We thank the editor and all three reviewers for their thoughtful consideration of our manuscript. We have submitted a revised version where we address the comments made by all reviewers. The most significant changes that we have made in response to these comments is the inclusion of a large number of additional simulations in the supplement where we show that the results highlighted in the main text are generally robust to the specific assumptions we had made, a matter that was the subject of concern by the reviewers. Specifically we have repeated all of the simulations in the paper (in large number of replicates) under conditions where the following assumptions are relaxed :

- **Functional Additivity / Smooth landscape:** The new set of simulations include multiple non-additive community functions, including two where we model realistic scenarios: (1) selection for a community that optimizes the elimination of a specific metabolite (akin to bioremediation) and (2) selection for a community that will prevent invasion by an invasive species (akin to pathogen exclusion). As a pure exercise, we have also included selection for a function that is explicitly "epistatic", i.e. set by the sum of pairwise contributions and where there is no independent contribution from any one species.

- **Functional redundancy:** The new set of simulations include cases where only a small fraction (20%) of the species contribute to the function of interest, allowing us to test the effect of the original assumption of full functional redundancy.

- **Fitness neutrality:** The new set of simulations include the case where species contributions to function impose a cost on microbial growth (such cost lowers the amount of biomass that can be produced per unit of nutrient uptake). The cost grows linearly with the amount of contribution of a species to the collective function.

- **Type-III Functional response:** The new set of simulations include the case where species uptake rates follow a Type-II (Monod) and Type-I Linear functional response with nutrient concentration.

- **Pure competition for resources:** The new simulations include the scenarios where species can interact via cross-feeding based facilitation in addition to competition. This includes a new minimal-media scenario where there is only a single supplied resource and species coexistence is dependent on facilitation.

- **Species distribution in the regional pool**: The new simulations include additional species distributions in the species pool from where the inocula are

drawn (i.e., log-normal distribution vs power-law) as well as new forms of inoculation (i.e. fixed number of species at identical abundances in the starting community).

- **Distribution of per-capita contribution to function:** The new simulations include scenarios where the per-capita contribution to function is sampled from a normal distribution with mean 1 (eliminating the assumption that most species have a negative effect on function). We also include a scenario where per-capita contribution to function is sampled from a uniform distribution between 0 and 1 (eliminating the assumption that high functioning taxa are rare).

This new suite of simulations are presented in a new set of figures (Figures S11-S16) and are described in six additional pages of an expanded supplementary text. We have found that our main qualitative results hold for the majority of the new simulations with some interesting deviations that are discussed in the supplementary text (e.g. one such deviation is that the synthetic communities, as we had designed them, perform much worse when the function is not additive and that impacts the outcome of Fig. 4, though in a way that benefits artificial selection over bottom-up engineering).

That we were able to carry out all of these simulations in a relatively short time illustrates precisely what we believe is of the main strengths of our manuscript: that the modeling platform we introduced, and which is coded into the *ecoprospector* package, is very flexible and can be used to simulate the efficacy of a wide range of selection strategies for a very large range of possible ecological scenarios, beyond the ones we have added to the revised manuscript.

We have also carefully gone through each of the reviewers comments and re-written parts of our manuscript that were unclear before, addressing the concerns and suggestions made by the reviewers. We hope that these changes will clarify any errors or confusion in the manuscript and address the reviewers concerns. Below we **show all reviewer remarks in bold** and our responses in normal font. In the revised manuscript, all changes are presented in red font.

Reviewer #1 (Remarks to the Author):

The manuscript by Chang et al. presents a numerical study of the evolution of community function in multi-species communities, with the aim of identifying strategies that improve achievement of an optimal collective function. The central tool used for this study is the package Ecoprospector, an extension of the package MiCRM – that had been previously devised to study the ecology of complex communities – to a meta-community

structure. The new package adds the possibility of applying selection at the level of the communities, and to implement different rules for community 'reproduction'. As in other previously studied models, and as done in experiments, communities are periodically evaluated for their function, that depends on the abundance of the composing species, and a new generation of communities is seeded with probabilities that reflect the parental community performance. The simulation tool is used for two main purposes: to compare the effectiveness of different protocols (classified in 'migration pool' or 'propagule' classes) used in artificial selection experiments, and to devise new protocols to improve selection of community function. In examining existing protocols, the authors point out several practical aspects of community-level selection that have been overlooked in experiments. First, the need of appropriate control experiments. Second, the difference between improving average and maximal community performance. Third, the coexistence of multiple time scales (within-generation ecological changes and evolution of species-level parameters) in the time variation of community function.
Although these points, and in particular the latter, may not come as a surprise to modellers, they have not been taken explicitly into account in devising community evolution experiments. Based on these observations, the authors propose a set of different protocols that both preserve good-achieving communities – proper of propagule methods – and create sufficiently high inter-community variability for selection to work more efficiently. The protocols differ in how communities get perturbed from one generation to the next, with different methods being more or less easy to implement. The performance of these methods is evaluated based on the previously defined criteria, showing that they allow a significant, though apparently quantitatively modest, improvement of the selection outcome. Finally, the authors compare communities obtained by directed community selection to those assembled 'bottom-up' based on individual species performance, and show that if 'engineered' communities perform better, they are less resilient to invasion by new species.

We thank the reviewer for their clear summary of our contribution.

This study responds to a compelling and real need for efficient strategies able to make community-level selection efficient and its results robust. The question is of undeniable fundamental and applied relevance. Previous studies have used similar numerical methods to inquire about general principles of community evolution (Williams & Lenton, PNAS 2007), but most existing studies – especially experimental – have only explored a subset of very specific communities/protocols that responded to the researcher's intuition and empirical constraints, rather than to a reasoned choice among alternative scenarios. The strongest point of this work is in my opinion providing experimenters with a tool to compare the performance of alternative protocols (provided they have a sufficiently good idea of how to define an appropriate virtual community that corresponds to their experimental system).

We completely agree with the reviewer's comment.

As for the indications on what should be the preferred strategy, I am more dubious on the impact and generality of the conclusions. The more general among the results are actually not very surprising from a theoretical point of view. In a process that has two levels of competition (the species and the communities), maintenance of the adult phenotype has by definition to take the ecological dynamics into account. As a consequence, 'generational stability' is important, and it is easier to obtain when the community is at equilibrium.

We agree with the reviewer that the idea that communities should be generationally stable before applying community level selection may not be surprising in hindsight. Yet this has been ignored by all experimental studies that have attempted to carry out community level selection to date. It also helps explain why in Figure 1F all previously proposed selection schemes do worse than a no-selection control that has been absent from previous studies. We therefore think it is important to explicitly articulate this idea and we tried to communicate this to the community of experimentalists that have largely ignored it, by showing its impact in the very same protocols that they have used in the past. That was the rationale for that part of the paper.

Requirements of a balance between generation of variance and selection is a classic issue in optimization. However, there seem to be no qualitative feature that can generally guide an experimenter to choose the best strategy (although there can be in specific systems, as in Xie and Shou)

We agree the 'best strategy' will depend on the details of the experimental system. Despite this we expect that our proposed strategies will work in a wide range of systems specifically when compared to the previously used approaches. Whilst it would be impossible to prove this conclusively using a simulation based approach, we have sought to illustrate this more convincingly, in the revised manuscript by repeating our simulations across a broader range of different community functions and resource environments (Figure S12-S13). These include a scenario where species-level contributions to function are costly, a scenario where community function under selection is not additive, as well as two examples of functions that an experimentalist may want to optimize: i.e. designing a community that resists invasion by an undesirable species (e.g. pathogen resistance) and designing a community that eliminates an undesirable environmental molecule (e.g bioremediation). In all of these cases all of our proposed strategies are successful at improving maximum community function (panel D). In particular, we find that the combination of a stringent bottleneck (which randomly eliminates species from the community) with migration from the

regional pool (which randomly adds new species) is generically more successful across all functions and scenarios we tested that single-perturbation protocols. Although we do admit that this does not mean that these findings will be true for all applications and under all conditions, they worked for all of those we tried, suggesting some generality. But more importantly, the modeling methodology developed in our paper through the Ecoprospector package will make it possible for empiricists to easily model other specific scenarios and explore strategies that will be useful for their specific purposes.

What I missed most – but it is possibly unachievable without a more advanced theoretical understanding of the processes underpinning the observed evolutionary trajectories – is a discussion on the robustness of the conclusions. It is true that the authors repeat their simulations many times, but, if I understand correctly, the variation in the initial composition of the metapopulation is constrained by the fact that communities are sampled according to the same laws. It would be interesting to know how did the authors decide how to set the initial conditions (in the parameter and state spaces) and whether they tested different choices.

Our method for setting the initial composition of the inoculum species pool does not impose a strong constraint on the initial variation, though we think this was unclear in the previous version. To address this we have rewritten the second paragraph of the results section and the Initial Consumer Conditions section (lines 669-690) in the methods. We do not expect the exact species distribution in the species pools to be particularly important so long as one starts with diverse communities that show compositional and functional variability. To illustrate this we have repeated all of our simulations using a couple of alternative approaches to set the initial composition of the metacommunity (whilst maintaining roughly the same starting species richness). Specifically, we considered a scenario where the species abundance distribution of the pools followed a lognormal distribution as well as a scenario in which each community is seed with a fixed number of species with equal abundance. We find that our results are robust to the specific sampling strategy as shown in figure S15.

Another aspect that was not self-evident is why the authors expect that the function landscape is smooth as they represent it in Fig. 2.

As the reviewer points out, in the main text we had considered the simplest possible scenario where the function under selection is additive and smooth (hence the representation in Fig. 2). We agree that this will not be the case in general (see for instance previous work from our lab: Sanchez-Gorostiaga et al 2019). In the revised manuscript (see Figure S12 and Supplementary Text) we now report a new suit of

simulations using a number of non-smooth functional landscapes and alternative community functions. To address the specific point raised by the reviewer, we include a non-additive function in which there are pairwise functional interactions between species (analogous to epistasis in fitness landscapes) such that the structure-function landscape will be rugged. In addition, we explore other functional landscapes that are non-additive, where the function is the resistance to invasion from a “pathogen” or the elimination of a specific metabolite by the community.

Complex communities are known to have regimes where not only the attractor is not an equilibrium (see for instance Pearce et al. PNAS 2020), but the energy landscape is rugged (see for instance Biroli et al. New Journal of Physics 2018), or have multiple equilibria (Bunin PRE 2017). I guess that in these cases the strategies proposed in the manuscript may not be very helpful.

We agree with the reviewer and we thank them for raising these two interesting points and now include a discussion of this matter in the revised manuscript, and reference these papers in lines 543-552.

Some points where I think the manuscript needs to be improved:

I found the description of the initial metapopulation state (lines 124-130) unclear. I would place the definition of a species (lines 128-130) before that of the assembly. What remains the same within a single metacommunity evolution experiment, and what changes from one to the other of the 100 replicas should be made clearer here. I would have also liked a discussion of what changes – and what stays the same – from one to the other community in one specific initial metapopulation.

We thank the reviewer for bringing this up. After reading this comment we realized they were right and agree that this was unclear. We have incorporated the reviewers suggestion when rewriting paragraph 2 in the results section and switch the order as they suggest. We have also sought to clarify in lines 674-683 what parameters are constant across the replicates and which are resampled.

A particularly relevant thing to know – in light of the discussion on the transient community states, on the effect of bottlenecks and on comparison with synthetic communities made up of 12 species – is how many of the species initially pooled in each community survive under a purely ecological dynamic. I imagine that a fraction of the 2100 species probably goes extinct quite quickly. Did the authors measure it?

We do record the species richness for each community. The no selection lines shown in figure 1C and 1D communities started with 228± 19.4 (mean ± sd, n=96) species of which 26±5.5 (mean±sd, n=96) survive after 20 transfers in the absence of community-level selection. The value for the example given in Fig S3 are now reported in the caption.

Correspondingly, in the methods I found that several modelling choices were not discussed:

A. Eq. 1 and 2 are presented in a form that is more complex that what is then actually used (some parameters are set to 0, e.g. mortality). I would have preferred to see the equations that are actually simulated, and have their meaning described in more detail). Similarly, for lines 586-588: explain or remove.

The reviewer is right that in the main text the simulations we examined included a number of simplifications that render parts of the equations redundant. For some of the new simulations described in the supplementary text we do now vary these other parameters (e.g. $D_{\alpha\beta}$, l_α , $\sigma(c_{i\alpha}R_\alpha)$,) in Fig. S12,S13,S14). To address this we have now moved the full form of the model to the supplement and provide the simpler form in the main text. Per the reviewer's request, we have also removed lines 586 - 588.

B. It would be useful to have the range of variation of some of the indexes, such as u,v, U in eq. 5 and \nu in eqs. 8-9.

Because of a formatting issue during conversion to pdf at the time of submission and an error in equation numbering, we were not exactly sure which indexes the reviewer is referring to. We went through the paper and we believe they refer to index i in (now with correct numbering) equation 8 and index u,v in equations 10-11. We have now specified the ranges for these indexes in the revised manuscript.

Also the meaning of n and N do not seem to be consistent in the Methods and the main text (does the combination of line 125 and line 630 mean that communities are started with 10^12 cells?).

The reviewer is correct, and we appreciate them noticing this. The confusion arose from the incorrect usage of N in line 125. The parameter n should reflect the number of cells whereas N should denote the abundance of a species in the consumer resource model. Each community within a metacommunity starts with 10^6 cells. We have corrected this and ensured consistency throughout in the revised manuscript.

C. Line 651: it would be useful to have stated here that the weights \phi_i are held constant along every single evolution experiment.

We have now started this on line 134 in the result section.

D. The distribution eq. 5 should be discussed: why choosing a power-law, why with such a small exponent, why keeping the same rank of species in all communities in the initial metacommunity?

The power law distribution was chosen based on previous work that shows that natural microbial communities have an abundance distribution that are power-law like (Shoemaker et al 2017). The exponent was chosen to ensure that when sampling 10^6 individuals we start with ~200 species which is in line with typical laboratory experiments in our own group (Goldford et al 2018). We mention this in lines 679-681 and also illustrate this in figure S11 where we show the rarefaction curves for our initial communities and by comparison, we also show empirical rarefaction curves from the data in Goldford et al 2018. We do not keep the same rank of species in each community in the initial metacommunity and have now clarified this in the Initial Consumer Conditions section in method (lines 669-675). Importantly, in the revised manuscript we have repeated our simulations using alternative sampling distributions (figure S15). Our main results all hold.

E. I do not understand the sum in eq. 8: do you consider here that each community has multiple parents? In that case, what about the multinomial sampling? v and \nu are used a bit liberally, and this makes it difficult to understand the equations. In eq. 9, what is \psi? Is a sum over \nu missing?

We thank the reviewer for pointing this out and we see that our description was unclear. As the reviewer guessed, in equation 8 (now equation 10) each community can have multiple parents (for example in migrant-pool strategies a community can have more than one parent). We have clarified this and explained the components of this equation in lines 738-745. We did not do multinomial sampling for resources because these equations correspond to the resource abundance and resources are treated as continuous (this is justifiable when the number of molecules of each resource is large relative to the number of cells).

In contrast for species abundances we do use multinomial sampling. In equation 9 (now equation 11) psi is the conversion factor which transforms abundance into cell counts (defined previously in lines 684-686, but now repeated here in lines 754-755). We have rewritten lines 746-750 to make equation 9 (now equation 11) clearer.

F. What distributions are c_{i,\alpha} taken from?

We explain how c_ia are sampled in the Uptake Rate section in the methods (lines 635-661). Briefly c_ia is the product of two random variables, one sampled from a gamma distribution and the other from a Bernoulli distribution. The latter controls the degree of generalism/specialization (i.e. if species i can or cannot uptake resource alpha), whereas the former controls the quantitative level of resource uptake. We have justified this sampling approach in the Methods section.

G. Lines 701-702: if this means that the community function varies across runs, it should be stated explicitly in the main text.

The reviewer is correct, this means that the per capita community functions of all species in each independent run differ (as well as the media composition and resource uptake rates). We have now stated this in the main text (line 179-182), and clarified it in the Methods section (lines 710-711)

H. Line 918: explain why do you need two different measures of the overall resistance

F* quantifies the community function after perturbation while R quantifies the change in community function after a perturbation on a consistent scale. Which measures are of interest may depend on the objectives of the experimenter.

I think the results of the ecological prospection deserve more comment, for instance on the fact that propagule protocols seem to be more effective than migration protocols, and on the role of sufficient initial sampling of the abundance space in order for no selection experiments to fare better than when selection is imposed.

The propagule protocol is more effective than the pooling protocol because it preserves high performing communities. When pooling low functioning species from lower functioning communities outcompete high functioning species in the top community. This is discussed in the supplementary text and shown in figure S6.

Since there are different concepts of stability discussed here (ecological stability of the community dynamics and 'generational stability' that can also occur when the ecology is not stable, or if the community is out of equilibrium), the authors

should check carefully that their use of 'stable' means the same thing everywhere (for instance, in lines 259, 264 and 365).

We have checked this carefully and specifically make the use of "generationally stable" clear throughout when we mean that the successions are identical across community generations (including line 259,264 and 364).

Lines 183-185: it would be appropriate here to refer to the alternative propagule protocol depicted in Supp. Fig. 1.

The alternative propagule protocol is now depicted in Figure 1 and the caption has been updated.

Line 279: what is the statement 'None of these details are critical …' based on?

We have removed this statement.

<u>Reviewer #2 (Remarks to the Author):</u>

Chang and colleagues present a simulation study that aims at finding ecosystem selection schemes that maximize some desirable community function. The most outstanding of their findings are: 1) that traditional selection schemes, meaning propagule-selection, migrant-pool selection, and including all sub-variants in every publication they can find, result in 'selected' communities that are inferior than simply picking the best starting community. 2) A simple trick to change that depressing finding is to impose a very stringent (but not catastrophically stringent) bottleneck during the selection transfers; a number of other alterations also work, especially a strong bottleneck + introduction of migrants. Finally, 3) while a bottom-up method that puts together high performing strains out-does any selection scheme in terms of community function, these 'synthetic' communities are very fragile in terms of resilience against invasion, whereas lower-performing (but still well-performing) selected communities tend to be much more resilient.
The authors also emphasize that successional dynamics causes problems, because measured functions after one growth cycle are likely to change due to succession after multiple transfers, which can alter community function and adds a lot of noise to initial community ranking.
This is an interesting, well-written, and well-illustrated manuscript that I think is deserving of publication and that is likely to appeal to readers in a number of

subdisciplines, from basic microbial ecologists to applied microbiome researchers. I very much enjoyed reading it. That said, there are some improvements which I think could be made.

We thank the reviewer for their clear summary of our contribution and for their interest in this paper.

My main recommended improvement is to expand the discussion, and possibly explore some more data, related to limitations. The authors commendably spend a large paragraph in the discussion on the limitations of using the model they chose (the paragraph starting line 508). However, it needs to be mentioned that not only have they not explored all possible ecological scenarios, but more importantly that their main conclusions may be entirely dependent on choosing a pure-competition model with resource substitutability.

For example, I would hypothesize that the bottleneck method, which is one of their main successful selection mechanisms, may perform much less well when community function requires inter-specific facilitation or signaling by rare species. I don't think it would be fair to ask to explore an entirely new ecological scenario. However, I do think it would be useful to understand a bit more about the terminal communities, which may help us indirectly understand what the ecological scenario they chose may be imposing on the study. For instance, are they always ending up with a similar looking ecology, such as a group of species that have high phi, and relatively non-overlapping uptake rates (i.e. high-performing specialists)? Such an exploration could also help inform us on why the synthetic communities went wrong. If they found that high-performing terminal communities always had a consistent community structure such as I hypothesized above(or some other structure), they could ask whether creating synthetic communities based upon this joint understanding (choosing high performing specialists, as opposed to just high-performing strains) resulted in synthetic communities which do not suffer the same downfalls as the ones from their method of bottom-up construction. Even the latter set of simulations are not something I would wish to require, but the authors should at least spend time talking about the limitations to their method of choosing the synthetic community, as it may have biased against more informed bottom-up strategies. To summarize this paragraph, I think the authors need to discuss more about how their ecological scenario, and their synthetic community construction, might have causally affected their results and conclusions, and explore the ecological details of their high-performing terminal communities as a way to answer that question.

We thank the reviewer for these suggestions. In the new Figure S13 we show simulations for a couple of additional ecological scenarios including those where co-existence is maintained purely through cross-feeding (illustrating that our results do not depend on a pure competition-model). The reviewer is correct that we have not explored the issue of resource-substitutability and we now make note in line 521. We have also expanded the discussion on the limitations of our specific strategy to choose a synthetic community in the revised Discussion section (lines 552-555)

A second recommendation is about Fig. 1B. First of all, I think the authors have in general done a great job using illustrations to aid the reader's understanding of the various selection regimes. That said, to fully understand the situation, the reader must rely on Fig. 1B, Supp. Fig. 11, and a somewhat technical, matrix-inspired mathematical description in the methods. (I understand why the authors chose that technical description, as it directly correlates to their code, but it isn't strictly necessary to understand what occurred). I think it would be possible to incorporate some details from S11 into Fig. 1B to simplify understanding. First of all, Fig. 1B could have another row for the propagule method, which is key to understanding part of their results but is missing from 1B. Secondly, looking at S11, it is unclear why many offspring wells would receive the same inocula, until one realizes that Poisson sampling may result in different offspring well species distributions despite a common inoculum. This information is implicit in Fig. 1B, because two wells from one tube have different cells in them, but could be made explicit, for example by stating on those arrows "random sample," or showing a representation of dice, or something better the authors can think of.

We thank the reviewer for these two suggestions. We have incorporated the propagule method (previously in Figure S1 now into Figure 1B). We have also added 'stochastic sampling' to the arrows to make clear that different offspring wells will receive different taxa. We do worry that including the matrix inspired implementation in Figure 1 will overcrowd an already dense figure. However we have however replaced the previous Figure S1 (now redundant), with a simplified version of Figure S11, This new Figure S1 is referenced in the caption for Figure 1.

The authors went to great lengths to gather detail from many selection papers, which is to be applauded. One detail I wished to see, however, was the degree of bottleneck each of these papers used, since the authors found that to be such an important parameter. Presumably these papers applied some sort of bottleneck,

was it always above the threshold the authors found to be required to get community improvements?

We have added the bottleneck sizes (dilution factors) to TableS1 and made note of this in lines 841-843 in Methods. Briefly, most bottleneck sizes applied in these papers are much lower (less bottleneck stress) than the level of dilution factor that improves community function.

Fig 2D-I: The main information here is the distance above the y = x line. Right now, it is hard to compare across selection regimes. I recommend altering these plots to show differences (analogous to the Q used elsewhere, except offspring minus parents as opposed to selected minus unselected). This would help the reader see which mechanism is better. Especially because from this figure, I would have thought bottlenecking was superior to a migration method, but the opposite is what seems to occur in Fig 3 (although my interpretation there may be clouded because they show Q, not offspring – parent). Explanation for that apparent discrepancy is also warranted.

We have introduced a new Fig. S12 where we present the requested comparison across strategies (panel C). The reviewer is right that the manner in which we have displayed these results in the main text makes it difficult to compare across different types of perturbation. This is a deliberate choice; as we do not believe we can make a general claim about which strategy is superior in Fig. 2. Which of the 6 proposed strategies is superior will depend on the magnitude of the perturbation (dilution factor in bottleneck, amount of resource being shifted, number of cells introduced during migration etc) which we have not systematically explored. As we now make clear on lines 930-934 we choose parameter values so that the effect sizes in Figures 2D-2I are qualitatively similar in magnitude, to avoid misleading comparisons. In addition, which strategy works best will depend on the function under selection and the ecological scenario (as can now be seen in Fig. S12-13). Instead we simply propose these 6 perturbations as methods to generate random compositional variants and show that all can do better than a simple performance screen because they allow for the exploration of new regions of the structure-function landscape. We have now emphasized this point in lines 335-336 of the revised manuscript.

In Figure 3 we do not seek to make any statement about whether migration will generally be more effective than bottlenecking as again this may depend on the magnitude of the perturbation. Rather we are claiming that combining the two will be generally more effective than either method alone, as the two allow for simultaneous

additions and deletion of species. We have edited the figure removing this statistical comparison to make this point clearer.

Finally the apparent discrepancy between figure 3 and figure 2 is due to a different choice of value for both n_{mig} and d_{bot} across these two figures. This is now explained in lines 957-960.

Model definitions: Since crossfeeding is not used, I would argue they should simplify equations 1,2 to reflect this, as at the moment term 1 in equation 2 (the larger term) is always == 0, which can be confusing. Similarly, there is no reason to have variables for the hill coefficient or m when they are fixed values.

We have simplified this equation in the main text and moved the full version of the equation to the supplement, where we have also carried out a new set of simulations that do including cross feeding (Fig. S13).

Why was this sigmoidal function chosen and not an arguably more typical function with hill = 1 (i.e. Monod)?

This was an arbitrary choice on our part. We do now show in Figure S14 that the choice of functional response for example either using hill = 1 (type-II), or using a linear type-I functional response does not qualitatively change any of our results.

The community function of interest could, in theory, be contributed to by all species. Each cell had a per-capita (φ) contribution to the function. Interestingly, they allow species to negatively affect community function, because phi was sampled from a normal distribution with mean zero. This also implies that it is rare for species to be strong contributors—most species will not have any contribution to the community function. I think this could be better described in the text (lines 120-122), because while it is technically additive, it requires more explanation how it could represent e.g. total biomass of the community. Furthermore, while they note in the discussion that this is a limitation because it can't represent functions such as host fitness, I think it is also worth discussing whether other distributions of phi (e.g. uniform from 0-1, or normal distributions centered around 1 rather than 0) may change things.

In Figure S16 we have explored multiple additional additive functions and show that our results hold true for all of them. This includes a scenario in which only a subset of species contribute to the function addressing the first part of the reviewers comment. We have also repeated the simulations with a phi sampled uniformly in the interval [0,1],

as well as from a normal distribution centered around 1. This does not qualitatively change any of our results.

One of the most important results is that all typical selection schemes result in worse outcomes than a performance screen. In my opinion, this is important enough to belong in the abstract.

We thank the reviewer for encouraging us to add this to the abstract. It is included in the revised version.

Line 659: the use of “cells” is confusing. Maybe this should actually be “wells”?

We agree with this suggestion and have changed this accordingly.

Eq 8: I get that S_uv is a dilution factor, but when it is defined (line 659) it is referred to as a fraction of cells

We have reworded this as a fraction of parents well.

Eq 6 didn’t render well in the pdf. I am pretty sure it is just F = phi * N, hopefully that is correct.

This is correct and has been updated in the new pdf.

The result about the importance of allowing for community stabilization is important. However, it has two drawbacks. Practically, doing 20+ transfers between each selection may not be feasible in a real-world application.

With respect to the feasibility, we agree 20 no-selection transfers may be a-lot, though we note that it is still far less work than doing 20 rounds of selection for a worse outcome (as in figure 1F). Experimenters may wish to optimize the number of transfers and the time between transfers to ensure rapid equilibration, though this is beyond the scope of our paper. We have made note of this in the methods in lines 857-861.

More critically, evolution is inevitable during such transfers, potentially rendering the conclusion moot. Please discuss.

We have added a paragraph where we discuss our results in the context of species-level evolution, drawing from a number of recent experimental studies. We agree that incorporating evolution into this framework is a critical next step for this work

and we have explicitly highlighted this as a future direction both theoretically and experimentally.

Reviewer #3 (Remarks to the Author):

The authors have conducted modeling experiments addressing two interesting questions concerning the directed evolution of microbial communities and obtained intriguing results. The major weakness concerns the major and unrealistic assumptions: that all species contribute to the function of interest and that all such contributions are fitness neutral. While I appreciate the suggestive references provided by the authors, neither of these seems likely to hold in practice for the vast majority of potential applications. As such, I view the manuscript as primarily theoretical, with the goal of exploring the rules governing community diversification and adaptation in a highly unrealistic setting. The results could plausibly guide real-world experiments or be tested in models without such assumptions, but should not claim to be relevant to real-world applications. This is admittedly challenging since directed evolution is itself mostly relevant to applications, but the notion of exploring community traits under group selection is interesting in and of itself.

We thank the reviewer for raising this important point. In response to these concerns, we have repeated all of the simulations in the paper for additional scenarios where these two assumptions have been relaxed, i.e. we consider the cases where (1) there is a fitness cost associated with the species' contribution to the community function; and (2) where only a fraction of species in the species pool contribute to community function. In addition, we have also repeated all of our simulations for a couple of biologically motivated complex functions where we do not a-priori assume any structure-function landscape: In the first one, we optimize the collective degradation/consumption of a given resource by the community (e.g. mimicking bioremediation); In the second case study, we optimize the community resistance to invasion by a given species (e.g. mimicking the development of consortia that will exclude a pathogen from the habitat). In Figures S12 and S16 we show that the main results reported in the paper still hold under all of these scenarios where the assumptions of additivity, redundant contributions and fitness neutrality have been eliminated.

We wish to note that, in our opinion, a major contribution of our paper is the development of a modeling package for artificial community-level selection

(*ecoprospector*) that would allow one to explore these as well as many other scenarios. Because the *ecoprospector* platform is flexible, one can use it to model different real-life scenarios (such as the two we added to the revised paper and discuss above) and also to address theoretical questions about the structure of the functional landscape, the importance of species interactions, etc. and how these would impact selection strategies.

In the first page of this document we summarize all of the new simulations we have carried out to explore the limitations and possible generalizability of our findings, which we did by relaxing the following assumptions (we copy and paste from the first page of this document for the convenience of the reviewer, as we do not know if the full document will be made available to all referees or only the part that addresses their specific concerns):

- ***Functional Additivity / Smooth landscape:*** *The new set of simulations include multiple non-additive community functions, including two where we model realistic scenarios: (1) selection for a community that optimizes the elimination of a specific metabolite and (2) selection for a community that will prevent invasion by an invasive species. As a pure exercise, we have also included selection for a function that is explicitly "epistatic", i.e. set by the sum of pairwise contributions and where there is no independent contribution from any species.*

- ***Functional redundancy:*** *The new set of simulations include cases where only a small fraction (20%) of the species even contribute to function, allowing us to test the effect of the original scenario where all species contributed.*

- ***Fitness neutrality:*** *The new set of simulations include the case where species contributions to function impose a cost on microbial growth (such cost lowers the amount of biomass that can be produced per unit of nutrient uptake). The cost grows linearly with the amount of contribution of a species to the collective function.*

- ***Type-III Functional response:*** *The new set of simulations include the case where species uptake rates follow a Type-II Hill and Type-I Linear functional response with nutrient concentration.*

- ***Pure competition for resources:*** *The new simulations include the scenarios where species can interact via cross-feeding based facilitation in addition to competition. This includes a new minimal-media scenario where there is only a single supplied resource and species coexistence is dependent on facilitation.*

- ***Species distribution in the regional pool****: The new simulations include additional species distributions in the species pool from where the inocula are drawn (i.e., log-normal distribution vs power-law) as well as new forms of inoculation (i.e. fixed number of species at identical abundances in the starting community).*

- ***Distribution of per-capita contribution to function:*** *The new simulations include scenarios where the per-capita contribution to function is sampled from a normal distribution with mean 1 (eliminating the assumption that most species have a negative effect on function. We also include a scenario where per-capita contribution to function is sampled from a uniform distribution between 0 and 1 (eliminating the assumption that high functioning taxa are rare).*

This new suite of simulations illustrates the flexibility of our modeling approach, and the outcome is presented in a new set of figures (Figures S11-S16) and described in six additional pages of an expanded supplementary text. Whilst we cannot feasibly explore all possible ecological and functional scenarios one could conceive, as they are so vast, we hope that the new batch of simulations we have carried out will convince the reviewer that our methods and simulation package can be extended well beyond the particular set of assumptions we have focused on in this paper, and also may be applied to model concrete real-world applications (of course doing this properly for each individual case would require additional thought on the part of the interested investigator).

In the revised manuscript, we introduce the flexibility of our modelling approach in lines 130-133 referring to the new simulations in Supplements and explicitly discuss this point in the first paragraph of Limitations section in Discussion (lines 513-529).

Organizationally, the major flaws in the paper include an overall lack of clarity and the initial section in the results, which in my opinion shouldn't be present in the main text at all. My first thought when reading the protocol outline was to wonder: is it long enough for the communities tos stabilize under the default conditions? One cannot tell from Fig. 1, and we only learn that this was a hypothesis to be tested (possibly - it's unclear) later on. It also means almost two full pages of the draft (three including Figure 1) is devoted to discovering and correcting what is essentially a flaw in the initial silico approach – not unimportant to future modeling efforts, but definitely not related to the stated thrust of the manuscript. This section should be condensed or moved; Fig. 1C-F could be in the supplement.

We thank the reviewer for pointing this out. This has been a miscommunication by our part, and we have attempted to make the writing clearer to avoid the confusion we created.

The first part of the results section seeks to provide a plausible explanation for why previous empirical attempts to do artificial ecosystem level selection (in papers published over the past twenty years) have largely been underwhelming.

To do that, we show that the selection protocols that have been used in those previous empirical studies are intrinsically flawed, and also show that key controls were lacking. We do this by applying those empirical protocols in a systematic fashion to an in silico community-selection problem, where, if they fail, we may understand what went wrong and why. None of these previously used protocols work for the in silico problem we posed (which is also one of the simplest: an additive function that is also redundant; now we show they do not work either for a sample of additional optimization problems).

However, this is not a flaw of the initial in silico approach. On the contrary, it is a fundamental flaw of the empirical protocols themselves. Similarly we think it is important to walk the reader through the need to use F_{max} as the appropriate metric for success rather than the mean, again, because previous studies have all considered an improvement in mean function as the metric for 'success'. We demonstrate that the growth in mean function is a severely flawed metric for engineering purposes (though it is valid if what you want to ask are theoretical questions regarding the levels of selection). This important point has never, to our knowledge, been made before in the Artificial *Ecosystem* Selection field.

We are open to having a discussion with the editor and the reviewer about the inclusion of this first section of the Results in the main text. In our opinion, this section is important, as it anchors our work into the existing artificial ecosystem level selection literature. This is a field that has been fumbling in the dark with no theory, and we believe our paper will help this community understand the failures of previous papers and design better protocols. We do want to communicate with researchers in this field, and for that reason we believe it is important to keep this section.

We would also like to note that *reviewer 2 has suggested that we include the key findings from this section in the abstract* due to their importance to that field.

To improve the clarity of our presentation, we have edited the abstract and added lines 83-88 towards the end of the introduction to make the structure of the paper clearer.

In contrast, the description and rationale for iteratively combining bottlenecks and migrations is somewhat clearer, as is the section on robustness to invasion. Both describe intriguing and powerful results, so much so that I wish the assumptions were not so unrealistic. The exception is the "top-down" vs "bottom-up" language, which although making sense from an engineering perspective, seems to be a new terminology intended to replace the standard language contrasting directed evolution with rational design. This seems confusing and unnecessary, at least from the perspective of the molecular evolution literature.

We thank the reviewer for pointing out that it parallels with the rational design vs directed evolution terms in the molecular evolution literature. In order to make our paper accessible to both audiences we have stated this parallel in the introduction. We have also removed the reference to top-down in the first line of the abstract and from the title of the paper. We note, however, that the "top-down vs bottom-up" language is used commonly in the synthetic microbial ecology literature and especially in the Artificial Ecosystem Selection literature. The terms "bottom up" vs "top-down engineering of microbial communities" were used by Sloan Wilson and co-authors in the very first (and seminal) paper published on Artificial Ecosystem Selection (Swenson, PNAS 2000), so we adopted it here for that reason. We elaborate a bit more on this in the response to a later comment by the reviewer.

Overall, the paper has several merits. It raises two interesting hypotheses (1: different diversification algorithms will vary in their effectiveness when applied to the directed evolution of microbial consortia; 2: apply directed evolution strategies that include migrations will generate communities that are more robust to invasion than bottom-up assembly and stabilization) and tests them computationally, yielding clean and convincing results. Major weaknesses include the exceedingly large simplifying assumptions in the model, about which little can be done, and persistent clarity problems that could be fixed.

We thank the reviewer for their interest in this approach and hope the new simulations address the simplifying assumptions (shown in Figures S12-S16) they were concerned about.

In particular, the abstract and introduction could be much clearer with less linguistic overreach about the purely computational findings, lines 98-240 should be outright deleted or at least moved to the supplement as they actively obscure the key messages of the manuscript, and a bunch of language would best be

modulated (e.g. microbial ecosystems being "dynamic" whereas evolving genes are not and other miscommunications – see below).

With respect to lines 98-240 (the first part of the results section) please see our previous comment. In terms of language we agree that there may be some confusion and have sought to clarify them. See below where we addressed the specific issues that the reviewer raised.

In short, there's a lot of cleaning up to do, but there does seem to be something worthwhile here, even if it appears too high-level to pin down at present – unless, of course, the findings would change dramatically when there is a fitness cost to performing the relevant trait. I'd appreciate a chance to see a revised draft.

We are grateful to the reviewer for their thoughtful comments and hope that by incorporating the new simulations and rewriting substantial parts of the paper, we will have addressed their primary concerns.

Sincerely,
Kevin Esvelt

Specific communication/accuracy issues:

Title: "Top-down" directed evolution – as far as I know, there is no such thing as bottom-up directed evolution. The strain-by-strain "bottom up" approach presented is what we would call rational design. I recommend either naming this something other than directed evolution, or removing the top-down language in favor of directed evolution vs rational design.

As we discuss above, the terms "top-down" vs "bottom up community engineering" have been commonly used in the Artificial Ecosystem Selection field since its inception two decades ago. Since this paper has that field as a primary audience, we believe it would be good to maintain that language. At the same time, we do want to speak to the molecular directed evolution community as well. To find a middle ground between the language commonly used in both fields, we have eliminated the words Top-down from the title and the first sentence of the abstract (which refers to the history of directed evolution, where that language would not be appropriate), but we keep them in the text when we make references to artificial ecosystem selection (where the term is commonly used) to provide the appropriate context. We hope that this solution will be acceptable as a compromise.

Abstract: Presenting the findings as though they were general and relevant to applications is misleading. I highly recommend explaining the no-fitness-cost and all-species-contribute assumptions, to clarify for the reader and to make it clear that there's nothing wrong with using a simple model to explore dynamics.

We now show in Figures S12 and S16 that neither of these and other assumptions are required for our results to hold true. Throughout the discussion we highlight that although our findings are robust when we relax those assumptions, they are not necessarily general under all conditions and for all functions. Rather, we emphasize that the methods we have developed are broadly applicable and can be used to explore other regimes and scenarios.

To further emphasize this point, we have created a new subsection of the discussion titled Limitations, where we explicitly discuss the many assumptions that we have not yet studied and the limits this imposes to any claims of generality.

Line 90: "Unlike molecular fitness landscapes, however, the ecological structure-function landscape is dynamic..."

Molecular fitness landscapes are highly dynamic: they are primarily shaped by competition from other mutants in the population. To phrase it similarly to the language used, the composition of variants in a given region of the fitness landscape changes over time due to competitive interactions; fitness is relative. As some have suggested, "seascape" would be more appropriate than "landscape"… at any rate, the suggested difference does not seem remotely as straightforward as the language indicates. Moreover, the more-dynamic nature of the ecological results seems an artifact of the way that they were generated, which seems unlikely to be the way one would go about an actual directed ecological evolution experiment.

We have revised that sentence in the introduction and first paragraph of the discussion in light of the comment by the reviewer.

Line 215: How is the process by which the composition of the community changes as it adapts to the batch environment from the random initial seeding not evolution? The individual species are all under selection pressure to utilize the 90 limiting nutrients. That seems no different from the cells of Lenski's LTEE.

The LTEE is initialised from a single clonal population and new genotypes are generated via mutation. Selection then acts on this variation (at the strain level). In community assembly experiments (and in the simulations in this manuscript), an extremely large amount of variation is introduced at the start of the experiment by including multiple different genotypes (or species) and this standing variation can be replenished by the methods we outline in the text. Selection then acts on this standing genetic variation to produce communities on short timescales. There is no 'within species' evolution in our simulations. It is important to note that group-level evolution does not require the evolution of the species within, and can be reached through purely ecological processes. This has been noted numerous times in the multi-level selection literature (e.g. Wilson (1972) PNAS, Wilson (1995) Ecology, Williams and Lenton (2007) PNAS)

Line 263: the nature of "compositional variants" is quite vague. Adding "using one of a variety of different possible methods" would clarify that there are options, e.g. varying proportions of the same species, adding new ones, removing species, etc. For reasons of flow, I recommend explicitly discussing the options to be tested later on, then noting that you began with bottlenecking.

We thank the reviewer for raising this point and agree that we did not precisely define a compositional variant and have now clarified this on line 273-275. We have also added a sentence to line 309 which we believe improves the flow of this section.

Line 327: knocking down a species with a phage is easy; knocking it out is exceptionally difficult. Seems like it would be easier to sequence the community and reconstitute in similar relative proportions without one or another target species. Similarly, narrow-spectrum antibiotics are extremely blunt instruments; describing them as a method for targeted removal doesn't seem reasonable.

We have re-worded this sentence and it now reads: In practice, entirely knocking out a species from a natural habitat is challenging, but tools exist for the depletion or knock-down of species from natural and synthetic communities (Ting et al. 2020; Sheth et al. 2016; Lemon et al. 2012; Harcombe and Bull 2005; Chan et al. 2018).