Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AminoGrapes: Structure-Weighted Evolutionary Scoring for Viral Protein Fitness Prediction

MUHAMMAD Saad

memaliksaad17@gmail.com

Afrium https://orcid.org/0009-0002-0910-0684

Method Article

Keywords: protein fitness prediction, deep mutational scanning, viral proteins, ProteinGym, multiple sequence alignment, ESM2

Posted Date: October 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11236060/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: Muhammad Saad is the founder of Afrium, where the AminoGrapes model family is being developed. The authors declare no other financial or non-financial competing interests.

AminoGrapes: Structure-Weighted Evolutionary Scoring for Viral Protein Fitness Prediction

Muhammad Saad Khan
Afrium, Islamabad, Pakistan | October 2026

Abstract. Zero-shot prediction of mutation effects on viral proteins is hard because viral families evolve fast and are sparsely sampled, and protein language models are weaker on them. We evaluate AminoGrapes, an independent-site model from Afrium that scales a sequence-weighted log-odds conservation score from a homolog alignment by residue burial computed from a predicted structure. The formulation follows RSALOR [7]. The AminoGrapes score has no fitted parameters. On the 31 viral assays of ProteinGym v1.3, AminoGrapes reaches a mean Spearman correlation of 0.430, compared with 0.269 for Meta's ESM2-650M (+0.162, 95% CI 0.078 to 0.245), and is better on 25 of 31 assays. A blend with ESM2-650M, with the blend weight chosen only on non-viral assays, reaches 0.376 (+0.108, 95% CI 0.078 to 0.140) and is better on 27 of 31. AminoGrapes does not match the original RSALOR scores (0.480) or the strongest published methods on these assays, it adds nothing to an ensemble of the best public models, and its design was settled while viral results were visible (Section 4.2). This is an evaluation on a public benchmark, not an official leaderboard entry.

Keywords: protein fitness prediction, deep mutational scanning, viral proteins, ProteinGym, multiple sequence alignment, ESM2

1. Introduction

Deep mutational scanning (DMS) measures the functional effect of thousands of variants of one protein in a single experiment. ProteinGym [1] collects these assays into a standard benchmark for zero-shot fitness prediction and ranks methods by the Spearman correlation between predicted scores and measured fitness. Viral proteins are an important subset: knowing which mutations preserve function matters for vaccine design and for anticipating immune escape. They are also a hard subset, since homologs are often few and divergent, and protein language models such as ESM2 [2] tend to be weaker on them. In the ProteinGym score files, ESM2-650M has a plain mean Spearman of 0.269 on the 31 viral assays and 0.466 on the 186 non-viral assays (different assays, so this is indicative only).

Two ingredients are cheap, interpretable and complementary. The first is evolutionary conservation read from a multiple sequence alignment (MSA), which captures what the protein family tolerates. The second is burial: a substitution at a buried position is more likely to disrupt folding or packing than one at an exposed position. AminoGrapes multiplies the two in a single closed-form score. This is the RSALOR formulation of Tsishyn et al. [7], who showed that conservation scaled by solvent accessibility is competitive with far more complex predictors. AminoGrapes is an independent implementation of that idea and makes no claim to a new scoring formula. It is the first model from Afrium in the AminoGrapes model family; Afrium plans further models in the family for other protein families.

This report (i) specifies the score and every constant it uses; (ii) evaluates it on the 31 viral assays of ProteinGym v1.3 against ESM2-650M, a family of ESM2 sizes, and published methods for context; and (iii) states what the results do and do not show, including which design decisions were made after viral results had been seen and a failed attempt to add AminoGrapes to an ensemble of public models.

2. Method

Figure 1 gives the overview. Sections 2.1 to 2.4 describe the AminoGrapes score and Section 2.5 the optional ESM2 blend.

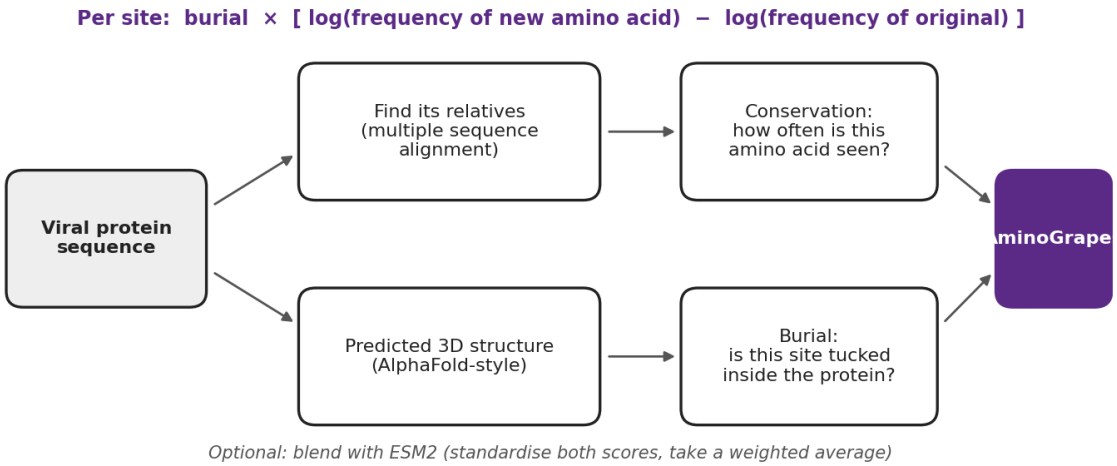


Figure 1. Overview of AminoGrapes. A viral protein sequence is turned into a homolog alignment, giving per-site conservation, and into a predicted structure, giving per-site burial. The two are multiplied for each substituted site and summed over the mutation. An optional blend with ESM2 standardises both scores and takes a weighted average.

2.1 Sequence weighting

MSAs of related viruses are heavily redundant because a few outbreak lineages are sequenced thousands of times. Without correction, such a clade dominates every site frequency. Each sequence k receives a weight equal to the inverse of the number of sequences within identity distance θ of it (itself included), the standard clustering scheme. Gaps are treated as a 21st state during this step. The sum of weights is the effective sequence count N_{eff}.

$$w_k = |\{\, l : \mathrm{id}(x_k, x_l) \geq 1 - \theta \,\}|^{-1}, \quad \theta = 0.2 \qquad (1)$$

When ProteinGym provides a precomputed weight file of matching length, it is used directly; this was the case for 16 of the 31 viral assays. Otherwise (15 assays) the weights are computed by the routine above, on GPU in half precision when available. Sequences with non-standard letters or more than 50% gaps are removed first, and insertion columns are dropped so the alignment is indexed by the query.

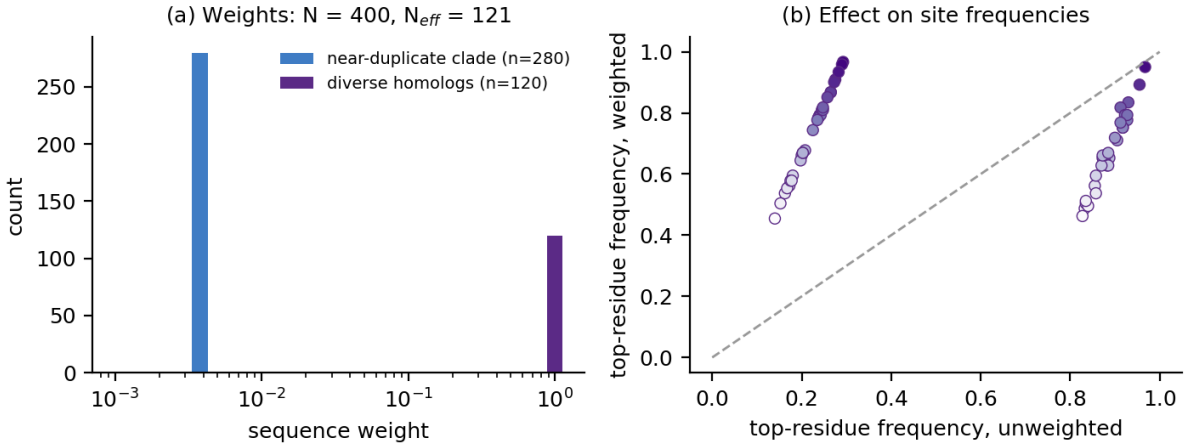


Figure 2. Effect of weighting on a synthetic MSA (illustrative, not a real viral family). (a) A 280-sequence near-duplicate clade receives very small weights, reducing 400 sequences to an effective count of about 121. (b) Sites where the clade disagrees with the wider family move off the diagonal, which is where unweighted frequencies would mislead.

2.2 Site frequencies

Weighted counts are normalised per site and smoothed with a fixed pseudocount ε so that no amino acid has zero probability, where $C_i(a) = \Sigma_k\, w_k\, 1[x_{ki} = a]$:

$$f_i(a) = (1 - \varepsilon) \cdot C_i(a) / \Sigma_b\, C_i(b) + \varepsilon/20, \quad \varepsilon = 0.01 \qquad (2)$$

2.3 Structural burial

Relative solvent accessibility (RSA) is computed from the AlphaFold2 structure [4] that ProteinGym provides for each target, using the Shrake-Rupley algorithm [5] at residue level (as implemented in Biopython [10]), divided by the theoretical maximum accessible surface area of that residue type from Tien et al. [6] and capped at 1.

$$RSA_i = \min(1, SASA_i / MaxASA(a_i)) \qquad (3)$$

The structure sequence is aligned to the target with a global BLOSUM62 aligner (gap open −10, extend −0.5, free end gaps). A structure is used only if at least 30% of target positions map and at least 80% of mapped residues are identical. Unmapped positions take the mean RSA. If no structure passes, every site receives the same constant weight, which leaves the ranking of mutations unchanged and reduces AminoGrapes to the plain alignment (conservation) score for that assay. This happened for 2 of the 31 viral assays (POLG_HCVJF_Qi_2014 and A0A140D2T1_ZIKV_Sourisseau_2019), where the best-matching structure covered 8% and 15% of the target positions.

2.4 The AminoGrapes score

For a mutation *m* consisting of one or more substitutions, the score is the sum over substituted sites of the burial weight times the log-odds of the mutant against the wild-type residue. Higher means more likely to be tolerated. Multi-mutants are scored additively, so epistasis is not modelled. Positions that cannot be mapped between the DMS target and the MSA query contribute zero. This is the RSALOR form [7]; details such as the weighting scheme, the pseudocount and the structure handling may differ from the RSALOR package.

$$s_{AminoGrapes}(m) = \Sigma_{i \in m}\, (1 - RSA_i)\, [\, \log f_i(a_i^{mut}) - \log f_i(a_i^{wt}) \,] \qquad (4)$$

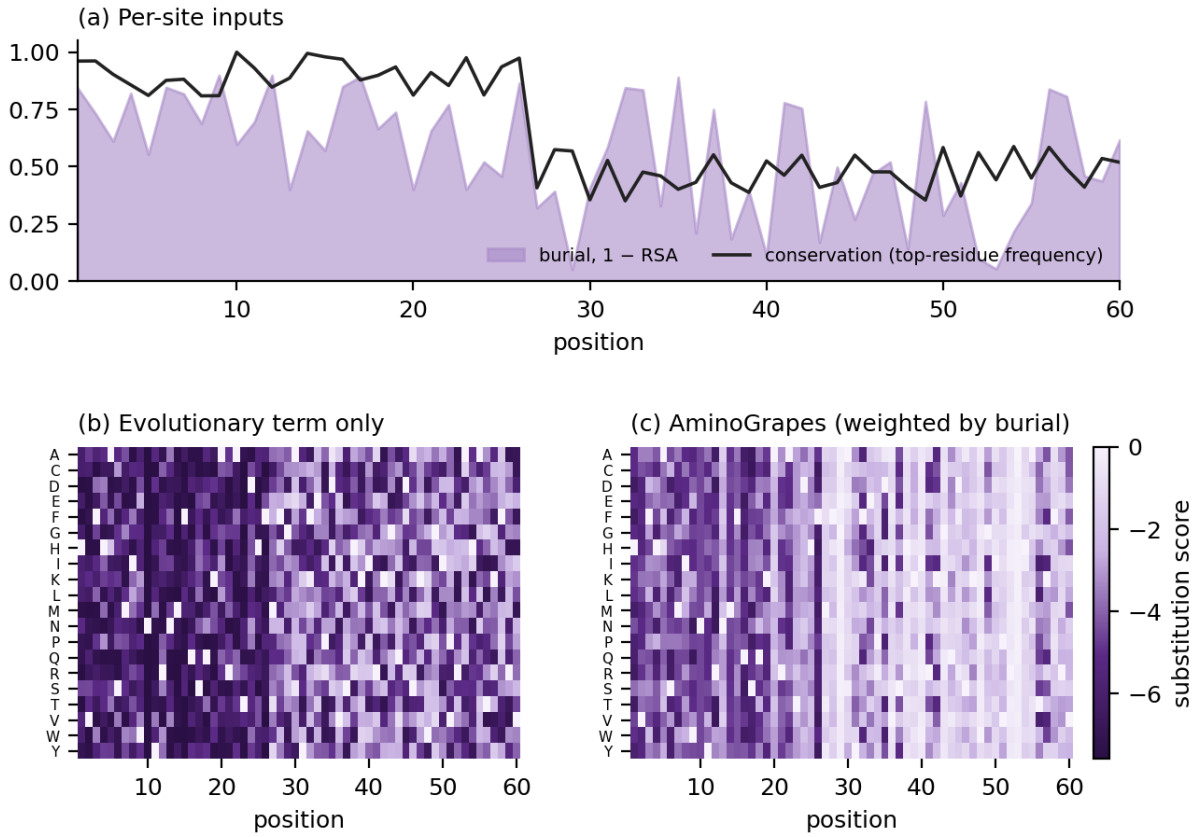


Figure 3. Per-substitution scores on a synthetic 60-residue protein (illustrative). (a) Conservation and burial profiles. (b) Evolutionary log-odds alone. (c) The same scores after multiplying by burial: substitutions at exposed sites are shrunk toward zero while substitutions at buried, conserved sites keep their full penalty. Rows are the mutant amino acid; darker means a larger predicted penalty.

2.5 ESM2 masked-marginal baseline and blend

As a language-model reference we score every substitution with ESM2-650M (esm2_t33_650M_UR50D) using the masked-marginal method [3]: the site is masked, and the score is the log-probability of the mutant minus that of the wild type. Proteins longer than 1,022 residues are scored on a window centred on the site. Because AminoGrapes and ESM2 scores live on different scales, they are z-scored per assay before mixing; $\lambda = 0$ is AminoGrapes alone and $\lambda = 1$ is ESM2 alone.

$$s_{\mathrm{ESM2}}(m) = \Sigma_{i \in m} \left[\log p_i(a_i^{\mathrm{mut}} \mid x_{\setminus i}) - \log p_i(a_i^{\mathrm{wt}} \mid x_{\setminus i}) \right] \qquad (5)$$

$$s_{\mathrm{blend}}(m) = \lambda\, z(s_{\mathrm{ESM2}}) + (1 - \lambda)\, z(s_{\mathrm{AminoGrapes}}), \quad \lambda = 0.7 \qquad (6)$$

3. Implementation

AminoGrapes is a single Python file (afrium.py, kept under that name for compatibility with the accompanying notebook) depending on NumPy, SciPy, pandas and Biopython, and optionally PyTorch and Hugging Face Transformers for the weighting speed-up and ESM2. Every constant is fixed in code and listed in Table 1. The AminoGrapes score itself has no fitted parameters; only the optional blend weight was chosen, and it was chosen on non-viral assays.

Constant	Value	Role
θ (clustering)	0.2	Sequences more than 80% identical share weight
ε (pseudocount)	0.01	Smooths site frequencies
Gap filter	50%	Drops sequences with more than half gaps

Constant	Value	Role
Structure mapping	≥ 30% mapped, ≥ 80% identical	Condition for using a structure
No-structure fallback	constant 0.5	Reduces the score to plain conservation (ranking unchanged)
MaxASA table	Tien et al. theoretical, 20 values	Normalises SASA to RSA
ESM2 window	1,022 residues	Maximum input length
Blend weight λ	0.7 (grid 0 to 1, steps of 0.1)	Chosen on 30 non-viral assays, then frozen (Section 4.1)

Table 1. Fixed constants in the AminoGrapes implementation.

4. Evaluation protocol

4.1 Data, metric and statistics

Data are ProteinGym v1.3 [1]: the reference file, substitution assay files, MSAs, precomputed MSA weights, AlphaFold2 structures and the official zero-shot score files. Assays are filtered to taxon Virus, giving 31 assays. For each assay we compute the Spearman correlation between each score and the DMS fitness and report the plain unweighted mean over assays. This is not the aggregate used on the official ProteinGym leaderboard, which averages within groups first and covers all 217 assays, so numbers here should not be mixed with leaderboard numbers. Scores for all other models are taken from the official ProteinGym score files and evaluated the same way. Our own ESM2-650M rerun reproduces the official scores (per-assay agreement r = 1.000; means 0.2693 here against 0.2687 official, with small differences on individual assays such as POLG_HCVJF_Qi_2014, 0.131 against 0.116, a long protein scored with windowing); all comparisons use the official ESM2-650M scores.

Differences are given as the mean of per-assay differences with a 95% percentile bootstrap interval (10,000 resamples of assays, fixed seed), the number of assays won, and a Wilcoxon signed-rank p-value. Assays are treated as independent, but several share a protein (two influenza assays on the same protein, two lambda repressor assays, two SARS-CoV-2 spike assays), so the intervals are somewhat optimistic.

The blend weight λ was chosen on 30 non-viral assays and never on viral data. These 30 were sampled at random (fixed seed) from the non-viral assays with at most 400 residues and at most 6,000 variants, not from all 186. Experiments ran on a single Colab T4 GPU. The full command and a ready-to-run Colab notebook are described in Appendix C.

4.2 Development history: what was seen, and when

AminoGrapes was developed with the viral results in view. We list the decisions made after seeing viral scores so that readers can judge the risk of fitting this benchmark. (i) **Weights and position mapping.** The first implementation fell back to uniform sequence weights where the weight file did not match and read the alignment offset from the reference file; its alignment-only mean was 0.149. Correcting the weights and using a best single offset gave 0.330. Under that offset, nine assays still showed wild-type match rates between 10% and 85%, which flags a position-mapping error without reference to any fitness score. Replacing the offset with a global alignment (Section 2.3) fixed these (identity at aligned positions became 1.00 for all nine) and gave 0.393. (ii) **Burial.** The burial term was added after the conservation-only score was known, to follow RSALOR; it was run once, in one form, with no variants. (iii) **Blend weight.** It was selected on non-viral assays because a weight selected on viral assays would be optimistic. No numeric parameter was fitted to viral data and no constant was changed after the final burial run. Even so, because the pipeline was built and debugged on these assays, the numbers below should be read as development-set results until they are replicated on new data. Appendix A lists every stage with its result.

5. Results

5.1 Main comparison

AminoGrapes reaches a mean Spearman correlation of 0.430 over the 31 viral assays against 0.269 for ESM2-650M: a gain of +0.162 (95% CI 0.078 to 0.245, Wilcoxon p = 0.0008), with AminoGrapes ahead on 25 of 31 assays. Three of those 25 wins are by less than 0.015 (A0A2Z5U3Z0_9INFA_Wu_2014, RDRP_I33A0_Li_2023 and POLG_CXB3N_Mattenberger_2021). The blend with ESM2 at λ = 0.7 scores 0.376, a gain of +0.108 over ESM2-650M (95% CI 0.078 to 0.140), ahead on 27 of 31. The blend scores lower than AminoGrapes alone on viral assays because its weight was set on non-viral proteins, where ESM2 is stronger (Section 5.5).

Method	Mean Spearman	Gain vs ESM2-650M (95% CI)	Assays won
ESM2-650M (official)	0.269	–	–
AminoGrapes, conservation only (no burial)	0.393	+0.124 (0.029 to 0.216)	21 / 31
AminoGrapes	**0.430**	**+0.162 (0.078 to 0.245)**	**25 / 31**
AminoGrapes + ESM2 blend, λ = 0.7	0.376	+0.108 (0.078 to 0.140)	27 / 31

Table 2. Main results on the 31 viral assays of ProteinGym v1.3. Means are plain means of per-assay Spearman correlations.

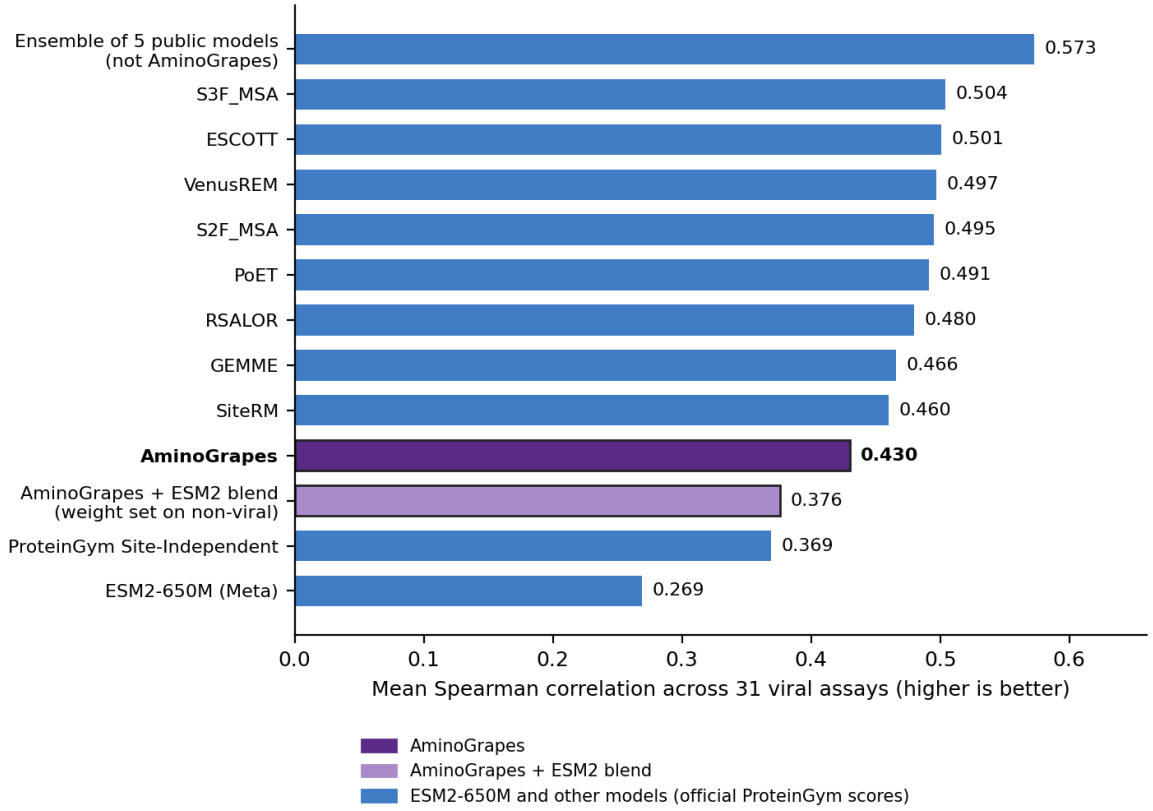


Figure 4. Mean Spearman correlation across the 31 viral assays. AminoGrapes is shown in grape purple (the blend in a lighter purple); ESM2-650M and all other models are shown in blue and are official ProteinGym scores, plain means, not rerun here. This is a selection, not a complete ranking: about 18 of the 95 official models score above AminoGrapes on these assays. The ensemble bar is a combination of other groups' models and is not AminoGrapes (Section 5.6).

Method (as named in ProteinGym)	Mean Spearman, 31 viral assays
S3F_MSA	0.504
ESCOTT	0.501
VenusREM	0.497
S2F_MSA	0.495

Method (as named in ProteinGym)	Mean Spearman, 31 viral assays
PoET	0.491
RSALOR	0.480
GEMME	0.466
SiteRM	0.460
AminoGrapes	**0.430**
ProteinGym Site-Independent	0.369
ESM2-650M	0.269

Table 3. Published methods that score above or near AminoGrapes on the 31 viral assays, from ProteinGym's official score files (plain means, not rerun here). See the ProteinGym repository for the references for each method.

5.2 Comparison with larger language models

ESM2 improves steadily from 8M to 15B parameters, reaching 0.336 at 15B, still below AminoGrapes at 0.430 and below the blend at 0.376. AminoGrapes has access to homolog sequences and a predicted structure, which ESM2 does not see, so the comparison is between different information sources as well as different models. These are comparisons of means; no significance test was run against the larger ESM2 models.

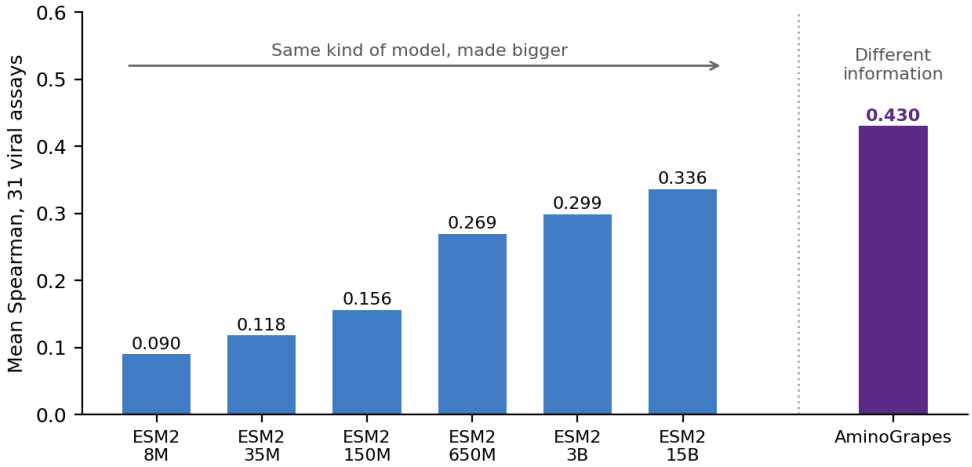


Figure 5. Official ESM2 scores at six sizes (blue) versus AminoGrapes (purple), mean Spearman on the 31 viral assays.

5.3 Where the gain comes from

Starting from ESM2-650M, conservation from the alignment accounts for +0.124 (95% CI 0.029 to 0.216; ahead on 21 of 31; p = 0.015) and weighting by burial adds a further +0.038 (95% CI 0.009 to 0.064; ahead on 21 of 31; p = 0.016), ending at 0.430. The split depends on the order in which components are added and is a rough attribution. Burial does not help everywhere: it lowered the score on four influenza assays (PA_I34A1_Wu_2015, A4D664_9INFA_Soh_2019, I6TAH8_I68A0_Doud_2015 and NCAP_I34A1_Doud_2015) by 0.09 to 0.12 each, and the official RSALOR scores are also low on those assays. For reference, the conservation-only score is 0.024 above ProteinGym's own site-independent model (95% CI 0.005 to 0.045; ahead on 19 of 31; p = 0.066).

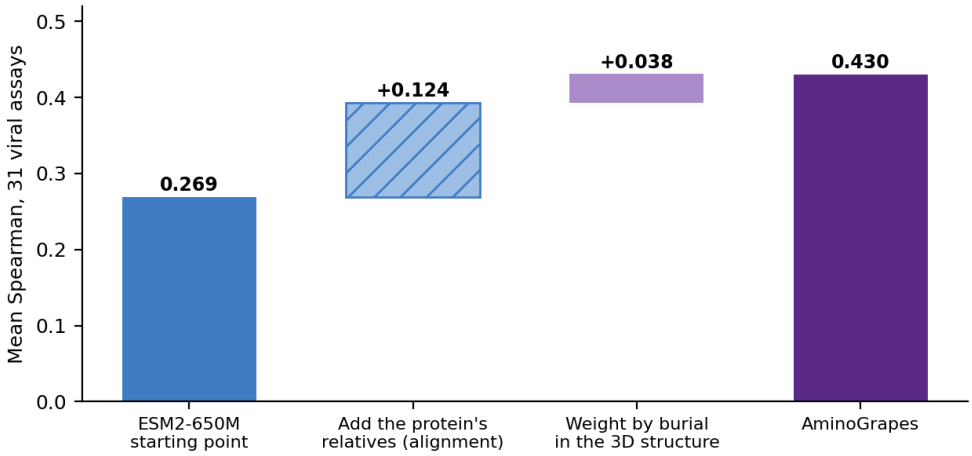


Figure 6. Step-wise gain from ESM2-650M to AminoGrapes on the 31 viral assays.

5.4 Per-assay behaviour

AminoGrapes is better on 25 assays and ESM2-650M on 6: HCP_LAMBD, RCRO_LAMBD, both RPC1_LAMBD assays, VG08_BPP22 and VRPI_BPT7. All six are bacteriophage proteins (lambda, P22 and T7), although not every phage assay favours ESM2 (RPC1_BP434 does not). We have not tested why. The largest gains are on R1AB_SARS2_Flynn_2022 (0.105 to 0.641), POLG_HCVJF_Qi_2014 (0.116 to 0.661), Q2N0S5_9HIV1_Haddox_2018 (0.044 to 0.525) and POLG_DEN26_Suphatrakul_2023 (0.153 to 0.588). A full per-assay table is in Appendix B. With 31 assays, a handful can move the mean, so the confidence intervals matter more than the mean alone.

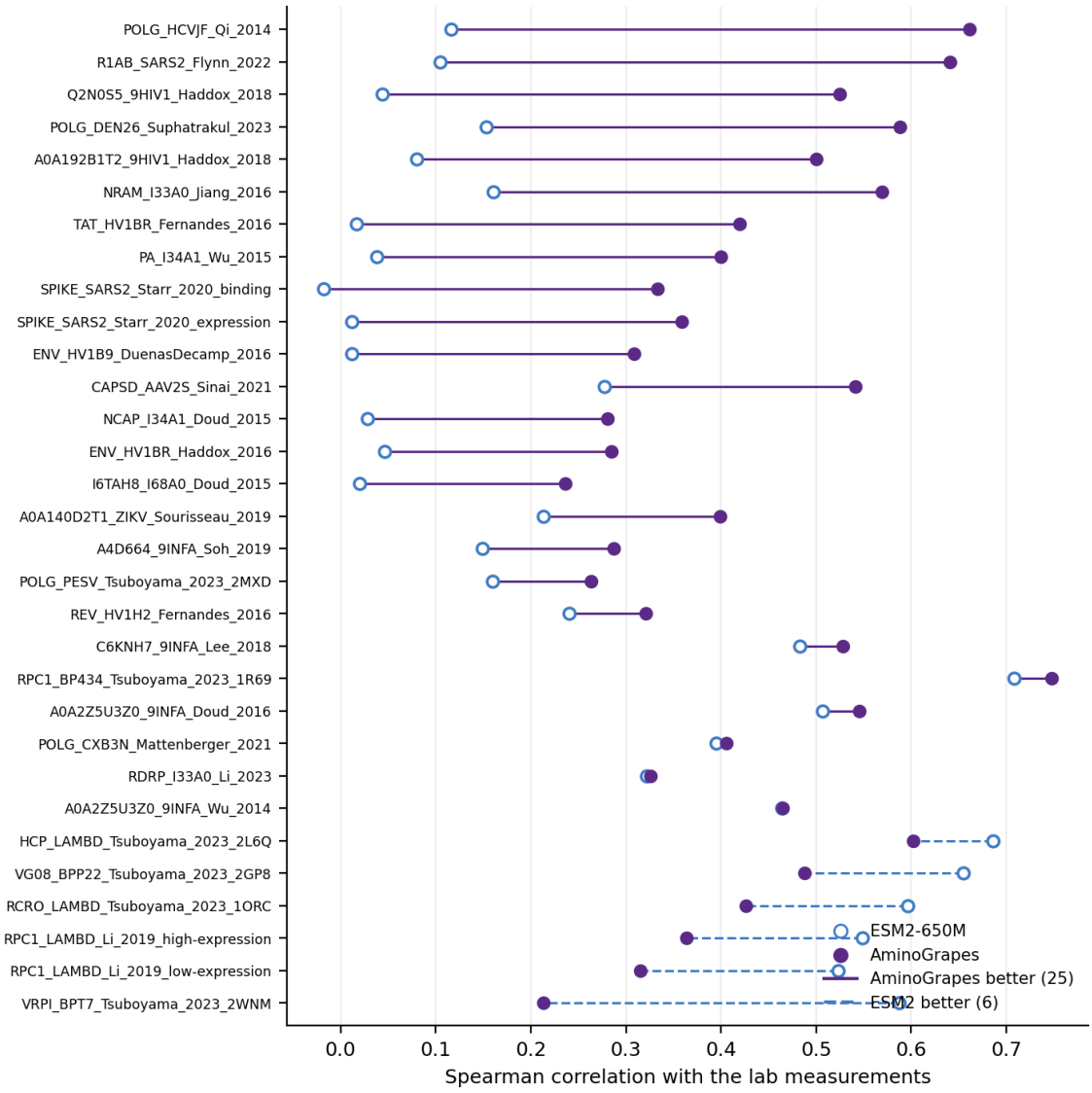


Figure 7. Per-assay Spearman correlation for ESM2-650M (open blue circles) and AminoGrapes (filled purple circles), sorted by the difference. Solid lines mark the 25 assays where AminoGrapes is better, dashed lines the 6 where ESM2-650M is better.

5.5 Blend weight

The best blend weight differs between protein groups. On the 30 non-viral tuning assays ESM2 alone (0.540) is stronger than AminoGrapes alone (0.449), and the blend peaks near λ = 0.7 (0.555). On the 31 viral assays it peaks near λ = 0.3 (about 0.445), slightly above AminoGrapes alone, and falls steadily as the ESM2 weight grows. To avoid tuning on the evaluation set, the headline blend uses λ = 0.7, chosen on non-viral assays. The viral optimum is shown for transparency and is not reported as a result. As a sensitivity check, choosing the weight on the other viral proteins (leaving out one protein at a time) gives 0.441. AminoGrapes is therefore a viral-protein result and not a general replacement for ESM2.

Mean Spearman	λ = 0 (AminoGrapes only)	λ = 0.3	λ = 0.7	λ = 1 (ESM2 only)
31 viral assays	0.430	about 0.445 (peak)	0.376	0.269
30 non-viral tuning assays	0.449	–	0.555 (peak)	0.540

Table 4. Blend weight: values reported in the text; intermediate grid points are omitted. λ = 0.7 was chosen on the non-viral set.

5.6 Relation to published methods

RSALOR. AminoGrapes is not competitive with the best published methods on these assays (Table 3). RSALOR is the same formula and scores 0.480 in the official ProteinGym files; AminoGrapes is 0.049 lower (95% CI −0.083 to −0.019) and is ahead on only 7 of 31 assays. The largest shortfalls are POLG_PESV_Tsuboyama_2023_2MXD (0.263 against 0.600), RCRO_LAMBD (0.426 against 0.644), VRPI_BPT7 (0.213 against 0.453), both SPIKE_SARS2_Starr_2020 assays (0.333 against 0.482 and 0.359 against 0.553) and both RPC1_LAMBD assays (0.364 against 0.476 and 0.315 against 0.487). We have not found the cause; possible sources are differences in alignment weighting, structure choice or accessibility calculation, and they remain untested. We therefore do not claim parity with RSALOR.

An ensemble of public models. As a context check we built an ensemble from the official ProteinGym score files of other groups' models. Each model's scores were standardised per assay and summed. Models were added greedily using 186 non-viral assays only (candidate pool of 95 models, non-viral assays subsampled to 3,000 mutants), and the result was frozen before testing on viral assays. The selected models are VenusREM, ProSST-4096, S3F_MSA, ProSST-1024 and RSALOR. On the 31 viral assays this ensemble scores 0.573. It is ahead of ESM2-650M (+0.305, 95% CI 0.230 to 0.382, 31 of 31 assays), RSALOR (+0.094, 27 of 31), PoET (+0.083, 95% CI 0.036 to 0.131, 22 of 31) and the best single official model S3F_MSA at 0.504 (+0.069, 25 of 31). This is an ensemble of other groups' models and is not AminoGrapes. We then added the AminoGrapes score to the ensemble with the weight chosen on the 30 non-viral assays. The chosen weight was zero, and on viral assays the score fell as AminoGrapes' weight rose (0.573 at weight 0, 0.570 at weight 1, 0.550 at weight 3). AminoGrapes therefore does not add information beyond what the strongest public models already provide. Its value is as a simple, cheap and interpretable model that beats ESM2 on viral proteins. (An earlier three-model ensemble built from a smaller candidate list of 8 models scored 0.540 and is superseded.)

5.7 What these results do and do not show

They show that, on these 31 viral assays, a conservation-times-burial model has a higher mean than ESM2-650M, a statistically clear difference, and a higher mean than ESM2 at every size tested up to 15B (means only). They do not show that AminoGrapes is competitive with the best published methods (GEMME 0.466, RSALOR 0.480, PoET 0.491 and several newer models score higher on the same assays, and an ensemble of public models reaches 0.573), that it is a new scoring formula, or anything about the other 186 assays in ProteinGym v1.3. AminoGrapes is not on the ProteinGym leaderboard and cannot be looked up there; submission requires predictions for all 217 assays, which is outside this report.

6. Limitations

AminoGrapes treats sites independently, so it cannot capture epistasis, and its multi-mutant scores are sums of single-site terms. It depends on MSA depth: for viruses with few homologs the frequencies are dominated by the pseudocount. For 15 assays the sequence weights were recomputed rather than taken from ProteinGym's files, so weighting differs from the official pipeline there; in half precision, identity counts above 2,048 are not exactly representable, so the clustering threshold is slightly approximate for the longest proteins. Burial depends on the quality of the predicted structure and on whether the assayed molecule is a monomer, so burial computed on one chain can mislabel interface residues of complexes and oligomers as exposed. Two assays had no usable structure. Fitness in a DMS is defined by the assay, which for some viral proteins measures replication or binding rather than stability, and conservation need not track either. The comparison with ESM2 mixes model and information source, since AminoGrapes sees an alignment and ESM2 sees only the query sequence. The blend weight was chosen on 30 small non-viral assays, not all 186. The 31-assay benchmark is small, several assays share a protein, and results come from a single evaluation on assays that were also used during development (Section 4.2), so the confidence intervals are optimistic and should not be read as out-of-sample. The gap to RSALOR is unexplained. Results are restricted to viral assays.

7. Reproducibility

All data are public ProteinGym v1.3 files, which are not redistributed here (check the ProteinGym repository for its license before reusing them). ESM2-650M weights are public. The model is the single file afrium.py and a Colab notebook that runs the whole comparison is provided with this report; Appendix C explains it step by step. Per-assay reference values are in Appendix B. The alignment, mapping and structure functions were unit-tested on synthetic data; the reference numbers in this report come from the interactive notebook runs, and a clean rerun of the notebook should reproduce them to about three decimals.

Appendix A. Analysis log

Stage	Mean Spearman (31 viral assays)	Note
ESM2-650M, our rerun	0.2693	Official: 0.2687. Per-assay r = 1.000.
First implementation, alignment only	0.149	Uniform-weight fallback, offset from reference file
Alignment only, corrected weights, best single offset	0.330	9 assays had low wild-type agreement
50/50 average with ESM2, offset mapping	0.3448	Superseded
Alignment only, BLOSUM62 mapping	0.3926	9 assays remapped (criterion: wild-type agreement)
50/50 average with ESM2, BLOSUM62 mapping	0.3796	Superseded
Alignment x burial (AminoGrapes)	0.4304	Structure used for 29 of 31 assays
AminoGrapes + ESM2 blend, λ = 0.7	0.3764	λ chosen on 30 non-viral assays. An earlier run without burial chose 0.8.
Ensemble of 5 public models	0.5732	Selected on 186 non-viral assays. Earlier 3-model version: 0.5399.
Ensemble + AminoGrapes	0.5732 at weight 0	Chosen weight 0. Weight 1: 0.5699. Weight 3: 0.550.

Appendix B. Per-assay Spearman correlations

Spearman correlation with the lab measurements for each of the 31 viral assays, transcribed to three decimals from the notebook outputs of the runs reported here. ESM2-650M and RSALOR values are from ProteinGym's official score files. AminoGrapes is alignment x burial. Structure: Y if a structure passed the mapping check, N if the constant-weight fallback was used. Per-assay values of the blend are not tabulated. Means are computed from the rounded values and match Table 2 to within 0.001.

Assay	ESM2-650M (official)	AminoGrapes, conservation only	AminoGrapes	RSALOR (official)	Structure
A0A140D2T1_ZIKV_Sourisseau_2019	0.213	0.399	0.399	0.423	N
A0A192B1T2_9HIV1_Haddox_2018	0.080	0.471	0.500	0.532	Y
A0A2Z5U3Z0_9INFA_Doud_2016	0.507	0.493	0.545	0.562	Y
A0A2Z5U3Z0_9INFA_Wu_2014	0.464	0.452	0.465	0.466	Y
A4D664_9INFA_Soh_2019	0.149	0.403	0.287	0.322	Y
C6KNH7_9INFA_Lee_2018	0.483	0.417	0.528	0.533	Y
CAPSD_AAV2S_Sinai_2021	0.277	0.447	0.541	0.555	Y
ENV_HV1B9_DuenasDecamp_2016	0.012	0.319	0.309	0.276	Y
ENV_HV1BR_Haddox_2016	0.046	0.307	0.285	0.281	Y

Assay	ESM2-650M (official)	AminoGrapes, conservation only	AminoGrapes	RSALOR (official)	Structure
HCP_LAMBD_Tsuboyama_2023_2L6Q	0.686	0.473	0.602	0.646	Y
I6TAH8_I68A0_Doud_2015	0.020	0.355	0.236	0.248	Y
NCAP_I34A1_Doud_2015	0.028	0.373	0.281	0.255	Y
NRAM_I33A0_Jiang_2016	0.161	0.498	0.569	0.536	Y
PA_I34A1_Wu_2015	0.038	0.522	0.400	0.347	Y
POLG_CXB3N_Mattenberger_2021	0.395	0.416	0.406	0.425	Y
POLG_DEN26_Suphatrakul_2023	0.153	0.507	0.588	0.610	Y
POLG_HCVJF_Qi_2014	0.116	0.661	0.661	0.629	N
POLG_PESV_Tsuboyama_2023_2MXD	0.160	0.208	0.263	0.600	Y
Q2N0S5_9HIV1_Haddox_2018	0.044	0.485	0.525	0.555	Y
R1AB_SARS2_Flynn_2022	0.105	0.562	0.641	0.654	Y
RCRO_LAMBD_Tsuboyama_2023_1ORC	0.596	0.377	0.426	0.644	Y
RDRP_I33A0_Li_2023	0.322	0.371	0.326	0.334	Y
REV_HV1H2_Fernandes_2016	0.240	0.239	0.321	0.330	Y
RPC1_BP434_Tsuboyama_2023_1R69	0.708	0.726	0.748	0.760	Y
RPC1_LAMBD_Li_2019_high-expression	0.549	0.284	0.364	0.476	Y
RPC1_LAMBD_Li_2019_low-expression	0.523	0.243	0.315	0.487	Y
SPIKE_SARS2_Starr_2020_binding	-0.018	0.172	0.333	0.482	Y
SPIKE_SARS2_Starr_2020_expression	0.012	0.203	0.359	0.553	Y
TAT_HV1BR_Fernandes_2016	0.017	0.361	0.420	0.428	Y
VG08_BPP22_Tsuboyama_2023_2GP8	0.655	0.373	0.488	0.471	Y
VRPI_BPT7_Tsuboyama_2023_2WNM	0.587	0.055	0.213	0.453	Y
Mean (31 assays)	**0.269**	**0.393**	**0.430**	**0.480**	**29 Y / 2 N**

Appendix C. Reproduce it yourself in Google Colab

No programming is needed; you only click through a notebook. You need a free Google account.

Step	What to do
1	Open colab.research.google.com and sign in. Choose File, Upload notebook, and select Afrium_reproduce_in_Colab.ipynb.
2	Choose Runtime, Change runtime type, pick T4 GPU, and press Save.
3	Run the cells from top to bottom (the play button on each cell, or Runtime, Run all). Cell titles are numbered Step 1 to Step 8.
4	Wait. Downloading ProteinGym takes a few minutes. Step 5 (scoring all 31 viral assays with AminoGrapes and ESM2) is the longest. Allow 30 to 60 minutes in total.
5	At Step 6 the notebook prints the means, the gains, the confidence intervals and the win counts. Compare them with Table 2. Small differences in the third decimal are normal.
6	If Colab disconnects, reconnect and run Steps 2 to 5 again. Steps 7 and 8 are optional (check the official ESM2 numbers yourself; draw Figure 7).

What you should see (mean Spearman over the 31 viral assays): ESM2-650M about 0.269; AminoGrapes about 0.430; AminoGrapes + ESM2 blend ($\lambda = 0.7$) about 0.376. The command that the notebook runs is:

```
python afrium.py --ref DMS_substitutions.csv --data DMS_ProteinGym_substitutions --msa DMS_msa_files
--weights DMS_msa_weights --struct ProteinGym_AF2_structures --taxon Virus --out afrium_out --esm2
```

Plain-language glossary. *ESM2-650M*: a protein language model released by Meta AI that reads one protein sequence and estimates how plausible each mutation is (650 million parameters); the baseline here and one component of the optional blend. *AminoGrapes*: the scoring method in this report, how often nature uses a given amino acid at a position among the protein's relatives multiplied by how buried that position is; the main score does not use ESM2. *ProteinGym*: a public benchmark of real lab experiments (deep mutational scans). *Spearman correlation*: a number from −1 to 1 measuring how well two rankings agree. *Assay*: one lab experiment on one protein; the viral set has 31. *Alignment (MSA)*: the protein lined up against its evolutionary relatives. *RSA / burial*: how exposed an amino acid is on the protein surface; buried positions are usually less tolerant of change.

References

[1] Notin P, Kollasch AW, Ritter D, et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. NeurIPS Datasets and Benchmarks, 2023.

[2] Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379:1123-1130, 2023.

[3] Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. NeurIPS, 2021.

[4] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 596:583-589, 2021.

[5] Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. J Mol Biol 79:351-371, 1973.

[6] Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilities of residues in proteins. PLoS ONE 8:e80635, 2013.

[7] Tsishyn M, Hermans P, Rooman M, Pucci F. Residue conservation and solvent accessibility are (almost) all you need for predicting mutational effects in proteins. Bioinformatics 41(6):btaf322, 2025. doi:10.1093/bioinformatics/btaf322.

[8] Laine E, Karami Y, Carbone A. GEMME: a simple and fast global epistatic model predicting mutational effects. Mol Biol Evol 36:2604-2619, 2019.

[9] Truong T Jr, Bepler T. PoET: a generative model of protein families as sequences-of-sequences. NeurIPS, 2023.

[10] Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422-1423, 2009.