nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00931-6

Evaluating generalizability of artificial intelligence models for molecular datasets

Received: 31 March 2024

Accepted: 21 October 2024

Published online: 6 December 2024

Check for updates

Yasha Ektefaie[1] ✉, Andrew Shen[1,2], Daria Bykova[3], Maximillian G. Marin[1], Marinka Zitnik[1,4,5,6,8] ✉ & Maha Farhat[1,7,8] ✉

Deep learning has made rapid advances in modelling molecular sequencing data. Despite achieving high performance on benchmarks, it remains unclear to what extent deep learning models learn general principles and generalize to previously unseen sequences. Benchmarks traditionally interrogate model generalizability by generating metadata- or sequence similarity-based train and test splits of input data before assessing model performance. Here we show that this approach mischaracterizes model generalizability by failing to consider the full spectrum of cross-split overlap, that is, similarity between train and test splits. We introduce SPECTRA, the spectral framework for model evaluation. Given a model and a dataset, SPECTRA plots model performance as a function of decreasing cross-split overlap and reports the area under this curve as a measure of generalizability. We use SPECTRA with 18 sequencing datasets and phenotypes ranging from antibiotic resistance in tuberculosis to protein–ligand binding and evaluate the generalizability of 19 state-of-the-art deep learning models, including large language models, graph neural networks, diffusion models and convolutional neural networks. We show that sequence similarity- and metadata-based splits provide an incomplete assessment of model generalizability. Using SPECTRA, we find that as cross-split overlap decreases, deep learning models consistently show reduced performance, varying by task and model. Although no model consistently achieved the highest performance across all tasks, deep learning models can, in some cases, generalize to previously unseen sequences on specific tasks. SPECTRA advances our understanding of how foundation models generalize in biological applications.

Understanding generalizability—how well a machine learning model performs on unseen data—is a fundamental challenge for the broad use of computation in biological discovery. In living cells, information flows from DNA to RNA to protein and dictates cell phenotypes. To model phenotypes, deep learning models are trained to predict biological relationships between and within sequences and the phenotype. This approach has been successfully implemented through a broad array of machine learning models, including convolutional neural networks (CNNs)[1–4], recurrent neural networks[5–7], graph neural networks[8–11] and large language models[12–16]. However, model evaluation is challenging because (1) available molecular sequencing data often capture a small fraction of all possible sequences[4,17], and (2) sequences evolve and acquire new mutations that are not present in existing datasets. This results in differences in the distribution of sequences and their aggregate properties between datasets, known as distribution shifts, that can lead to degradation of model performance[18–20].

A full list of affiliations appears at the end of the paper. ✉e-mail: yasha_ektefaie@hms.harvard.edu; marinka@hms.harvard.edu; maha_farhat@hms.harvard.edu

Although distribution shifts are a well-recognized challenge in machine learning more generally[21,22], they are less well characterized in biology due to the lack of approaches that measure model performance in the context of distribution shifts. Although numerous benchmarks have been developed to assess model performance across datasets[16,23–26], there are still large gaps between model performance during benchmarking and real-world use[27–31] (Fig. 1a). This gap in generalizability must be addressed before machine learning models can be broadly used in biology.

While helpful, the central shortcoming of existing benchmarks is the approach to model evaluation. Existing methods for model evaluation split input molecular sequencing datasets into train and test sets in metadata-based (MB) or similarity-based (SB) splits (Fig. 1c). MB splits ensure specific metadata properties are not shared across splits. One example is a temporal split of COVID-19 viral sequences in which a vaccine escape model is trained on sequences collected before a specific time and tested on sequences evolved after that time[32–34]. A random split is also an MB split where the metadata property is a sample identity. SB splits ensure no two samples across splits share sequence similarity beyond a predefined threshold, with the exact threshold being problem specific[35–39]. However, as we show in this study, MB splits cannot guarantee that high performance on the test set will transfer to a new molecular sequencing dataset. This is because MB partitioning does not control sequence similarity between data subsets. Model generalizability can be overestimated when training sequences are more similar to sequences in the test set than sequences in a new dataset. In SB splits, sequence similarity can be controlled. However, the generalizability of the model at similarity thresholds different from a handful of those tested during model benchmarking remains unknown, resulting in an incomplete evaluation of the model. Further, SB splits rely on limited summary metrics such as sequence distance to quantify similarity between sequences that may not capture the full range of similarities. This lack of understanding about model generalizability can lead to several important issues. First, there could be catastrophic degradation of model performance on new datasets, which means that predictions made on unseen data may be highly inaccurate. This inaccuracy can mislead biological research, causing wasted resources on false leads or overlooking potential discoveries. Second, models that perform well on training data but poorly on unseen data can contribute to overfitting, where the model captures noise rather than the underlying biological processes. This overfitting can severely limit the applicability of computational tools in new or broader biological contexts, potentially leading to erroneous conclusions about biological mechanisms and phenotypes[40].

Here we introduce the spectral framework for model evaluation (SPECTRA), a framework for evaluating generalizability of machine learning models for molecular sequences. The term spectral here refers specifically to the evaluation framework and should not be confused with its use in matrix analysis. Given a model, a molecular sequencing dataset, and a spectral property definition, SPECTRA generates a series of train–test splits with decreasing overlap, that is, a spectrum of train–test splits. SPECTRA then plots the model's performance as a function of cross-split overlap (Fig. 1b,d), generating a spectral performance curve (SPC). We propose the area under this curve (AUSPC) as a new, more comprehensive metric for model generalizability. As opposed to MB or SB splits, SPECTRA can incorporate multiple similarity definitions in the spectral property definition, such as sequence distance and structural similarity of protein sequences. We apply SPECTRA to 18 molecular sequencing datasets from three prominent benchmarks (PEER[24], ProteinGym[16], TAPE[23]) and find (1) existing SB and MB splits have large amounts of cross-split overlap, (2) SPECTRA generates splits with similar levels of cross-split overlap compared to existing SB and MB splits, and (3) existing SB and MB splits represent single points in the SPC, leaving the rest of the SPC uncharacterized. To demonstrate the need to characterize model SPCs, we apply SPECTRA to 11 state-of-the-art machine learning models, including pretrained and finetuned large language models, CNNs, graph neural networks, variational autoencoders and diffusion generative models. None of the machine learning models tested achieves a high AUSPC across all tested tasks. We show that examining the SPC can help identify unconsidered spectral properties that influence model generalizability in molecular sequencing datasets. By applying SPECTRA to pretrained protein language models, we demonstrate how SPECTRA can be used to evaluate foundation models in biology. SPECTRA is a new paradigm for model evaluation in its ability to more comprehensively characterize model generalizability and uncover shortcomings of existing machine learning models. These capabilities will guide the development of machine learning models for molecular sequencing data.

Results

Overview of SPECTRA

In contrast to the prevailing approach to machine learning model evaluation using MB and SB splits, the spectral framework (SPECTRA) provides a more comprehensive overview of model performance by examining a model's spectral performance curve (SPC) for a given molecular sequencing dataset. The approach focuses on one or more characteristics of input molecular sequences or molecular sequence properties (MSPs) (for example, GC content of a gene). We define a spectral property as an MSP expected to affect model generalizability for a specific task (for example, three-dimensional (3D) protein structure for protein binding prediction; Methods, 'PDBBind dataset'). The definition of the spectral property is task-specific and, together with the molecular sequence dataset and model, are the only inputs to SPECTRA (Fig. 1d, Supplementary Note 5 and Supplementary Table 5). First, SPECTRA compares the spectral property for all pairs of sequences in the dataset, identifying pairs that share the spectral property. The procedure is used to construct a spectral property graph (SPG) from which adaptive train–test splits are generated with decreasing cross-split overlap or the proportion of samples in the test set that share a spectral property with the train. SPECTRA controls cross-split overlap by ranging an internal spectral parameter (SP) from SP = 0 to SP = 1 for maximal and minimal cross-split overlap, respectively (Supplementary Note 6). Last, the model is trained and tested on each split to generate a plot of the model's performance against the SP, the model's SPC, for the molecular sequencing dataset. The area under the SPC (AUSPC) summarizes model performance across all levels of cross-split overlap. It can compare model generalizability to other models within and across tasks.

SPECTRA unifies model evaluation and benchmarking approaches

We propose that SB and MB splits used in existing machine learning benchmarks represent individual points on the SPC and that SPECTRA provides a more comprehensive view of model generalization than prevailing dataset splits. To evaluate this proposition, we applied SPECTRA to six molecular sequencing datasets from two widely used protein sequence benchmarks, TAPE[23] and PEER[24], as well as a protein structural biology benchmark, PDBBind[41], and a protein mutational benchmark, ProteinGym[16] (Methods, sections 'PEER benchmark datasets' to 'PoseBusters dataset'). First, we were interested in what specific points on the SPC each dataset split in these benchmarks corresponds to. Specifically, we calculated the cross-split overlap in the MB and SB splits implemented in these benchmarks and identified the SP that generates splits with similar cross-split overlap. For example, our analyses using SPECTRA revealed that the MB family split from the remote homology dataset in TAPE had a cross-split overlap of 97% and can be instantiated at SP = 0.025. In another example, the temporal-based MB split of PDBBind used to train the Equibind model had a cross-split overlap of 76% and was generated at SP = 0.55 (Fig. 2a)[34].

https://doi.org/10.1038/s42256-024-00931-6

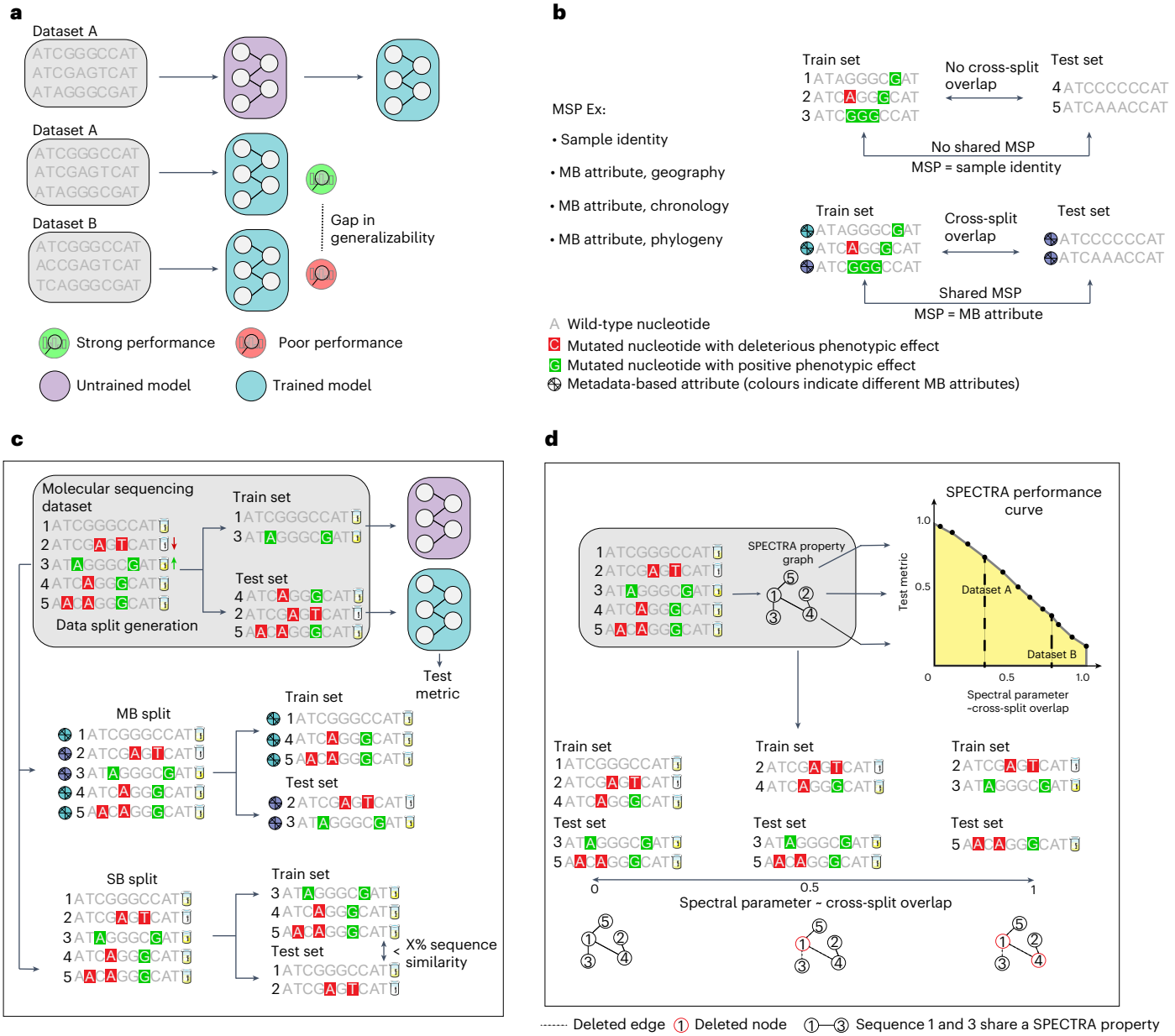

Fig. 1 | The spectral framework for model evaluation (SPECTRA). a, Machine learning models for molecular sequencing data struggle to generalize across datasets. **b**, Every train–test split partitions samples based on a chosen molecular sequence property (MSP). Cross-split overlap exists between train and test sets when samples share MSPs. Shown are two examples of train–test splits. In the first, samples are split based on their identity, shown by the sample number, and there is no cross-split overlap as no two samples share identity across the train and test set. In the second, samples are split based on metadata-based (MB) attributes with cross-split overlap as two samples share an attribute across the train and test set. **c**, Traditional approaches for model evaluation create train–test splits either based on MB attributes (MB split) or sequence similarity (similarity based or SB split). **d**, SPECTRA, the spectral framework for model evaluation, generates train–test splits with a spectrum of cross-split overlap. It does so by constructing a spectral property graph where nodes are samples and edges are between samples that share a spectral property. The spectral property is a MSP that influences model generalizability. It then iteratively deletes nodes and edges based on the spectral parameter, an internal parameter that scales with cross-split overlap, to generate train and test sets. After training and evaluating an input model to each generated split, SPECTRA generates a spectral performance curve (SPC) that plots model test performance versus the spectral parameter. The SPC shows the gap in generalizability depicted in panel a is due to differing levels of cross-split overlap between splits generated in dataset A and those of dataset B. The area under the spectral performance curve (AUSPC) (highlighted in yellow) summarizes model performance across all levels of cross-split overlap and is a new metric for model generalizability. Note that the use of the word spectral here refers only to the framework for model evaluation and should not be confused with other uses of the term in matrix analysis.

We find that model performance tends to decrease with decreasing cross-split overlap. The accuracy of long short-term memory (LSTM)[42] and CNN[43] models decreased by 50% between family and superfamily splits in the TAPE benchmark's remote homology dataset. The cross-split overlap was lower for the superfamily (71%) compared to the family split (97%) (Fig. 2b). We observed a similar pattern when using SPECTRA to study models for predicting secondary protein structure (Fig. 2c) and protein–ligand binding affinity (Fig. 2d). These findings show that existing molecular benchmarks capture only a few points on the SPC, providing a myopic assessment of model generalizability. Further, the observation that model performance diminishes when cross-split overlap decreases identifies an overestimation of model performance by existing benchmarks, which would lead to suboptimal performance when the models are implemented in the real world.

https://doi.org/10.1038/s42256-024-00931-6

a

Task	Benchmark	Molecular sequence property	Cross-split overlap (%)	Spectral parameter
Remote homology	TAPE	Structural features: protein fold	81	0.275
Remote homology	TAPE	Homology: protein family	97	0.025
Remote homology	TAPE	Homology: protein superfamily	71	0.475
Secondary structure	TAPE, PEER	Dataset identity (CASP12)	48	0.5
Secondary structure	TAPE, PEER	Dataset identity (CB513)	52	0.475
Secondary structure	TAPE, PEER	Dataset identity (TS115)	54	0.45
Amyloid beta protein aggregation	ProteinGym	Sample identity	50	0
RRM domain activity	ProteinGym	Sample identity	53	0
Subcellular localization	PEER	Sample identity	6.8	0.325
Protein–ligand binding	PDBBind	Sample collection date (Equibind)	76	0.55
Protein–ligand binding	PDBBind	Dataset identity: astex diverse set	82	0.45
Protein–ligand binding	PDBBind	Sequence similarity* (LPPDBind)	91	0.275
Protein–ligand binding	PDBBind	Sample collection date* (Posebusters)	70	0.575

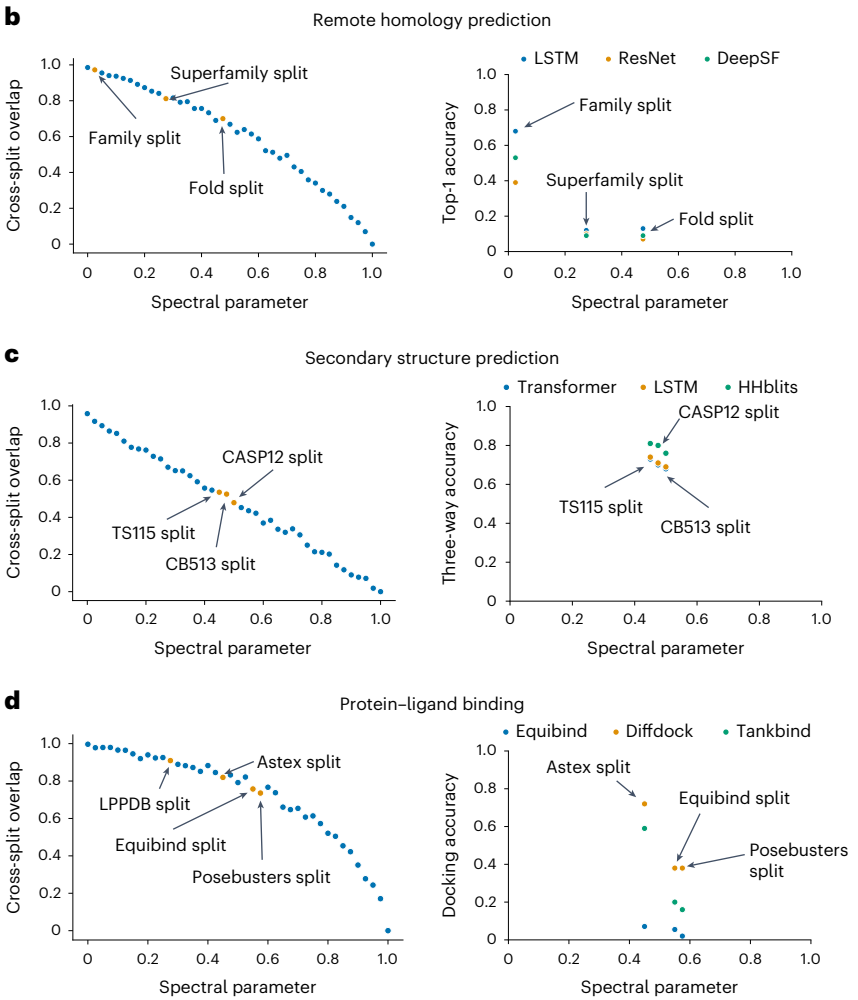

Fig. 2 | Applying SPECTRA to existing molecular sequencing benchmarks. **a**, Results of applying SPECTRA on datasets from the PEER, TAPE, ProteinGym and PDBBind benchmarks. For every dataset, we report the corresponding task and benchmark, the MSP that was used to split the dataset into train and test, cross-split overlap and the spectral parameter in SPECTRA that generated a split with similar levels of cross-split overlap. If a specific name is used to refer to the split in literature, the name is provided in parentheses. (*Posebusters and LPPDBind use multiple manually defined quality control criteria to create splits). **b**, Left, cross-split overlap as spectral parameter increases in SPECTRA for the TAPE remote homology benchmark. Labelled are the points on the curve where the family, superfamily and fold splits have similar levels of cross-split overlap. Right, a partial SPC for an LSTM[42], ResNet[98] and DeepSF[43] model. **c**, Left, cross-split overlap as spectral parameter increases in SPECTRA for the TAPE secondary structure benchmark. Labelled are the points on the curve where the CASP12 (ref. 99), TS115 (ref. 100) and CB513 (ref. 101) splits have similar levels of cross-split similarity. Right, a partial SPC for a Transformer[49], LSTM[42] and HHblits[102] model. **d**, Left, cross-split overlap as spectral parameter increases in SPECTRA for the PDBBind benchmark. Labelled are the points on the curve where the LPPDBind[66], Astex diverse set[88], Equibind[34] and Posebusters[65] splits have similar levels of cross-split similarity. Right, a partial SPC for Equibind[34], Diffdock[84] and Tankbind[103] models.

SPECTRA reveals generalization gaps in molecular ML models

To demonstrate the use of SPECTRA in characterizing the full model SPC, we evaluate six models in five molecular sequencing datasets. Specifically, we generate SPECTRA splits for each dataset, train and test models on each split, generate the SPC and calculate the AUSPC for each model.

Our data span three diverse problems: antibiotic resistance in *Mycobacterium tuberculosis*[1] (tuberculosis), vaccine escape in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[44] and fluorescence prediction in the green fluorescent protein (GFP) of *Aequorea victoria*[45]. Antibiotic resistance in tuberculosis is a whole organism phenotype where inputs are the nucleic acid sequences of genes and non-coding regions causally linked to resistance to the specific drug. The output is resistance binary phenotype determined using a culture-based assay. We consider tuberculosis resistance to the antibiotics rifampicin (RIF), isoniazid (INH) and pyrazinamide (PZA). For fluorescence prediction of *A. victoria*, we rely on the amino acid sequence of GFP protein and its variants. Vaccine escape in SARS-CoV-2 maps mutations in the receptor binding domain (RBD) of the spike protein to a continuous value that represents antibody escape (Methods, sections 'Tuberculosis dataset' to 'SARS-CoV-2 dataset').

We evaluated six approaches to modeling phenotype from molecular sequence data. We generated SPCs for a logistic regression model, a CNN[1], a pretrained (GearNet[8]) and finetuned structure-based graph neural network (GearNet-Finetuned), a pretrained (Evolutionary Scale Modelling or ESM2, ref. 12) and a finetuned large language model (ESM2-Finetuned), a multiple sequence alignment-based generative model (an evolutionary model of variant effect prediction, EVE[32]) and an alignment-free generative model (SeqDesign[46]) (Methods, 'Training models').

As was seen for the protein benchmarks above, all evaluated models demonstrate a decrease in performance as cross-split overlap decreases (Fig. 3a). Logistic regression decreases performance from

https://doi.org/10.1038/s42256-024-00931-6

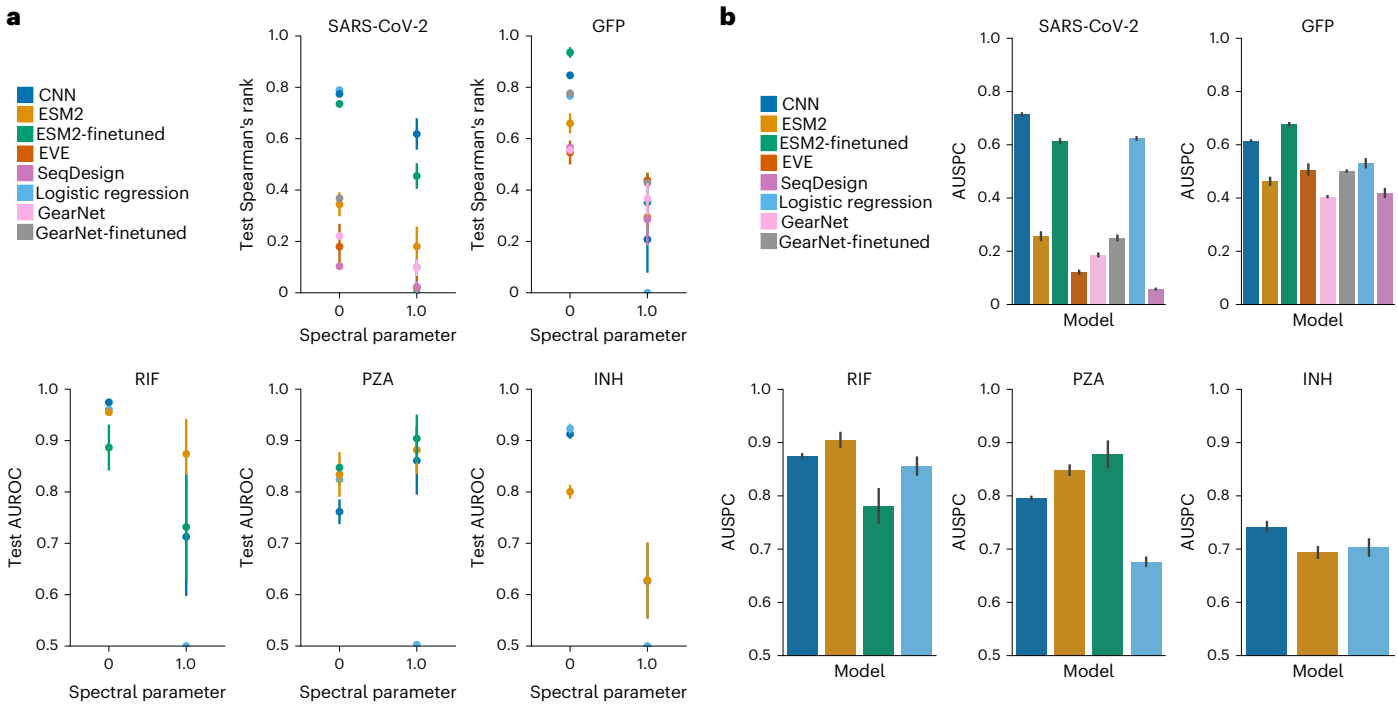

Fig. 3 | Application of SPECTRA to eight machine learning models across five molecular sequencing datasets. a, Test metric change between the spectral parameter of 0 (maximum cross-split overlap) and 1 (no cross-split overlap) across all tested models and tasks. **b**, Average AUSPC across all models and tasks. The *y* axis begins at random AUSPC. Error bars in both plots indicate standard deviation.

an area under the curve greater than 0.9 for RIF and INH resistance prediction to 0.5 (RIF AUSPC = 0.86, INH AUSPC = 0.74). ESM2-Finetuned decreases in performance for GFP from a Spearman's rank correlation greater than 0.9 to one less than 0.4. Although all models demonstrate decreased performance with decreased cross-split overlap, some models continue to perform well at minimum cross-split overlap (SP = 1). In RIF and PZA, ESM2, ESM2-Finetuned and CNN maintain areas under the curve greater than 0.7 at SP = 1 (Fig. 3a). No single model outperforms others across all tasks by AUSPC (Fig. 3b).

SPECTRA identifies critical spectral properties

Every SPC represents a hypothesis that model performance will vary as a function of the defined spectral property. Any additional deviation of model performance from the expected performance under the defined SPC may indicate the presence of unconsidered spectral properties. By identifying these spectral properties, we can better understand what models learn from molecular sequencing datasets and identify shortcomings of existing models.

CNN model performance demonstrates a high variance across splits with the same SP, suggesting the presence of an unconsidered spectral property (Fig. 4a). Three splits in RIF at SP = 0.9, 0.95 and 1.0 have an AUROC standard deviation (and total performance decrease) of 0.09 (26%), 0.10 (31%) and 0.08 (23%), respectively (Fig. 4a). RIF resistance is caused by missense mutations in an 81 base pair region of the RNA polymerase beta-subunit (*rpoB*) gene, the RIF resistance-determining region (RRDR) (Fig. 4b)[47,48]. Models able to learn the association of RRDR to resistance most comprehensively will achieve high test performance. We propose that as SP increases the genetic distance between observed RRDR mutations in the train and test increases. On such splits, models may associate partial regions of the RRDR to resistance that do not align with RRDR regions in the test, resulting in poor generalizability.

To investigate this hypothesis, we calculated the difference in the range of positions for RRDR mutations observed in the train and the test splits (diff-RRDR) (Fig. 4c). Diff-RRDR explains the variance in model performance observed at SP = 0.9 (Spearman's rank correlation of $\rho = -0.51$, $P = 1.79 \times 10^{-5}$, between diff-RRDR and AUROC, Fig. 4d). We observe similar patterns for the SP = 1.0 split and SP = 0.95 split (Fig. 4d). This correlation is model specific; ESM2 performance experiences no degradation in performance as diff-RRDR increases in the SP = 0.9, 0.95 and 1.0 splits (Fig. 4e). ESM2 considers a larger range of input positions (512) than the CNN (12) when making a prediction. We expect this to make ESM2 more invariant to increases in diff-RRDR[49]. Our results improve our understanding of *M. tuberculosis* resistance prediction where state-of-the-art performance in this field is currently achieved by CNN models[1]. The generalizability of *M. tuberculosis* resistance prediction models can be improved by potentially considering longer contexts of DNA sequence. We characterize the difference in the length of genetic context (diff-RRDR) as a spectral property relevant for phenotype prediction for proteins where functionally impactful mutations concentrate along the DNA sequence surrounding the active site.

SPECTRA evaluates generalizability of foundation models

Many machine learning models are pretrained on molecular sequencing datasets and then trained and tested on usually smaller task-specific datasets not encountered in pretraining[12,15,50]. These foundation models have the potential to offer better flexibility and adaptability to a wide range of tasks as they have done in computer vision[51] and natural language processing[52–54]. Despite the potential of these models, their generalizability is unknown as they have rarely been tested prospectively on non-overlapping datasets[55,56]. Current approaches to benchmark foundation models report the average performance across multiple task-specific datasets. However, this approach exaggerates foundation model generalizability by failing to consider cross-split overlap in the task-specific data splits and the overlap between the pretraining and task-specific datasets. Running SPECTRA with foundation models on task-specific datasets evaluates generalizability by measuring AUSPCs. We can then assess the effect of overlap between the pretraining and task-specific datasets on these AUSPCs to understand foundation model generalizability.

https://doi.org/10.1038/s42256-024-00931-6

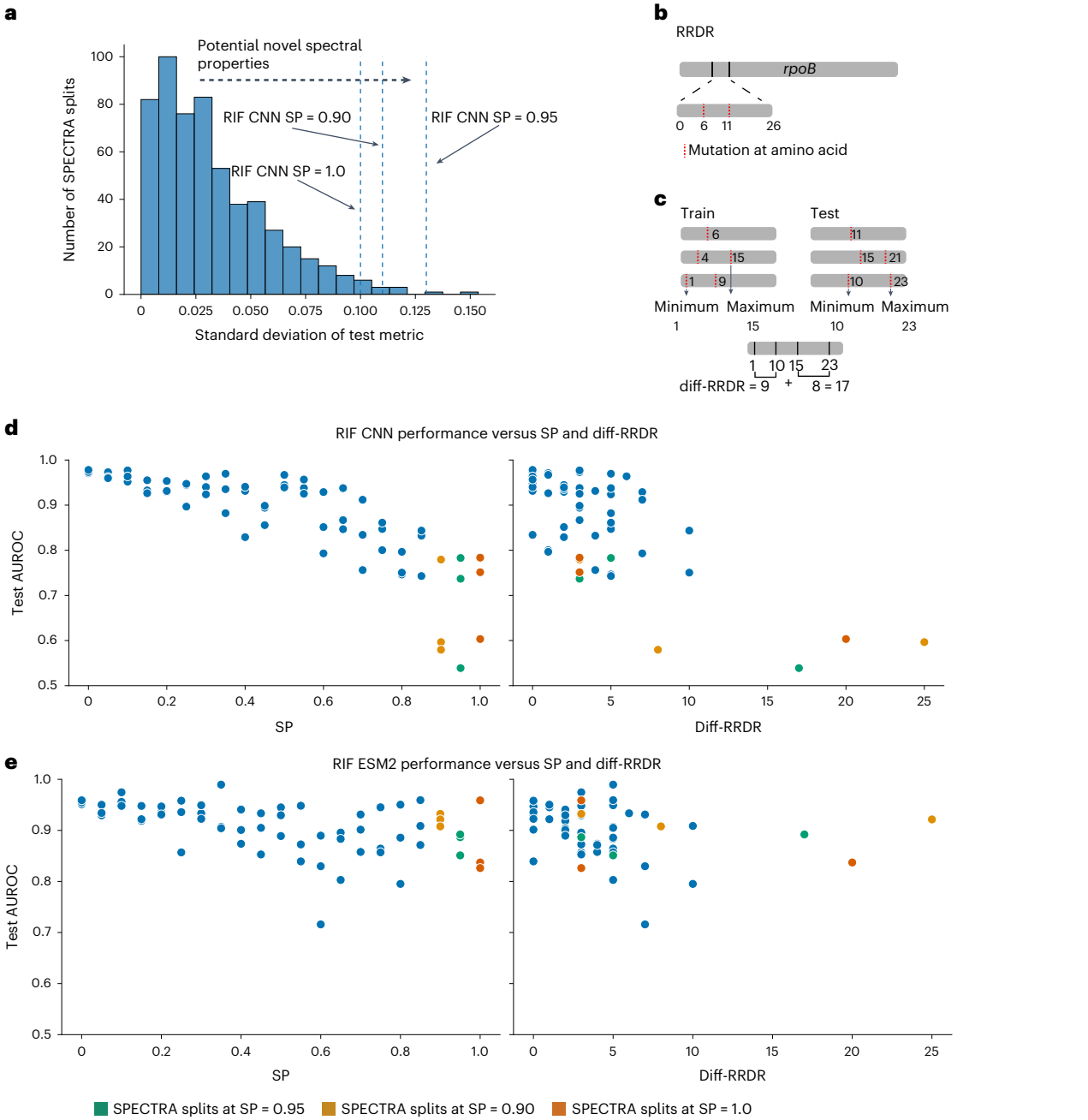

Fig. 4 | SPECTRA identifies spectral properties critical for model performance that might otherwise be overlooked. a, Distribution of standard deviation of test metric performance across splits generated with the same spectral parameter for the RIF, PZA, INH, GFP and SARS-CoV-2 datasets across all tested models. Splits with a high standard deviation could have new spectral properties. RIF CNN SP = 1.0, SPECTRA splits generated for RIF dataset at SP 1.0. **b**, The RRDR of the *rpoB* gene in Tuberculosis is a 26 amino acid region strongly associated with RIF resistance. **c**, Diff-RRDR is a metric that measures the difference in the minimum and maximum positions of observed RRDR mutations in the train and test. **d**,**e**, The SPC for CNN (**d**) and ESM2 (**e**) in the RIF dataset (shown left). Test AUROC versus diff-RRDR for CNN (**d**) and ESM2 (**e**) in the RIF dataset (shown right). Highlighted are points representing splits generated at a spectral parameter of 0.90, 0.95 and 1. All test metrics are averaged across three independent runs.

To demonstrate this capability of SPECTRA, we investigate the generalizability of several protein foundation models. From our previous analysis, the AUSPC of pretrained ESM2 for RIF, PZA, INH, SARS-CoV-2 and GFP phenotype prediction varied widely from 0.91 in RIF to 0.26 in SARS-CoV-2 (Fig. 3b). We calculated the overlap between these task-specific datasets and UniRef50, the pretraining dataset used by ESM2 (ref. 57) (Fig. 5a; Methods, 'Using SPECTRA to evaluate foundation models in biology') finding a significant correlation between this overlap and the AUSPC of ESM2 (Spearman rank correlation 0.9, $P = 1.4 \times 10^{-27}$, Fig. 5a). Finetuning ESM2 on downstream tasks improves ESM2 AUSPC for PZA, SARS-CoV-2 and GFP phenotypic prediction (Fig. 3b, ESM2-Finetune). The effect of overlap between pretraining and task-specific data holds when evaluating protein foundation models other than ESM2, including Transception, MSATransformer, ESM1v and Progen on five molecular sequencing datasets in the ProteinGym benchmark[16] (Supplementary Note 3 and Fig. 5b; Spearman rank correlation 0.9, $P = 0.04$).

Discussion

Understanding how well molecular machine learning models perform on unseen data is a fundamental problem for protein design[58,59], defence against emerging pathogens[33,60,61] and therapeutic

https://doi.org/10.1038/s42256-024-00931-6

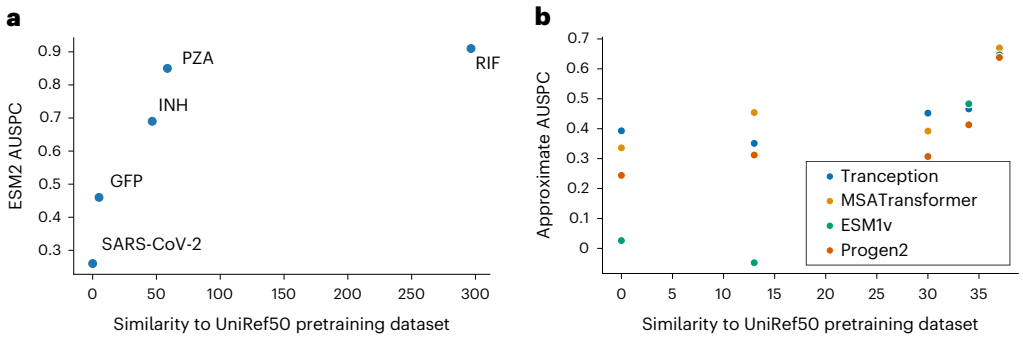

Fig. 5 | Influence of pretraining set similarity on protein foundation model generalizability. a, ESM2 AUSPC versus similarity to UniRef50, the pretraining dataset of ESM2, for the RIF, INH, PZA, GFP and Covid tasks. Each point is labelled with the corresponding molecular sequencing dataset. **b**, Approximate AUSPC versus UniRef50 similarity for four protein foundation models, Transception, MSATransformer, ESM1v and Progen2, in five molecular sequencing datasets from ProteinGym (dataset names in order from left to right: A4D664-9INFA-Soh-CCL141-2019, A0A140D2T1-ZIKV-Sourisseau-growth-2019, A4-HUMAN-Seuma-2021, AACC1-PSEAI-Dandage-2018, A4GRB6-PSEAI-Chen-2020). All correlations shown are significant ($P < 0.05$).

science[62–64]. Increasingly, studies are demonstrating how reported benchmarking model performance is overly optimistic about model generalizability[56,65,66]. We show this generalizability gap between model performance on benchmarks and external datasets is due to inadequate assessment of overlap between training, test and external datasets. We demonstrate that model performance decreases in existing benchmarks as cross-split overlap decreases. To address this challenge, we introduce SPECTRA to evaluate, compare and understand model generalizability by explicitly controlling cross-split overlap when constructing train–test splits to generate SPCs and calculate the AUSPC. We demonstrate that SPECTRA summarizes model test performance more comprehensively than in previous benchmarks. We show SPECTRA can help identify new and relevant spectral properties that influence model generalizability. We also demonstrate SPECTRA's ability to be used as a tool to evaluate foundation models in biology. After applying SPECTRA to four protein foundation models across 11 molecular sequencing datasets, we found that foundation models have better generalizability in task-specific datasets with greater similarity to the pretraining dataset. These findings corroborate existing literature, which indicates single-cell foundation models struggle to generalize in task-specific datasets dissimilar from pretraining datasets[55].

Existing SB and MB splits are attractive because they are computationally inexpensive. However, as shown in this paper, the time saved in computation leads to a mischaracterization of model generalizability. Model evaluation should be seen as a step in model development as computationally expensive as model training, not an afterthought to generate a performance metric. The purpose of this study is not to rank models by their AUSPCs but to demonstrate SPECTRA on state-of-the-art models. To rank models requires averaging over many AUSPCs calculated across various tasks in a SPECTRA benchmark for molecular sequencing data.

Our framework makes urgent the need to define and then refine MSP benchmarks for generalizability. In the case of *M. tuberculosis*' resistance to RIF, previous work has found high AUROC using logistic regression models, which supported the development of commercial genotypic assays for RIF resistance prediction. SPECTRA found logistic regression has high AUROC at SP = 0 (maximal cross-split overlap), but degrades to near-random performance at SP = 1 (no cross-split overlap). This corresponds to the current understanding that resistance to RIF is encoded by a handful of very common mutations concentrated in the active site and a long tail of rare mutations in a more distant proximity that can be more common in specific regions of the world. As a result, these RIF resistance commercial tests were later recognized to misclassify resistance in large populations of patients in specific geographic areas[67]. SPECTRA could have provided a more realistic understanding of the sensitivity of these tests and the mutations they are based on before deployment. This case of *M. tuberculosis*' resistance to RIF has implications for any protein where mutations with a functional effect on phenotype concentrate along the protein's length or structure, including, for example, the concentration of immune escape mutations in the SARS-CoV spike RBD.

SPECTRA can be used with models beyond the molecular sequencing domain. For example, in machine learning for small molecule therapeutics, the chemical structure is the spectral property that separates molecules into structurally dissimilar train–test splits[68]. In inverse protein folding problems, when models associate the amino acid sequence that folds into a particular protein structure, the ontological label of protein folds can be used as the spectral property[69]. Future directions include applying SPECTRA to other molecular sequence modalities such as RNA sequences, metabolomics or epigenetic sequences and considering metrics beyond the area under the curve to summarize the SPC.

In biology, there is a general belief that observations made in the past can only predict similar future observations or that models can only generalize to sequences similar to previously encountered sequences. Our results challenge this belief by demonstrating that models can, in some cases, perform well on sequences with never-before-seen mutations, motifs or sequence identity. Our work identifies MSPs that allow models to generalize in these scenarios, such as sequence similarity to a pretraining set. The SPECTRA framework for model evaluation represents a new frontier, generating a robust understanding of model performance beyond a metric on a MB or SB split.

Methods

This section describes the details for (1) the SPECTRA framework, (2) the curation of all molecular sequencing datasets used in this study, (3) model architecture and training and (4) characterizing distribution shifts in molecular sequencing datasets.

The SPECTRA framework for model evaluation

For an input molecular sequencing dataset and model, SPECTRA consists of three steps: (1) spectral property definition, (2) spectral property graph (SPG) construction and split generation and (3) model SPC generation.

Defining spectral properties. An MSP is a property that is either given or calculated inherent to a molecular sequence. Spectral properties are MSPs that influence model generalizability. Spectral properties are task and data specific: what may be a spectral property for one data type and task is not one in another data modality and task. For example, for predicting DNA binding motifs, the number of adenines in an input nucleic acid sequence is an MSP but not a spectral property. For secondary

structure prediction, the 3D structure of an amino acid sequence is an MSP, which is also an spectral property because the structural motifs present in a train set versus a test set will influence model generalizability.

The choice of how to define spectral properties is dataset and problem specific. The datasets used in this study can be divided into two categories: (1) mutational scan datasets (MSDs), which comprise a single set of sequences with different mutations and their effect on phenotype and (2) sequence-to-sequence datasets, which comprise different sequences and their properties.

In sequence-to-sequence datasets, the spectral property of interest is sequence identity. To calculate whether two sequences share sequence identity, we perform a pairwise alignment between input sequences and calculate the proportion of aligned positions to the length of the pairwise alignment. If this proportion is greater than 0.3, then the two sequences share this spectral property[70]. We use Biopython[71] to align with a match score of 1, a mismatch score of −2 and a gap score of −2.5. We used heuristics to define the comparator for larger datasets when exhaustive pairwise comparison of all sequences was not computationally feasible (Methods, 'Using SPECTRA to evaluate foundation models in biology'). In MSDs, phenotypically meaningful differences are in the scale of single mutations. Thus, using the definition of the spectral property from sequence-to-sequence datasets would underestimate differences between samples. To address this, we represent samples in MSDs by their sample barcode or a string representation of the mutations present in the sample. The spectral property of a sample is its sample barcode. Two samples share this property if their sample barcodes share at least one mutation.

Constructing SPECTRA property graphs and splits. After the spectral property is defined, a SPG is constructed where nodes are samples in the input dataset, and edges are between samples that share a spectral property (Supplementary Figs. 1 and 2). Finding a split such that no two samples share a spectral property is the same as finding the maximal independent set of the SPG or the maximum set of vertices such that no two nodes share an edge[72]. Finding the maximal independent set is NP-Hard[73], we approximate it via a greedy random algorithm where we (1) randomly order SPG vertices, (2) choose the first vertex and remove all neighbours and (3) continue until no vertices remain in the graph. To create an overlap in generated splits, we introduce the SP to the algorithm. Instead of deleting every neighbour, we delete each neighbour with a probability equal to the SP. If the SP is 1, we approximate the maximal independent set; if it is 0, we perform a random split. Given a set of nodes returned by the independent set algorithm, we produce an 80–20 train–test split. Sample SPGs can be found in Supplementary Figs. 1 and 2, and statistics for all generated SPGs can be found in Supplementary Table 2.

This procedure is complicated in MSDs where sample barcodes map to multiple samples (that is, in MSDs where the number of unique sample barcodes is not equal to the number of samples). As a result, if split generation does not consider the number of samples, splits can be generated with a small or uneven distribution of samples (that is, a train set with 100 samples and a test set with 10,000 samples). To address this, we applied two changes: (1) weighing nodes in the SPG by the number of samples represented by the sample barcode and biased the algorithm to choose these nodes, and (2) when splitting the nodes into train–test splits, we ran a subset sum algorithm to ensure train and test splits had 80 and 20% of samples, respectively (Supplementary Note 1). Statistics for all generated SPECTRA splits can be found in the supplement (Supplementary Figs. 5 and 6).

Generating SPCs. To generate a SPC, we create splits with SPs between 0 and 1 in 0.05 increments. For each SP, we generate three splits with different random seeds. We then train and test models on generated splits and plot model test performance versus SP. The area under this curve is the AUSPC. We provide SPCs and AUSPCs for all relevant models in the GFP, SARS-CoV-2, RIF, PZA and INH datasets (Supplementary Figs. 7–11).

Datasets

This section outlines the datasets and processing performed for this study.

Tuberculosis dataset. Tuberculosis sequencing and antimicrobial screening data come from ref. 1. Paired-end reads are trimmed with trimmomatic[74], assembled using Spades[75] into contigs, and aligned to reference tuberculosis reference genome H37Rv via minimap2 (ref. 76). All tuberculosis datasets used in this study are MSDs. To generate sample barcodes for input tuberculosis isolates, we use Pilon[77] to generate VCF files. Bcftools[78] is then used to pull all variants identified in the regions of interest for a particular drug (Supplementary Table 1). From the output of Bcftools, each variant is summarized as a mutational barcode or a string representation of the position and nucleic acid change that defines the mutation. Each isolate is then summarized with a sample barcode or a concatenation of the mutational barcodes present in the isolate. We collect nucleic acid sequences from the contigs mapped to the regions of interest (Supplementary Note 2). For the RIF resistance prediction task, we have 17,474 *M. tuberculosis* clinical isolates where 4,963 are resistant and 12,511 are susceptible. There are 3,998 unique sample barcodes and 2,066 unique mutational barcodes. For PZA, there are 12,146 isolates, where 2,166 are resistant and 9,980 are susceptible. There are 2,571 unique sample barcodes and 2,742 unique mutational barcodes. For INH, there are 26,574 isolates, where 10,580 are resistant and 15,994 susceptible. There are 4,952 unique sample barcodes and 4,455 unique mutational barcodes.

GFP dataset. GFP dataset is an MSD where amino acid sequences, sample barcodes and phenotypes are obtained from ref. 45. The dataset maps amino acid sequences of the GFP of the *A. victoria* jellyfish to a value representing the fluorescence of the GFP protein. This is a regressive task where fluorescence values are log-transformed and min–max normalized. GFP has 54,024 samples with 54,024 unique sample barcodes and 1,880 unique mutation barcodes (Supplementary Fig. 3). The performance metric is Spearman's rank correlation between predicted and experimentally measured fluorescence values.

SARS-CoV-2 dataset. SARS-CoV-2 dataset is an MSD where amino acid sequences, sample barcodes and phenotypes are obtained from ref. 44. The dataset maps mutations of the amino acid sequences for the RBD of the SARS-CoV-2 spike protein to a value representing vaccine escape. This phenotype is measured by exposing an RBD sequence to a series of human antibodies and reporting the proportion of RBD sequences bound by each antibody. The higher the proportion, the less the mutation in the RBD domain is associated with vaccine escape. We take the smallest bound proportion for each mutated sequence to generate labels, log-transform and min–max normalize the values. This is a regressive task with 438,046 samples with 22,341 unique sample barcodes and 2,391 unique mutation barcodes (Supplementary Fig. 4). The performance metric is Spearman's rank correlation between predicted and ground truth escape values.

PEER benchmark datasets. PEER[24] is a benchmark consisting of 17 tasks spanning five task categories (protein function prediction, protein localization prediction, protein structure prediction, protein–protein interaction prediction and protein–ligand interaction prediction). From PEER, we run SPECTRA on the subcellular localization dataset for the protein localization prediction task from ref. 79, a sequence-to-sequence dataset with 13,949 samples, which maps protein sequences to one of ten labels that present subcellular locations. This dataset is a classification task reporting per-label and/or class accuracy as a performance metric.

ProteinGym benchmark datasets. ProteinGym[16] is a benchmark consisting of 94 deep MSDs assessing the effect of mutations on measured protein properties. From ProteinGym, we run SPECTRA on the amyloid beta protein aggregation dataset. This dataset from ref. 80 maps mutations of the amyloid beta peptide that aggregates in Alzheimer's disease to an enrichment score reflecting the ability of the mutated peptide to aggregate in a cell-based selection assay. This is a regressive task with 14,483 sample barcodes. The performance metric is Spearman's rank correlation between predicted and ground truth assay readouts. We also run SPECTRA on the RNA recognition motif dataset. This dataset from ref. 81 maps mutations in the RNA recognition motif-2 domain of the *Saccharomyces cerevisiae* yeast poly(A)-binding protein (Pab1) to an enrichment score. This score is calculated by taking the proportion of yeast strains with a specific mutation in an input population before and after selection. This is a regressive task with 37,708 sample barcodes. The performance metric is Spearman's rank correlation between predicted and ground truth enrichment values.

TAPE benchmark dataset. TAPE[23] is a benchmark comprising five tasks (secondary structure prediction, contact prediction, remote homology detection, fluorescence landscape prediction and stability landscape prediction). We run SPECTRA on a dataset from ref. 82 for the remote homology detection task, which maps input protein sequences to one of 1,195 different fold classifications. This dataset is a sequence-to-sequence dataset with 16,291 samples and is a sequence classification task, reporting average accuracy across labels. We also run SPECTRA on a secondary structure dataset for the secondary structure prediction task from ref. 83, a sequence-to-sequence dataset with 11,411 samples, which maps proteins to one of three classes representing different secondary structures. This dataset is a classification task reporting per-label and/or class accuracy as a performance metric.

PDBBind dataset. The Protein Data Bank (PDB) bind (PDBBind) dataset[41] is a collection of protein–ligand complexes along with their binding affinities. This is a generative task where models generate a protein–ligand complex from a protein structure and a ligand SMILES structure. The performance metric is the root mean square error between the predicted and actual protein–ligand complex[34,84]. We test two splits of the PDBBind dataset, one from ref. 66 with 14,993 protein–ligand complexes and another from ref. 34 with 16,742 protein–ligand complexes. To run SPECTRA, we first download the dataset from Stärk et al. from the provided source and gather protein sequences via the PDB REST API and use Open Babel[85] to convert ligand Mol2 files to SMI files containing ligand SMILES structural fingerprint. We use the procedure outlined in Li et al. to calculate ligand similarity. We use BLAST[86] to calculate sequence similarity and Foldseek[87] to calculate structural similarity. We then construct a SPG where every node represents a protein–ligand pair. Edges are between nodes with proteins more than 30% similar or ligands more than 99% similar.

Astex diverse dataset. The Astex diverse dataset[88] is a collection of 85 crystallized protein–ligand pairs used to benchmark binding models. We download the Astex diverse set from the Cambridge Crystallographic Data Centre[89]. To extract ligand smile structures, we use Open Babel to convert ligand Mol2 files to SMI files containing ligand smile structures. To extract protein sequences, we create a custom Python parser to pull protein sequences from protein Mol2 files.

PoseBusters dataset. The PoseBusters dataset[65] is a collection of 428 protein–ligand pairs used to benchmark binding models. It was created to uncover model performance when tested on dissimilar protein–ligand pairs according to PDBBind. We download the PoseBusters dataset, gather protein sequences via the PDB REST API and use Open Babel[85] to convert ligand Mol2 files to SMI files containing ligand SMILES structural fingerprints.

Training models

This section outlines the inputs, architecture and training details of the machine learning models used in this study.

Model architectures and inputs. *Logistic regression*. Logistic regression architecture and training is based on ref. 90. This model uses one-hot encoded vectors to represent samples, where vector positions indicate the presence of a specific mutational barcode found in training samples. A logistic regression model then fits onto one-hot encoded vectors to predict the interest phenotype.

CNN. CNN architecture and training are based on ref. 1. This model uses one-hot encoded vectors to represent nucleic acid and amino acid sequences where vector positions indicate the presence or absence of a base pair or amino acid. The architecture is modified to take in unaligned sequences where sequences are padded to the length of the longest input sequence.

ESM2. The ESM2 pretrained model is from ref. 12 (650 million parameter version) and is used to generate protein embeddings for input sequences. We chunk up input sequences longer than 512 amino acids, embed each chunk and average the embeddings. If the dataset has multiple protein sequences as input, we embed each input protein sequence and average before prediction. We convert nucleic acid sequences to protein. We tokenize sequences by amino acid identity before input into ESM2. Once input sequence embeddings are obtained, we train a linear probe[91], a logistic regression model on input embeddings to predict phenotype. To finetune ESM2, we freeze the first 30 layers of ESM2 and replace the masked language head with a linear layer to predict phenotype. We then train the modified ESM2 to predict phenotype.

EVE. EVE architecture is from ref. 32. To construct the multiple sequence alignments (MSAs) necessary for EVE, we use Jackhmmr[92] to pull sequences from UniRep100 (ref. 57) similar to the wild-type sequence for the GFP protein and resistance binding domain of the SARS-CoV-2 spike protein. We then use Muscle to align pulled sequences and the codebase from EVCouplings[93] and EVE[32] to process and filter MSAs. We then train EVE with default suggested parameters. The input for EVE is the input sequence aligned with pulled sequences from UniRep100. EVE then returns a low-dimensional representation of the input MSA, which is then used to predict phenotype via a linear probe[91].

SeqDesign. SeqDesign architecture is from ref. 46. Seqdesign input processing is the same as EVE except as input it takes in raw unaligned sequences of the input sequence with pulled sequences from UniRep100.

GearNet training. GearNet architecture is from ref. 8. The GearNet model is a graph neural network that learns protein representations from the 3D structure of the protein. We use a pretrained GearNet model to generate embeddings using protein structures. Structures are generated from protein sequences using ESMFold[12]. The structures are then passed into the pretrained GearNet model to create embeddings of size 512. We generate output predictions from the graph-level embeddings by training a linear probe[91] to predict the phenotype of interest. A finetuned GearNet model is trained on each dataset. Both pretrained and finetuned Gearnet models are trained and evaluated on the GFP and the SARS-CoV-2 datasets.

Training details. We use suggested hyperparameters from source studies to train all models except otherwise noted. Our objective function for models trained on the GFP and SARS-CoV-2 datasets is mean absolute error; for the RIF, INH and PZA datasets, it is binary cross-entropy.

All models were trained on one Tesla A10 except ESM2-Finetuned, which was trained on four TeslaA100s on an Azure cluster. When applicable, we leverage weights and biases[94] to select optimal hyperparameters via a random search for each model over learning rate. All code is written in PyTorch[95].

Uncovering spectral properties in molecular datasets

Ultimately, choosing a spectral property should capture domain-specific knowledge about the MSPs learned by models during training. However, SPECTRA can detect whether there exists an unconsidered spectral property. This occurs if large variations exist in model performance in splits generated with the same SP or if there is a positive slope in the shape of the SPC (that is, model performance improves with decreasing cross-split overlap). In our study, we focus on diff-RRDR to explain the variance observed in the SPC of the CNN in the RIF resistance prediction task in *M. tuberculosis*. To calculate diff-RRDR for a train–test split, we identify all positions in each split where a mutation occurred in the RRDR of the RNA polymerase beta-subunit (*rpoB*) gene. diff-RRDR is determined by finding the difference between the maximum position observed in the train set compared to the test set and likewise for the minimum positions, then adding these differences together, as shown below:

$$\begin{aligned}\text{diff-RRDR} = &(\max(\text{position}_{\text{train}}) - \max(\text{position}_{\text{test}}))\\ &+(\min(\text{position}_{\text{train}}) - \min(\text{position}_{\text{test}}))\end{aligned}$$

Using SPECTRA to evaluate biological foundation models

Beyond the cross-split overlap in evaluation datasets, the cross-split overlap between pretraining and evaluation datasets influences model performance for foundation models. The protein foundation model we evaluated with SPECTRA, ESM2, is pretrained with UniRef50 (ref. 57) with more than 60 million clusters of sequences. Each cluster in UniRef50 has a representative sequence that is at least 50% similar to all cluster sequences. To understand what level of similarity between two sequences is significant in UniRef50, we sample 100,000 random pairs of representative sequences and calculate the distribution of average random pairwise similarity (Supplementary Fig. 12). Two sequences are similar if the sequence similarity, or the proportion of aligned positions in a pairwise alignment, is greater than two standard deviations above mean random pairwise similarity or a sequence similarity of 0.4.

Calculating the sequence similarity between a sequence of interest and UniRef50 representative sequences is computationally infeasible. However, most representative sequences will not be similar to an input sequence. By finding clusters with annotations similar to the protein encoded by the input sequence, we can select the clusters most similar to the input sequences. Once similar clusters are identified, we calculate sequence similarity between the input and representative sequences of selected clusters and count the number of clusters with sequence similarity greater than 0.4. The number of similar clusters represents the similarity of the input sequence to UniRef50. For tasks with multiple input proteins, we average this number across sequences. The names and similarities calculated for all sequences in this study can be found in Supplementary Tables 3 and 4.

Data availability

All data used in this study are publicly available. The data used for the RIF, INH and PZA datasets can be found in Green et al.[1]. The data used for the GFP dataset come from Sarkisyan et al.[45]. The data used for the SARS-CoV-2 dataset are from Greaney et al.[44]. All other datasets were directly downloaded from their benchmark of origin. All data are also available on the project GitHub at https://github.com/mims-harvard/SPECTRA and on Harvard Dataverse at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/W5UUNN (ref. 96).

Code availability

The code to reproduce results, along with documentation and usage examples, is available on GitHub at https://github.com/mims-harvard/SPECTRA (ref. 97).

References

1. Green, A. G. et al. A convolutional neural network highlights mutations relevant to antimicrobial resistance in mycobacterium tuberculosis. *Nat. Commun.* **13**, 3817 (2022).
2. Kumar, V., Deepak, A., Ranjan, A. & Prakash, A. Lite-SeqCNN: a light-weight deep CNN architecture for protein function prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **20**, 2242–2253 (2023).
3. Sanderson, T., Bileschi, M. L., Belanger, D. & Colwell, L. J. Proteinfer, deep neural networks for protein functional inference. *eLife* **12**, e80942 (2023).
4. Bileschi, M. L. et al. Using deep learning to annotate the protein universe. *Nat. Biotechnol.* **40**, 932–937 (2022).
5. Griffith, D. & Holehouse, A. S. Parrot is a flexible recurrent neural network framework for analysis of large protein datasets. *eLife* **10**, e70576 (2021).
6. Liu, X. Deep recurrent neural network for protein function prediction from sequence. Preprint at https://arxiv.org/abs/1701.08318 (2017).
7. Hill, S. T. et al. A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic Acids Res.* **46**, 8105–8113 (2018).
8. Zhang, Z., Xu, M., Jamasb, A. R., Chenthamarakshan, V., Lozano, A., Das, P. & Tang, J. Protein representation learning by geometric structure pretraining. In *Proc. Eleventh International Conference on Learning Representations* (2023).
9. Somnath, V. R., Bunne, C. & Krause, A. Multi-scale representation learning on proteins. In *Advances in Neural Information Processing Systems* Vol. 34 (eds Ranzato, M. et al.) 25244–25255 (Curran Associates, 2021).
10. Jha, K., Saha, S. & Singh, H. Prediction of protein–protein interaction using graph neural networks. *Sci. Rep.* **12**, 8360 (2022).
11. Gao, Z. et al. Hierarchical graph learning for protein–protein interaction. *Nat. Commun.* **14**, 1093 (2023).
12. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
13. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
14. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
15. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
16. Notin, P. et al. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *Proc. 39th International Conference on Machine Learning*, Vol. 162 (eds Chaudhuri, K. et al.) 16990–17017 (PMLR, 2022).
17. Wright, B. W., Yi, Z., Weissman, J. S. & Chen, J. The dark proteome: translation from noncanonical open reading frames. *Trends Cell Biol.* **32**, 243–258 (2022).
18. Liu, J. et al. Towards out-of-distribution generalization: a survey. Preprint at https://arxiv.org/abs/2108.13624 (2023).
19. Ye, H. et al. Towards a theoretical framework of out-of-distribution generalization. In *Advances in Neural Information Processing Systems* (eds Beygelzimer, A. et al.) 1801 (2021).

20. Weber, M. et al. Certifying out-of-domain generalization for blackbox functions. In *Proc. 39th International Conference on Machine Learning* Vol. 162 (eds Chaudhuri, K. et al.) 23527–23548 (PMLR, 2022).
21. Koh, P. W. et al. Wilds: a benchmark of in-the-wild distribution shifts. Preprint at https://arxiv.org/abs/2012.07421 (2021).
22. Liang, P. et al. Holistic evaluation of language models. *Trans. Mach. Learn. Res.* 2835–8856 (2023).
23. Rao, R. et al. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems* **32**, 9689–9701 (2019).
24. Xu, M. et al. PEER: a comprehensive and multi-task benchmark for protein sequence understanding. In *Proc. Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2022).
25. Capel, H. et al. Proteinglue multi-task benchmark suite for self-supervised protein modeling. *Sci. Rep.* **12**, 16047 (2022).
26. Dallago, C. et al. FLIP: benchmark tasks in fitness landscape inference for proteins. In *Proc. Neural Information Processing Systems Track on Datasets and Benchmarks* Vol. 1 (eds Vanschoren, J. & Yeung, S.) (2021).
27. Hu, Y., Jacob, J., Parker, G. J. M. et al. The challenges of deploying artificial intelligence models in a rapidly evolving pandemic. *Nat. Mach. Intell.* **2**, 298–300 (2020)
28. Grazioli, F. et al. On TCR binding predictors failing to generalize to unseen peptides. *Front. Immunol.* **13**, 1014256 (2022).
29. Azizi, S. et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nat. Biomed. Eng.* **7**, 756–779 (2023).
30. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
31. Chatterjee, A. et al. Improving the generalizability of protein-ligand binding predictions with AI-bind. *Nat. Commun.* **14**, 1989 (2023).
32. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **601**, E7 (2022).
33. Thadani, N. N. et al. Learning from prepandemic data to forecast viral escape. *Nature* **622**, 818–825 (2023).
34. Stark, H., Ganea, O.-E., Pattanaik, L., Barzilay, R. & Jaakkola, T. Equibind: geometric deep learning for drug binding structure prediction. In *Proc. 39th International Conference on Machine Learning* Vol. 162 (eds Chaudhuri, K. et al.) 20503–20521 (PMLR, 2022).
35. Mahajan, S. P., Ruffolo, J. A. & Gray, J. J. Contextual protein and antibody encodings from equivariant graph transformers. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/2023.07.15.549154v2 (2023).
36. AlQuraishi, M. ProteinNet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics* **20**, 311 (2019).
37. Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).
38. Kroll, A., Ranjan, S., Engqvist, M. K. M. & Lercher, M. J. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nat. Commun.* **14**, 2787 (2023).
39. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **23**, 40–55 (2022).
40. Jambrich, M. A., Tusnady, G. E. & Dobson, L. How AlphaFold2 shaped the structural coverage of the human transmembrane proteome. *Sci. Rep.* **13**, 20283 (2023).
41. Wang, R., Fang, X., Lu, Y., Yang, C.-Y. & Wang, S. The PDBBind database: methodologies and updates. *J. Med. Chem.* **48**, 4111–4119 (2005).
42. Krause, B., Lu, L., Murray, I. & Renals, S. Multiplicative LSTM for sequence modelling. Preprint at https://arxiv.org/abs/1609.07959 (2017).
43. Hou, J., Adhikari, B. & Cheng, J. DeepSF: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics* **34**, 1295–1303 (2017).
44. Greaney, A. J. et al. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* **29**, 44–57.e9 (2021).
45. Sarkisyan, K. S. et al. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
46. Shin, J.-E. et al. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* **12**, 2403 (2021).
47. Forrest, G. N. & Tamura, K. Rifampin combination therapy for nonmycobacterial infections. *Clin. Microbiol. Rev.* **23**, 14–34 (2010).
48. Goldstein, B. P. Resistance to rifampicin: a review. *J. Antibiot.* **67**, 625–630 (2014).
49. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* Vol. 30 (eds Guyon, I. et al.) (Curran Associates, 2017).
50. Cui, H., Wang, C., Maan, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
51. Ramesh, A. et al. Zero-shot text-to-image generation. In *Proc. 38th International Conference on Machine Learning* Vol. 139 (eds Meila, M. & Zhang, T.) 8821–8831 (PMLR, 2021).
52. Brown, T. B. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* Vol. 33 (eds Larochelle, H. et al.) 1877–1901 (Curran Associates, 2020).
53. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).
54. Anil, R. et al. *Palm 2 Technical Report*. Preprint at https://arxiv.org/abs/2305.10403 (2023).
55. Kedzierska, K. Z., Crawford, L., Amini, A. P. & Lu, A. X. Assessing the limits of zero-shot foundation models in single-cell biology. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.16.561085 (2023).
56. Chekroud, A. M. et al. Illusory generalizability of clinical prediction models. *Science* **383**, 164–167 (2024).
57. Consortium, T. U. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2022).
58. Anand, N. et al. Protein sequence design with a learned potential. *Nat. Commun.* **13**, 746 (2022).
59. Guo, Z., Liu, J., Wang, Y. et al. Diffusion models in bioinformatics and computational biology. *Nat. Rev. Bioeng.* **2**, 136–154 (2024).
60. Youssef, A. et al. Rapidai: a framework for rapidly deployable ai for novel disease and pandemic preparedness. Preprint at *medRxiv* https://doi.org/10.1101/2022.08.09.22278600 (2022).
61. Morselli Gysi, D. et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc. Natl Acad. Sci. USA* **118**, e2025581118 (2021).
62. Huang, K. et al. Artificial intelligence foundation for therapeutic science. *Nat. Chem. Biol.* **18**, 1033–1036 (2022).
63. Gainza, P. et al. De novo design of protein interactions with learned surface fingerprints. *Nature* **617**, 176–184 (2023).
64. Wong, F. et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature* **626**, 177–185 (2024).
65. Buttenschoen, M., Morris, G. M. & Deane, C. M. Posebusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Preprint at https://arxiv.org/abs/2308.05777 (2023).
66. Li, J. et al. Leak proof PDBBind: a reorganized dataset of protein-ligand complexes for more generalizable binding affinity prediction. Preprint at https://arxiv.org/abs/2308.09639 (2024).
67. Sanchez-Padilla, E. et al. Detection of drug-resistant tuberculosis by xpert MTB/RIF in Swaziland. *New Eng. J. Med.* **372**, 1181–1182 (2015).

68. Dias, A. L., Bustillo, L. & Rodrigues, T. Limitations of representation learning in small molecule property prediction. *Nat. Commun.* **14**, 6394 (2023).
69. Hsu, C. et al. Learning inverse folding from millions of predicted structures. In *Proc. 39th International Conference on Machine Learning* Vol. 162 (eds Chaudhuri, K. et al.) 8946–8970 (PMLR, 2022).
70. Pearson, W. R. An introduction to sequence similarity ('homology') searching. *Curr. Protoc. Bioinformatics* **Chapter 3**, 3.1.1–3.1.8 (2013).
71. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
72. Krivelevich, M., Mészáros, T., Michaeli, P. & Shikhelman, C. Greedy maximal independent sets via local limits. Preprint at https://arxiv.org/abs/1907.07216 (2023).
73. Karp, R. M. *Reducibility among Combinatorial Problems* (ed. Bohlinger, J. D.) 85–103 (Springer, 1972).
74. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
75. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
76. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
77. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
78. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
79. Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**, 3387–3395 (2017).
80. Seuma, M., Lehner, B. & Bolognesi, B. An atlas of amyloid aggregation: the impact of substitutions, insertions, deletions and truncations on amyloid beta fibril nucleation. *Nat. Commun.* **13**, 7084 (2022).
81. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(a)-binding protein. *RNA* **19**, 1537–1551 (2013).
82. Hou, J., Adhikari, B. & Cheng, J. DeepSF: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics* **34**, 1295–1303 (2018).
83. Klausen, M. S. et al. NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins* **87**, 520–527 (2019).
84. Corso, G., Stark, H., Jing, B., Barzilay, R. & Jaakkola, T. Diffdock: diffusion steps, twists, and turns for molecular docking. In *Proc. Eleventh International Conference on Learning Representations* (2023).
85. O'Boyle, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).
86. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
87. van Kempen, M. et al. Fast and accurate protein structure search with foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
88. Hartshorn, M. J. et al. Diverse, high-quality test set for the validation of proteinligand docking performance. *J. Med. Chem.* **50**, 726–741 (2007).
89. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B* **72**, 171–179 (2016).
90. Chen, M. L. et al. Beyond multidrug resistance: leveraging rare variants with machine and statistical learning models in *Mycobacterium tuberculosis* resistance prediction. *EBioMedicine* **43**, 356–369 (2019).
91. Alain, G. & Bengio, Y. Understanding intermediate layers using linear classifier probes. In *Proc. Fifth International Conference on Learning Representations* (2018).
92. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
93. Hopf, T. A. et al. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584 (2018).
94. Wandb (Weights & Biases, 2020); https://wandb.com
95. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. In *Proc. 33rd International Conference on Neural Information Processing Systems* 721 (Curran Associates, 2019).
96. Ektefaie, Y. SPECTRA. *Harvard Dataverse*, vol. V2 https://doi.org/10.7910/DVN/W5UUNN (2024).
97. Ektefaie, Y. SPECTRA (the spectral framework of model evaluation) v.1.0.3. *GitHub* https://github.com/mims-harvard/SPECTRA (2024).
98. Yu, F., Koltun, V. & Funkhouser, T. Dilated residual networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 636–644 (2017).
99. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins* **86**, 7–15 (2018).
100. Yang, Y. et al. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Brief. Bioinform.* **19**, 482–494 (2018).
101. Cuff, J. A. & Barton, G. J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* **34**, 508–519 (1999).
102. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).
103. Lu, W. et al. Tankbind: trigonometry-aware neural networks for drug-protein binding structure prediction. In *Advances in Neural Information Processing Systems* Vol. 35 (eds Koyejo, S. et al.) 7236–7249 (Curran Associates, 2022).

Acknowledgements

Y.E. is supported by grant no. T32 HG002295 from the National Human Genome Research Institute and the NSDEG fellowship. Y.E. and M.Z. gratefully acknowledge the support of grant nos. NIH R01-HD108794, NSF CAREER 2339524 and US DoD FA8702-15-D-0001, awards from Harvard Data Science Initiative, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, Roche Alliance with Distinguished Scientists, Sanofi iDEA-iTECH Award, Pfizer Research, Chan Zuckerberg Initiative, John and Virginia Kaneb Fellowship award at Harvard Medical School, Biswas Computational Biology Initiative in partnership with the Milken Institute and Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Author contributions

Y.E. retrieved and processed all data. Y.E. performed all analyses except the SPC for GearNet and GearNet-finetuned, which was performed by A.S. and for ESM2 for the SARS-CoV-2 dataset, which was performed by D.B. M.G.M. assisted with the processing of the Tuberculosis data for the INH, PZA and RIF datasets. Y.E., M.F. and M.Z. designed the study. All authors contributed to writing the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00931-6.

Correspondence and requests for materials should be addressed to Yasha Ektefaie, Marinka Zitnik or Maha Farhat.

Peer review information *Nature Machine Intelligence* thanks Shaik Waseem Vali, Meng Yang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [2]Department of Computer Science, Northwestern University, Evanston, IL, USA. [3]Department of Biological Sciences, Columbia University, New York, NY, USA. [4]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Allston, MA, USA. [5]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [6]Harvard Data Science Initiative, Cambridge, MA, USA. [7]Division of Pulmonary and Critical Care, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [8]These authors contributed equally: Marinka Zitnik, Maha Farhat. ✉e-mail: yasha_ektefaie@hms.harvard.edu; marinka@hms.harvard.edu; maha_farhat@hms.harvard.edu