

Semantically Rich Local Dataset Generation for Explainable AI in Genomics



Pedro Barbosa

LASIGE, IMM
U. Lisbon



Alcides Fonseca

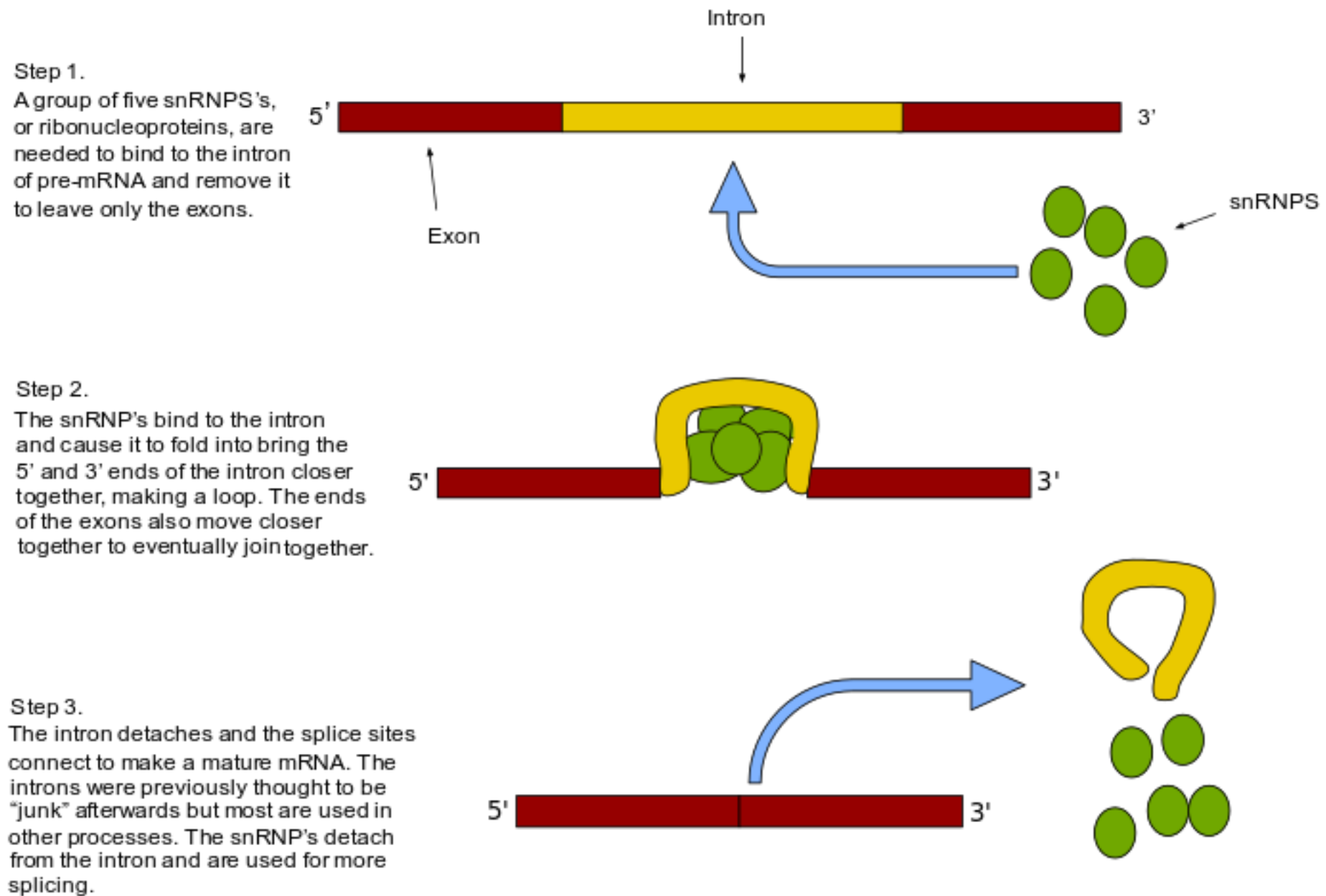
LASIGE
U. Lisbon



Rosina Savisaar

Mondego Science

Genomic mechanisms are not fully understood



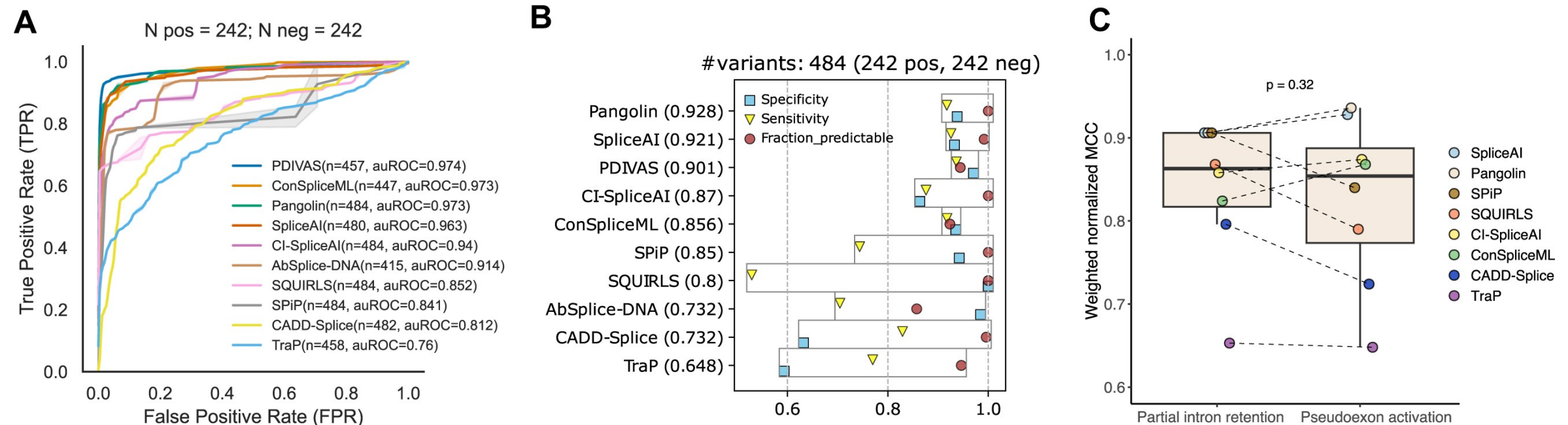
Deep Neural Networks can predict function

JOURNAL ARTICLE

Computational prediction of human deep intronic variation

Pedro Barbosa , Rosina Savisaar, Maria Carmo-Fonseca, Alcides Fonseca 

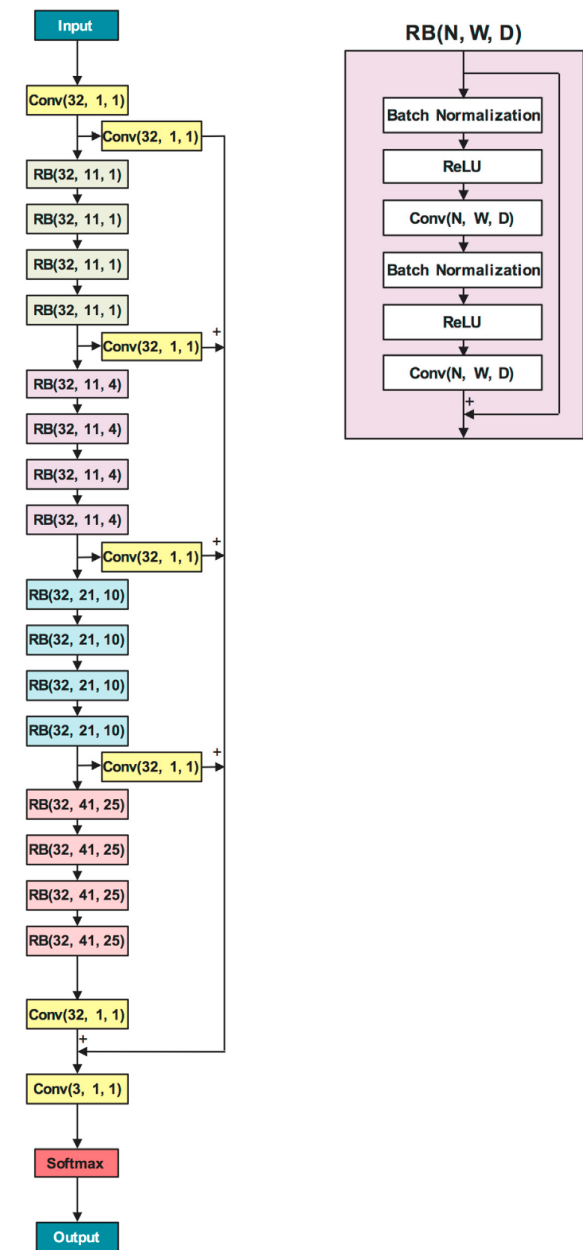
GigaScience, Volume 12, 2023, giad085, <https://doi.org/10.1093/gigascience/giad085>



General Problem

How can deep-neural networks reveal biology insight?

- Post-hoc interpretation techniques show patterns, but do not explain the semantic mechanisms
- Global interpretable surrogates do not scale to this complexity
- We need local interpretable surrogates!



Cell



Volume 176, Issue 3, 24 January 2019, Pages 535-548.e24

Article

Predicting Splicing from Primary Sequence with Deep Learning

Kishore Jaganathan^{1,6}, Sofia Kyriazopoulou Panagiotopoulou^{1,6}, Jeremy F. McRae^{1,6}, Siavash Fazel Darbandi², David Knowles³, Yang I. Li³, Jack A. Kosmicki^{1,4}, Juan Arbelaez², Wenwu Cui¹, Grace B. Schwartz², Eric D. Chow⁵, Efsthios Kanterakis¹, Hong Gao¹, Amirali Kia¹, Serafim Batzoglou¹, Stephan J. Sanders², Kyle Kai-How Farh^{1,7}  

Specific Problem

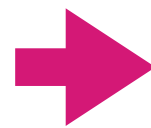
On which data should we train a local surrogate?

- Sequences should be syntactic neighbours
(biologically similar sequences)
- Sequences should be semantically diverse
(different outputs)

Specific Problem

On which data should we train a local surrogate?

- Sequences should be syntactic neighbours
(biologically similar sequences)
- Sequences should be semantically diverse
(different outputs)



Approach

Genetic Programming

- Custom Grammar, inspired by biological mutations
- Custom Fitness Function

Custom Grammar

Biological similarity limit

Biological mutations

Further restricted based on context

<i>Perturbation Sequence</i>	DiffSeq	::=	DiffUnit{1-6}
<i>Perturbation</i>	DiffUnit	::=	SNV Insertion Deletion
<i>SNV</i>	SNV	::=	<u>SNV</u> (Pos, Nuc)
<i>Insertion</i>	Insertion	::=	<u>Ins</u> (Pos, Nucs)
<i>Deletion</i>	Deletion	::=	<u>Del</u> (Pos, Size)
<i>Multiple Nucleotides</i>	Nucs	::=	Nuc+
<i>Nucleotide</i>	Nuc	::=	<u>A</u> <u>C</u> <u>G</u> <u>T</u>
<i>Position</i>	Pos	::=	<u>int</u>
<i>Size of deletion</i>	Size	::=	<u>int</u>

Original Sequence	Perturbation	Final Sequence
ATTCGCGTTA	[SNV(1,A)]	AATCGCGTTA
	[Ins(2,CG)]*	ATCGTCGCGTTA
	[Del(7,2)]*	ATTCGCGA
	[SNV(1,A), Ins(2,CG)]	AACGTCGCGTTA

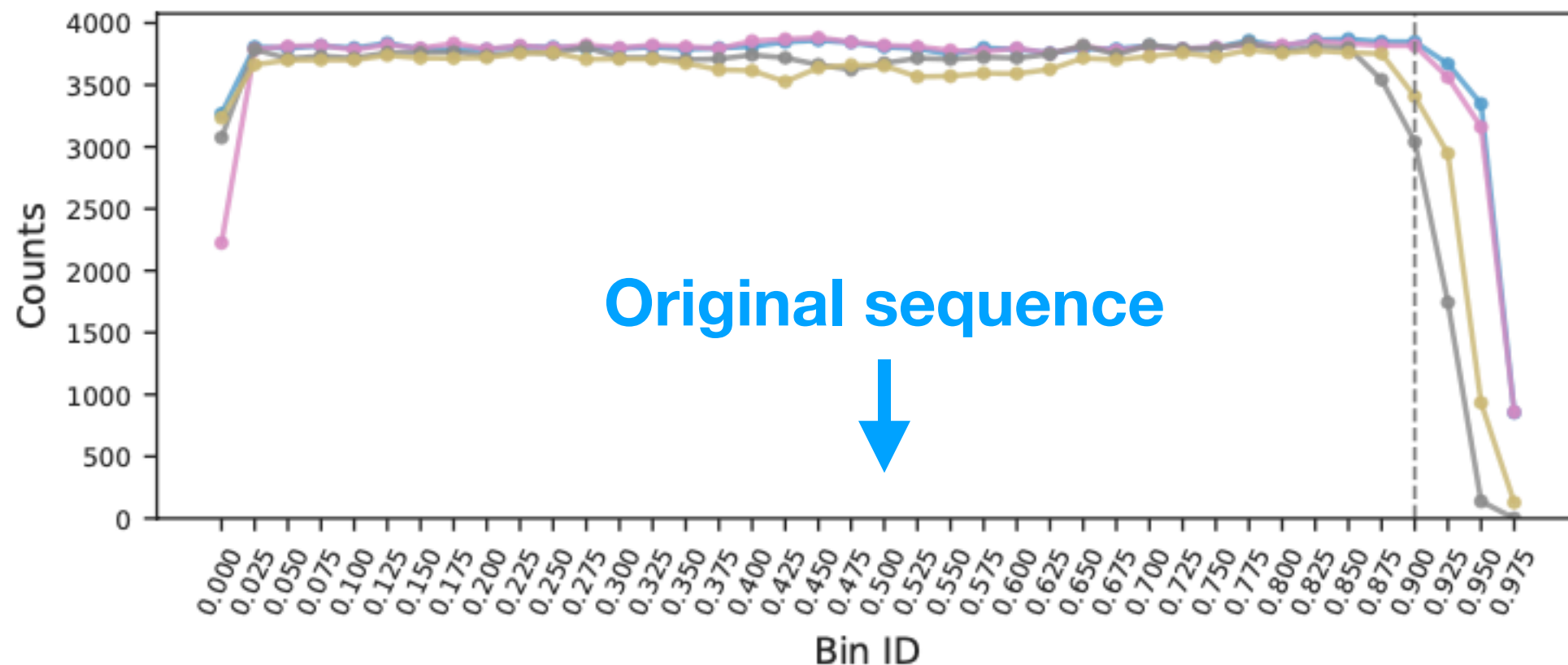
Custom Fitness Function

Bin Filler

Number of empty slots in the current bin

Increased Archive Diversity (IAD)

Increase in Inter-bin diversity



GP with Archive

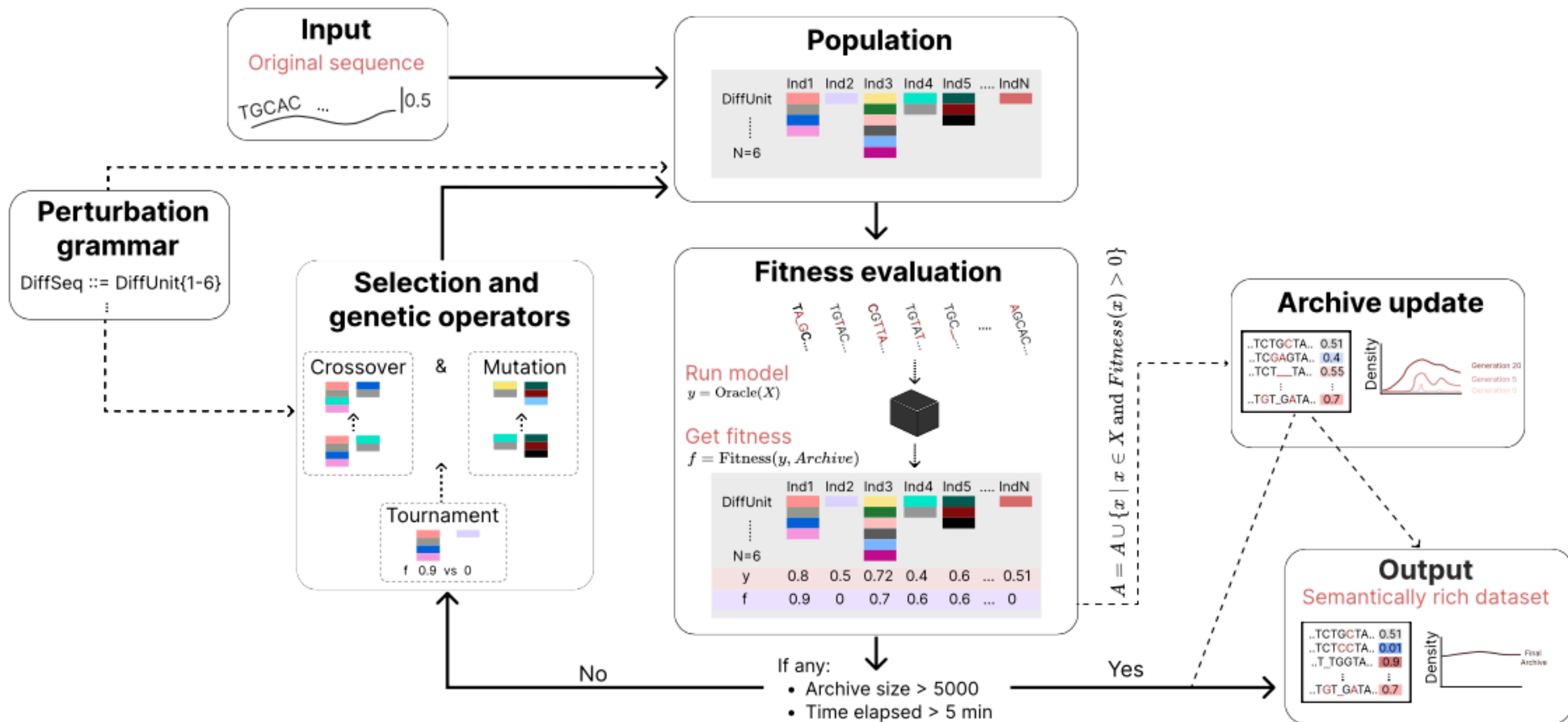
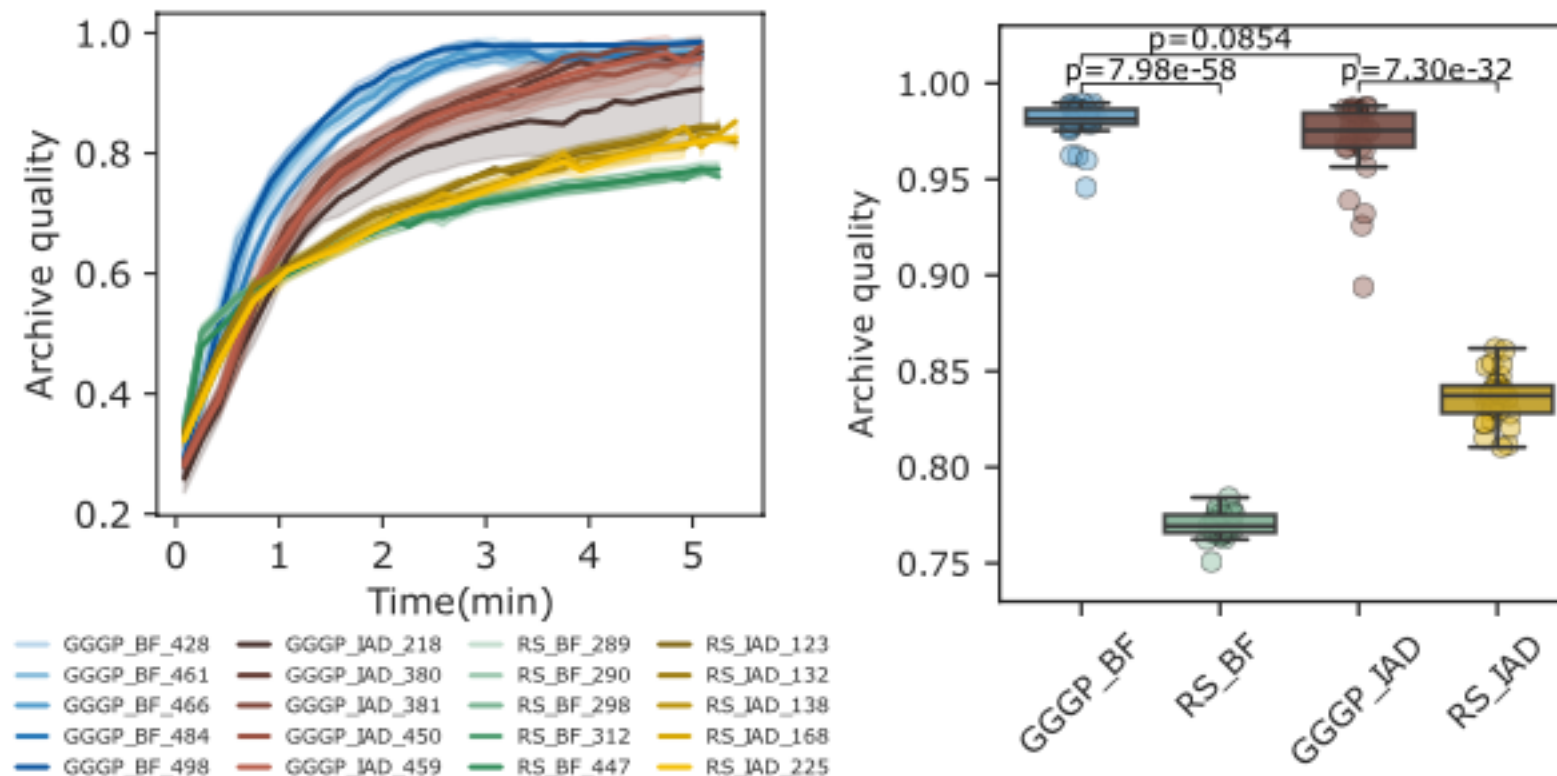


Figure 1: Summary of the proposed methodology.

Results

GP is better than Random Search

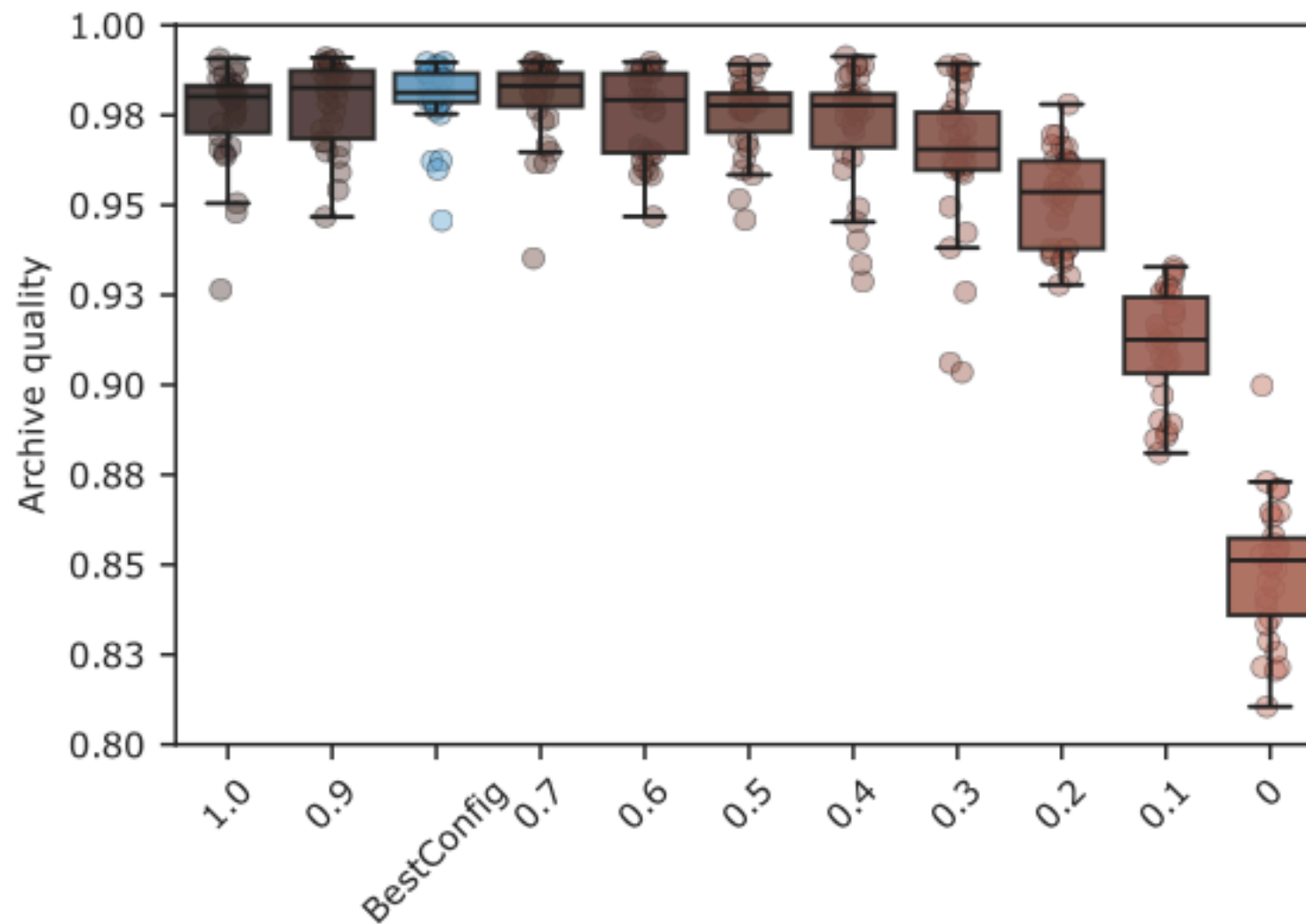
Archive Quality = 30% archive size + 30% inter-bin diversity + 20% intra-bin diversity + 20% <10 sequences bin count



- Enumerative approaches do not scale to this neighbourhood
- Random search does not take into account biology
- SpliceAI (and Splicing) are more sensitive to biological-inspired mutations

Results

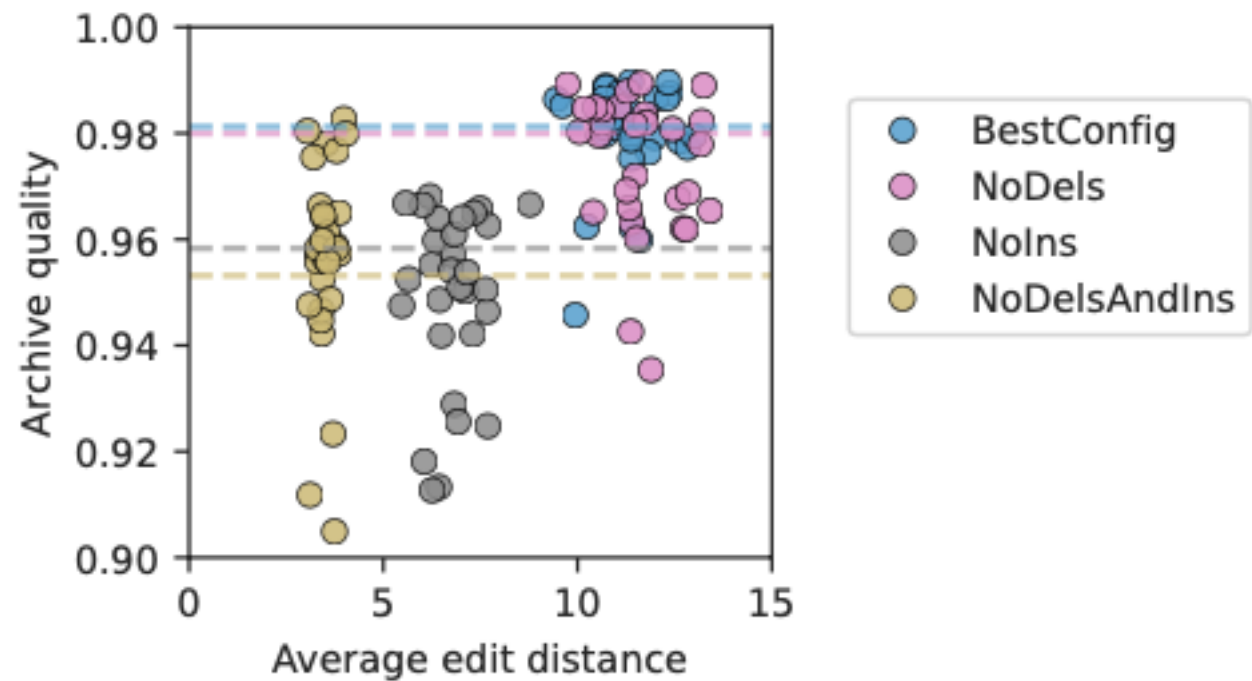
Locality is important



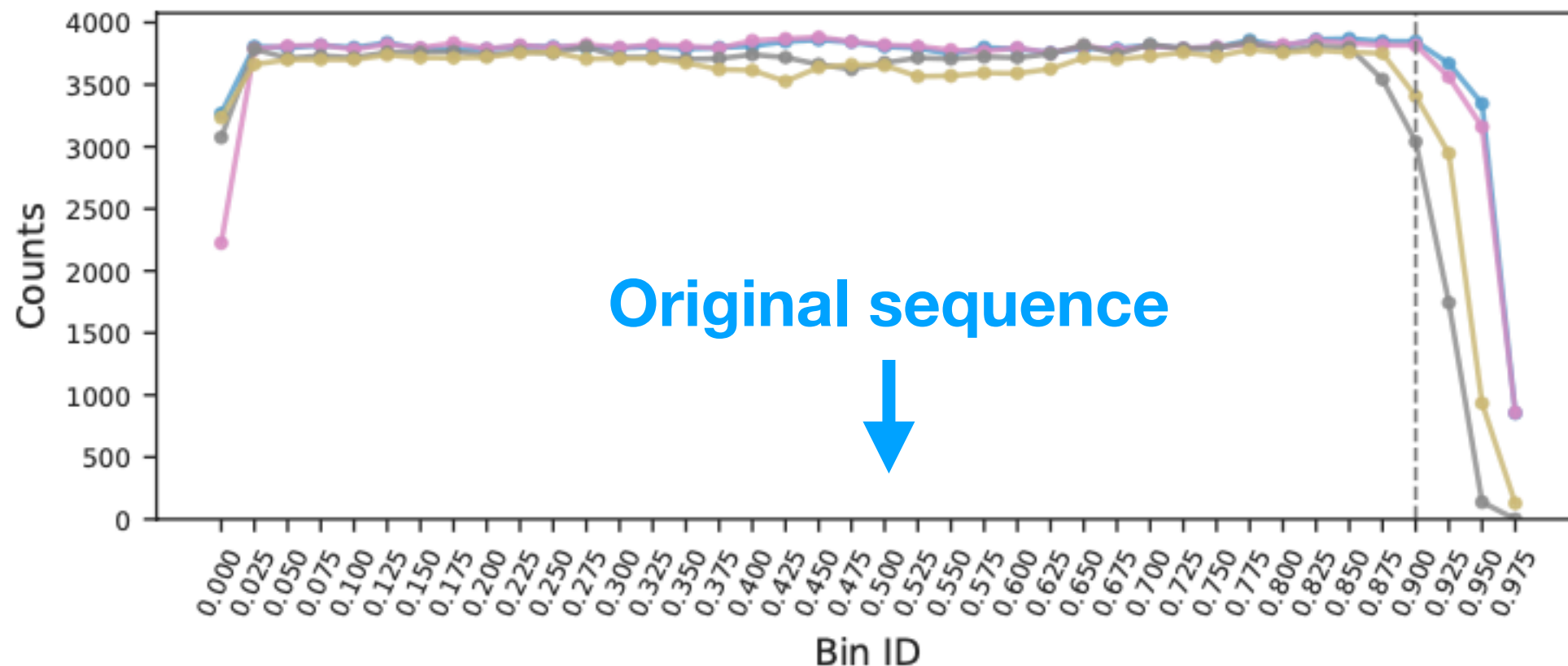
- A custom mutation operator keeps all mutations in the same sequence position + $\mathcal{N}(0,4)$
- High usage (80%) of the custom mutation operator creates better datasets than just using the syntax-oriented GGGP mutation operator.

Results

Insertions are important

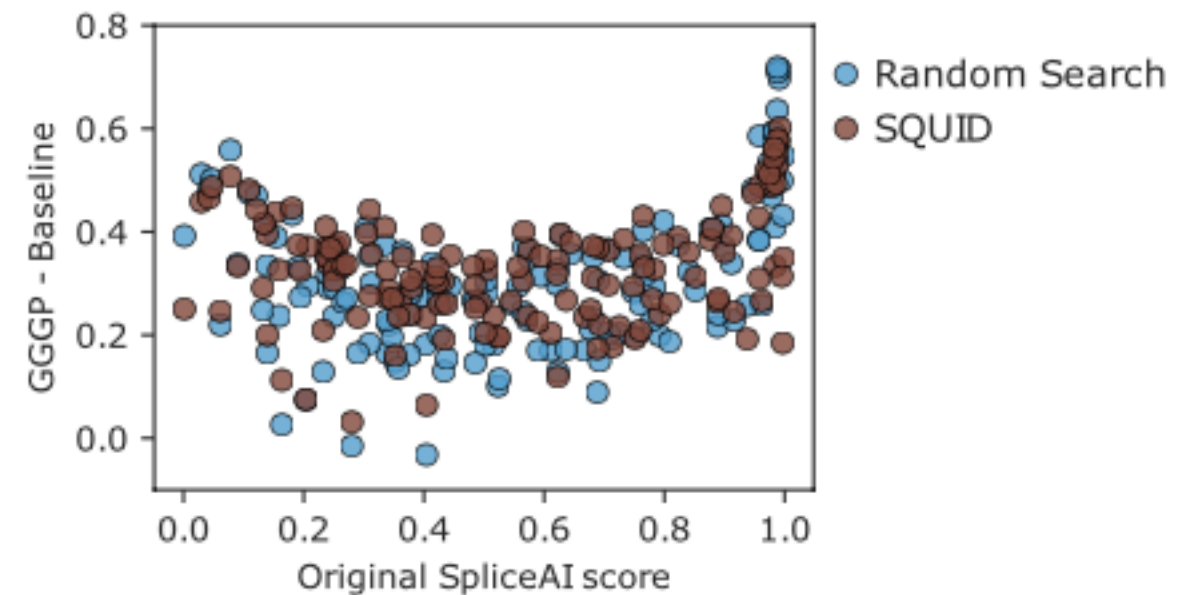
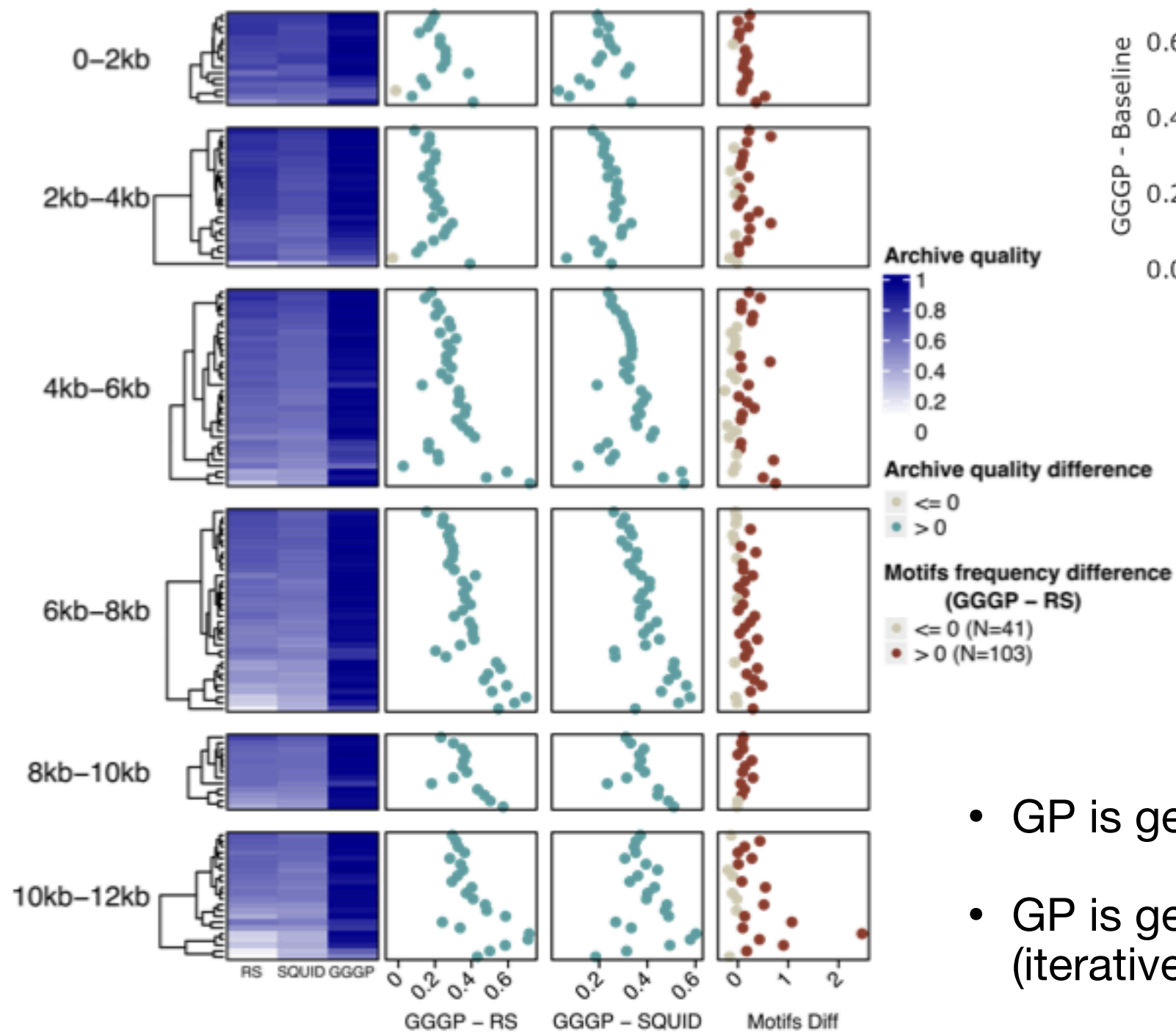


- Insertions increase the average archive quality at the cost of reducing the syntactic similarity, especially in the most difficult bins



Results

Performance scales to other sequences



- GP is generally better than RS
- GP is generally better than SQUID (iterative random perturbations)
- GP finds more known motifs than RS

Conclusion

- Grammar-guided GP can be used to generate datasets for local explainability
- Domain knowledge is needed, and can be encoded in the grammar, genetic operators, and fitness function
- Aligning GP operators with biological operators helps in exploring the rugged fitness landscape of genomics.

ACKNOWLEDGMENTS

This work was supported by FCT through a fellowship to P.B., ref. SFRH/BD/137062/2018, project RAP, ref.EXPL/CCI-COM/1306/2021 (<https://doi.org/10.54499/EXPL/CCI-COM/1306/2021>), project HPC, ref. 2022.15800.CPCA.A1, and the LASIGE Research Unit, ref. UIDB/00408/2020 (<https://doi.org/10.54499/UIDB/00408/2020>) and ref. UIDP/00408/2020 (<https://doi.org/10.54499/UIDP/00408/2020>).