Filtering viral data from protein models cuts viral performance, but fine-tuning restores it for under $80

Source copy kept by wheresthe.ai for the story “Filtering viral data from protein models cuts viral performance, but fine-tuning restores it for under $80” (https://wheresthe.ai/d/gb-filtering-viral-data-from-protein-models-cuts-viral-perfo-2026-10/).

Source: https://www.aisi.gov.uk/blog/pretraining-data-filtering-in-open-weight-biological-models

Text saved 26 Oct 2026 (the page as published).

SHA-256 of the saved text: d1ce5e262f54f113f347cef450d4b5d217f42772c2c7304aa79522fbc5c8d33f

---

Specialised biological AI models (or tools) are driving rapid advances in biotechnology. They are typically released as open-weight models and are smaller than frontier large language models (LLMs), as they are trained predominantly on biological data only. This makes them easier to train, share and modify. However, robustly safeguarding open-weight models is difficult. Once the weights and inference code are openly available, the effectiveness of safeguards is harder to guarantee. Post-training protections can often be cheaply undone through fine-tuning, and safety classifiers can often be removed, though how much effort this takes varies and is not always clear. To be effective, safeguards for open-weight models must be tamper-resistant. Of the safeguards developed to date, one of the most promising is stringent filtering of the pre-training data , since this often requires significantly more fine-tuning compute to undo, compared to post-training interventions. It has sometimes been pointed to by developers as a justification for openly releasing model weights. Prior work on LLMs (see here and here) suggests filtering can selectively degrade performance in a targeted domain, while leaving performance in other domains largely intact. This suggests it could, in principle, be used to remove information relevant to a dual-use capability from specialised scientific models, such as protein models. However, the training data for these models is much less varied than for LLMs, which makes a target domain harder to isolate: the data left behind after filtering may closely resemble the data removed, which could in turn affect how well filtering works in these contexts. AISI's Chem-Bio team conducts cutting-edge research to understand, evaluate and mitigate chemical and biological risks associated with frontier AI systems. This blog outlines the process and outcomes of our investigation into the effectiveness of pretraining data filtering on specialised scientific models, for which we used the open-weight ESM-2 family of protein models. Our findings show that data filtering can suppress a targeted capability in protein language models, without significantly impacting adjacent capabilities. However, it does not constitute a robust safeguard on its own, since the lost capability can be recovered at relatively low cost (under $80 of compute even for our largest model), and the data required to do so is often publicly available and easy to find. Therefore, for this to be an effective safeguard, it needs to come alongside other interventions, such as data access controls. How we tested it To measure the effects of data filtering, we pretrained ESM-2 protein language models from scratch. Protein language models are trained in a similar way to LLMs, but on amino acid sequences rather than words: the model sees millions of protein sequences and learns which amino acids tend to appear in which contexts. The sequence of these amino acids determines the shape the protein folds into, and this shape is what allows proteins to perform different functions within a living cell. The sequence of amino acids can be used to predict how a protein folds or how a mutation will affect its function. We conducted the training at several scales, ranging from 8M to 3B parameters. Parameter count is a standard measure of model size and complexity, and for comparison, frontier large language models (LLMs) exceed even our largest by orders of magnitude. Each model was trained on a dataset from which protein sequences that make up viruses (‘viral proteins’) had been removed. We refer to those excluded viral proteins as the ‘forget’ domain: the capability that filtering is intended to degrade. Performance on the ‘retain’ domain should be unaffected. In this case, the ‘retain’ domain refers to proteins that make up all other forms of life (‘non-viral proteins’). Filtering is considered successful only to the extent that it degrades performance on the former while leaving the latter intact. We then measured how much of the removed capability could be restored by fine-tuning: continuing to train each filtered model on the viral sequences that had been withheld during pre-training. To vary the extent of data removal, we defined a series of exclusion tiers , ranging from narrow (removing only a single strain of influenza) to broad (removing all viral proteins and their close homologues), with each tier a strict subset of the next. The tiers capture a central trade-off: narrow filtering removes less data and preserves more of the model's general capability, but may be easier to reverse, whereas broad filtering also removes data on closely-related proteins that could otherwise be used to reconstruct the targeted capability, offering stronger protection, but at a potentially greater cost to the model's capabilities in other ‘retain’ domains. Each model was then evaluated against three questions: 1. Does filtering degrade the targeted capability? Performance on the 'forget' domain (viral proteins) was compared with performance on the 'retain'

domain (non-viral proteins), using held-out protein data across three standard tasks: Masked language modelling: predicting amino acids that have been hidden from a protein sequence. Mutation-effect prediction: estimating how a change to a protein sequence alters its function. Contact prediction: identifying which amino acids of a protein are in physical contact within its 3D structure. 2. How tamper-resistant is the result? We attempted to restore the filtered capability by continuing to fine-tune each model on the excluded viral data, recording the compute required to achieve success. 3. How does the effect change with scale? The experiment was repeated across model sizes, since scaling behaviour is the question most relevant to real-world systems. Our key findings Filtering selectively degrades the targeted capability Excluding viral protein data from pretraining substantially reduced the model's performance on viral proteins across all three tiers, while performance on non-viral proteins was largely preserved. In other words, data filtering removed much of the targeted capability without reducing the model's usefulness for other types of proteins. The degradation was observed across all three of the tasks we used to measure viral protein capability (masked language modelling, mutation-effect prediction and contact prediction), although the extent varied by task. Degradation was clearest for masked language modelling, which was affected at every model size, while contact prediction was affected for models with over 150M parameters, and mutation-effect prediction for models with over 650M parameters. The effect was also stronger for larger, more capable models, because filtering holds back the viral protein performance that unfiltered models otherwise gain as they scale.

Figure 1. Filtered against unfiltered models of the same size, on the 'forget' domain (viral proteins) and the 'retain' domain (non-viral proteins). Filtering lowers performance on viral proteins while leaving non-viral performance largely intact. Filtering is effective but not durable Continued fine-tuning on the excluded viral data was able to completely restore performance on viral proteins, which shows that filtering raises the cost of recovering a capability rather than removing it permanently. At present, such viral data is routinely made publicly available, even if it is not used in model training. While the cost of fine-tuning models on this data increased with model size, the absolute cost remained modest: recovering the full capability even in the largest (3B-parameter) model required only about $80 of H100 compute.

Figure 2. Fine-tuning on the withheld viral data recovers the lost performance at every model size, and the compute required to do so increases with scale. These results are based on a single model family (ESM-2), so how well they generalise to other protein models is uncertain, and it is possible that filtering may be less effective for models that can retrieve additional data at inference time. Our measurements of recovery cost are also imprecise, which limits how confidently we can draw conclusions on how tamper-resistance changes with model size. What this means Overall, our findings indicate that data filtering can effectively and selectively suppress a targeted capability in protein language models. On its own, however, it does not constitute a robust safeguard, since the lost capability can be recovered at relatively low cost, and the data required to do so is routinely made publicly available. A durable approach is therefore likely to require additional measures alongside data filtering, including improved access controls for scientific models, or the biological data required to train them. You can read more in the full paper . AISI's Chem-Bio Team is hiring. If this work interests you, please apply here .