetic effects. For cases when sequence space is huge and most genotypes fall outside the dynamic range, sample sizes must be very large to gather a set of quantitatively informative genotypes to adequately characterize the terms of the model.

27. The authors apply a framework for understanding genetic architecture where genetic variation is analogous to change in energy, and functional vs. nonfunctional phenotypes are described in terms of the relative occupancy of a functional vs. nonfunctional state. For clarity, emphasizing that this is an apparent or effective free energy would be helpful. We have followed the reviewer's advice and modified this section to bring the concept of the apparent free energy out more clearly (especially lines 511-524 and within each case study).

28. Suggested citations: A publication by Buda, Miton and Tokuriki [PMID: 38129396] performed a similar comparative analysis of combinatorial landscapes and made somewhat different claims. Recent work from David McCandlish [PMIDs 32286265, 36129941, 35428271] has investigated genotype-phenotype map structures. In Tonner, et al. [PMID 35733251] the power of a number of alternative genotype-phenotype functional forms is compared, including linear and spline models.

    We appreciate these readings and have cited the papers by Buda et al. and Wong et al. (refs. 8 and 53 at line 45). We did not cite the others, because they did not use any of the methods we evaluate in this paper and did not assess the complexity or sparsity of genetic architecture (in the sense of quantifying the order of effects or the fraction of terms required to explain variance). Incorporating would therefore have required that we expand the scope of the text, which we are trying to keep tightly focused to comply with length limits.

Response to Reviewer 3

In their manuscript Park et al. describe a formalism to derive the genetic architecture of a protein, while accounting for unspecific epistatic effects. Different than other common methods their "reference free" approach sets the global average across all sequences as the zero order term /ground state and employs a sigmoid link function that is simultaneously fitted together with the epistatic coefficients for the model. The sigmoid link function is meant to capture all unspecific epistasis/non-linearities inherent to the data as e.g. caused by the measurement process. Using their method, they re-analyzed 20 previously published datasets using their and complementary frameworks. They show that heir framework appears to be more robust and furthermore suggests that the genetic architecture of protein function is rather simple, consisting mostly of additive terms, a few pairwise epistatic terms and almost no higher-order terms. This result generalizes recent observation made on e.g. protein stability (1) and tRNA functon(2) to the large variety of protein functions measured in

the 20 analyzed datasets.

There are few points though that presented themselves while reading the manuscript.

29. While an intuitive choice for a link function, no rational is given for choosing a sigmoid function. It would be interesting whether the authors have tested other sensible link functions with a similar result. We support the idea of unspecific epistasis e.g. non linearities introduced due to the involved measurement process and are intrigued by the authors' results. However, the function is used as a catch all term without validating or exploring this idea further by using a set of measured experimental values quantifying the relationship between measured and latent phenotype. Such measurements could deliver the upper and lower bounds the authors use as parameters for the link function.

    We have clarified the rationale and justification for using this link function by making textual changes and adding a new analysis. Specifically, we explained that diminishing effects of genetic variation near upper and lower phenotype bounds are likely to affect the majority of DMS datasets, and the sigmoid function provides a simple and easily estimable way of modeling this (lines 324-331). We modified our conclusion to state clearly that the causes of NSE are likely to be complex and to vary among datasets. Further, we added simulation-based analyses that show that 1) the sigmoid link function combined with RFA effectively captures the effect of phenotype bounds (Fig. 3c) and 2) when such bounds are absent, the sigmoid link function does not lead to false positive inferences, inferring minima and maxima well beyond the observed range of phenotypes and having no effect on the variance partition inferred for specific genetic effects (Supplementary Fig. 2). We acknowledge explicitly that the sigmoid relationship between genetic score and phenotype is not the only way that NSE could be modeled, and we suggest further research to examine the performance of other potential link functions (lines 623-632). Further discussion on this point is above at item 12.

30. Furthermore, in some cases the relation between assayed phenotype and function already follows a non-linear relationship e.g in cases were the impact of mutations on ddG binding itself is assayed, how would that be impacted by the unspecific link function?

    We added analyses that show that using the sigmoid link function does not affect the inferred specific genetic architecture when phenotypes are not affected by bounding– as might occur if phenotypes are transformed onto a ddG scale and all datapoints are within the dynamic range of measurement (Supplementary Fig. 4). However, we know of no measurement technique with an unlimited dynamic range, so bounding should be accounted for no matter what protein or phenotype is investigated. For example, several of the DMS studies that we analyzed expressed phenotypes on a ddG scale, which are transformations from estimates of occupancy (which by definition is bounded at 0 and 1); in most cases, occupancy is estimated using assays that measure fluorescence when proteins displayed on the cell surface bind to fluorescently-tagged ligands. These assays always have limited dynamic range over which changes in fluorescence (which decline near the maximum and minimum of occupancy) can be distinguished from noise. Indeed, in several of these ddG datasets, there is strong evidence of limited dynamic range, such as the CR9114-B dataset in which 99% of variants are at the apparent lower bound of measurement. We have discussed this point in the text at lines 82-88.

31. While the reference free approach is suggested by the authors to have clear advantages in terms of robustness to noise, completeness of genotypes and number of epistatic orders

studied (Fig.1), a recent publication by Dupic, Phillips and Desai (3) suggest that this is partly due to the way the epistatic effects were recalculated using the comparative methods. We would be interested in the authors response to that and would like to know how their recalculation compares to the original methods used in the studies they selected as datasets.

We have added a section and figures to address this issue. We show that RBA-regression produces biased estimates of the variance partition, inaccurate estimates of specific genetic effects, and is extremely sensitive to noise. For details, please see items 2 and 3 above.

32. Especially, since the authors themselves state that the epistatic coefficients derived via the different methods can transformed into each other by a simple linear remapping: “Although the coefficients of any of the four formalisms can be converted into those of the others using a linear re-mapping …” This would imply that the different results only differ in scale, but describe the same underlying genotype to phenotype mapping. As also noted in the paper cited above, this would imply that the difference lies in the interpretation of the coefficients.

We appreciate the impetus to clarify this point. The phrase “Linear remapping” invites misunderstanding of the relationship among the formalisms, so we have removed this phrase from our description and replaced it with a more precise description. In fact, the difference among the coefficients of the formalisms is not a matter of scale, except in the special case when the number of states $q = 2$. In all other cases, to “re-map” coefficients of one formalism into those of another requires a very complex set of equations involving a huge number of terms. We have addressed this point with changes to the text and figures, as detailed in item 10 above.

33. Using their method the authors find that the genetic architecture for almost all of the studied datastes is dominated by first and second order epistatic terms, if judged by the differences in out of sample R^2 (Fig2, Extended data Fig.1). However, this does not necessarily mean that the remaining higher order terms are not important/ crucial for the function. This could especially be true in cases where proteins are studied on a molecular level, were most of the variance might arise from mutant effects on protein folding /abundance rather than on e.g. the (indireclty) measured functional phenotype (e.g. binding).

We agree with the reviewer that the observation that high-order epistasis is generally unimportant does not mean that it *never* affects the phenotype. However, we show that outlier genotypes and large third-order effects are in fact quite rare (Fig. 4b). That said, there may still be a few lucky genotypes in which interactions at high orders may have a substantial effect on the phenotype. We have noted this in the text at lines 397-405.

Regarding the reviewer’s comment about epistasis at the level of protein stability relative to more complex functional phenotypes, we agree that it is an interesting question whether different kinds of protein phenotypes manifest different kinds of genetic architecture. The datasets that we have analyzed include both biophysical measurements such as free energy of binding and higher-level phenotypes such as cellular fitness, antibiotic resistance, and fluorescence. None of them manifest much epistasis above the second order. We also do not see any systematic trend along this axis (biological level of function) in the complexity of genetic architecture (expressed as the fraction of variance explained by first, second, or third-order models). Allosteric protein phenotypes may be different, however, and we now note this in the revised text, as detailed in item 16 above.

34. As stated by the authors, the choice of reference vs. reference free should be contextual to the data and the biological question. In cases were a molecular phenotype is directly targeted and the question is whether two mutations improve synergistically over the wildtype an approach using the wildtype as actual physically realized state should be advantageous over the use of a global hypothetical average.

    We absolutely agree that RFA and RBA have different goals, and further that RBA is the appropriate choice in cases like the one the reviewer points to. We have emphasized this distinction and the context-specific appropriateness of RBA at lines 634-641.

    In summary, the present manuscript of Park et al., is a valuable contribution to the analysis of genotype phenotype maps and while it is not necessarily superior to other methods it is important addition.

    1. Faure AJ et al., The genetic architecture of protein stability. bioRxiv (2023); https://doi.org/10.1101/2023.10.27.564339
    2. Domingo J, Diss G and Lehner B, Pairwise and higher-order genetic interactions during the evolution of a tRNA. Nature 558:117-121 (2018)
    3. Dupic T, Phillips AM and Desai MM. Protein sequence landscapes are not so simple: on reference-free versus reference-based inference. bioRxiv (2024); https://doi.org/10.1101/2024.01.29.577800

We are very grateful for the opportunity to respond to the reviewers' thoughtful comments. We hope that these changes are sufficient, and we look forward to hearing from you.

Sincerely,

Joseph Thornton, Yeonwoo Park, Brian Metzger

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

1. More on RBA vs. RFA
I appreciate the authors' clarification on the issues raised by me and reviewer 2. I think it is now clear which type of inference procedure the authors are performing from the text.
I have some follow up comments below.

1.1 On the authors responses.

To quote the authors comments on the two major issues of RBA on page 2 of the response letter.

"The first part of the RBA-regression procedure leads to a dramatically oversimplified estimate of the RBA genetic architecture, with greater variance attributed to low-order determinants and less to high-order terms than is true under the RBA formalism (Fig. 2)."

"The second bias arises in the last phase of the procedure, when the complete model is used to estimate the model terms. "

I think the second bias is a stronger point and that the examples are convincing.

Regarding the first bias, I think this is a weaker point. This is because regression using RBA does not actually bias the estimation of variance partitioning under the correct analysis (the one achieved using RFA or ANOVA-based variance component analysis see Zhou et al 2021 PMC9522415). The reason is that the the RBA and RFA models when fitted correctly will provide the same prediction (which the authors agree), which is why the red curves in figure 1c and figure 2d are seemingly identical.
My understanding is that the RBA only biases the variance estimation when compared with the variance partitioning under the local RBA model, which contains much less additive contribution in the authors' example in Fig 2d (although I'm not sure which reference genotype this RBA is based on).
I think the RFA or the ANOVA analysis are the only natural and widely accepted form of variance partitioning. Thus comparing the RBA prediction R^2 against the local RBA R^2 is somewhat irrelevant and that both methods can reveal the one ground truth. But I'm happy to learn about more references should the authors provide them.

1.2 More on the shortcomings of RBA (page 6-8 in revised manuscript).

First, I think on a conceptual level, the RBA and RFA represent two complementary philosophies of viewing sequence-function relationships and that no one method is conceptually superior to the other a priori.
And what is seemingly a shortcoming is only a logical consequence of the modeling choice.
For example, the authors point out that a shortcoming of local RBA is that the genetic architecture varies depending on the choice of WT genotype. However, this is merely a consequence of higher-order epistasis. In fact, how the local mutational effects (or epistatic interactions) varies across reference genotypes can be used to infer the extent of higher-order epistasis (Ferretti et al 2016 PMID: 26854875, Zhou et al 2021 PMID: 36129941, also one of the authors papers).
Additionally, I can also point out artificial characteristics of RFA (e.g. there is no biophysical basis for decomposing a protein's function for any possible genotype as sum of additive, pairwise, and higher-order terms, and this construction is merely out of mathematical convenience).
Therefore, I think the paper can benefit from a more balanced conceptual comparison between the two methods.

Second, the authors point out that another issue with RBA is that the true RBA cannot be estimated

with regression. The reason according to the authors is that the variance components estimated using RBA is the same wrt all reference genotypes. Again to mirror my earlier comments, I think there is only one true variance partitioning (under RFA or ANOVA) and that comparing the variance partitioning using RBA regression and local RBA is not relevant.

1.3 The RBA analysis by directly fitting local mutants and by regression is a nuanced point. I think the authors should spend more time on exposing the difference.

2. The role of RFA in this paper.

Similar to my comments in the first round of review, I am not quite certain about the importance of RFA in reaching the main conclusion of this paper.
Figure 4a seems to provide the strongest evidence for the claim of simplicity of protein SFs. But there are two issues with this figure. First, I can get the exact same R2 with any parameterization of a regression model such as RBA, which will also show that the landscape does not contain substantial epistasis beyond pairwise.
Second, the out-of-sample R^2 and the actual variance partitioning the authors introduced in the paper are not the same quantities. What is the relationship between these two metrics?
The authors do provide a variance partitioning formula using the RFA terms directly (Section 2.7 in the supplement), but this is not used in the main text. Wouldn't this formula provide a more direct estimate of the variance components?
The RFA model does provide some insight into the genetic architecture of protein landscapes, such as the sparsity of interaction in Figure 6. But there seems to be a general disconnect between the main conclusion of the paper and the main method the authors propose.

3. I think it is also important that the authors highlight some of the limitations of the RFA method. Specifically, the data used in the paper are all for relatively small sequence spaces where most of the possible genotypes have been measured. I think drawing a general conclusion about the lack of higher order epistasis based on this class of data is premature, as we do not know how much higher order epistasis determines the sequence-function relationship when more sites or all sites of a protein can be simultaneously mutated. It is fair to assume that in this regime there is more possibility for higher order epistasis, due to simple combinatorics. And there are in fact datasets for larger protein sequence-function relationships (e.g. the various GFP datasets generated by the Kondroshov group).
Given this somewhat limited scope, my question is - how easy it is to generalize the RFA framework to large protein sequence-function relationships.
I think there are some serious challenges. First, for large protein sequence-function relationships, the computational load to fit explicitly the RFA terms even for moderates order of interaction will be prohibitive.
Second, when the sequence-function relationships are very large, the sampling are inherently local and centered around a reference genotype. How does this nonrandom sampling of the genotype space bias the estimation of variance components in your framework?
I understand that it is unrealistic for the authors to fully generalize their method to this regime in this paper. But I think it is imperative that the authors discuss these limitations so that the field as a whole is aware of this other realm (full size protein landscapes) where things are largely unknown and do not immediately jump to conclusions.

4. The authors showed that the estimation of the RFA terms are unbiased in the Supplement. But this is only the case when no regularization terms are applied.
However, in the paper, L1 regularized regression was used to reduce model overfitting. This will certainly introduce bias. How does this affect the conclusions? This should be properly discussed in the paper.

Minor points:

1. Fig 2: what are the lambdas? Are these defined in the caption?
2. Line 176: "RFA is compatible with regression because its true terms minimize the sum of squared error across all genotypes". RBA regression also has this property.
3. Line 233-245: It is true that the RFA regression does not infer the true local RFA model with different reference sequences. But can't you recover the local RFA by first making predictions for all genotypes and convert this full landscape to local RFAs?
4. The y-ticks for Fig 2D (left) and Fig 1b right should be the same.

Reviewer #2 (Remarks to the Author):

We thank the authors for their thoughtful, extremely thorough response to our initial review as well as the substantial changes they have made to the manuscript. We recognize that the questions the reviewers raised are technical and sometimes subtle, and potentially difficult to address, and we appreciate the effort taken to do so. We emphasize again that this is an important, unusually rigorous investigation of genetic architecture writ broadly, and will be of general interest, and so these details are important.

With this said, we are happy with the current state of the manuscript and believe all of our issues have been addressed. The modified figures and text are clearer and directly address all issues we raised. We are happy with this current state and happy to recommend publication.

Christian Macdonald, Rosa Sanchez, Angelica Lam, and Willow Coyote-Maestas

Reviewer #3 (Remarks to the Author):

In their revised manuscript Park et al. sufficiently address the points we have raised in our initial review. Specifically, they extended the analysis of the sigmoid link function by adding additional sections in the main text and performing additional simulations to clarify the impact of the sigmoid link function. They show that the sigmoid link function does not lead to artifacts during the inference of the epistatic terms.
Furthermore, they re-worked and added additional sections clarifying how RFA differs in implementation and performance from alternative methods, specifically RBA.
In our opinion the improved manuscript is an important contribution to the field, and we endorse its publication.

We appreciate the reviewers' responses to our revision of the manuscript, "The simplicity of protein sequence-function relationships" (NCOMMS-24-07858). Reviewers 2 and 3 are fully satisfied and requested no further revisions. Here we address the remaining comments of Reviewer 1, which focus primarily on clarifying the difference between our method and the existing method of reference-based analysis (RBA) when coupled with regression. Many of these points were addressed in the prior revision, but we have made some additional revisions to bring out more clearly the issues raised by the reviewer. These changes have been tracked in the document.

Response to reviewer 1

1. More on RBA vs. RFA

I appreciate the authors' clarification on the issues raised by me and reviewer 2. I think it is now clear which type of inference procedure the authors are performing from the text.
I have some follow up comments below.

1.1 On the authors responses.

To quote the authors comments on the two major issues of RBA on page 2 of the response letter.

"The first part of the RBA-regression procedure leads to a dramatically oversimplified estimate of the RBA genetic architecture, with greater variance attributed to low-order determinants and less to high-order terms than is true under the RBA formalism (Fig. 2)."

"The second bias arises in the last phase of the procedure, when the complete model is used to estimate the model terms."

I think the second bias is a stronger point and that the examples are convincing.

Regarding the first bias, I think this is a weaker point. This is because regression using RBA does not actually bias the estimation of variance partitioning under the correct analysis (the one achieved using RFA or ANOVA-based variance component analysis see Zhou et al 2021 PMC9522415). The reason is that the the RBA and RFA models when fitted correctly will provide the same prediction (which the authors agree), which is why the red curves in figure 1c and figure 2d are seemingly identical.

My understanding is that the RBA only biases the variance estimation when compared with the variance partitioning under the local RBA model, which contains much less additive contribution in the authors' example in Fig 2d (although I'm not sure which reference genotype this RBA is based on).

I think the RFA or the ANOVA analysis are the only natural and widely accepted form of variance partitioning. Thus comparing the RBA prediction R^2 against the local RBA R^2 is somewhat irrelevant and that both methods can reveal the one ground truth. But I'm happy to learn about more references should the authors provide them.

We and the reviewer agree on almost all points here. We agree that RFA represents the correct and natural way to partition variance in genetic architecture. We agree that using regression to

estimate the RFA model will accurately estimate the true terms of the model and the variance explained at each order by that model. We agree that using regression to estimate the RBA model will not accurately estimate the terms and variance partition of that model. And we agree that in the latter case the estimates are systematically biased. The key point of departure seems to be the significance of the fact that regression will yield the RFA variance partition no matter what model it is implemented to estimate. We address this point further below in response to the reviewer's comments on the manuscript itself.

1.2 More on the shortcomings of RBA (page 6-8 in revised manuscript).

First, I think on a conceptual level, the RBA and RFA represent two complementary philosophies of viewing sequence-function relationships and that no one method is conceptually superior to the other a priori.

We agree with R1 that RBA and RFA serve distinct purposes, and both are legitimate: RBA examines how mutations act and interact on a particular sequence background, whereas RFA provides a bird's-eye view of genetic architecture. We addressed this in the manuscript but have clarified the text slightly to make this even more clear (lines 195-199 and 636-639):

> *RBA is useful in principle if one is interested in the effects and interactions of mutations when introduced into a particular sequence of interest[42,43]. Its structure is not suited, however, for understanding how sequence determines function across the space of possible variants.*
>
> *Our finding that RFA outperforms RBA in providing a compact and accurate description of the global sequence-function relationship does not mean that RBA is never useful. RBA is the appropriate approach to use if the object of interest is interactions among a few mutations in the background of a particular wild-type or ancestral protein.*

And what is seemingly a shortcoming is only a logical consequence of the modeling choice. For example, the authors point out that a shortcoming of local RBA is that the genetic architecture varies depending on the choice of WT genotype. However, this is merely a consequence of higher-order epistasis. In fact, how the local mutational effects (or epistatic interactions) varies across reference genotypes can be used to infer the extent of higher-order epistasis (Ferretti et al 2016 PMID: 26854875, Zhou et al 2021 PMID: 36129941, also one of the authors papers).

We agree with R1 that the dependence of variance explained on the reference genotype in RBA is not a shortcoming per se but is a reflection of the local, reference-dependent perspective that the RBA model entails. This becomes a shortcoming, however, if RBA is used to describe the global genetic architecture, and the fact is that RBA has been used in the literature for this purpose. We made clear in the text (lines 195-200) that the reference-dependence of RBA is problematic specifically in relation to that goal:

> *RBA is useful in principle if one is interested in the effects and interactions of mutations when introduced into a particular sequence of interest[42,43]. Its structure is not suited, however, for understanding how sequence determines function across the space of possible variants. First, the wild-type-centric view means that the genetic architecture varies depending on the choice of wild-type genotype; in the example of Fig. 2a, first-order effects may make zero contribution to phenotypic variance or explain most of it, depending on the reference sequence chosen, and the pairwise interaction switches in both magnitude and sign.*

Additionally, I can also point out artificial characteristics of RFA (e.g. there is no biophysical basis for decomposing a protein's function for any possible genotype as sum of additive, pairwise, and higher-order terms, and this construction is merely out of mathematical convenience). Therefore, I think the paper can benefit from a more balanced conceptual comparison between the two methods.

We agree that RFA model terms do not directly express biophysical quantities, but this is not a particular feature of RFA. The terms of RFA and RBA alike are mathematical decompositions of the genetic relationship between sequence variation and phenotypic variation, rather than direct biophysical quantities. The two sets of genetic terms have equal relevance to biophysical quantities. Consider a dataset that accurately measures the delta-G of binding to some ligand (dG). The zero-order term of RBA is the dG of binding by the reference protein, and the zero-order term in RFA is the average dG of binding by all genotypes. A first-order RBA term is the difference in dG between the protein carrying an amino acid mutation of interest and the wild-type, and the first-order RFA term is the average difference in dG between the set of genotypes carrying that state and all genotypes. The RFA terms are thus no less or more biochemically meaningful than the RBA terms.

We explain this structure when we define RFA at lines 134-143. Also, the section "Understanding genetic architecture" shows how the terms of the RFA model, applied to three real-world datasets, can be understood in terms of biophysical parameters such as ligand binding. The situation becomes more complicated for RBA and RFA alike when phenotypes are nonlinear transforms of biophysical quantities; our manuscript (including the case studies) shows how this gap from phenotype to biophysics can be partially bridged.

Second, the authors point out that another issue with RBA is that the true RBA cannot be estimated with regression. The reason according to the authors is that the variance components estimated using RBA is the same wrt all reference genotypes. Again to mirror my earlier comments, I think there is only one true variance partitioning (under RFA or ANOVA) and that comparing the variance partitioning using RBA regression and local RBA is not relevant.

There are two senses in which the true RBA model cannot be accurately estimated by regression: the parameters of the RBA model (the effects of mutations) are misestimated, and so is the variance explained by the model terms at each order. This is explained in the manuscript at lines 237-274 and is treated in further depth in the supplement 1.1. The reviewer is correct that the RFA variance partition is the correct one, and that it can be estimated by regression coupled to any model (including RFA, RBA, or any other model with the same degrees of freedom per order). However, the variance partition inferred by regression is accurate, meaningful and interpretable only in light of the RFA formalism. In the manuscript, we explain why RBA-regression yields a variance partition identical to that of RFA in the supplement in section 1.1, just below figure S1.

1.3 The RBA analysis by directly fitting local mutants and by regression is a nuanced point. I think the authors should spend more time on exposing the difference.

We agree that we do not want readers to miss this point. We revised the description to more clearly distinguish RBA from RBA-regression (lines 234-236):

To cope with this limitation of exact estimation of RBA models, an alternative approach has been to use least-squares regression: a series of truncated RBA models are fit to the data to estimate the variance explained by the RBA model at each order, and the complete model is then used to estimate the individual effects[7,18,19].

2. The role of RFA in this paper.

Similar to my comments in the first round of review, I am not quite certain about the importance of RFA in reaching the main conclusion of this paper.

Figure 4a seems to provide the strongest evidence for the claim of simplicity of protein SFs. But there are two issues with this figure. First, I can get the exact same with any parameterization of a regression model such as RBA, which will also show that the landscape does not contain substantial epistasis beyond pairwise.

R1 is correct that we would have obtained the same variance partition in Fig. 4a using RBA-regression instead of RFA, and this is discussed in supplement section 1.1. As discussed above, this correspondence had not been established prior to our work, and this variance partition becomes interpretable and meaningful only with the development of RFA.

Second, the out-of-sample R^2 and the actual variance partitioning the authors introduced in the paper are not the same quantities. What is the relationship between these two metrics?
The authors do provide a variance partitioning formula using the RFA terms directly (Section 2.7 in the supplement), but this is not used in the main text. Wouldn't this formula provide a more direct estimate of the variance components?

We appreciate this question. As the size of the dataset increases, the out-of-sample R^2 is expected to converge on the true RFA parameter values and on the true proportion of variance explained at each order. At small sample sizes, however, out-of-sample R^2 may underestimate the variance explained because of noise in each test set. One advantage of RFA is that that this bias is expected to be weak, especially at low orders, because the model terms are averaged over large numbers of genotypes. The alternative way of estimating variance explained – to fit the model to all data and directly estimate variance explained from the model terms – would have risked overfitting model terms to the data, yielding a potential overestimate of variance explained. We preferred to be conservative and therefore used out-of-sample R^2. We have revised the methods section at lines 785-787 to note this point:

To estimate variance explained using truncated models, we used ten-fold cross-validation, which may slightly underestimate accuracy, but this bias is expected to be weakbecause RFA uses many genotypes to estimate each model term at low orders.

The RFA model does provide some insight into the genetic architecture of protein landscapes, such as the sparsity of interaction in Figure 6. But there seems to be a general disconnect between the main conclusion of the paper and the main method the authors propose.

We are glad that the reviewer agrees that Fig. 6 on the sparsity of genetic architecture depends directly on the RFA formalism (and presumably Fig. 7, as well, which directly follows and clarifies the findings of Fig. 6). But we do not understand the disconnect that the reviewer refers

to. As we described in our first revision letter, virtually none of our conclusions could have been reached – or they would have been invalid and uninterpretable – if we had not developed RFA. Figures 1-3 demonstrate the desirable properties of RFA relative to existing formalisms, as well as the anomalies that arise when regression is coupled with RBA but not with RFA. Figure 4 demonstrates the simplicity of genetic architecture when modeled using the RFA formalism, which differs dramatically from the complexity of the architecture under an accurate estimate of the RBA model. (We did not use RBA-regression because of its problems, but if we had done so, we would have concluded that the RBA genetic architecture is simple, which is false; moreover, when we examined the estimated terms of the RBA model (Fig. S5), we would have been led to the anomalous conclusion that high-order RBA terms are widespread and large while simultaneously explaining very little variance; this would rightly have called into question the variance partition we inferred.) Figure 5 shows how coupling RFA with a simple global transformation explains the vast majority of phenotypic variance as the result of low-order effects and a simple nonlinearity – which contrasts dramatically with the architecture under RBA. Figures 6 and 7 establish the sparsity of genetic architecture under the RFA parameterization and its implications for estimation using small sample sets– which again would not be apparent using the RBA formalism, which is far less sparse. And Figure 8 provides case studies of three particular proteins of the application of RFA and the interpretation of its terms. None of this work on the global genetic architecture of proteins could have been accomplished without the RFA formalism.

3. I think it is also important that the authors highlight some of the limitations of the RFA method. Specifically, the data used in the paper are all for relatively small sequence spaces where most of the possible genotypes have been measured. I think drawing a general conclusion about the lack of higher order epistasis based on this class of data is premature, as we do not know how much higher order epistasis determines the sequence-function relationship when more sites or all sites of a protein can be simultaneously mutated. It is fair to assume that in this regime there is more possibility for higher order epistasis, due to simple combinatorics. And there are in fact datasets for larger protein sequence-function relationships (e.g. the various GFP datasets generated by the Kondroshov group). Given this somewhat limited scope, my question is - how easy it is to generalize the RFA framework to large protein sequence-function relationships.

I think there are some serious challenges. First, for large protein sequence-function relationships, the computational load to fit explicitly the RFA terms even for moderates order of interaction will be prohibitive.

Second, when the sequence-function relationships are very large, the sampling are inherently local and centered around a reference genotype. How does this nonrandom sampling of the genotype space bias the estimation of variance components in your framework?

I understand that it is unrealistic for the authors to fully generalize their method to this regime in this paper. But I think it is imperative that the authors discuss these limitations so that the field as a whole is aware of this other realm (full size protein landscapes) where things are largely unknown and do not immediately jump to conclusions.

We agree with R1 and want to be clear in the paper that there are limitations in the datasets we examined and that analyzing global genetic architecture is challenging. On R1's first point, it is true that understanding global genetic architecture requires sampling across sequence space as a whole rather than in a single local neighborhood (although we find that exhaustive sampling is not required, Fig. 6). This is not a limitation of RFA per se but a challenge of the problem itself:

learning the global genetic architecture requires global sampling, no matter what model is used to analyze the data. Sampling low-order mutants around a designated wild-type is suitable for understanding the local genetic architecture, but it offers very limited information about the global architecture (unless the genetic architecture lacks epistasis, which we can never know in advance).

We agree with R1 that caution is required concerning the generality of our findings when extending from the relatively small sequence spaces analyzed thus far to larger spaces. The text discusses this (lines 615-620):

> *A possible explanation [for the lack of high-order epistasis] is that these datasets held most sites in the protein constant and therefore presumably maintained the overall conformation (or caused it to unfold entirely). High-order interactions that specify a protein's fold might be revealed in a library large enough to contain variants with multiple folds, or if phenotypes involving multiple conformations within a single fold were measured.*

To be appropriately cautious, we also modified the abstract's final line so that the conclusion is now specific to the datasets we analyzed: "The sequence-function relationship in these datasets is therefore far simpler than previously thought, opening the way for new and tractable approaches to characterize proteins' genetic architecture."

In the last paragraph of the discussion, we discuss at length the practical issues that arise when extending our approach to larger datasets. We address the reviewer's concern that sampling is intrinsically local and suggest a strategy for overcoming this. We also note that an implication of our finding is that the experimental and computational load of sampling high-order combinations and computing high-order effects will be dramatically lightened.

4. The authors showed that the estimation of the RFA terms are unbiased in the Supplement. But this is only the case when no regularization terms are applied. However, in the paper, L1 regularized regression was used to reduce model overfitting. This will certainly introduce bias. How does this affect the conclusions? This should be properly discussed in the paper.

In Supplementary Fig. 4, we computed the out-of-sample R^2 for every dataset under a wide range of regularization strengths. We found that regularization strength has almost no impact on the out-of-sample R^2 of the first and second-order RFA models. This is because the number of parameters in these low-order models is much smaller than the number of genotypes and therefore overfitting is negligible. This is noted at lines 392-394 of the main text and shown in Fig. S4.

Minor points:

1. Fig 2: what are the lambdas? Are these defined in the caption?

The lambdas refer to the RBA coefficients. We thank the reviewer for catching this, and we have now defined them in the figure legend.

2. Line 176: "RFA is compatible with regression because its true terms minimize the sum of squared error across all genotypes". RBA regression also has this property.

We changed this to “The RFA model can be accurately estimated by regression because its true terms minimize the sum of squared error across all genotypes”. The RBA model does not have this property.

3. Line 233-245: It is true that the RFA regression does not infer the true local RFA model with different reference sequences. But can’t you recover the local RFA by first making predictions for all genotypes and convert this full landscape to local RFAs?

We believe R1 meant to say RBA here, not RFA. If so, the reviewer would be asking whether the RBA model could be estimated by using regression to fit the complete nontruncated model across all genotypes and exactly fitting RBA model terms to the predicted phenotypes. This is true only if there is no measurement noise. When noise is present, using RBA-regression using the full model is exactly the same as directly computing the terms without regression. This procedure yields the same model estimates as exact RBA, which is highly sensitive to measurement noise and missing data. These points are address at lines 249-255 and Fig. 2d, as well as in Fig. S3 and the surrounding text.

4. The y-ticks for Fig 2D (left) and Fig 1b right should be the same.

We appreciate this suggestion and implemented it to make the comparison of the two graphs more straightforward.

Response to reviewers 2 and 3

(R2) We thank the authors for their thoughtful, extremely thorough response to our initial review as well as the substantial changes they have made to the manuscript. We recognize that the questions the reviewers raised are technical and sometimes subtle, and potentially difficult to address, and we appreciate the effort taken to do so. We emphasize again that this is an important, unusually rigorous investigation of genetic architecture writ broadly, and will be of general interest, and so these details are important.

With this said, we are happy with the current state of the manuscript and believe all of our issues have been addressed. The modified figures and text are clearer and directly address all issues we raised. We are happy with this current state and happy to recommend publication.

(R3) In their revised manuscript Park et al. sufficiently address the points we have raised in our initial review. Specifically, they extended the analysis of the sigmoid link function by adding additional sections in the main text and performing additional simulations to clarify the impact of the sigmoid link function. They show that the sigmoid link function does not lead to artifacts during the inference of the epistatic terms.

Furthermore, they re-worked and added additional sections clarifying how RFA differs in implementation and performance from alternative methods, specifically RBA.

In our opinion the improved manuscript is an important contribution to the field, and we endorse its publication.

We are very happy that we are now fully on the same page with reviewers 2 and 3.

We thank all the reviewers for helping us to clarify our material and improve its presentation. We appreciate their deep engagement with our manuscript. We hope that the current responses and revisions are sufficient to address any remaining issues.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

I appreciate the authors' very thorough responses to my comments and the updates they have made in the revision. I think the manuscript is appropriate for publication.