

Deep indel mutagenesis reveals the impact of amino acid insertions and deletions on protein stability and function

Corresponding Author: Professor Ben Lehner

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Nature Communications.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The authors have addressed most of my comments and included some interesting new analyses (e.g. of human pathogenic variants, etc). I have no further substantive critiques.

(Remarks on code availability)

The GitHub repo looks complete and README file is adequate, but I didn't try to download/run the code. Since predictions are pre-computed for the human genome, I don't feel that publication should hinge on how well the code runs.

Reviewer #3

(Remarks to the Author)

(Remarks on code availability)

The code and data are available to reproduce the figures.

I have not installed and run the code.

Reviewer #4

(Remarks to the Author)

The authors satisfied all my concerns and those of my co-reviewers. We particularly appreciated that the figures were heavily edited for clarity and that more context is given for specific structural comparisons. This is a majorly revised manuscript and is suitable for publication.

Up to the discretion of the authors, but I suggest potentially considering citing a recent preprint (Tenthorey, Biorxiv 2024) and published manuscript (Bakhache, Nature Microbiology 2024) in this exciting space exploring the role of indels in the evolution of massive viral capsid proteins and host immune receptor scans.

Willow Coyote-Maestas

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Revision “Deep indel mutagenesis reveals the impact of amino acid insertions and deletions on protein stability and function”.

The authors greatly improved their work, the manuscript text and the presentation of the results/figures. These are only few issues left that are important to include or consider:

We thank the reviewers for their enthusiasm and additional helpful suggestions.

1. The authors use histograms to show Pearson correlation coefficient among different groups (insertions, deletions). These could be complemented with boxplots or violin plots to show differences among groups more clearly. Pearson correlation coefficient assumes that the data is distributed normally which does not seem the case with this data, thus a non-parametric correlation coefficient spearman should be preferred.

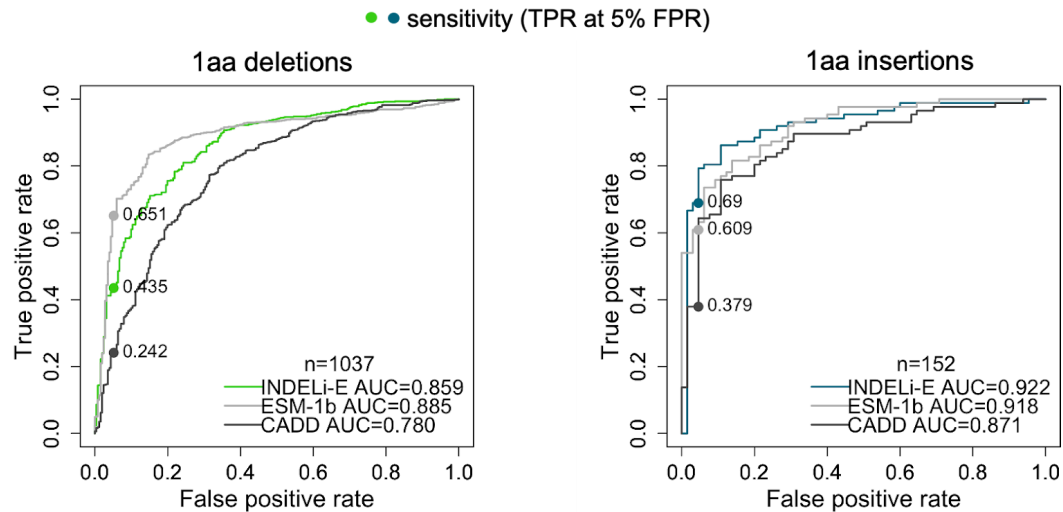
We thank the reviewers for highlighting this important point. In response, we have calculated both the Pearson and Spearman correlation coefficients to evaluate the performance of both the INDELi models and the existing predictors. The results indicate highly similar performance. For example, the mode of the correlation distributions for predicting deletion tolerance with INDELi-X changes only slightly, from 0.620 (Pearson) to 0.616 (Spearman). There are arguments for using both metrics of correlation. For example, when distributions are bimodal, the Spearman rank correlation coefficient is not a good metric to use.

2. The authors may want to check again throughout the text that the word correlation is used appropriately and not confused with prediction (comments made already in the first revision of the manuscript and more detailed comments below). The authors can use the word prediction when more rigorous analyses are discussed. For example, when comparing predictive models such as VEPs and INDELi models.

We have carefully reviewed the text again to address this point. Importantly, a predictive model does not require establishing causation. For example, substitution scores correlate with indel tolerance and they are predictive of indel tolerance.

3. The authors claim ‘excellent’ performance of INDELi models. This is a bold statement especially with 43% sensitivity on clinical variants. Moreover, a comparison on clinical variants with ESM1b could be useful since it is already used for other comparisons.

We thank the reviewers for the suggestion of including ESM-1b in our evaluation of INDELi-E as a pathogenicity classifier. We have included ESM-1b into the analysis and additionally included the performance of CADD, a trained pathogenicity classifier, as another reference point. Our findings show that CADD exhibits poorer performance i for classifying 1-amino acid deletions and insertions using ClinVar indels. By contrast, ESM-1b, a predictor of evolutionary fitness, demonstrates higher performance for classifying deletions, but lower performance for insertions when compared to INDELi-E.



We have reworded the section referring to the performance of INDELi, as suggested by the reviewer. However, we would like to emphasize that the reported sensitivity represents only one point on the AUC curve. Sensitivity measures how effectively the model identifies true cases of pathogenic indels at a specified false-positive rate, and hence it is distinct from accuracy. For deletion classification, a sensitivity of 43% is observed at a false-positive rate of 5%, reflecting a trade-off between sensitivity and specificity. Importantly, the overall performance of the model, as indicated by an AUC of 0.859, is highly competitive with state-of-the-art methods, such as ESM-1b (AUC = 0.885) and CADD (AUC = 0.780) for 1aa deletions.

Overall, the manuscript improved a lot compared to the previous version. There is an impressive amount of data and more rigorous analyses. I recommend it for publication.

Once again we thank the reviewers for making helpful suggestions and helping us improve our manuscript and analyses.

Minor comments:

Since the manuscript text does not have lines' numbers, the authors will find below the comments grouped by the relevant sections.

Supplementary Files

It would be great to add in the Supplementary PDF the description of each supplementary file provided (e.g., file name 510864_0_data_set_9092481_sdwy48.csv contains this information).

Note that file name 510864_0_data_set_9092481_sdwy48.csv is corrupted and cannot be opened.

We apologise for the inconvenience. We have added a description of the supplementary files in the compiled document (Supplementary Information) and made sure all files are accessible and can be opened in the resubmitted version.

Figures

Fig. 1

Panel a has a typo: TCT is actually TGT. The authors illustrate the 1-3aa insertions with comments, they could do the same illustration for the 1-3aa deletions.

Thank you for catching the typo. We have corrected it and added an illustration for the 1-3aa deletions to Figure 1.

Fig. 2

Move the legend on the top of the plot. The +32aa on the heatmap of protein domain PSD95-PDZ3 is confusing, can you remove it or make it clearer somehow?

We have removed the +32aa from the heatmap and instead left the description in the figure legend.

“For PSD95-PDZ3 and GRB2-SH3 only, we mutated the first 52 residues, leaving, respectively, 32 and 4aa in the C-terminal as wild type. ”

Fig. 4

Panel on the left has a typo: TCT is actually TGT.

Thank you, we have corrected the typo.

Figure legends

Fig. 1b

Add correlation metric used and add ddG stability (x-axis) and aPCA (y-axis). The names used in the legend are different than the ones used on the plot.

Thank you, we have corrected this in the Figure legend 1.

Figure 1. Deep indel mutagenesis of diverse protein domains. **a.** Library design for the 9 protein domains, including single and multi-aa indels, delSubs and single aa substitutions. **b.** Correlation between *in vitro* measured ddG stability (x-axis) and the raw aPCA scores (growthrate; y-axis) of selected substitution mutants. Selection of mutants was based on available *in vitro* ddG scores in Protherm²⁹.

Fig. 2a and Fig. 3

Which p-value threshold was considered significant? Please add this information and specify which two-tailed test was used.

Thank you, we have clarified the test directionality and the p-value significance threshold in the figure legends.

Figure 2. Patterns of indel tolerance within and between domains. **a.** Heat maps of protein domain abundance effects of all mutation types with significant changes in

abundance, as determined by a Bonferroni multiple testing correction of z-scores (calculated using a two-tailed test with a significance threshold of $FDR \leq 0.05$), are marked with “*”. The two-tailed test compares the z-statistic for each mutation to the null hypothesis (mean = 0). We show the 2-dimensional structural representation of the domains (SSDraw⁶²) under each heatmap. For PSD95-PDZ3 and GRB2-SH3 only, we mutated the first 52 residues, leaving, respectively, 32 and 4aa in the C-terminal as wild type.

Figure 3. Contrasting the impact of insertions and deletions to substitutions. a.

Heatmaps of 19 single aa substitutions, 20 aa insertions, 1-3aa insertion repeats and 1-3aa deletion effects on protein abundance. Changes in abundance significantly different from the weighted mean of synonymous variants, determined using a Bonferroni multiple testing correction of z-scores (two-tailed test, $FDR \leq 0.05$), are marked with “*”. The two-tailed test compares the z-statistic for each mutation to the null hypothesis (mean = 0). The y-axis shows the identity, type, or length of mutation (multi-insertion repeats refer to duplicating the preceding amino acid 1-3 times; CCC indicates repeating the 1-3 amino acid sequence preceding the insertion site), and the x-axis shows the mutated position and amino acid sequence of the domain. We show the 2-dimensional structural representation of the domains (SSDraw⁶²) under each heatmap. For PSD95-PDZ3 and GRB2-SH3, we mutated the first 52 residues, leaving, respectively, 32 and 4aa in the C-terminal as wild type.

“For PSD95-PDZ3 and GRB2-SH3 only, we mutated the first 52 residues, leaving, respectively, 32 and 4aa in the C-terminal as wild type.”

The rationale is unclear. Could you add an explanation why.

We have added an explanation in the Methods section.

For PSD95-PDZ3 and GRB2-SH3, we mutated the first 52 residues, leaving, respectively, 32 and 4 amino acids in the C-terminal as wild type. This approach was due to the length constraints of the commercially available oligos.

Fig. 3

In panel a, the rectangle illustration is misleading and redundant. The authors could add the next to the relevant heatmap the corresponding illustrations. For example, the first heatmap is about residue substitution therefore on the lefthand side of the plot you could add the illustration of a substitution.

We have removed the legend and instead added a more detailed explanation in the figure legend.

Figure 3. Contrasting the impact of insertions and deletions to substitutions. a.

Heatmaps of 19 single aa substitutions, 20 aa insertions, 1-3aa insertion repeats and 1-3aa deletion effects on protein abundance. Changes in abundance significantly different from the weighted mean of synonymous variants, determined using a Bonferroni multiple testing correction of z-scores (two-tailed test, $FDR \leq 0.05$), are marked with “*”. The two-tailed test compares the z-statistic for each mutation to the null hypothesis (mean = 0). The y-axis shows the identity, type, or length of mutation (multi-insertion repeats refer to duplicating the preceding amino acid 1-3 times; CCC indicates repeating the 1-3 amino acid sequence preceding the insertion site), and the x-axis shows the mutated position and amino acid

sequence of the domain. We show the 2-dimensional structural representation of the domains (SSDraw⁶²) under each heatmap. For PSD95-PDZ3 and GRB2-SH3, we mutated the first 52 residues, leaving, respectively, 32 and 4aa in the C-terminal as wild type.

The illustration for the heatmap of the insertion remains unclear to me. What does each cell in the heatmap represent? How did you summarize the 1-3aa insertions?

The second heatmap represents the abundance score for the effects of 1aa insertions. To clarify, we have added “1aa insertions” to the legend. The heatmap summarising the effects of 1-3aa insertions is below the “1aa insertions” heatmap. We have added the title “multi-insertions” for clarification.

Typo: the last panel letter is d not c.

Corrected.

Extended Data Figure 2. f

Are those scores still abundance scores?

Yes, this has now been highlighted in the figure and figure legend.

Extended Data Figure 4. d

What is on the x and y axis?

The y-axis is the measured DMS score, while the x-axis is the predicted INDELi-E score for the set of full-length proteins from ProteinGym. We have changed the x- and y-axis titles for clarification.

Introduction

Typo: “Indels are a fundamentally different type of perturbation to substitutions.”
Indels are a fundamentally different type of perturbation than substitutions.

Thank you, corrected.

Typo: “most often due polymerase slippage during DNA replication”
most often due to polymerase slippage during DNA replication

Thank you, corrected.

Typo: “Here we use this dataset, as well as new deep indel mutagenesis quantifying in vivo protein abundance across structurally diverse domains to test the generality of previously proposed and new hypotheses for indel tolerance.”

Thank you, corrected.

“We also use the data to develop a series of models – INDELi - that perform well for predicting the effects of indels on protein stability. These models also perform well for predicting indels that are pathogenic in humans, consistent with reduced protein stability being the major mechanism by which indel variants cause genetic diseases.”

Missing reference on the last part of the sentence.

Thank you for pointing this out, we have rephrased this paragraph.

We also use the data to develop a series of models - INDELi - that perform well for predicting the effects of indels on protein stability. These models also perform well for predicting indels that are pathogenic in humans, [suggesting that](#) reduced protein stability [is a](#) major mechanism by which indel variants cause genetic diseases.

Results

“In general, 1aa CCC insertions are slightly better tolerated (Kolmogorov-Smirnov test, $p\text{-value}=0.0343$),[...]

As well as the authors did in other parts of the text, they should specify the directionality of the test (e.g., two tailed).

Thank you, directionality of the test was added.

In general, 1aa CCC insertions are slightly better tolerated (Kolmogorov-Smirnov test, [one-tailed](#), $p\text{-value}=0.0343$), with 64% strongly reducing abundance ([Supplemental Fig. 1c](#)).

“For PSD95-PDZ3, which tolerates indels throughout its tertiary structure (Fig. 3a), the average effect of substitutions predicts reasonably well the effect of deleting the residue ($r=0.534$, Fig. 3b) as well as CCC insertions after ($r=0.441$, Fig. 3b) and before ($r=0.313$, Extended Data Fig. 2b) the site.”

According to Fig. 3a, for the domain PSD95-PDZ3 there are regions, e.g., loop 1 and 2, there are tolerant to indels, and regions that don't, e.g., residues 25-40, 44-53. Thus, it seems that the tolerance varies based on the tertiary structure element.

Furthermore, substitute the word ‘predict’ with the word ‘correlate’ for what regards Fig. 3b.

Thank you, we have reworded this section.

For PSD95-PDZ3, which tolerates indels throughout [its secondary structure elements](#) (Fig. 3a), the average effect of substitutions [correlates](#) reasonably well [with](#) the effect of deleting the residue ($r=0.534$, Fig. 3b) as well as CCC insertions after ($r=0.441$, Fig. 3b) and before ($r=0.313$, [Supplemental Fig. 2b](#)) the site.

“For GRB2-SH3, which is highly intolerant of indels (Fig. 3a), substitution tolerance is reasonably predictive of CCC insertions after the substitution site ($r=0.397$) but less predictive of the tolerance for insertions before a site ($r=0.0617$ [$p=0.667$], Extended Data Fig. 2b) and deletion tolerance ($r=0.208$ [$p=0.143$], Fig. 3b).”

The same comment as above, substitute the word ‘predict’ with the word ‘correlate’.

Thank you, we have reworded this section.

For GRB2-SH3, which is highly intolerant of indels (Fig. 3a), substitution tolerance is reasonably [correlated to that of](#) CCC insertions after the substitution site ($r=0.397$) but less [correlated to](#) the tolerance for insertions before a site ($r=0.0617$ [$p=0.667$], [Supplemental Fig. 2b](#)) and deletion tolerance ($r=0.208$ [$p=0.143$], Fig. 3b).

“Filtering the high-confidence subset of this dataset for positions with measurements of 1aa deletions, 1 Ala insertions and 19 substitutions, results in a dataset of substitutions and indels at 6,288 positions across 181 protein domains (see Methods and Fig. 3c).”

Hard to read, please rephrase.

Thank you, we have rephrased this section.

[By](#) filtering the high-confidence subset of this dataset [to include](#) positions with measurements [for](#) 1aa deletions, 1 Ala insertions and 19 substitutions, [we obtained](#) a dataset of [comprising](#) substitutions and indels at 6,288 positions across 181 protein domains (see Methods and Fig. 3c).

“Across 171 protein domains (domains with >9 indel measurements) , substitution tolerance predicts Ala insertion tolerance similarly well after (mode $r=0.263$) and before a residue ($r=0.253$) and better predicts 1 aa deletion tolerance (mode $r=0.450$, Fig. 3c and Extended Data Fig. 2c). “

Substitute ‘predict’ with ‘correlate’ in the entire paragraph. Clarify phrasing and avoid the use of modifiers like well and just report the correlation metric. The substitution tolerance and the 1aa deletion tolerance histogram could be added to the histogram plots in Extended Data Fig. 2c, as they all related to statements in this sentence. The two histograms in Fig. 3c are only mentioned in the next section and one is not mentioned at all.

[As stated above, the relationship between substitution scores and indel tolerance is only a correlation, but it does mean that substitution scores predict indel tolerance.](#)

[Regarding the suggestion to add substitution and 1aa deletion tolerance histograms to Extended Data Fig. 2c, this information is already included in main text Figure 3c.](#)

“Interestingly, substitution tolerance is a better predictor of insertion and deletion tolerance in loops than in secondary structure elements (mode $r=0.510$ and $r=0.544$ for insertions and deletions in loops, mode $r=0.219$ and $r=0.333$ for insertions and deletions in secondary structure elements, Extended Data Fig. 2d).”

Substitute ‘predict’ with ‘correlate’ in the entire paragraph.

[We respectfully maintain our rationale for the use of the term “predictive”, as described above.](#)

“we quantified the effects of 325 across the nine structurally diverse proteins.”
Add 325 delSubs.

Thank you, this typo has been corrected.

“However, analyzing substitutions, 1 Ala insertions and 1aa deletions across 171 domains shows that solvent exposure is a much less important determinant of indel tolerance than of substitution tolerance.”

Please rephrase the sentence, very hard to read.

Thank you, we have rephrased this sentence.

However, an analysis of substitution and 1 Ala insertion and 1aa deletion tolerance across 171 domains, reveals that solvent exposure plays a much smaller role in determining indel tolerance compared to substitution tolerance.

“This suggests substitution scores could be repurposed for indel variant effect prediction. Moreover, although the identity of the wildtype aa only poorly predicts deletion and insertion tolerance (model m2, mode $r=0.226$ and 0.182 , respectively, Fig. 7b), a regression model using secondary structure features (strand/helix/loop/termini, length of secondary structure, neighboring structures, and the position of the indel) provides quite good deletion (model m3, mode $r=0.538$) and insertion (mode $r=0.573$, Fig. 7b) prediction, when evaluated by leave-one-out cross validation (See Methods, Fig. 7a). This performance is similar to that of the state-of-the-art ESM1b protein language model (Fig. 6a), despite the model containing only 33 parameters.”

This paragraph jumps into the models developed without introducing the need of a model, what is the input and output of the model, how many models were trained, what architecture was used, what data was used to train and test the models, a reference table to the features used. Fig. 7 contains all the information to enrich and develop the paragraph.

Thank you for pointing this out. We have added an additional paragraph to this section to provide a clearer introduction to the models and features used.

Accurate indel variant effect prediction using INDELi

Across 171 domains, the average effect of substitutions correlates with deletion tolerance (mode $r=0.450$) and more weakly with insertion tolerance (mode $r=0.253$, Fig. 3c and model m1 in Fig 7a-b for comparison). This suggests substitution scores could be repurposed for indel variant effect prediction.

We therefore trained and evaluated a series of regression models to predict deletion and insertion tolerance. The models were trained on the Tsuboyama et al. dataset using regularised multiple linear regression to prioritise simplicity and interpretability and were evaluated using a leave-one-domain-out approach (Fig. 7a).

“Strikingly, this model has predictive performance similar to state-of-the-art substitution variant effect predictors, with mode $r=0.620$ for deletions and mode $r=0.608$ for insertions (Fig. 7c).”

Fig. 7c only shows performance of INDELi-X and m4, perhaps the comparison with the state-of-the-art tools is somewhere else?

We thank the reviewer for pointing this out. This information is available in Supplemental Fig. 4e and has now been added to the main text.

Strikingly, this model has predictive performance similar to state-of-the-art substitution variant effect predictors, with mode $r=0.620$ for deletions and mode $r=0.608$ for insertions (Fig. 7c and Supplemental Fig. 4e).

“Moreover, the predictive performance is consistent across domains, with a narrower distribution of correlation coefficients compared to ESM1b (Fig. 7c and Extended Data Fig. 4e, interquartile range from $r=0.512$ to 0.686 for deletions and $r=0.393$ to 0.646 for insertions).”

The same comment as above, the comparison with ESM1b is not present in the plot. You have the performance of ESM1b in Fig. 6 but the reader can't compare directly INDELI-X and ESM1b in Fig. 7c and Extended Data Fig. 4e. Add the tool in the plot to aid the comparison.

Apologies, same as above, the reviewer is right and we have now added the correct reference to the text.

Moreover, the predictive performance is consistent across domains, with a narrower distribution of correlation coefficients compared to ESM1b (Fig. 7c and Supplemental Fig. 4e, interquartile range from $r=0.512$ to 0.686 for deletions and $r=0.393$ to 0.646 for insertions).

“Although the performance of these three models is similar, INDELI-E has the major advantage that it uses ESM-1v substitution predictions, which can be made genome-wide.” Please rephrase.

Thank you, we have rephrased the paragraph.

Although the performance of these three models is similar, INDELI-E has the major advantage that it uses ESM-1v substitution predictions, which can be efficiently run genome-wide, making INDELI-E highly scalable.

“Importantly for clinical applications, INDELI-E also has high sensitivity when making predictions with a low false positive rate (high specificity): for example at a 5% false positive rate, the sensitivity (true positive rate) for identifying pathogenic deletions and insertions is 43.4% and 68.2%, respectively (Fig. 7e).”

Claiming that a sensitivity of 43.5% and 68.2% is high for clinical applications is too strong. A sensitivity of 43.4% means that the predictor will fail to predict a pathogenic mutation in 56.6% of cases which is more than half, thus a random classifier will predict just as good. In addition, ESM1b could be added to check its performance on ClinVar variants.

We thank the reviewer for the suggestion of including ESM-1b in our evaluation of INDELI-E as a pathogenicity classifier. In response, we have incorporated ESM-1b into the analysis and additionally included the performance of CADD, a trained pathogenicity classifier, as

another reference point. Our findings show that CADD exhibits poorer performance for classifying 1aa deletions and insertions using ClinVar indels. By contrast, ESM-1b, a predictor of evolutionary fitness, demonstrates better performance for classifying deletions, but lower performance for insertions when compared to INDELi-E.

We have reworded the section referring to the performance of INDELi, as suggested by the reviewer. However, we would like to emphasize that the reported sensitivity represents only one point on the AUC curve. Sensitivity measures how effectively the model identifies true cases of pathogenic indels at a specified false-positive rate, and hence it is distinct from accuracy. For deletion classification, a sensitivity of 43% is observed at a false-positive rate of 5%, reflecting a trade-off between sensitivity and specificity. Importantly, the overall performance of the model, as indicated by an AUC of 0.859, is highly competitive with state-of-the-art methods, such as ESM-1b (AUC = 0.885) and CADD (AUC = 0.780) for 1aa deletions.

“This excellent performance is consistent with fold stability being the major mechanism by which aa indels cause human genetic diseases.”

Same comment as above regarding “excellent performance”, that should be toned down.

Thank you, we have rephrased this section.

This [accurate](#) performance is consistent with fold stability being the major mechanism by which aa indels cause human genetic diseases.

“Thus, although insertions and deletions - like substitutions - primarily have loss-of-function molecular phenotypes, insertions in two loops of PSD95-PDZ3 cause strong gain-of-function phenotypes.”

By strong do you mean pathogenic?

No, strong gain-of-function here, refers to the strong positive (higher-than-wt) effect of insertions on binding between PSD95-PDZ3 and CRIPT. We thank the reviewer for pointing this out and have rephrased this sentence for clarification.

Thus, although insertions and deletions - like substitutions - primarily have loss-of-function molecular phenotypes, insertions in two loops of PSD95-PDZ3 cause strong gain-of-function phenotypes, [enhancing binding to CRIPT](#).

Discussion

“Here we have used deep indel mutagenesis to quantify, understand and learn how to predict

the effects of insertions and deletions on protein fold stability . Indels and substitutions are fundamentally different changes to proteins, and this is reflected in their different patterns of tolerance. Our data show that tolerance varies both among different proteins and different regions of the same protein, as does the relative tolerance to insertion and deletion.”

Add references to figures in the text in support of the statements.

Thank you for pointing this out, a reference to figures in the text has been added.

Our data show that tolerance varies both among different proteins and different regions of the same protein, as does the relative tolerance to insertion and deletion (Fig. 2 and Supplemental Fig. 1c).

“Together with previous work on substitutions 41–45 , this suggests that fold destabilization is the most important mechanism by which aa variants cause human diseases.”

I don't see the evidence for this statement. INDELi-E does not have an accuracy, sensitivity, specificity close to 90% which is the common performance of VEPs, especially for clinical applications. This suggests that the missing performance is due to other mechanisms important to predict the pathogenicity of indels that are not included in the INDELi-E.

Since the INDELi-E model is trained specifically on stability changes and still it performs well at predicting pathogenicity (AUC-ROC=0.859 and 0.922 for classifying 1aa deletions and insertions respectively), it suggests that stability changes are likely a major mechanisms of pathogenicity for indels. If stability changes were not a key mechanism of indel pathogenicity, the model would struggle to differentiate pathogenic from non-pathogenic indels resulting in a much lower AUC-ROC. The high AUC supports the hypothesis that destabilisation is a primary driver of pathogenicity for indels. Hence, we respectfully disagree with the reviewer, especially based on the improved Figure 7e, where INDELi-E performs similarly or better than state-of-the-art VEPs for indel effect prediction.

“We designed indel libraries and performed selections for 13 domains but observed weak selection for 4 resulting in high measurement error and weak separation of nonsense and synonymous variants, likely due to insufficient growth time of the pool. We excluded these four domains from the analyses.”

It would be useful to report the names of the 4 domains that were excluded from the analysis.

Thank you for the suggestion, we have added the names of the domains in the methods section.

We designed indel libraries and performed selections for 13 domains but observed weak selection for 4 resulting in high measurement error and weak separation of nonsense and synonymous variants, likely due to insufficient growth time of the pool. We excluded these four domains (*arc-1*, *cro-1*, *MeCp2-MBD* and *FYN-SH3*) from the analyses.

References:

1. Dias, M., Orenbuch, R., Marks, D. S. & Frazer, J. Toward trustable use of machine learning models of variant effects in the clinic. *Am J Hum Genet* **111**, 2589–2593 (2024).

Revision “Deep indel mutagenesis reveals the impact of amino acid insertions and deletions on protein stability and function”.

The authors greatly improved their work, the manuscript text and the presentation of the results/figures. These are only few issues left that are important to include or consider:

1. The authors use histograms to show Pearson correlation coefficient among different groups (insertions, deletions). These could be complemented with boxplots or violin plots to show differences among groups more clearly. Pearson correlation coefficient assumes that the data is distributed normally which does not seem the case with this data, thus a non-parametric correlation coefficient spearman should be preferred.
2. The authors may want to check again throughout the text that the word correlation is used appropriately and not confused with prediction (comments made already in the first revision of the manuscript and more detailed comments below). The authors can use the word prediction when more rigorous analyses are discussed. For example, when comparing predictive models such as VEPs and INDELi models.
3. The authors claim ‘excellent’ performance of INDELi models. This is a bold statement especially with 43% sensitivity on clinical variants. Moreover, a comparison on clinical variants with ESM1b could be useful since it is already used for other comparisons.

Overall, the manuscript improved a lot compared to the previous version. There is an impressive amount of data and more rigorous analyses. I recommend it for publication.

Minor comments:

Since the manuscript text does not have lines’ numbers, the authors will find below the comments grouped by the relevant sections.

Supplementary Files

It would be great to add in the Supplementary PDF the description of each supplementary file provided (e.g., file name 510864_0_data_set_9092481_sdwy48.csv contains this information).

Note that file name 510864_0_data_set_9092481_sdwy48.csv is corrupted and cannot be opened.

Figures

Fig. 1

Panel a has a typo: TCT is actually TGT. The authors illustrate the 1-3aa insertions with comments, they could do the same illustration for the 1-3aa deletions.

Fig. 2

Move the legend on the top of the plot. The +32aa on the heatmap of protein domain PSD95-PDZ3 is confusing, can you remove it or make it clearer somehow?

Fig. 4

Panel on the left has a typo: TCT is actually TGT.

Figure legends**Fig. 1b**

Add correlation metric used and add ddG stability (x-axis) and aPCA (y-axis). The names used in the legend are different than the ones used on the plot.

Fig. 2a and Fig. 3

Which p-value threshold was considered significant? Please add this information and specify which two-tailed test was used.

“For PSD95-PDZ3 and GRB2-SH3 only, we mutated the first 52 residues, leaving, respectively, 32 and 4aa in the C-terminal as wild type.”

The rationale is unclear. Could you add an explanation why.

Fig. 3

In panel a, the rectangle illustration is misleading and redundant. The authors could add the next to the relevant heatmap the corresponding illustrations. For example, the first heatmap is about residue substitution therefore on the lefthand side of the plot you could add the illustration of a substitution.

The illustration for the heatmap of the insertion remains unclear to me. What does each cell in the heatmap represent? How did you summarize the 1-3aa insertions?

Typo: the last panel letter is d not c.

Extended Data Figure 2. f

Are those scores still abundance scores?

Extended Data Figure 4. d

What is on the x and y axis?

Introduction

Typo: “Indels are a fundamentally different type of perturbation to substitutions.”

Indels are a fundamentally different type of perturbation than substitutions.

Typo: “most often due polymerase slippage during DNA replication”

most often due to polymerase slippage during DNA replication

Typo: “Here we use this dataset, as well as new deep indel mutagenesis quantifying *in vivo* protein abundance across structurally diverse domains to test the generality of previously proposed and new hypotheses for indel tolerance.”

“We also use the data to develop a series of models – INDELi - that perform well for predicting the effects of indels on protein stability. These models also perform well for predicting indels that are pathogenic in humans, consistent with reduced protein stability being the major mechanism by which indel variants cause genetic diseases.”

Missing reference on the last part of the sentence.

Results

“In general, 1aa CCC insertions are slightly better tolerated (Kolmogorov-Smirnov test, p-value=0.0343),[...]”

As well as the authors did in other parts of the text, they should specify the directionality of the test (e.g., two tailed).

“For PSD95-PDZ3, which tolerates indels throughout its tertiary structure (Fig. 3a), the average effect of substitutions predicts reasonably well the effect of deleting the residue ($r=0.534$, Fig. 3b) as well as CCC insertions after ($r=0.441$, Fig. 3b) and before ($r=0.313$, Extended Data Fig. 2b) the site.”

According to Fig. 3a, for the domain PSD95-PDZ3 there are regions, e.g., loop 1 and 2, there are tolerant to indels, and regions that don't, e.g., residues 25-40, 44-53. Thus, it seems that the tolerance varies based on the tertiary structure element.

Furthermore, substitute the word ‘predict’ with the word ‘correlate’ for what regards Fig. 3b.

“For GRB2-SH3, which is highly intolerant of indels (Fig. 3a), substitution tolerance is reasonably predictive of CCC insertions after the substitution site ($r=0.397$) but less predictive of the tolerance for insertions before a site ($r=0.0617$ [$p=0.667$], Extended Data Fig. 2b) and deletion tolerance ($r=0.208$ [$p=0.143$], Fig. 3b).”

The same comment as above, substitute the word ‘predict’ with the word ‘correlate’.

“Filtering the high-confidence subset of this dataset for positions with measurements of 1aa deletions, 1 Ala insertions and 19 substitutions, results in a dataset of substitutions and indels at 6,288 positions across 181 protein domains (see Methods and Fig. 3c).”

Hard to read, please rephrase.

“Across 171 protein domains (domains with >9 indel measurements) , substitution tolerance predicts Ala insertion tolerance similarly well after (mode $r=0.263$) and before a residue ($r=0.253$) and better predicts 1 aa deletion tolerance (mode $r=0.450$, Fig. 3c and Extended Data Fig. 2c). “

Substitute ‘predict’ with ‘correlate’ in the entire paragraph. Clarify phrasing and avoid the use of modifiers like well and just report the correlation metric. The substitution tolerance and the 1aa

deletion tolerance histogram could be added to the histogram plots in Extended Data Fig. 2c, as they all related to statements in this sentence.

The two histograms in Fig. 3c are only mentioned in the next section and one is not mentioned at all.

“Interestingly, substitution tolerance is a better predictor of insertion and deletion tolerance in loops than in secondary structure elements (mode $r=0.510$ and $r=0.544$ for insertions and deletions in loops, mode $r=0.219$ and $r=0.333$ for insertions and deletions in secondary structure elements, Extended Data Fig. 2d).”

Substitute ‘predict’ with ‘correlate’ in the entire paragraph.

“we quantified the effects of 325 across the nine structurally diverse proteins.”

Add 325 delSubs.

“However, analyzing substitutions, 1 Ala insertions and 1aa deletions across 171 domains shows that solvent exposure is a much less important determinant of indel tolerance than of substitution tolerance.”

Please rephrase the sentence, very hard to read.

“This suggests substitution scores could be repurposed for indel variant effect prediction. Moreover, although the identity of the wildtype aa only poorly predicts deletion and insertion tolerance (model m2, mode $r=0.226$ and 0.182 , respectively, Fig. 7b), a regression model using secondary structure features (strand/helix/loop/termini, length of secondary structure, neighboring structures, and the position of the indel) provides quite good deletion (model m3, mode $r=0.538$) and insertion (mode $r=0.573$, Fig. 7b) prediction, when evaluated by leave-one-out cross validation (See Methods, Fig. 7a). This performance is similar to that of the state-of-the-art ESM1b protein language model (Fig. 6a), despite the model containing only 33 parameters.”

This paragraph jumps into the models developed without introducing the need of a model, what is the input and output of the model, how many models were trained, what architecture was used, what data was used to train and test the models, a reference table to the features used. Fig. 7 contains all the information to enrich and develop the paragraph.

“Strikingly, this model has predictive performance similar to state-of-the-art substitution variant effect predictors, with mode $r=0.620$ for deletions and mode $r=0.608$ for insertions (Fig. 7c).”

Fig. 7c only shows performance of INDELi-X and m4, perhaps the comparison with the state-of-the-art tools is somewhere else?

“Moreover, the predictive performance is consistent across domains, with a narrower distribution of correlation coefficients compared to ESM1b (Fig. 7c and Extended Data Fig. 4e, interquartile range from $r=0.512$ to 0.686 for deletions and $r=0.393$ to 0.646 for insertions).”

The same comment as above, the comparison with ESM1b is not present in the plot. You have the performance of ESM1b in Fig. 6 but the reader can’t compare directly INDELi-X and ESM1b in Fig. 7c and Extended Data Fig. 4e. Add the tool in the plot to aid the comparison.

Although the performance of these three models is similar, INDELi-E has the major advantage that it uses ESM-1v substitution predictions, which can be made genome-wide.

Please rephrase.

Importantly for clinical applications, INDELi-E also has high sensitivity when making predictions with a low false positive rate (high specificity): for example at a 5% false positive rate, the sensitivity (true positive rate) for identifying pathogenic deletions and insertions is 43.4% and 68.2%, respectively (Fig. 7e).

Claiming that a sensitivity of 43.5% and 68.2% is high for clinical applications is too strong. A sensitivity of 43.4% means that the predictor will fail to predict a pathogenic mutation in 56.6% of cases which is more than half, thus a random classifier will predict just as good.

In addition, ESM1b could be added to check its performance on ClinVar variants.

This excellent performance is consistent with fold stability being the major mechanism by which aa indels cause human genetic diseases.

Same comment as above regarding “excellent performance”, that should be toned down.

Thus, although insertions and deletions - like substitutions - primarily have loss-of-function molecular phenotypes, insertions in two loops of PSD95-PDZ3 cause strong gain-of-function phenotypes.

By strong do you mean pathogenic?

Discussion

Here we have used deep indel mutagenesis to quantify, understand and learn how to predict the effects of insertions and deletions on protein fold stability . Indels and substitutions are fundamentally different changes to proteins, and this is reflected in their different patterns of tolerance. Our data show that tolerance varies both among different proteins and different regions of the same protein, as does the relative tolerance to insertion and deletion.

Add references to figures in the text in support of the statements.

Together with previous work on substitutions ^{41–45}, this suggests that fold destabilization is the most important mechanism by which aa variants cause human diseases.

I don't see the evidence for this statement. INDELi-E does not have an accuracy, sensitivity, specificity close to 90% which is the common performance of VEPs, especially for clinical applications. This suggests that the missing performance is due to other mechanisms important to predict the pathogenicity of indels that are not included in the INDELi-E.

“We designed indel libraries and performed selections for 13 domains but observed weak selection for 4 resulting in high measurement error and weak separation of nonsense and synonymous variants, likely due to insufficient growth time of the pool. We excluded these four domains from the analyses.”

It would be useful to report the names of the 4 domains that were excluded from the analysis.

REVIEWERS' COMMENTS

Reviewer #2 (Remarks to the Author):

The authors have addressed most of my comments and included some interesting new analyses (e.g. of human pathogenic variants, etc). I have no further substantive critiques.

[We thank the reviewer for the enthusiasm around our new analyses and helpful suggestions.](#)

Reviewer #2 (Remarks on code availability):

The GitHub repo looks complete and README file is adequate, but I didn't try to download/run the code. Since predictions are pre-computed for the human genome, I don't feel that publication should hinge on how well the code runs.

Reviewer #3 (Remarks on code availability):

The code and data are available to reproduce the figures.

I have not installed and run the code.

[We thank the reviewer for the comments.](#)

Reviewer #4 (Remarks to the Author):

The authors satisfied all my concerns and those of my co-reviewers. We particularly appreciated that the figures were heavily edited for clarity and that more context is given for specific structural comparisons. This is a majorly revised manuscript and is suitable for publication.

[We thank the reviewer for the enthusiasm around our edited figures and analyses.](#)

Up to the discretion of the authors, but I suggest potentially considering citing a recent preprint (Tenthorey, Biorxiv 2024) and published manuscript (Bakhache, Nature Microbiology 2024) in this exciting space exploring the role of indels in the evolution of massive viral capsid proteins and host immune receptor scans.

[Thank you, these references have now been added.](#)

Willow Coyote-Maestas