Training data composition determines machine learning generalization and biological rule discovery

Eugen Ursu[a,b,1], Aygul Minnegalieva[c,1], Puneet Rawat[d], Maria Chernigovskaya[d], Robi Tacutu[a,b], Geir Kjetil Sandve[e], Philippe A. Robert[d,f], Victor Greiff[d,2]

Affiliations:
[a] Systems Biology of Aging Group, Institute of Biochemistry of the Romanian Academy, Bucharest, Romania, 060031
[b] Cellfabrik S.R.L., Bucharest, Romania, 060512
[c] Okinawa Institute of Science and Technology, Okinawa, Japan, 904-0497
[d] Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, Norway, 0313
[e] Department of Informatics, University of Oslo, Oslo, Norway, 0313
[f] Department of Biomedicine, University of Basel, Basel, Switzerland, 4001

[1] E.U. and A.M. contributed equally to this work.
[2] To whom correspondence may be addressed.

Email: victor.greiff@medisin.uio.no

Author Contributions: V.G. conceived the study. V.G., A.M., and E.U. designed analyses and visualizations. A.M., E.U. and P.R. performed analyses and visualizations. M.C. consulted on statistical matters. V.G., A.M., and E.U. wrote the first draft of the manuscript. All authors revised the manuscript and approved its contents. Given entirely equal contributions, A.M. and E.U. are free to list their respective names first in their respective CV. LLMs were used for text drafting.

Competing Interest Statement: V.G. declares advisory board positions in aiNET GmbH, Enpicom B.V, Absci, Omniscope, and Diagonal Therapeutics. V.G. is a consultant for Adaptive Biosystems, Specifica Inc, Roche/Genentech, immunai, LabGenius, and FairJourney Biologics.

Classification: Biological Sciences (major), Biophysics and Computational Biology (minor).

Keywords: antigen binding prediction, simulated and experimental data, robustness, out-of-distribution machine learning model generalization, biological rule discovery.

Abstract

Supervised machine learning models rely on training datasets with positive (target class) and negative examples. Therefore, the composition of the training dataset has a direct influence on model performance. Specifically, negative sample selection bias, concerning samples not representing the target class, presents challenges across a range of domains such as text classification and protein-protein interaction prediction. Machine-learning-based immunotherapeutics design is an increasingly important area of research, focusing on designing antibodies or T-cell receptors (TCRs) that can bind to their target molecules with high specificity and affinity. Given the biomedical importance of immunotherapeutics, there is a need to address the unresolved question of how negative training set composition impacts model generalization and biological rule discovery to enable rational and safe drug design. We set out to study this question in the context of the antibody-antigen binding prediction problem by varying the negative class, encompassing a binding affinity gradient. We based our investigation on large synthetic datasets that provide ground truth structure-based antibody-antigen binding data, allowing access to residue-wise binding energy on the binding interface. We found that both out-of-distribution generalization and binding rule discovery depended on the type of negative dataset used. Importantly, we discovered that a model's capacity to learn the binding rules of the positive dataset is not a trivial correlate of its classification accuracy. We confirmed our findings with real-world relevant experimental data. Our work highlights the importance of considering training dataset composition for achieving optimal out-of-distribution performance and rule learning in machine-learning-based research.

Significance Statement

The effectiveness of supervised machine learning models hinges on the composition of their training datasets, particularly the inclusion of negative examples. This bias in negative sample selection can greatly impact model performance. As the development of immunotherapeutic agents using machine learning is becoming increasingly crucial in biomedicine, understanding the impact of negative training set composition is imperative. Our study, focused on the antibody-antigen binding prediction problem, reveals that the choice of negative dataset significantly affects both out-of-distribution generalization and binding rule discovery across synthetic and experimental data. These findings underscore the necessity of carefully considering training dataset composition in machine-learning-driven biomedical research for optimal performance, robustness and meaningful rule acquisition.

Main Text

Introduction

Supervised machine learning and deep learning models require the collection or construction of training datasets, which are composed of positive and negative examples (samples). Training datasets represent a fundamental part of a machine learning (ML) model's success as it determines learned rules and prediction accuracy when applied to other tasks and domains (*out-of-distribution*, *OOD*) (Figure 1A) (1–5), and may introduce biases into the model's decision-making process (6, 7). Specifically, one important case of sample selection bias centers on negative samples, that is, samples that do not represent the target class. It should be noted that in binary settings, negative and positive class definitions are mostly not independent. Negative sample bias has been discussed in different domains and contexts, such as text classification (7–9), protein-protein interaction prediction (e.g., immune receptor-antigen prediction) (10–13), and antimicrobial peptide prediction (14). While numerous strategies to address negative data sampling and data imbalance have been proposed (15–17), there is still a notable lack of research on how the negative dataset quality impacts the ML model's outcome. More broadly, the influence of training set composition on generalization capacity and on the extraction of domain-specific knowledge (explainability) remains largely unexplored. In particular, within the sequence-level antibody-antigen binding problem, there is a wide array of qualitatively different negative datasets, offering an insightful terrain for evaluating the impact of negative data on model behavior.

ML methods recently have demonstrated high performance in predicting immune receptor (antibody and T-cell receptor (TCR)) features such as their binding capability as well as pharmaceutically important biophysical properties (18–24) and, therefore, have become a promising approach in speeding up the development of immunotherapeutics. Regarding immunotherapeutics design, previous studies have demonstrated that the choice of negative data can impact prediction accuracy and generalization in antibody-antigen binding (25–27) and TCR specificity prediction (13, 28–31). The influence of the negative dataset for antibody binding prediction was studied not only at the sequence level but also structurally, using simulated (27) and experimental data (25). Here, we explored the impact of training set composition on prediction accuracy and generalizability, with a particular emphasis on the discovery of binding rules (Figure 1B). We define learning of binding rules as identification of sequence-level binding energies and residue-level contributions to binding (energies).

The binding of immune receptors to their corresponding antigen is mainly determined by sequentially diverse complementarity-determining regions (CDRs) (32, 33). Particularly, the CDR3 regions present in the heavy chain (CDRH3) of antibodies or beta chain (CDR3β) of TCRs with an average length of ~16 amino acids contribute significantly to immune receptor-antigen binding (32–34). Therefore, in the current study, CDRH3 regions were used as a proxy to the full-length immune receptor sequences to train machine learning models. The three-dimensional interaction interface of the immune receptor-antigen complex is called a paratope on the immune receptor side and epitope on the antigen side. Existing state-of-the-art immune receptor-antigen binding databases contain $<10^4$ data points (35–40) with known sequence/structures but, mostly, without the binding affinity of the paratope or epitope. To address this gap, synthetic data has emerged as a viable solution to benchmark machine learning methodologies (1, 41, 42). Synthetic data enables unconstrained access to ground truth data, enabling arbitrary composition of training datasets and, thus, a more controlled analysis of generalization and rule discovery. We have previously developed the antibody-antigen binding simulation framework “Absolut!” (27), enabling the generation of synthetic antibody-antigen complex structures at an unconstrained scale, where it was observed that accuracy-based rankings of ML methods trained on experimental data hold for ML methods trained on Absolut!-generated data, for both sequence- and structure-based tasks. These findings established the real-world relevance of ML benchmarking studies performed on Absolut!-generated data (27).

Here, we utilized the Absolut! framework (27) to create synthetic training datasets with various negative data compositions mimicking those commonly used and most readily available in immune receptor-antigen prediction studies (i.e., binders versus non-binders) (Figure 1C). We trained simple neural networks on each type of dataset. Our results demonstrated that ML models generalized to unseen data of the same type. However, not all definitions of negative data in the training set led to models that could generalize to datasets containing different types of negative samples (Figure 1D,F). We also investigated the effect of dataset composition on the discovery of binding rules by using model explainability methods (namely DeepLIFT (43–46)) (Figure 1E). Explainability aims to reveal the reasoning underlying the model's predictions, including which features and characteristics of the data have the most influence on those predictions. Our analysis demonstrated that biological rule discovery was a function of training dataset composition. Importantly, biological rule discovery was not trivially associated with either in-distribution (*ID*) or out-of-distribution (*OOD*) model performance (Figure 1F). The conclusions of our work on synthetic data were also confirmed using recently published "real-world"-relevant experimental large-scale antibody sequence data that was affinity-labeled for the HER2 breast cancer antigen (47) (Figure 1C).

Lastly, since our work relies on sequence-based data, its relevance reaches beyond the antibody and TCR field, to any sequence-based classification (48). This is especially useful in the case of sequence-based classifications where a sequence to fitness mapping that is relevant for learning can be defined. Our findings suggest that in such cases it is crucial to carefully consider the composition of ML training datasets to extract biological knowledge from sequence data.

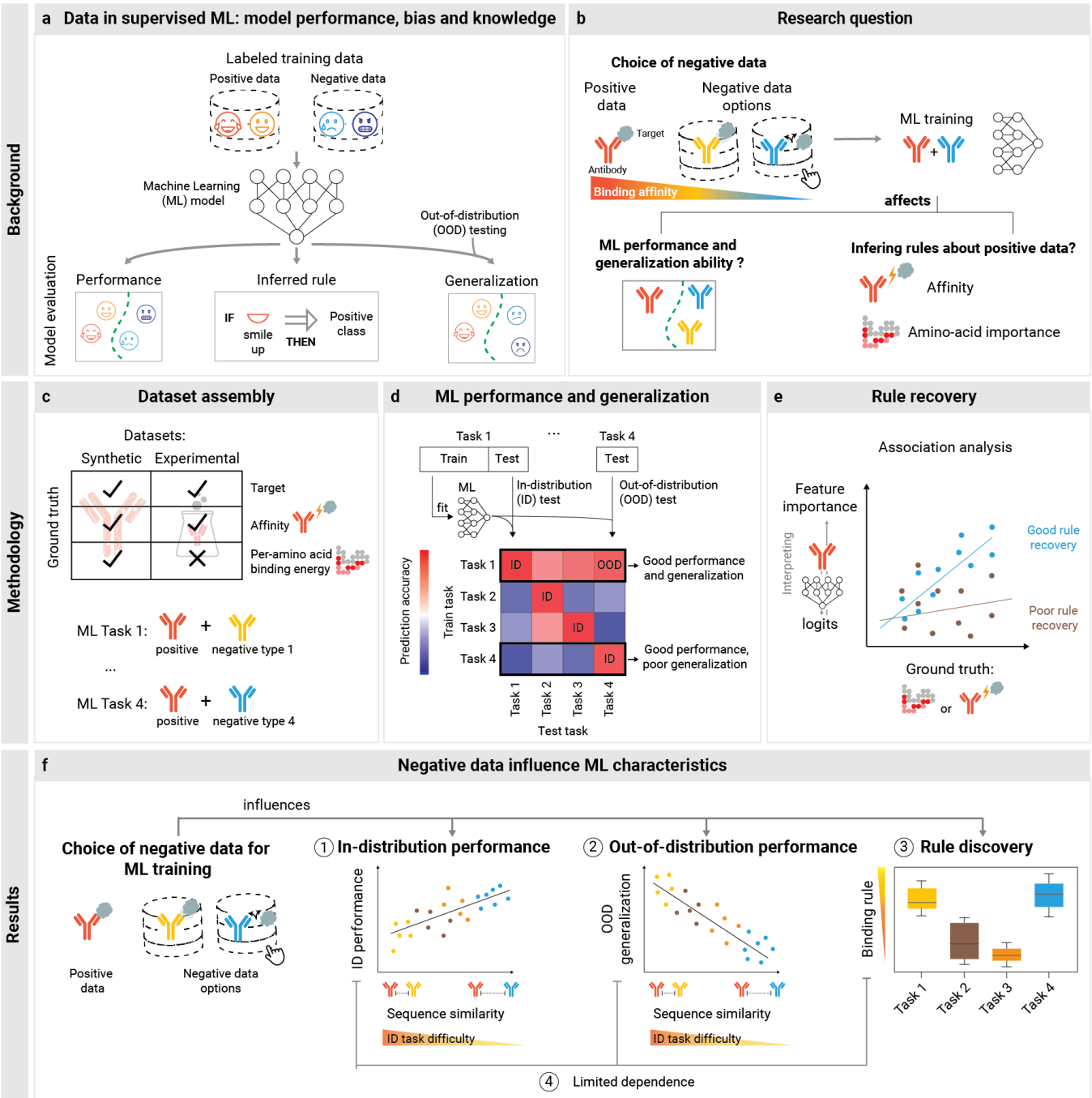


Figure 1. **Training data composition determines machine learning generalization and biological rule discovery. (a)** In supervised machine learning, particularly in binary classification, models are trained on labeled data divided into two categories: positive instances that align with the target concept and negative instances that do not. For illustration, positive emotion emojis represent the positive class, and all others represent the negative class. The standard way to evaluate an ML model is by testing it on a hold-out test dataset. Less frequently, the decision-making principles – rules – are examined, and even more rarely, the model's generalization is assessed on an independent dataset where the data distribution may vary, referred to as an out-of-distribution (*OOD*) test. Notably, in cases where negative data is abundant and diverse, ML models can face significant challenges when tested with real-world data, which may exhibit shifts in the distribution of negative data. **(b)** Our research question explores how the selection of negative data for training the ML model affects its performance, generalization and rules in the context of antibody-antigen binding prediction. Specifically, we aimed to understand how negative data influences the identification of features associated with positive class antibodies. **(c)** Using both synthetic and experimental datasets with accessible ground truth information on affinity and target specificity, we constructed various binary classification tasks. These tasks shared the same positive data, while the negative data varied based on either affinity or target specificity. **(d)** After training the ML model on these tasks, we evaluated its performance on a hold-out test dataset, where the negative data had the same distribution as in training (in-distribution test). We assessed the model's generalization ability by testing it on datasets with different distributions of negative data (out-of-distribution test). **(e)** The acquisition of affinity-related rules was estimated using explainability methods on trained models and comparing their outputs to ground truth information. **(f)** As a result, we found that the type of the negative dataset, and particularly its similarity to the positive dataset, correlates with both in-distribution performance and out-of-distribution generalization. Our results also

revealed that negative data can either enhance or restrict the learning of affinity-based rules. These capabilities must be evaluated separately, as the in-distribution and out-of-distribution performance provide limited insights non-trivially about each other and about rule discovery.

Results

1. Training dataset sequence composition influences prediction performance in ID and OOD binary classification tasks

1.1 Machine learning setup on synthetic and experimental data

We first investigated the influence of negative data composition on the performance and generalization of antibody-antigen binding predictors within the context of supervised binary classification. We define four distinct tasks and their corresponding datasets, with all datasets sharing the same positive CDRH3 sequences (binders), but with different types of negatives. These tasks broadly cover most scenarios of negative dataset definition (Table 1, Figure 2A). For a comprehensive evaluation of training dataset composition on ML performance, we employed synthetic CDRH3 sequences annotated with their binding energy to each of ten different antigens generated using the "Absolut!" framework (27). For each antigen, each of the four datasets shares the top 1% high-affinity percentile CDRH3 sequences as positives (Table 1). (i) The first dataset, "vs Non-binder", contains the bottom 95% of low-affinity sequences to the same antigen, used as negative sequences; (ii) The second dataset, "vs Weak", uses antibodies in the weak affinity range, from the top 1% to 5% percentile, as negatives. The "vs Non-binder" and "vs Weak" tasks are disjoint. (iii) The third dataset, "vs 1", contains CDRH3 sequences that do not bind to the positive class antigen but that bind (top 1% affinity) to another predefined antigen. (iv) The "vs 9" dataset extends the "vs 1" by having binders to any other of the 9 antigens in equal amounts. Of note, datasets are balanced with 20 000 positive and 20 000 negative sequences, among which 10 000 are held-out as test data. We named the associated machine learning tasks by the name of the antigen (defining the positives) and the name of the dataset.

Table 1. **Binary dataset composition and generation for binding prediction in adaptive immune receptors**

Dataset type	Positive dataset	Negative dataset	Real-world relevance
vs Non-binder	Binder sequences with high affinity to the target antigen (range: 0–1% percentiles).	Sequences with low to no affinity (5–100% percentile) to positive class antigen – non-binders.	A common type of experimental dataset where binders and non-binders are identified by testing against a specific antigen and separated by a margin of binding affinity, ensuring that non-binders exhibit no considerable levels of binding. This dataset can be generated from experiments using fluorescent sorting or display selection. Alternatively, it can be acquired from broad-spectrum sequencing experiments without specificity information. In the TCR field it is similar to "negatives from background data" (13).
vs Weak		Sequences with moderate affinity (1–5% percentile) to positive class antigen – weak binders.	A dataset in which positive and negative data are close in affinity. It can be acquired through the same selection experiments mentioned above but by selecting adjacent bins along the binding affinity gradient.
vs 1		Sequences with high (affinity) binders to another antigen (range: 0-1% percentiles to another antigen).	A dataset consisting only of high-affinity binders to two antigens. Easily assemblable from public databases. In the TCR field, it is similar to "negatives by shuffling" where the same epitopes and TCR from positive binding pairs are reused, but each TCR is paired with a different epitope (13).
vs 9		Sequences from high (affinity) binders of 9	Assembled from individual binders to 10 antigens. It requires less data from each antigen and presumably

		other antigens.	covers a wider landscape of negative data. Also analogous to "negatives by shuffling" in the context of TCRs (13).

On each dataset, we trained a fully connected neural network with 10 hidden neurons and one neuron in the final layer (the model is hereafter termed SN10, see *Methods*). We utilized this relatively simple network architecture in our ML tasks as we previously showed that it outperforms classical ML models such as SVM and logistic regression, and more complex architectures such as convolutional neural networks or transformers do not yield any significant improvement in performance in similar ML tasks within Absolut! (27). To ensure the robustness of all results, the training process was replicated by varying both the train-test split composition and seed (influencing the model's weights initialization, see Methods) (Figure 2A).

We repeated the analysis on synthetic data using recently published experimental data by Porebski et al. (47) that included affinity-labeled binding data for 24,790 CDRH3 sequences against HER2, further referred to as "the experimental dataset" (see Methods). In short, the authors developed an experimental method, which enables annotation of antibody sequence data with affinity-like information, replicating the data structure of the synthetic data we used.

1.2 ID prediction accuracy is a function of training dataset composition

To evaluate the capacity of trained models to learn generalizable rules, we compared their performance on *ID* test datasets, with the same positive/negative distribution as used in training (Figure 2A). Across all four types of *ID* test datasets, trained models achieved median prediction accuracies of over 0.85. Notably, we observed that the prediction accuracy decreased from the "vs Non-binder" task (range: 0.97–1.00, median: 0.99) to the "vs 1" task (range: 0.94–1.00, median: 0.98), followed by the "vs 9" (range: 0.91–0.98, median: 0.94), and finally the "vs Weak" task (range: 0.85–0.98, median: 0.92) (Figure 2B). Thus, the "vs Weak" task was the most challenging to achieve high accuracy on. Performance on shuffled control datasets where sequence labels were randomly rearranged reached as expected 0.5 (Supp. Table 1). Thus, there are learnable CDRH3-based patterns that allow reaching high accuracies, which are lost due to label shuffling. In addition, we asked whether the SN10 model offers any advantage over logistic regression models (LR) and found that SN10 is on average 12% more accurate in more difficult tasks, such as "vs Weak" (Supp. Figure 3C). This suggests that SN10 might exploit inter-positional dependencies, which are not covered by LR. Overall, SN10 and LR show a similar trend of performance across task types, with decreasing performance from "vs Non-binder" to "vs 1", "vs 9" and "vs Weak".

1.3 The distance between positive and negative datasets is linked to ID prediction accuracy

Given the gradation in *ID* performance across tasks, we asked whether the observed decrease in *ID* prediction accuracy may be attributed to the sequence dissimilarity between the positive and negative training datasets. We calculated the Jensen-Shannon distance (JSD) between the position-weight matrices (PWM) of the positive and negative class sequences (Figure 2A). The JSD exhibited an increasing trend from "vs Weak" to "vs 9", then "vs 1" and "vs Non-binder", suggesting increasing sequence divergence in this particular ordering. A corresponding increase in prediction accuracy accompanied this. Notably, the JSD between the positive and negative class sets demonstrated a stronger and significant ($p<0.05$) association with prediction accuracy for the "vs 9" ($r=0.94$), "vs 1" ($r=0.73$), and "vs Non-binder" ($r=0.77$) tasks, however not for the "vs Weak" task ($r=0.30$, p-value ≥ 0.05) (Figure 2C) and not for the shuffled controls (p-value ≥ 0.05 for all negative controls across each task type, Supp. Table 1). The trend of increased prediction accuracy with higher JSD values remained largely consistent across all antigens (Supp. Figure 3A). We excluded that the diversity of the overall training dataset, quantified by Shannon entropy, was significantly associated with *ID* prediction accuracy (p-value ≥ 0.05) (Supp. Figure 1C). These results show that the SN10 model is more effective at discerning positive and negative training data when

their sequence PWM-based Jensen-Shannon distance is greater. Furthermore, this suggests that binding prediction tasks with higher dissimilarity between positives and negatives are easier to learn when evaluated in an *ID* context.

1.4 OOD prediction accuracy is a function of training dataset composition

Having established that *ID* predictive performance depends on the task, we set out to investigate the *OOD* performance of trained models by taking models trained on one task and testing them on tasks types different then those of the training ones. For example, a model trained on the "vs 9" task would be tested using the "vs Non-binder" test data (Figure 2D, Supp. Figure 3D). Expectedly, *OOD* performance of models were lower compared to the respective *ID* performance. The difference was highest for models trained on "vs Non-binder", from 0.99 *ID* to 0.82 on the "vs 1" *OOD* task, 0.81 on the "vs 9" *OOD* and 0.72 on the "vs Weak" *OOD* tasks (Supp. Figure 3D). The models trained on the "vs 9" task demonstrated the best *OOD* performance on the "vs 1" task, achieving an accuracy of 0.94. This result may be explained by the fact that "vs 9" includes binders to the same antigen as the "vs 1" task. Also, "vs 9" models exhibited generalization to the "vs Non-binder" separation task for 50% of antigens, resulting in a median accuracy of 0.91. We observed a high variation in the "vs 1" *OOD* performance depending on which other antigen was used to draw the negative sequences (Supp. Figure 5). Generally, models trained on "vs 1" task demonstrated better *OOD* performance on the "vs 9" task, with an accuracy of 0.78, but performed relatively worse on the "vs Non-binder" task accuracy of 0.72 (Supp. Figure 3D). The "vs Weak" task proved to be the most difficult *OOD* test task, as indicated by a median accuracy of 0.58 for models trained on "vs 1", and 0.71 for models trained on "vs Non-binders" and "vs 9". Importantly, models trained on the "vs Weak" task demonstrated the best overall generalization in the considered *OOD* tasks, with median prediction accuracies of 0.96 on the "vs Non-binder" task, 0.91 on the "vs 1" task, and 0.90 on the "vs 9" task although accuracies for certain antigens declined to 0.80 (Figure 2D). When testing *OOD* generalization of logistic regression models, the same trend held, but overall, SN10 outperformed logistic regression in terms of *OOD* generalization (Supp. Figure 3E). A negative control with shuffled labels ensured results were not due to chance: SN10 *OOD* on controls dropped to an accuracy of 0.5, (Supp. Table 2).

We did not find a correlation between *ID* and *OOD* performance when considering all tasks together. However, when investigating *OOD* performances per task type, we observed associations between *ID* and *OOD* performance for "vs Weak" (r range: 0.8–0.98, p-value < 0.05) and "vs Non-binder" (r range: 0.37–0.57, p-value ≥ 0.05), however not for "vs 9" (r range: 0.09–0.11, p-value ≥ 0.05) (Supp. Figure 4A).

In summary, dataset composition impacts the generalization capacity of models. Importantly, *ID* performance is not a reliable predictor of generalizability. This is exemplified by the "vs Non-binder" task. It yielded high *ID* prediction accuracies, but models trained on it yielded the lowest *OOD* accuracies. In contrast, the "vs Weak" task yielded the lowest *ID* accuracies but the best generalization ability.

1.5 OOD performance is associated with positive-negative similarity in training datasets

Next, we asked whether a stronger generalization ability of trained models to *OOD* tasks could be explained by a larger distance between the positive and negative datasets between the *ID* and *OOD* settings. To answer this question, we examined the correlation between the change in the distances between negative and positive datasets between the *ID* and corresponding *OOD* tasks, measured as the gap in Jensen-Shannon distance (ΔJSD=JSD(*ID*)-JSD(*OOD*)), and the corresponding performance gap in prediction accuracy between *OOD* and *ID* (Δacc=acc(*ID*)-acc(*OOD*)) (Figure 2E). We observed that models trained on "vs Weak" task had negative ΔJSD, which means that positive and negative data in the *OOD* test tasks were more distant compared to the *ID* test task ("vs Weak"). "Vs Weak" models demonstrated relatively stable accuracy, and no correlation was observed between ΔJSD and Δacc ($r = 0.13$, p-value ≥ 0.05). Approximately half (52%) of Δacc values were less than or equal to zero. For the other training tasks ("vs Non-binder", "vs 1", "vs 9"), we observed a significant correlation between ΔJSD and Δacc ranging from 0.47 to 0.70 (p-value < 0.05). This indicates that if the positive and negative data in

the training (and *ID* test) dataset are more distant than in the *OOD* test dataset, the *OOD* performance is likely to decrease. The extent of this decrease depends on the difference in the distance between positive and negative data in *ID* and *OOD* tests. However, the regression between ΔJSD and Δacc explained only 0.21–0.34 of the variance in the data, which may be partially due to the heterogeneity between antigens (Supp. Figure 3B) but also suggests the existence of additional factors that contribute to model robustness. The shuffled controls corresponding to the *OOD* setting showed no significant correlation between ΔJSD and Δacc (Supp. Table 2).

In summary, when models are trained on tasks with more distinct positive and negative datasets – as indicated by a larger JSD – they exhibit a decision boundary that, while effective within the *ID* context, may not be ideally configured for *OOD* tasks where the data distributions are closer. Conversely, training models on datasets characterized by lower JSD, where positive and negative data are inherently more similar, appear to foster a more adaptable and tighter decision boundary, leading to better generalization (*OOD* performance).

1.6 Experimental dataset replicates synthetic data findings on ID and OOD performance

We showed that negative data composition affects *ID* and *OOD* (Figure 2B,D) in synthetic datasets. To evaluate the persistence of this key finding in experimental data, we leveraged a recently published experimental dataset (47), which resembles the "Absolut!" dataset as both contain CDRH3 sequences annotated with an affinity proxy (see Methods). We generated “vs Non-binder” and “vs Weak” tasks from the experimental dataset following the same strategy as in the synthetic data. Additionally, we created a “vs Randomized” task by randomly reordering the amino acids within the binder sequences while maintaining their global amino-acid composition. Then, we analyzed the *ID* test-set performance and the *OOD* generalization performance of SN10 in these experimental datasets (Figure 2F). Although the model trained on the “vs Randomized” dataset achieved 100% accuracy on its test set, in *OOD* scenarios, it failed to discriminate between sequences (50% on “vs Non-binder”, 50% on “vs Weak”). Models trained on “vs Weak” and on “vs Non-binder” performed well on “vs Randomized”, with “vs Weak” achieving 95% accuracy compared to “vs Non-binder” (100%). Models trained on “vs Non-binder” achieved 98% on its test sample, but could not perform well on “vs Weak” (68%). However, training on “vs Weak” resulted in comparable performance on both its test set (91%) and “vs Non-binder” (88%). Therefore, the key finding on synthetic datasets, which showed that models trained on “vs Weak” tasks generalized best to *OOD* tasks, was recapitulated on the experimental data.

To summarize, our results indicate that the composition of the negative dataset can impact predictive performance and generalization in binary classification tasks in synthetic and experimental data. No direct definitive association between *ID* and *OOD* was observed (Supp. Figure 4A). Specifically, we found that predictive performance for *ID* and generalization capacity for *OOD* was associated with the similarity between positive and negative datasets.

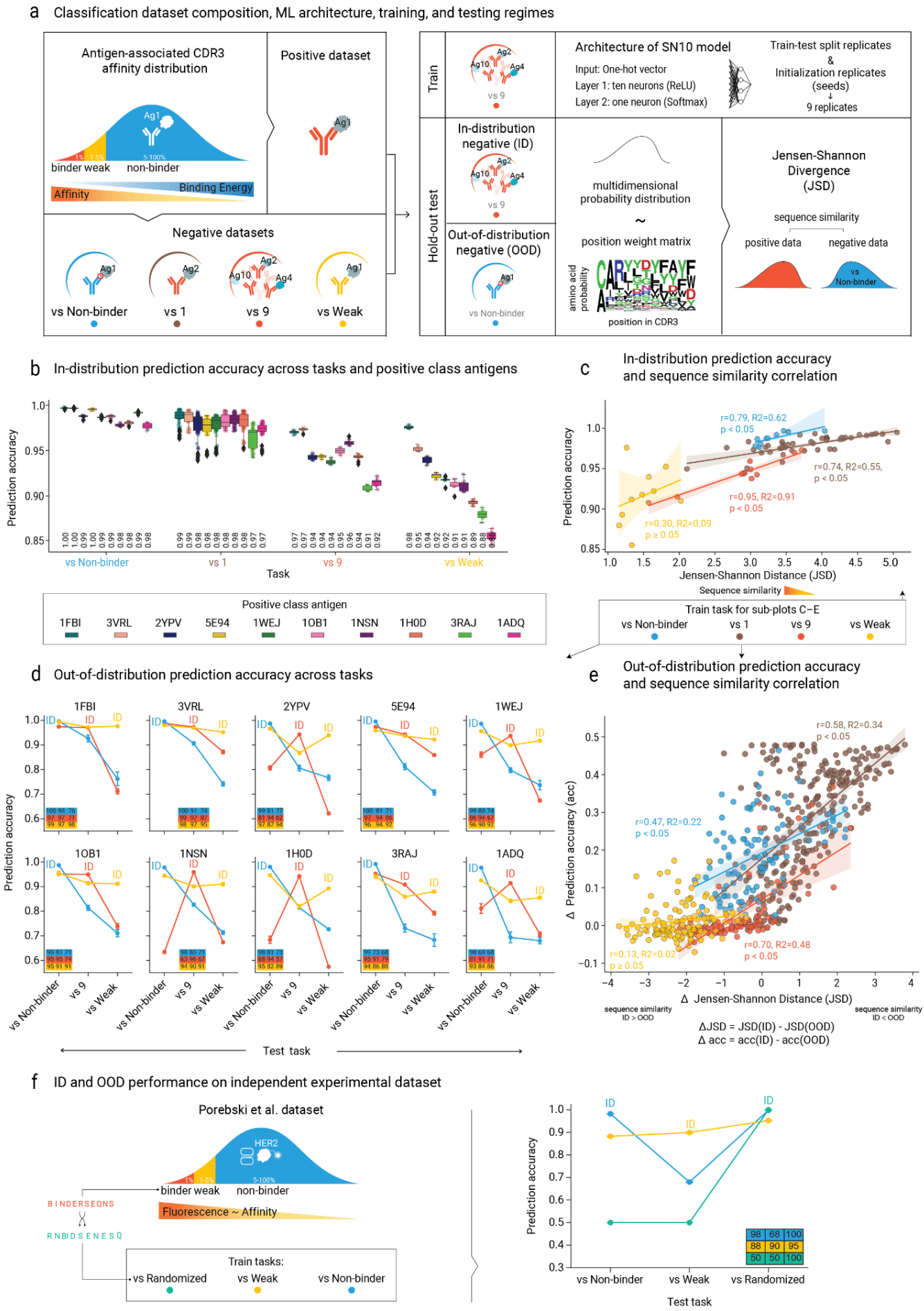

a Classification dataset composition, ML architecture, training, and testing regimes
Antigen-associated CDR3 affinity distribution
Positive dataset
binder weak
non-binder
Affinity
Binding Energy
Negative datasets
vs Non-binder
vs 1
vs 9
vs Weak
Train
Hold-out test
In-distribution negative (ID)
Out-of-distribution negative (OOD)
Architecture of SN10 model
Input: One-hot vector
Layer 1: ten neurons (ReLU)
Layer 2: one neuron (Softmax)
Train-test split replicates & Initialization replicates (seeds)
9 replicates
multidimensional probability distribution
position weight matrix
amino acid probability
position in CDR3
Jensen-Shannon Divergence (JSD)
sequence similarity
positive data
negative data
b In-distribution prediction accuracy across tasks and positive class antigens
Prediction accuracy
Task
Positive class antigen
1FBI
3VRL
2YPV
5E94
1WEJ
1OB1
1NSN
1H0D
3RAJ
1ADQ
c In-distribution prediction accuracy and sequence similarity correlation
r=0.79, R2=0.62
p < 0.05
r=0.74, R2=0.55,
p < 0.05
r=0.95, R2=0.91
p < 0.05
r=0.30, R2=0.09
p ≥ 0.05
Jensen-Shannon Distance (JSD)
Sequence similarity
Train task for sub-plots C–E
d Out-of-distribution prediction accuracy across tasks
Test task
e Out-of-distribution prediction accuracy and sequence similarity correlation
r=0.58, R2=0.34
p < 0.05
r=0.47, R2=0.22
p < 0.05
r=0.13, R2=0.02
p ≥ 0.05
r=0.70, R2=0.48
p < 0.05
Δ Prediction accuracy (acc)
Δ Jensen-Shannon Distance (JSD)
sequence similarity ID > OOD
sequence similarity ID < OOD
ΔJSD = JSD(ID) - JSD(OOD)
Δ acc = acc(ID) - acc(OOD)
f ID and OOD performance on independent experimental dataset
Porebski et al. dataset
HER2
binder weak
non-binder
BINDERSEQNS
RNBIDSENESQ
Fluorescence ~ Affinity
Train tasks:
vs Randomized
vs Weak
vs Non-binder
Test task

Figure 2. **Classification performance varies across antigens, binding prediction tasks and the similarity between positive and negative data.** **(a)** Antibodies related to a specific antigen can be categorized into affinity groups based on their binding energy: Binder, Weak, and Non-binder. The classification datasets are composed of a positive class of Binders to the target antigen and a negative class containing sequences that either bind to other antigen(s) ("vs 1" and "vs 9"), exhibit Weak binding ("vs Weak"), or are Non-binders ("vs Non-binder") with respect to the same antigen (see Methods for details, Table 1). Low binding energy translates to high affinity and vice-versa. Each dataset was divided into train and test sets in five different ways (split replicates). We trained a machine learning model, SN10, consisting of a single layer with 10 neurons, employing four different initializations for one of the splits (seed replicates). To optimize computation time, we worked with nine replicates per task, which included one seed replicate combined with all split replicates and all seed replicates with one fixed split replicate. Evaluation was performed on the same type of task used during training – referred to as *in-distribution* (*ID*) testing, as well as on different tasks, termed *out-of-distribution* (*OOD*) testing. Additionally, we quantified the dissimilarity between the positive and negative classes in the test datasets using Jensen-Shannon distance (JSD) between the position-weight matrices of the corresponding sequences. **(b)** The *in-distribution* (*ID*) prediction accuracy of the SN10 model is grouped by negative dataset type and colored by positive class antigen. The boxes represent data between the 25th and 75th percentile. Numbers displayed indicate median prediction accuracies between replicates. **(c)** Relationship between the *ID* performance and the dissimilarity of the positive class and negative class sets of CDRH3 sequences (JSD). There are 10 data points for "vs 9", "vs Non-binder" and "vs Weak" tasks and 45 points for "vs 1" task, each representing the average among replicates for respective combinations of positive and negative classes. The coefficient of determination (R2) and the Pearson correlation coefficient (r) were used to assess the relationship between Jensen-Shannon distance and prediction accuracy. Two-sided t-test p-values were below 0.05 for all tasks except for the 'vs Weak' task. **(d)** *OOD* test performance. Each subplot corresponds to a positive class antigen. The hue indicates the training task (indicated by the "ID" label as well), and the x-axis reflects the test task. The plot includes tables that display the precise mean prediction accuracies. Error bars represent one standard deviation from the mean, capturing 68% of the variation in *OOD* performances. **(e)** The relationship between the change in Jensen-Shannon distance (JSD) values and the corresponding change in prediction accuracy (acc) between *ID* and *OOD* tasks. The ΔJSD and Δacc are calculated by taking the corresponding values for the *ID* test dataset and subtracting *OOD* test dataset values. The analysis includes 110 data points for the "vs 9", "vs Non-binder", and "vs Weak" tasks, and 270 points for the "vs 1" task. Two-sided t-test p-values were below 0.05 for all but the "vs Weak" task. **(f)** Results on Porebski et al. experimental dataset (47). The experimental dataset consists of CDRH3 sequences annotated with binding affinity, conceptually analogous to the synthetic dataset. The left panel illustrates the dataset generation strategy: the "vs Weak" and "vs Non-binder" datasets are defined identically to the Absolut! dataset (subplot **a**), while the "vs Randomized" dataset was created by randomly reordering the amino acids in the 'Binder' sequences. The right panel shows the SN10 test prediction accuracy for these datasets, where color denotes the training task and the x-axis represents the test task (similarly to subplot **d**). Points are marked as 'ID' to indicate *in-distribution* performance; all other points represent *OOD* performance. Error bars signify one standard deviation from the mean. The inset table numerically presents mean *ID* and *OOD* prediction accuracies.

2. Training dataset composition determines the accuracy of biological rule recovery

Having shown that training negative data composition influences prediction accuracy and generalization on supervised binary sequence-based ML tasks, we set out to investigate whether negative data composition also impacts the capability to identify antibody-antigen binding rules. In this work, we define binding rules as whole-sequence (per-sequence) binding energies or per-residue amino-acid-based contributions to binding energies.

2.1 Training dataset composition impacts the learning of CDRH3 binding energy

First, we investigated whether trained models have learned the energy properties of sequences beyond just classifying them as positives and negatives. To investigate to what extent the SN10 model could recover a map from sequence to corresponding binding energy, we quantified the correlation of output logits and binding energy on a per sequence level. Output logits are raw, pre-sigmoidal activation scores, measuring the model's confidence in predicting the positive class (Figure 3A,B). We found that the Pearson correlation of output logits and binding energy varied by the task type (one-way ANOVA, p-value=8.7e-58) and positive antigen type (one-way ANOVA, p-value=4.6e-25) (Figure 3B). An example 2D distribution per task is provided (3VRL “vs Weak” task, Figure 3B inset). Secondly, the task types “vs Weak” (range: -0.33 to -0.90, median: -0.62) and “vs Non-binder” (range: -0.29 to -0.85, median: -0.51) exhibited notably stronger correlations than “vs 1” and “vs 9”, overall, with no significant correlations being observed for the shuffled controls (Supp. Table 3). This implies that both the antigen itself and the task type influence the model's ability to learn binding energies. Furthermore, we quantified how the dependence of logit and binding energy correlates with *ID* prediction accuracy (Figure 3C) and found significant correlations only for “vs Non-binder” (r=-0.83) and for “vs Weak” (r=-0.77). Prediction accuracy in *OOD* tasks appeared however to be significantly associated with logit-based rule discovery in all cases, except when training on “vs Non-binder” and testing on “vs Weak” (Supp. Figure 4B).

Generally, in identifying binding or non-binding antibody-antigen pairs, a fundamental criterion relies on establishing a threshold on the binding energy, which is intrinsically linked to the binding probability, thus binarising sequences into a positive and negative class. The logit-output correlation with binding energy demonstrates the model's ability to capture the relationship between sequence and binding energy across a range of binding energies, thereby indicating its capacity to acquire meaningful knowledge about the binding process beyond the somewhat arbitrary definition of positive and negative class. The efficiency of such rule recovery depends on the choice of the negative class.

2.2 Training dataset composition impacts learning of position-wise contribution to the binding energy

Having established that the model implicitly acquired knowledge about binding energies for specific tasks and antigens on the sequence level, we further aimed to investigate the model's ability to acquire knowledge regarding the contribution of individual CDRH3 amino acids to binding — a crucial aspect of the binding process. To explore learning on the residue level, we obtained the energy contribution per amino acid to the binding energy from the Absolut! simulation suite (27). Our hypothesis was that if the model successfully acquired binding rules as amino acid contributions to binding, there would be associations between the trained model's attention towards specific amino acids and their contributions to binding energies. For the SN10 model, we analyzed attribution values per residue to measure the model's attention towards specific amino acids using an explainable deep learning technique called DeepLIFT (46). The correlations with per-residue energies are computed for positive class CDRH3 sequences within each dataset (Figure 3D). An example 2D distribution per task is provided (5E94 “vs 9” task, Figure 3D inset). Consistent with the learning of per-sequence binding energies, which were presented in the previous section, the “vs Weak” with -0.69 and “vs Non-binder” with -0.71 tasks exhibited stronger correlations than “vs 1” with -0.01 and “vs 9” with -0.36, showing thus a varying degree of correlation across task types (one-way ANOVA, p-value=3.5e-43), while the negative controls exhibited no significant correlations (Supp. Table

3). Overall, the model's capacity to learn binding energies on a per-residue basis (attribution – per-residue energy, e.g. “vs Non-binder”: median -0.71) appeared to be more effectively acquired compared to on a per-sequence basis (logits – per-sequence energy, e.g. “vs Non-binder”: median -0.51). We found no significant association between the prediction accuracy and the correlation between per-residue energy and attributions, across all datasets (Figure 3E), providing another example of a negative dataset-dependent property of a trained network that is not significantly affected by the *ID* prediction performance. Although no significant relation between *ID* prediction accuracy and attribution-based rule discovery could be established (Figure 3E), we found that *OOD* prediction accuracy was associated with attribution-based rule discovery for all “vs 1” *OOD* scenarios and for all “vs 9” *OOD* tasks except “vs 9” trained models tested on “vs 1” (Supp. Figure 4B). This relation is influenced by both the choice of positive dataset (i.e. the antigen), as well as the negative dataset type.

We noticed, however, an asymmetry in the case of “vs 1”, namely, the correlations with energies, both per-sequence and per-residue, were strongly inversely related in the “vs 1” datasets upon switching the role of the positive and negative antigen, e.g. attribution – per-residue energy correlation for 5E94 vs 2YPV (r=-0.8) and 2YPV vs 5E94 (r=0.78) (Supp. Figure 6A,C). The allocation of positive and negative status to the two antigens involved in a “vs 1” task does not impact performance but it does critically influence the sign of the correlation with the energy of the positive dataset, and not so much the magnitude (Supp. Figure 6A, C). This could be a bias in the learning process of SN10, in which the model tends to learn binding rules towards the sequences of the same antigen, irrespective of the allocation to the positive or negative dataset, e.g. SN10 might be biased to 5E94, both in the cases of 5E94 vs 2YPV and of 2YPV vs 5E94 (in the latter leading to a strong positive correlation, higher logits corresponding to higher binding energy to 5E94 of the 2YPV sequences, rather than to lower binding energy to 2YPV). To account for this effect, we picked from each pair the dataset that led to the strongest negative correlation between energy and logits or attribution, and after this correction (represented by “vs 1” *(adapted)*) we observed the associations for “vs 1” *(adapted)* to be comparable to “vs 9” (energy-logit median r=-0.26, energy-attribution median r=-0.4, Supp. Figure 6B,D).

When comparing the attributions acquired in the different task types on the same positive dataset, we generally observed high correlations between the attributions of “vs Weak” with “vs Non-binder”, on one hand, and “vs 1” with “vs 9”, on the other hand, although we could also find high correlations between all task types for some antigens (Supp. Figure 8A, B). This shows that some tasks are more similar in terms of the rules they induce in the trained model.

2.3 The majority of binding rules recovered are linear in nature

To obtain further insight into rule discovery by SN10, we investigated the associations with energies of weights of logistic regression models trained in identical settings on synthetic data, as baselines for comparisons. Logistic regression is fundamentally SN10 without the middle hidden layer. The weights of the logistic model can be similarly interpreted as the attributions of SN10, and namely as contributions to the logits. We showed that the correlations per-sequence (logits – total energy) and per-residue (weights – per-residue energies) are similar in magnitude to the ones based on SN10 (Supp. Figure 7A,B), e.g., median per-sequence correlation for “vs Weak” in SN10 (-0.62) is similar in logistic regression (r=-0.58). Moreover, the correlation between the SN10 attributions and the weights of the logistic regression model is, on average, high across the tasks and antigens (Supp. Figure 7C), with median correlations range 0.6–0.84 across all considered task types. The high correlations suggest that rules discovered by SN10 and revealed by the explainability methods used ((46), see *Methods*) are predominantly driven by linear combinations of features, and thus discovered by logistic regression as well. However, the higher SN10 performance in certain task types (Supp. Figure 3C) suggests that SN10 might also capture nonlinear rules.

2.4 Training dataset composition drives per sequence binding energy recovery from experimental data

We investigated whether the rule discovery conclusions on synthetic datasets would be replicated on experimental datasets by applying the same methodology to the Porebski experimental dataset (47), as in the previous section. Note that, due to the nature of the experimental data, we had access only to a proxy of per sequence binding energies, without any available proxy for per amino acid contributions, in contrast to the synthetic data. Therefore, we examined the associations between the logits of the models trained on the experimental dataset and the fluorescence-based binding intensity measure (further called intensity), as previously defined (47) (see Methods). We observed high correlations between model logits and intensity ("vs Weak" r=0.84, p-value < 0.05; "vs Non-binder" r=0.82, p-value < 0.05), in line with the results from the synthetic dataset (Figure 3F). The models trained on negative datasets with randomized sequences did not learn the same associations ("vs Randomized" r=-0.05), despite the 100% test set accuracy, which might suggest that the model trained on "vs Randomized" could achieve high accuracy in distinguishing between the positive and negative class examples by leveraging simpler rules unrelated to binding, such as ones based on amino-acid frequency. Although learning binding energies, which are continuous quantities, from binary classification was a non-trivial result, it has been recently shown in a different context (49). More broadly, our analysis of synthetic data highlighted dataset compositions with favorable properties in terms of *OOD* generalizability and rule discovery, which were found to behave similarly on independent experimental data.

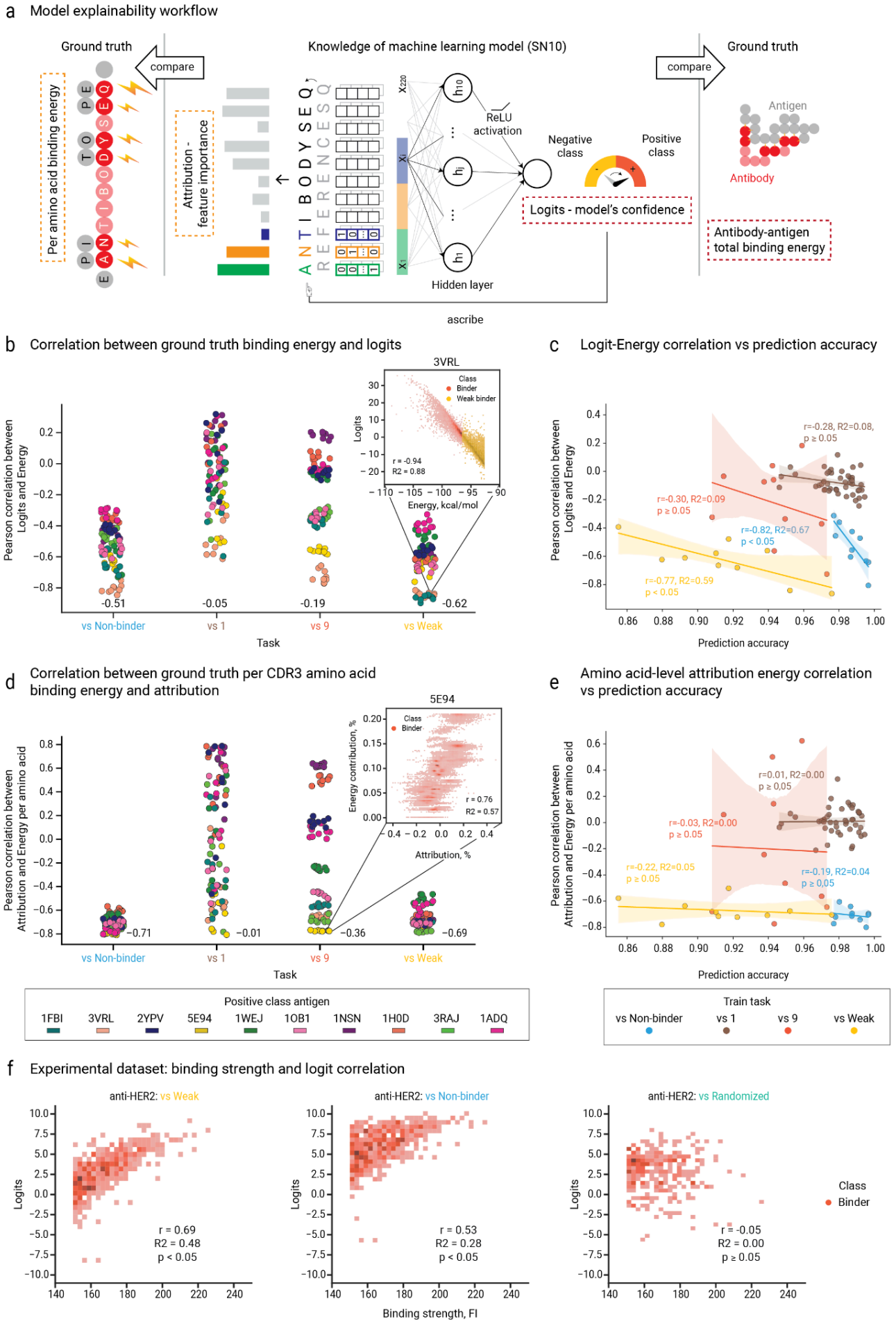

a Model explainability workflow
Ground truth
compare
Per amino acid binding energy
Attribution - feature importance
Knowledge of machine learning model (SN10)
ReLU activation
Hidden layer
Negative class
Positive class
Logits - model's confidence
ascribe
Ground truth
compare
Antigen
Antibody
Antibody-antigen total binding energy
b Correlation between ground truth binding energy and logits
Pearson correlation between Logits and Energy
3VRL
Class
Binder
Weak binder
Logits
r = -0.94
R2 = 0.88
Energy, kcal/mol
-0.51
-0.05
-0.19
-0.62
vs Non-binder
vs 1
vs 9
vs Weak
Task
c Logit-Energy correlation vs prediction accuracy
Pearson correlation between Logits and Energy
r=-0.28, R2=0.08, p ≥ 0.05
r=-0.30, R2=0.09 p ≥ 0.05
r=-0.82, R2=0.67 p < 0.05
r=-0.77, R2=0.59 p < 0.05
Prediction accuracy
d Correlation between ground truth per CDR3 amino acid binding energy and attribution
Pearson correlation between Attribution and Energy per amino acid
5E94
Class
Binder
Energy contribution, %
r = 0.76
R2 = 0.57
Attribution, %
−0.71
−0.01
−0.36
−0.69
vs Non-binder
vs 1
vs 9
vs Weak
Task
e Amino acid-level attribution energy correlation vs prediction accuracy
Pearson correlation between Attribution and Energy per amino acid
r=0.01, R2=0.00 p ≥ 0.05
r=-0.03, R2=0.00 p ≥ 0.05
r=-0.22, R2=0.05 p ≥ 0.05
r=-0.19, R2=0.04 p ≥ 0.05
Prediction accuracy
Positive class antigen
1FBI
3VRL
2YPV
5E94
1WEJ
1OB1
1NSN
1H0D
3RAJ
1ADQ
Train task
vs Non-binder
vs 1
vs 9
vs Weak
f Experimental dataset: binding strength and logit correlation
anti-HER2: vs Weak
r = 0.69
R2 = 0.48
p < 0.05
anti-HER2: vs Non-binder
r = 0.53
R2 = 0.28
p < 0.05
anti-HER2: vs Randomized
r = -0.05
R2 = 0.00
p ≥ 0.05
Logits
Binding strength, FI
Class
Binder

Figure 3. **The choice of negative datasets can enhance the learning of binding rules of the positive class sequences. (a)** To understand which antibody characteristics beyond their binding class models can acquire, we compared ground truth information available from Absolut! with the model's interpretative quantities. As ground truth information, we used total antibody-antigen binding energy and per amino-acid binding energy. The model's interpretive components were of two types: logits – the model's output prior to the final activation, representing the model's confidence in the predicted class; and feature attribution – the influence of a feature on logits. We employed the DeepLIFT attribution method to calculate how each feature in the input contributes to the prediction when compared to the reference sequence. In our study, we utilized DeepLIFT with ten distinct shuffles of the original sequence, subsequently averaging the results, as explained in the Methods section **(b)** Pearson correlation coefficients between logits and total binding energy in the positive class examples across different dataset types and antigens. We limited our analysis to the positive class examples from a dataset, since in practical scenarios, such as those corresponding to the "vs 1" and "vs 9" tasks, data regarding the binding of negative class sequences to the antigen of the positive class will be unavailable. The numbers at the bottom of the scatter plots indicate the medians of Pearson correlation coefficients within the tasks. Each dot in the scatter plots represents a correlation coefficient calculated for the relationship between logits and total binding energy values corresponding to a model-antigen pair. An example is provided for the "3VRL" antigen in the inset 2D-histogram (bins=100). **(c)** The relationship between SN10 model accuracy and logit-energy correlation coefficient was described for the preceding subplot. Pearson correlation coefficient (r) and coefficient of determination (R^2) provide numerical descriptions of this relationship. The two-sided p-value is less than 0.05 for "vs Weak" and "vs Non-binder" tasks, and greater or equal for "vs 9" and "vs 1" task. For the tasks "vs 9", "vs Non-binder", and "vs Weak", there are 10 data points in each, and for the "vs 1" task, there are 45 points. Each data point is an average across model train-test split and seed replicates. **(d)** Pearson correlation coefficients between energy contribution per CDRH3 amino acid and the SN10 amino acid attribution in the positive class samples across all dataset types and antigens. The median values are presented at the bottom. An example of a strong association is shown in the inset 2D histogram (bins = 100) for the 5E94 "vs Weak" task. The x-axis and y-axis values are expressed as signed percentages: each value represents the absolute attribution/energy for a specific amino acid in a CDRH3, divided by the sum of the absolute values of the attribution/energy for all amino acids in the CDRH3, with the sign preserved. **(e)** The relationship between SN10 model accuracy and per amino acid energy – attribution correlation coefficient, is described in the previous subplot. Similar to subplot (c). All p-values > 0.05. **(f)** The co-distribution of binding strength, measured as fluorescence signal intensity (FI), where higher fluorescence indicates stronger binding, and SN10 logit values for anti-HER2 scFv datasets is illustrated in the 2D histogram with 30x30 bins. The Pearson correlation coefficient, indicated on each plot, is calculated solely for the sequences identified as binders.

Discussion

The definition of "negative" and "positive" data is crucial in supervised machine learning settings, but how the definition of negative and positive data changes ML model behavior (prediction accuracy, generalization, rule inference) has not been thoroughly investigated. Here, we showed that *ID* and *OOD* performance, as well as the rule discovery vary as a function of training dataset composition in the antibody-antigen binding prediction task. As mentioned throughout the text, there is considerable discussion in the TCR-epitope prediction field about the choice of negative data (13, 28–31). Given that TCR-epitope prediction settings are very similar to this work's ML setup (supervised classification, several classes of sequences in the datasets), this work provides guidance on immune receptor dataset construction, benefiting both the BCR and TCR fields. More generally, our work shows that ML dataset construction is far from trivial and needs to be planned carefully for robust *ID* and *OOD* performance and rule discovery.

ID performance does not predict OOD predictive performance

Briefly, we found that for both synthetic and experimental data, in the *ID* setting, prediction performance was high (with "vs Weak" having the lowest performance, Figure 2B) but varied across training data compositions. In contrast, in the *OOD* setting, only models trained on "vs Weak" performed consistently well (Figure 2D, Supp. Figure 3). Thus, in line with previous observations in an entirely different domain (50, 51), we found a trade-off between *ID* and *OOD* settings across tasks. While within "vs 9" task *ID* performance did not predict *OOD* performance, "vs Non-binder" indicated a non-significant positive trend in the *ID-OOD* relationship, and "vs Weak" exhibited a positive correlation (Supp. Figure 4A), suggesting that these ML models learned invariant (robust) features of the positive data, making *ID* informative of *OOD*. In cases of negative or no correlation, the model likely relied mostly on spurious features (50). Specifically, we found that both *ID* and *OOD* performance were antigen-dependent. This is an observation we also made in another context (generative design of antibodies (52)). We believe that this may be due to a variety of factors such as antigen shape, antigen-specific composition of training datasets, as well as epistatic considerations. As previously mentioned (52), this is beyond the scope of this work and requires further investigation in a separate study. That said, we found that sequence similarity (measured by JSD) between positive and negative data correlated with *ID* and *OOD* performance. The smaller the distance, the lower was *ID* performace but the more stable *OOD* was. The challenging aspect of the "vs Weak" datasets may be due to the higher incidence of sequences where a single amino-acid substitution changes the binding class (Supp. Figure 2A). Thus for *ID* and *OOD* performance, the distance between positive and negative datasets is important.

Rule discovery is not correlated with ID prediction performance, but sometimes with OOD prediction performance

We found that the composition of the negative dataset also impacted the model's ability to identify binding energies (rules) at both the sequence and residue levels (Figure 3). Specifically, (i) the correlation between output logits and antibody binding energy was particularly pronounced in the "vs Weak" and "vs Non-binder" tasks, (ii) as was the correlation between individual amino acid contribution to binding energy and its contribution to model's prediction (attribution). Notably, the capacity of rule recovery (per-residue energy contributions correlation with model attributions) appears to be independent of the overall prediction accuracy (Figure 3E). This finding contrasts with the hypothesis presented in a TCR-field study suggesting that models with higher performance are more likely to rely on biological rules (53). Our work demonstrates that while *OOD* performance is worse for "vs Non-binder" and "vs 9" compared to "vs Weak", "vs Non-binder" did not perform worse on rule discovery (Figure 3B,D). Thus, a model's ability to learn detailed binding information is an additional aspect of its performance, not merely a byproduct of its classification accuracy and not trivially related to *OOD* generalization. It appears that performance is necessary but not sufficient for rule discovery. The *OOD* performance can sometimes serve as a predictive

indicator for rule discovery, as demonstrated by our findings, with particular consistency in the case of logits-based rule discovery (with the exception of one scenario, *OOD* is always correlated with rule discovery). It is noteworthy that *OOD* performance may also predict attribution-based rule discovery (Supp. Figure 4B), including the assessment of generalization of "vs 1" on "vs 9" test datasets (both datasets consisting of high-binders only). Importantly, a higher *OOD* accuracy can imply better rule discovery when comparing a dataset to a different dataset evaluated on the same *OOD* task, e.g. the "vs 1" task of "1ADQ vs 5E94" evaluated on "vs 9" compared to "1ADQ vs 3VRL" evaluated on "vs 9" as well. Thus *OOD* performance might offer a way to measure in a relative way and compare among datasets the potential for rule discovery. However, the *OOD* performance cannot be used as an absolute measure, because the inference of rule discovery from the *OOD* is highly specific to the *OOD* task and a high absolute *OOD* accuracy does not necessarily imply reliable rule discovery, e.g. for "vs 1" tasks, an *OOD* accuracy of >90% on "vs 9" implies a rule discovery correlation coefficients in the range of -0.25 – -0.5, while an *OOD* accuracy of even >70% on "vs Weak" implies rule discovery coefficients in the range of -0.5 – -0.75. More broadly, measuring *OOD* on "vs Weak" and "vs Non-binder" appears to be most robust when inferring rule discovery from *OOD*, since both tasks retain the highest correlation coefficients (Supp. Figure 9C).

Speculatively, high performance on the "vs 1" and "vs 9" tasks (Figure 2B) could be obtained by leveraging sequence biases that distinguish the positive and negative dataset, but that are unrelated to the binding phenomenon itself (lower correlations with energies, Figure 3B, D). The progression in the correlation strength observed across the sequence of tasks "vs 1", "vs 9", and "vs Non-binder" suggests that augmenting the number of binders in the negative dataset, surpassing the current limit of nine, holds the potential to create a negative dataset that may induce comparable learning effects as that of the "vs Non-binder" negative dataset.

Definition of binding rules via position-wise affinity contributions

In this work, the term "binding rule" describes the binding energies per CDRH3 amino acid for a given antibody-antigen pair. The generality of this work's binding rule definition is an advantage over previous definitions where more importance was given to distinguishing binding residues from non-binding residues while omitting energies (34). Indeed, residues that do not contribute to binding energy might still contribute to binding indirectly via structural constraints. The strength of our definition lies in its contextual nature, emphasizing that the binding rules for an antibody are intricately linked to the specific antigen it interacts with. This approach contrasts with antigen-independent rules, which are typically based on broader characteristics such as the charge, hydrophobicity, or aromatic nature of amino acids (19, 54). That said, in the future, it would be of interest to widen the definition of binding rules to interpositional interactions, also known as epistatic relationships (55–57), since we suspect that interpositional interactions are essential to identifying binding motifs.

However, there are still unanswered questions regarding attributions and their interpretation. For example, we note that the magnitude of attributions, evaluated by L2 Norm, was variable across task types, the attributions from "vs Non-binder" appearing to have consistently higher complexity (Supp. Figure 8C, D)

Another pertinent concern revolves around the learning of rules, particularly concerning the contribution of energy at the residue level. In our investigation, we did not observe scenarios where the energy contributions at the residue level exhibited correlations better than -0.6 to -0.7. Possible explanations include errors in rule learning, limitations of rule learning (e.g. DeepLIFT does not account for feature interactions), errors in energy computations, incomplete energy computations (e.g. missed interactions by analyzing only CDRH3), and, in general, the higher complexity of per-residue binding energy. It cannot be ruled out that the model is learning and utilizing rules of a different nature, which are not reflected in correlations with residue-level energy contributions, with epistatic interactions being a prime example.
A different perspective highlighting the complexity of rule learning in form of per-residue binding energies stems from the analysis based on epitope-based datasets, in which the positive class consists of only sequences binding to a single epitope as opposed to a single antigen (Supp. Text 2, Supp. Figure 9). Although the use of epitope-based

datasets in comparison to the use of antigen-based datasets resulted in significant improvements in logits-based rule discovery, no significant and consistent difference was noted for the attribution-based rule discovery (Supp. Figure 9D,E). Thus, although both per-sequence and per-residue energies appear to be discoverable, per-residue energies appear harder to be disentangled.

Synthetic antibody-antigen binding data enables a detailed study of the impact of training dataset composition on ML performance

In this work, we used synthetic antibody-antigen binding data that was previously shown to help guide ML considerations on experimental data (27). The use of synthetic data helped us explore how training data compositions impacts ML performance, including the configuration of datasets by affinity. Although the affinity information allowed us to construct the datasets, we intentionally did not train the ML model to predict it. Our aim was to create an information bottleneck by designing the task as a binary classification, aligning with the most accessible experimental data for training immune receptor ML models. Current experimental data, as was used in this work, is only starting to become available in single sequence affinity-annotated format (47, 58). However, these datasets may be generated with different technologies potentially compromising the comparability of affinity values. Furthermore, most publicly available datasets are not annotated or have affinity information with different cutoffs for binding/non-binding thereby increasing heterogeneity and raising challenges prediction performance and rule discovery. Even if ignoring the mentioned limitations of current experimental data, the available experimental dataset size is one to two orders of magnitude lower (our synthetic dataset has ~700k sequences in comparison to ~25k for experimental) and the synthetic dataset covers more complexity, e.g. as evaluated by the higher number of covered antigens. The use of synthetic data enabled the testing of various sampling strategies and performance covariates (e.g., sequence composition, sequence entropy, antigen), which would not have been possible in that breadth with experimental data. Indeed, the testing of several hypotheses may be used for experimental design. Furthermore, synthetic data has the advantage of avoidance of confirmation bias in explainability studies (41, 59, 60) as ground truth is available. Indeed, experimental data is distributed over an array of features, some relevant and others irrelevant for a given prediction task. These features may be structural properties, affinity, k-mer distributions, amino acid composition, and length of the CDRH3. It is generally not known in advance which features are irrelevant and thus which should be balanced in the training data (both positive and negative). Once identified, the irrelevant features could be balanced between the positive and negative datasets to prevent the model from learning spurious correlations. Towards this goal, here, as a first step, we explored this avenue by designing different negative classes. The large difference in performance among the different training data compositions supports the conclusion that there exist irrelevant features that are not equally distributed between different types of negatives, which can lead to high test set accuracy based on irrelevant features, a condition that leads to non-robust models with low generalization capabilities. In follow-up work, it may be interesting to investigate more granularly how the model behavior is impacted after certain features are balanced (e.g., forcing the distribution of accessible features such as k-mer distribution, resampling hidden features such as structural binding properties).

Of note, the access to ground truth enabled the verification of generalizability of our results across antigen- and epitope-based datasets (Supp. Text 2, Supp. Figure 9). In our main analysis we included sequences that bind an antigen irrespective of epitope, although we know from our previous work that in the synthetic data, antigens often have multiple epitopes (27). We performed a supplementary analysis which included epitope-based datasets for 3 antigens (Supp. Text 2, Supp. Figure 9), in which we compared the performance (*ID* and *OOD*), as well as the rule discovery of epitope-based and antigen-based datasets. Large differences in *ID* and *OOD* performance and in attribution-based rule discovery were not observed. This might be explained by the fact that even in antigen-based datasets, most of the sequences stem from one or a few epitopes, thus leading to highly skewed distributions of the number of sequences per epitope within an antigen. A significant improvement was seen in logit-based rule discovery for epitope-based datasets, the cause of which could lie in a simpler, less heterogeneous positive dataset, which might lead to more consistent rule discovery and an easier evaluation set.

Machine learning considerations

In any kind of ML setting, information leakage (also called: data leakage) may occur where information from the testing or evaluation dataset inadvertently leaks into the training dataset. This may lead to overfitting, signifying that the ML model performs well on the evaluation data but fails to generalize to new unseen data. In biological sequence data settings, it is therefore usually advised to avoid sequences in train and test that are close in sequence similarity (1, 61, 62). However, this approach does not account for the fact that very similar sequences may have entirely different functions (e.g., binding profiles), which is commonly observed in immune receptors (63, 64) and which constitutes a great challenge. In our study, we did not observe a significant contribution to *ID* performance from sequences with the same label and 90% sequence identity between the training and testing sets. Sequences that are similar but have different targets or affinities are challenging to predict and should therefore be distributed across train and test datasets (Supp. Text 1, Supp. Figure 2).

In this study, we used the SN10 model because it was previously identified as optimal for learning Absolut! data (27). Our aim was to employ a neural network capable of accounting for interpositional dependencies. Consequently, we compared the SN10 model to logistic regression (Supp. Figure 2C, Supp. Figure 7) and showed that the SN10 model's performance is superior suggesting that it can take advantage of the detection of higher-order non-linear sequence dependencies. However, we could not identify better rule discovery in SN10 in comparison to logistic regression, which underscores the need for appropriate ground truth and explainability methods that can reliably identify feature interactions that could potentially have been learned by SN10, as suggested by better *ID* performance in "vs Weak". In the future, it would be of interest to study the association of training dataset composition and ML architecture. Studies investigating in more detail the biases induced by the model architecture and data type (sequence vs structure) in the learning of the antibody sequence-binding affinity landscape have started to emerge (65, 66). These studies reveal that simpler models tend to learn more about binding interactions, whereas complex models capture broader biophysical properties related to protein folding.

Regarding observed biases, we noted that both the SN10 and LR models tend to learn rules associated with antigens that are easier to predict, irrespective of how high-binders of an antigen are labeled as positive or negative. Specifically, we observed that reversing the roles of positive and negative datasets in 'vs 1' task can lead to a loss or identification of associations between energy and logits or attributions for positive data (Supp. Figure 6, Supp. Figure 7). Given the fact that switching the role of positive and negative class in case of "vs 1" doesn't impact prediction accuracy, it is noteworthy to highlight that in the case of "vs 1" the rule discovery for the positive class depends on which antigen serves as the positive class.

We demonstrated that effective rule extraction relies not only on the mining of positive data but also on the careful selection of negative data. Our findings reveal that "vs Weak" and "vs Non-binder", establish decision boundaries based on features directly related to the binding phenomenon, such as total binding energy and per-amino acid energy contributions, notably both datasets containing sequences along a spectrum of binding energies across both the positive and negative parts. In contrast, "vs 9" and "vs 1" appear to focus more on spurious features. However, this hypothesis warrants further investigation.

Conclusions

Our work provides evidence that training models on datasets where positive and negative data distributions are close (affinity and sequence-wise), leads to models that can generalize better in an *OOD* setting. We find that the closer the positive and negative training datasets are, the better the *OOD* performance is. Therefore, there is a requirement for computational techniques that allow choosing negative data that are close to the positive data (such negative examples are also called "hard negatives" in the literature (15)). Indeed, we see it as a future challenge to study more closely ideal cutoffs for dataset "closeness". Our analysis of synthetic data revealed dataset compositions with favorable properties for *OOD* generalizability and rule discovery, aligning with observations on

independent experimental data. Experimentally, deep mutational scanning (DMS) methods allow creation of positive and negative datasets that are close in sequence and affinity (64, 67, 68). Our work shows that DMS datasets will be crucial to teaching ML models sequence to function maps where very small sequence differences can have large functional effects (Supp. Figure 2). By showcasing the added value of datasets with close positive and negative examples and the ability to learn energies only from sequence data, we suggest fruitful avenues of further research into more detailed analyses of hard negatives and near-misses. Our discovery of learning continuous binding energies from binary classification, although not trivial, complements similar findings in a recent study (49). Furthermore, given the current rise in sequence-based language models (69–71), and evidence that some protein language models (72, 73) have demonstrated the ability to capture affinity information, it would be of considerable interest to investigate to what extent our findings apply to unsupervised settings and whether including similar sequences may help encode binding energy into embeddings more efficiently. Finally, our work adds to previous work (27, 74), which showed that sequence-based data may be used to infer structural features (or rules) via, for instance, explainability methods. This is significant as sequence-based data are comparably easier to generate at high throughput than structural data.

Materials and Methods

Synthetic antibody-antigen binding dataset generated using the antibody-antigen simulation framework Absolut!

We used a semi-synthetic dataset of antibody-antigen binding as previously generated by Robert et al. and Akbar et al. (27, 52). Briefly, this dataset comprises real-world antigen and antibody (CDRH3) proteins with antibody sequences from mouse repertoires (75) and with the antigens' 3D structures sourced from the Protein Data Bank (PDB) and transformed into a 3D lattice format. The authors of the dataset applied an exhaustive docking method to evaluate the binding of discretized rigid antigens with flexible CDRH3s of length 11 from the antibody side. The result included the acquisition of binding energy values, binding positions that minimize free energy, and the identification of interacting amino acids between antigens (epitope) and CDRH3s (paratope). Each antigen was associated with a set of 11-mers derived from CDRH3s and ranked according to their binding energy. Those in the lowest 1% percentile, with an average of 51,814.7 sequences, were labeled as high-affinity binders (or simply 'binders'). The 1-5% percentile, averaging 205,069.3 sequences, were labeled as weak-affinity binders. Finally, the highest 95% percentile, with an average of 391,498.8 sequences, were labeled as non-binders. This labeling scheme aligns with the methodology used in the study by Robert et al. (27). For our study, we used antibody binding data from 10 antigens, with PDB-IDs 3VRL, 1NSN, 3RAJ, 5E94, 1H0D, 1WEJ, 1ADQ, 1FBI, 2YPV, 1OB1.

Prediction tasks, training, and test datasets

We used balanced CDRH3 datasets (30k antibody sequences for training [15k positive class, 15k negative class], 10k antibody sequences for testing [5k positive class, 5k negative class]), which varied by the positive class antigen and the negative dataset type (Table 1). The positive class of the dataset consisted of binder sequences, i.e., sequences having binding energies between 0–1% percentiles.
The negative class of each dataset was built depending on the binding prediction task type, as follows:

- vs 1 – sequences that are binders for another antigen but not the positive class antigen.
- vs 9 – sequences that bind to one of the 9 other antigens and do not bind to the positive class antigen, represented in equal proportions. To illustrate, in the training set of a total of 15,000 sequences, each of the 9 antigens has approximately 1,666 (+/- 1) associated sequences (~15,000/9).
- vs Weak – sequences that bind the positive class antigen, but with binding energies in the 1–5% percentile.
- vs Non-binder – sequences with binding energies in the >5% percentile for the positive class antigen.

We generated replicates to ensure the reproducibility of the experiments. Specifically, six different train-test splits were generated for each prediction task and positive antigen. For one of the splits, we trained and evaluated the model four times with various seeds, therefore controlling for the randomness in model initialisation and training. The positive class of the datasets is kept constant across the tasks and in each replicate experiment. For example, the positive class data in *1ADQ Binder* "vs Weak" is the same as the positive class data in *1ADQ Binder* "vs 9". The sequences corresponding to each positive and negative dataset were randomly sampled to achieve the previously mentioned dataset sizes. Throughout our work, we employed negative controls by randomly shuffling the class labels of sequences for each task. The prediction performance was evaluated on hold-out test datasets using accuracy metrics, calculated as the number of correct predictions divided by the total number of predictions. The *out-of-distribution (OOD)* performance was evaluated by calculating the performance on the test sets of other tasks, e.g. "vs 9" performance on a "vs Weak" test set. The out-of-distribution estimations were performed over the same set of replicates as in the in-distribution case.

Machine learning model architecture, training, and evaluation

We trained a simple neural network, previously (27) and henceforth called SN10 (27), for binding prediction tasks using CDRH3 sequences. The sequences, each consisting of 11 amino acids, are one-hot encoded, resulting in a dimensionality of 220. The neural network architecture consisted of 220 neurons in the input layer, 10 neurons in the hidden layer (ReLU as activation), and 1 output neuron (sigmoidal activation). For training, after exploration of the parameter space, we used binary cross-entropy loss with the Adam optimizer with a batch size of 64, a learning rate of 0.001, and a momentum of 0.9, with no weight decay. We trained the model for 50 epochs and inspected the loss curves, which did not show signs of overfitting. To further enhance the model's performance, we implemented the technique of stochastic weight averaging (SWA) (76, 77), which involves averaging multiple sets of weights obtained during the training process to reduce the variance and improve the generalization performance of the model. We apply SWA after epoch 3/50, using the SWA implementation available in Pytorch v1.12.1 (78). A logistic regression-based model was further implemented for binary classification as a baseline for comparison. The model was trained and evaluated on the same datasets using one-hot encoded CDRH3 sequences, as previously described for the SN10 (27).

Information theory-based calculations

Jensen-Shannon distance (JSD) is a measure of dissimilarity between two probability distributions. It is derived from the Kullback–Leibler divergence (KL) and calculated using the following formula:

$$JSD(P||N) = \sqrt{\frac{KL(P||M) + KL(N||M)}{2}}$$

$$KL(P||M) = \sum_{x \in X} P(x) \log \frac{P(x)}{M(x)}$$

$$M = \frac{P + N}{2}$$

In our study, P and N represent the probability distributions for the positive and negative data sequences, respectively. The JSD was computed using the scipy.special.distance package. As a measure of the total differences in marginal distributions for positive versus negative data sequences (i.e., excluding differences in statistical interactions), we utilized position weight matrices (PWMs) that represent the probabilities of each amino acid occurring at specific positions within binder or non-binder sequences. The JSD in our setting is the sum of JSDs across each position. PWMs were calculated using the bio.motifs package from Biopython (79).

Sequence Shannon entropy calculations were performed to measure antibody sequence diversity for given sets of sequences (e.g. 5E94 binders, 3VRL weak binders) (see Supp. Figure 1). The Shannon entropy was computed using the scipy.stats Python module to quantify the information content based on the PWM (computed as previously described) of the set of sequences.

Attribution analysis

To better understand the rules learned by a SN10, we employed interpretability techniques and applied them on test set examples in order to gauge the importance of input features for each sequence, importance in this case meaning contribution to the prediction. We used the DeepLIFT (46) implementation from the Captum library v0.5.0 (80) as a fast method that approximates Integrated Gradients (81). The computation requires baseline sequences, which were obtained by randomly reordering input sequences, a technique known in literature as the shuffling approach (82). Although using a zero input vector as baseline is common and plausible in certain applications, in the case of sequences it is problematic, since a zero input vector does not represent any sequence, i.e., there is no sequence that would be one-hot encoded as a zero input vector. The shuffling approach is based on random sequences and has been the preferred approach for biological sequences in the recent literature (82, 83). We generated ten shuffled

versions for each CDRH3 sequence and ran DeepLIFT ten times per sequence, once for each of the shuffled baselines. The final attribution values were obtained by averaging the results from these ten runs. The attributions were estimated with respect to the predicted logits, not expits. We used local attributions (84), since we were interested in the impact to the prediction of each amino acid in each sequence, and the implementation of this calculation in Captum (80) involved setting *multiply_by_inputs* to True. The attribution analysis was applied to both synthetic and experimental data. However, since there was no binding energy data per amino acid in the experimental dataset, we could compare attributions to ground truth per-residue energies only in the synthetic dataset. Attributions were computed across all replicates except for "vs 1", where due to the much larger number of datasets we restricted the computation to once per dataset.

Epitope-based analysis

For the epitope-based analysis, we generated new datasets while restricting, in each case, the positive datasets to consist only of sequences that bind to the major, i.e. most frequent, epitope of an antigen. For this analysis we selected the antigens that contain a sufficient number of epitope-specific sequences to allow for a training set size with the same size as in the antigen-based analysis (N_{train} = 30000). However the test set size, due to the lower availability of sequences, had to be reduced w.r.t. antigen-based analysis (N_{test} = 6000). Due to the training size restriction we had to limit the epitope-based analysis to 3 antigens: 1H0D, 1WEJ, 1OB1. In all our workflows and results, to distinguish between antigen- and epitope-based, we depicted the epitope-based datasets with the same antigen id, followed by "E1" (indicating the most frequent epitope): 1H0DE1, 1WEJE1, 1OB1E1. To generate the training and test sets, we filtered and adjusted the positive datasets by keeping the sequences that bind only to the major epitope and kept the same negative datasets as for the corresponding antigen-based analyses. After generating the corresponding datasets, we performed identical computations as for the antigen-based analyses that were shown in the main text.

Absolut! binding energy calculations

We refer the reader to Absolut! (27) for a detailed explanation on the computation of binding energies. Briefly, in the 3D lattice only non-covalently bound neighboring amino acids are considered to interact. The binding energies are then computed by summing the Miyazawa-Jernigan potential for each interacting pair of residues. By summing the Miyazawa-Jernigan potential for all interactions in which a CDRH3 residue is involved, the energy contribution of that residue to CDRH3:antigen binding energy is computed.

Experimental antibody-antigen binding datasets

The experimental dataset. We identified a publicly available experimental dataset (47), generated for high-throughput antibody discovery using next-generation sequencing, ribosome display and affinity screening techniques, referred to throughout the text as the "Porebski dataset" or "Experimental dataset". The technique generates binding kinetics data for an antibody library, which can be used to derive measures of antibody affinity and thus to perform screening. Fluorescently-labeled antigen accumulates around higher-affinity binders and images acquired at different antigen concentrations are processed to derive a per-sequence spatially integrated signal based on the fluorescence intensity, which is associated with affinity. The authors of the study performed screening of 2 anti-HER2 antibody libraries, one library (HER2affmat) generated through affinity maturation of G98A, a low affinity anti-HER2 scFv, and another library (HER2mllib) with higher-affinity antibodies, that was generated using a protein language model trained on the HER2affmat. To operate on higher affinity antibodies, we used the HER2mllib.

Processing. We kept the unique sequences that had ≥ 12 replicates and that contained no unresolved amino acids and no stop codons, resulting in 24,790 sequences. All CDRH3 sequences had a length of 21 amino acids. Leveraging the results of the study, we used the *intensity 8* measure (the fluorescence-derived signal at a specific antigen concentration from the study) as a proxy for affinity (higher intensity ~ higher affinity), as it has been used

in the original paper to classify antibodies based on affinity. Using 2 thresholds we classified the sequences in *Binders* (*intensity 8* > 150, 2638 sequences, 10%), *Weak* (122 < *intensity 8* < 150, 3574 sequences, 15%) and *Non-binders* (*intensity 8* < 122, 18578 sequences, 75%). We used thresholds with different percentiles compared to Absolut in order to adapt to the smaller dataset size and obtain more sequences per antibody class for more reliable training and testing. Based on this classification we built "vs Weak" and "vs Non-binder" datasets. In addition we included a negative dataset consisting of randomly reordered CDRH3 sequences from the Binders – "vs Randomized".

Model training and attributions. We trained SN10 models with replicates and computed attributions using exactly the same methodology as described above for the Absolut dataset, by adapting the input size to 21 amino acids and the size of the training set (N=4000) and of the test set (N=1000). Inspection of train and test loss by epochs showed no signs of train set overfitting in our setup. All settings, including performance metrics, were identical to the ones used with the synthetic data.

Code availability

Code is available at https://github.com/csi-greifflab/negative-class-optimization.

Data availability

All the analysis stems from the Absolut Database, available on https://greifflab.org/Absolut/. We make all the processed datasets, trained models, attribution values, and results available at https://doi.org/10.5281/zenodo.11191740. The experimental Porebski dataset is available at https://doi.org/10.5281/zenodo.8241732.

Acknowledgments

Funding

The Leona M. and Harry B. Helmsley Charitable Trust (#2019PG-T1D011, to VG), UiO World-Leading Research Community (to VG), UiO: LifeScience Convergence Environment Immunolingo (to VG and GKS), EU Horizon 2020 iReceptorplus (#825821) (to VG), a Norwegian Cancer Society Grant (#215817, to VG), Research Council of Norway projects (#300740, #331890 to VG), a Research Council of Norway IKTPLUSS project (#311341, to VG and GKS), and Stiftelsen Kristian Gerhard Jebsen (K.G. Jebsen Coeliac Disease Research Centre) (to GKS). This project has received funding (to VG) from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 101007799 (Inno4Vac). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. This communication reflects the author's view and neither IMI nor the European Union, EFPIA, or any Associated Partners are responsible for any use that may be made of the information contained therein. Funded by the European Union (ERC, AB-AG-INTERACT, 101125630, to VG). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801133 (to PR).

References

1. A. M. Hummer, C. Schneider, L. Chinery, C. M. Deane, Investigating the Volume and Diversity of Data Needed for Generalizable Antibody-Antigen ΔΔG Prediction. *BioRxiv* (2023) https:/doi.org/10.1101/2023.05.17.541222.
2. T. Wang, P. Isola, Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. *arXiv* (2020) https:/doi.org/10.48550/arxiv.2005.10242.
3. R. Yang, J. Mao, P. Chaudhari, Does the data induce capacity control in deep learning?. *International Conference on Machine Learning*, 25166 (2022).
4. D. Alvarez-Melis, N. Fusi, Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems* **33**, 21428–21439 (2020).
5. T. Wang, P. Isola, Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *International conference on machine learning*, 9929 (2020).
6. J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence, Eds., *Dataset shift in machine learning* (The MIT Press, 2008) https:/doi.org/10.7551/mitpress/9780262170055.001.0001.
7. X. L. Li, B. Liu, S. K. Ng, Negative training data can be harmful to text classification. *Proceedings of the 2010 conference on empirical methods in natural language processing*, 218 (2010).
8. H. Wunsch, S. Kübler, R. Cantrell, Instance Sampling Methods for Pronoun Resolution in *Proceedings of the International Conference RANLP-2009*, (Association for Computational Linguistics, 2009), pp. 478–483.
9. M. Saeidi, R. Kulkarni, T. Togia, M. Sama, The Effect of Negative Sampling Strategy on Capturing Semantic Similarity in Document Embeddings in (Association for Computational Linguistics, 2017), pp. 1–8.
10. A. Ben-Hur, W. S. Noble, Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics* **7 Suppl 1**, S2 (2006).
11. F.-E. Eid, M. ElHefnawi, L. S. Heath, DeNovo: virus-host sequence-based protein-protein interaction prediction. *Bioinformatics* **32**, 1144–1150 (2016).
12. S. Tsukiyama, M. M. Hasan, S. Fujii, H. Kurata, LSTM-PHV: prediction of human-virus protein-protein interactions by LSTM with word2vec. *Brief. Bioinformatics* **22** (2021).
13. C. Dens, K. Laukens, W. Bittremieux, P. Meysman, The pitfalls of negative data bias for the T-cell epitope specificity challenge. *Nat. Mach. Intell.* (2023) https:/doi.org/10.1038/s42256-023-00727-0.
14. K. Sidorczuk, *et al.*, Benchmarks in antimicrobial peptide prediction are biased due to the selection of negative data. *Brief. Bioinformatics* **23** (2022).
15. L. Xu, *et al.*, Negative Sampling for Contrastive Representation Learning: A Review. *arXiv* (2022) https:/doi.org/10.48550/arxiv.2206.00212.
16. V. Werner de Vargas, J. A. Schneider Aranda, R. Dos Santos Costa, P. R. da Silva Pereira, J. L. Victória Barbosa, Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowl. Inf. Syst.* **65**, 31–57 (2023).
17. E. Loffredo, M. Pastore, S. Cocco, R. Monasson, [2405.09535] Restoring balance: principled under/oversampling of data for optimal classification (2024) (June 17, 2024).
18. R. Akbar, *et al.*, Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *MAbs* **14**, 2008790 (2022).
19. V. Greiff, G. Yaari, L. G. Cowell, Mining adaptive immune receptor repertoires for biological and clinical

information using machine learning. *Current Opinion in Systems Biology* **24**, 109–119 (2020).

20. W. Wilman, *et al.*, Machine-designed biotherapeutics: opportunities, feasibility and advantages of deep learning in computational antibody discovery. *Brief. Bioinformatics* **23** (2022).

21. R. Khetan, *et al.*, Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics. *MAbs* **14**, 2020082 (2022).

22. M. L. Fernández-Quintero, *et al.*, Assessing developability early in the discovery process for novel biologics. *MAbs* **15**, 2171248 (2023).

23. L. Chinery, *et al.*, Baselining the Buzz Trastuzumab-HER2 Affinity, and Beyond. *BioRxiv* (2024) https:/doi.org/10.1101/2024.03.26.586756.

24. R. A. Ehling, *et al.*, Synthetic coevolution reveals adaptive mutational trajectories of neutralizing antibodies and SARS-CoV-2. *BioRxiv* (2024) https:/doi.org/10.1101/2024.03.28.587189.

25. C. Schneider, A. Buchanan, B. Taddese, C. M. Deane, DLAB: deep learning methods for structure-based virtual screening of antibodies. *Bioinformatics* **38**, 377–383 (2022).

26. L.-M. Krützfeldt, M. Schubach, M. Kircher, The impact of different negative training data on regulatory sequence predictions. *PLoS ONE* **15**, e0237412 (2020).

27. P. A. Robert, *et al.*, Unconstrained generation of synthetic antibody-antigen structures to guide machine learning methodology for antibody specificity prediction. *Nat. Comput. Sci.* **2**, 845–865 (2022).

28. A. Montemurro, L. E. Jessen, M. Nielsen, NetTCR-2.1: Lessons and guidance on how to develop models for TCR specificity predictions. *Front. Immunol.* **13**, 1055151 (2022).

29. F. Grazioli, *et al.*, On TCR binding predictors failing to generalize to unseen peptides. *Front. Immunol.* **13**, 1014256 (2022).

30. L. Deng, *et al.*, Performance comparison of TCR-pMHC prediction tools reveals a strong data dependency. *Front. Immunol.* **14**, 1128326 (2023).

31. Y. Gao, Y. Gao, K. Dong, S. Wu, Q. Liu, Reply to: The pitfalls of negative data bias for the T-cell epitope specificity challenge. *Nat. Mach. Intell.* **5**, 1063–1065 (2023).

32. J. L. Xu, M. M. Davis, Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* **13**, 37–45 (2000).

33. M. M. Davis, P. J. Bjorkman, T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).

34. R. Akbar, *et al.*, A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Rep.* **34**, 108856 (2021).

35. SAbDab: The Structural Antibody Database (July 30, 2023).

36. M. Shugay, *et al.*, VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* **46**, D419–D427 (2018).

37. N. Tickotsky, T. Sagiv, J. Prilusky, E. Shifrut, N. Friedman, McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).

38. M. B. Swindells, *et al.*, abYsis: Integrated Antibody Sequence and Structure-Management, Analysis, and Prediction. *J. Mol. Biol.* **429**, 356–364 (2017).

39. M. I. J. Raybould, A. Kovaltsuk, C. Marks, C. M. Deane, CoV-AbDab: the coronavirus antibody database. *Bioinformatics* **37**, 734–735 (2021).

40. S. Mahajan, *et al.*, Epitope specific antibodies and T cell receptors in the immune epitope database. *Front. Immunol.* **9**, 2688 (2018).

41. V. Chen, *et al.*, Best practices for interpretable machine learning in computational biology. *BioRxiv* (2022) https:/doi.org/10.1101/2022.10.28.513978.

42. G. K. Sandve, V. Greiff, Access to ground truth at unconstrained size makes simulated data as indispensable as experimental data for bioinformatics methods development and benchmarking. *Bioinformatics* **38**, 4994–4996 (2022).

43. M. Sundararajan, A. Taly, Q. Yan, [1611.02639] Gradients of Counterfactuals. *arXiv* (2016).

44. Md. R. Karim, *et al.*, Explainable AI for Bioinformatics: Methods, Tools, and Applications. *arXiv* (2022) https:/doi.org/10.48550/arxiv.2212.13261.

45. M. R. Karim, *et al.*, Explainable AI for bioinformatics: methods, tools and applications. *Brief. Bioinformatics* **24** (2023).

46. A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences. *International conference on machine learning*, 3145 (2017).

47. B. T. Porebski, *et al.*, Rapid discovery of high-affinity antibodies via massively parallel sequencing, ribosome display and affinity screening. *Nat. Biomed. Eng.* **8**, 214–232 (2024).

48. Z. Xing, J. Pei, E. Keogh, A brief survey on sequence classification. *SIGKDD Explor. Newsl.* **12**, 40 (2010).

49. M. Case, M. Smith, J. Vinh, G. Thurber, Machine learning to predict continuous protein properties from binary cell sorting data and map unseen sequence space. *Proc Natl Acad Sci USA* **121**, e2311726121 (2024).

50. D. Teney, Y. Lin, S. J. Oh, E. Abbasnejad, Id and ood performance are sometimes inversely correlated on real-world datasets. *Advances in Neural Information Processing Systems* **36** (2024).

51. J. P. Miller, *et al.*, Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. *International conference on machine learning*, 7721 (2021).

52. R. Akbar, *et al.*, In silico proof of principle of machine learning-based antibody design at unconstrained scale. *MAbs* **14**, 2031482 (2022).

53. I. Papadopoulou, A.-P. Nguyen, A. Weber, M. R. Martínez, DECODE: a computational pipeline to discover T cell receptor binding rules. *Bioinformatics* **38**, i246–i254 (2022).

54. N. Thomas, *et al.*, Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics* **30**, 3181–3188 (2014).

55. R. M. Adams, J. B. Kinney, A. M. Walczak, T. Mora, Epistasis in a Fitness Landscape Defined by Antibody-Antigen Binding Free Energy. *Cell Syst.* **8**, 86-93.e3 (2019).

56. T. N. Starr, J. W. Thornton, Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).

57. S. Cocco, L. Posani, R. Monasson, Minimal epistatic networks from integrated sequence and mutational protein data. BioRxiv (2023) https:/doi.org/10.1101/2023.09.25.559251.

58. A. Shanehsazzadeh, *et al.*, Unlocking de novo antibody design with generative artificial intelligence. *BioRxiv* (2023) https:/doi.org/10.1101/2023.01.08.523187.

59. D. Leben, Explainable AI as evidence of fair decisions. *Front. Psychol.* **14**, 1069426 (2023).

60. D. Yang, A. Singh, H. Wu, R. Kroe-Barrett, Comparison of biosensor platforms in the evaluation of high affinity antibody-antigen binding kinetics. *Anal. Biochem.* **508**, 78–96 (2016).

61. D. T. Jones, Setting the standards for machine learning in biology. *Nat. Rev. Mol. Cell Biol.* **20**, 659–660 (2019).

62. I. Walsh, G. Pollastri, S. C. E. Tosatto, Correct machine learning on protein sequences: a peer-reviewing perspective. *Brief. Bioinformatics* **17**, 831–840 (2016).

63. P. Dash, *et al.*, Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).

64. D. M. Mason, *et al.*, Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* (2021) https:/doi.org/10.1038/s41551-021-00699-9.

65. S. A. Fahlberg, C. R. Freschlin, P. Heinzelman, P. A. Romero, Neural network extrapolation to distant regions of the protein fitness landscape. *BioRxiv* (2023) https:/doi.org/10.1101/2023.11.08.566287.

66. A. V. Kulikova, *et al.*, Two sequence- and two structure-based ML models have learned different aspects of protein biochemistry. *Sci. Rep.* **13**, 13280 (2023).

67. S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, G. M. Church, Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).

68. D. M. Fowler, S. Fields, Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).

69. M. H. Vu, et al., Advancing protein language models with linguistics: a roadmap for improved interpretability. arXiv (2022) https:/doi.org/10.48550/arxiv.2207.00982.

70. M. H. Vu, et al., ImmunoLingo: Linguistics-based formalization of the antibody language. arXiv (2022) https:/doi.org/10.48550/arxiv.2209.12635.

71. R. Singh, *et al.*, Learning the language of antibody hypervariability. *BioRxiv* (2023) https:/doi.org/10.1101/2023.04.26.538476.

72. E. Nijkamp, J. A. Ruffolo, E. N. Weinstein, N. Naik, A. Madani, ProGen2: Exploring the boundaries of protein language models. *Cell Syst.* **14**, 968-978.e3 (2023).

73. S. Chu, K. Wei, Generative Antibody Design for Complementary Chain Pairing Sequences through Encoder-Decoder Language Model. *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop* (2023).

74. J. M. Schmiedel, B. Lehner, Determining protein structures using deep mutagenesis. *Nat. Genet.* **51**, 1177–1186 (2019).

75. V. Greiff, *et al.*, Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Rep.* **19**, 1467–1478 (2017).

76. P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, A. G. Wilson, Averaging Weights Leads to Wider Optima and Better Generalization. *arXiv* (2018) https:/doi.org/10.48550/arxiv.1803.05407.

77. B. Athiwaratkun, M. Finzi, P. Izmailov, A. G. Wilson, There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average. *arXiv* (2018) https:/doi.org/10.48550/arxiv.1806.05594.

78. torch.optim — PyTorch 2.0 documentation (August 2, 2023).

79. P. J. A. Cock, *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

80. N. Kokhlikyan, *et al.*, Captum: A unified and generic model interpretability library for PyTorch. *arXiv* (2020) https:/doi.org/10.48550/arxiv.2009.07896.

81. M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks. *International conference on machine learning*, 3319 (2017).

82. G. Novakovsky, N. Dexter, M. W. Libbrecht, W. W. Wasserman, S. Mostafavi, Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat. Rev. Genet.* **24**, 125–137 (2023).

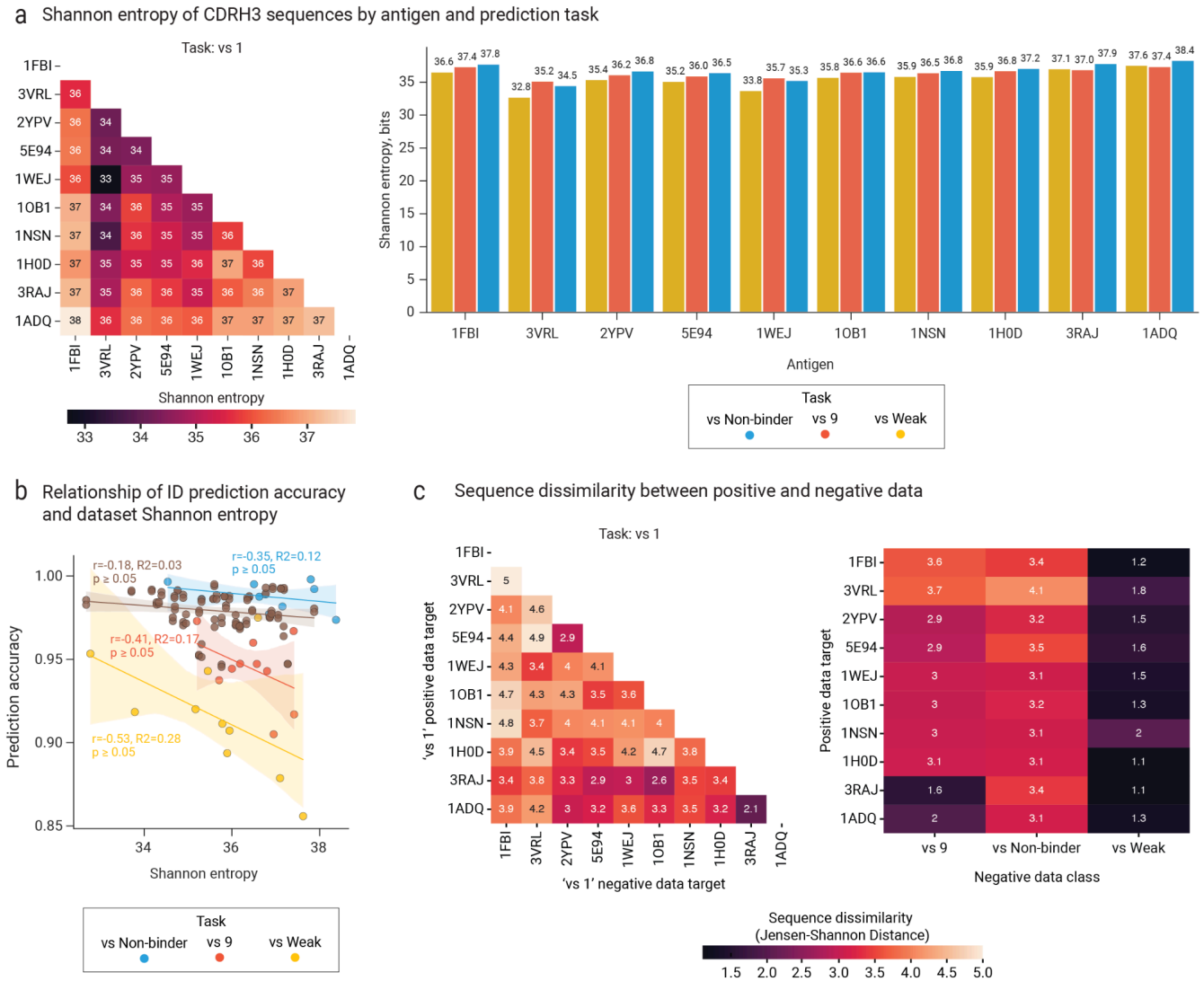
83. D. Huang, *et al.*, Weakly supervised learning of RNA modifications from low-resolution epitranscriptome

data. *Bioinformatics* **37**, i222–i230 (2021).

84. M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for Deep Neural Networks. *arXiv* (2017) https:/doi.org/10.48550/arxiv.1711.06104.

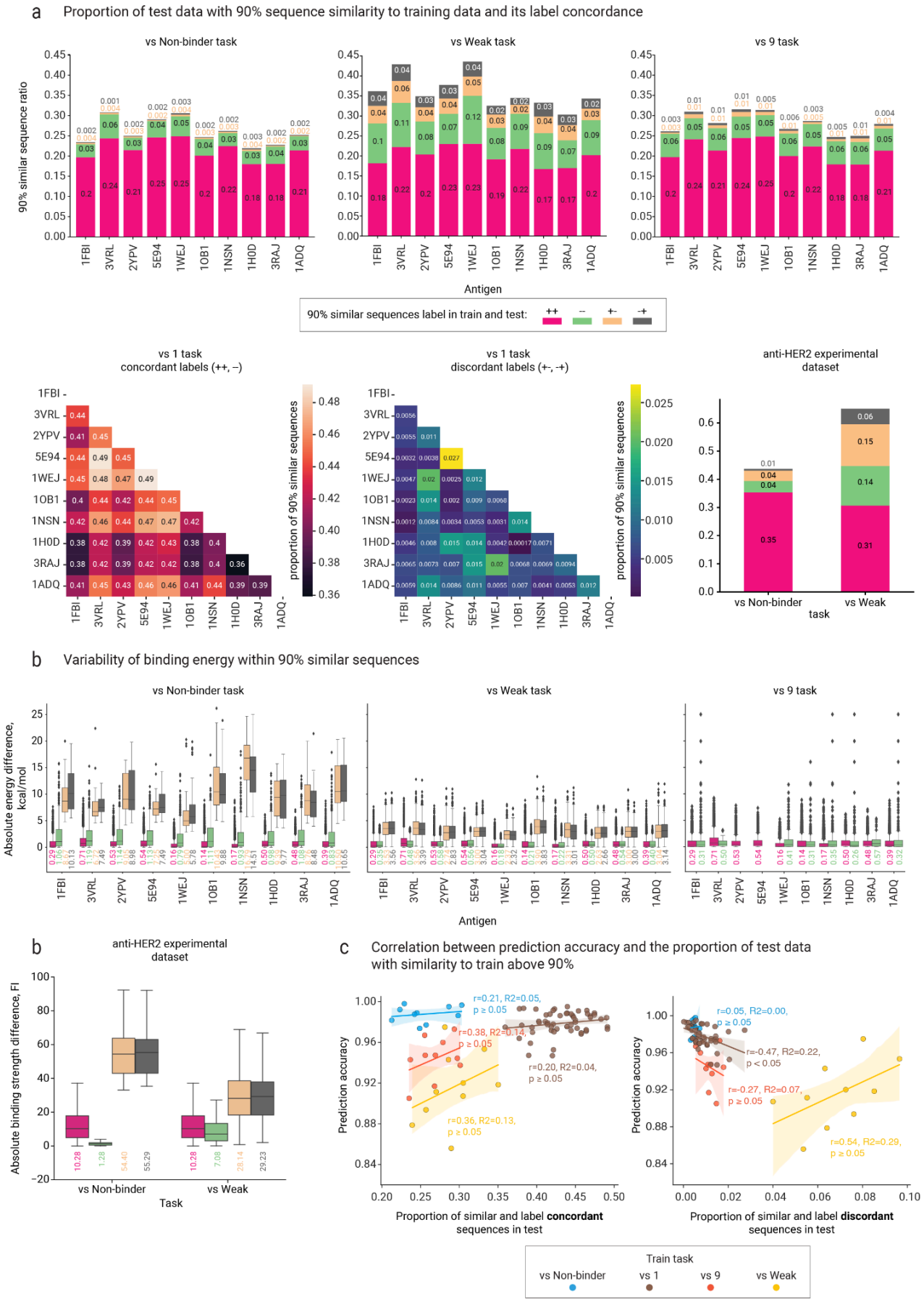
Supplementary Material

Supplementary Figure 1



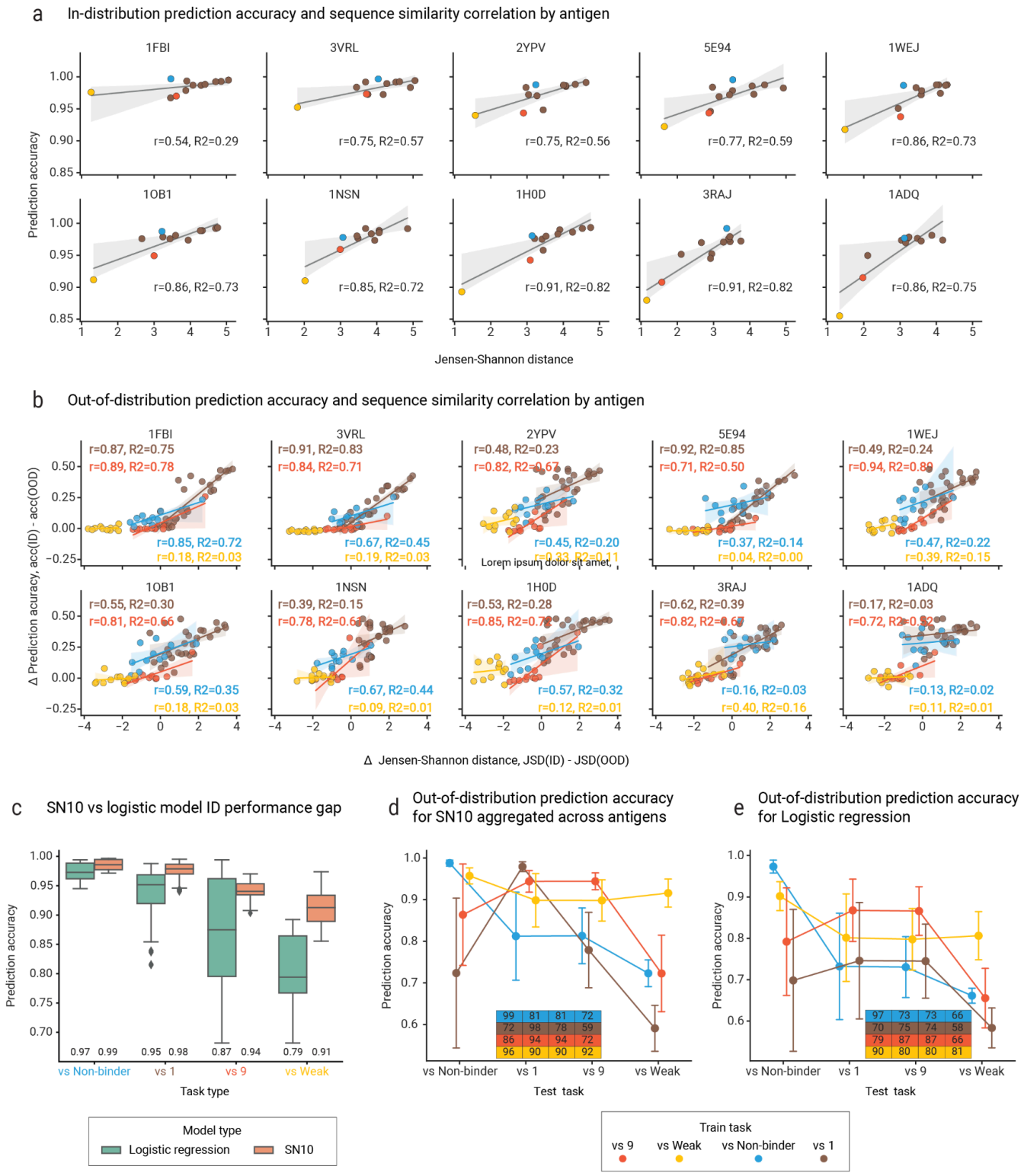
Supp. Figure 1. **The diversity of CDRH3 sequences within the entire dataset provides less accurate information on *ID* performance than the similarity between the positive and negative data. (a)** Shannon entropy calculated for CDRH3 sequences from training datasets (30K sequences). The left plot corresponds to the 'vs 1' task, and the right plot corresponds to all other tasks. **(b)** Correlation between Shannon entropy of the whole train dataset with *ID* prediction accuracy. There are 10 data points for "vs 9", "vs Non-binder" and "vs Weak" tasks and 45 points for "vs 1" task. **(c)** Jensen-Shannon distance between positive and negative data by antigen. The left plot corresponds to the 'vs 1' task, and the right plot corresponds to all other tasks. Relates to Figure 2C.

Supplementary Figure 2



Supp. Figure 2. **Single amino-acid substitution in CDRH3 can lead to antibody binding class shift, challenging SN10 model accuracy.** (**a**) Proportion of sequences in the test dataset that are 90% identical to the training dataset by type of task. The color indicates a data point class in train and test sets, e.g., '+-' signifies positive (+) in train but negative (-) in test (and, e.g., "-+" vice versa). The bottom subplots present the 90% identical sequences from "vs 1" and experimental datasets. (**b**) Binding energy absolute difference in groups of sequences with 90% sequence similarity, each subplot corresponding to one task. The boxes span the interquartile range (Q1 to Q3) and the median is indicated by the middle line. Whiskers extend up to 1.5*IQR from the quartiles, and points outside this range are plotted as outliers. The median value is numerically shown under each boxplot. Plots for synthetic datasets are shown in the upper row, and one plot for the experimental dataset is shown in the following row. (**c**) Pearson correlation between *ID* prediction accuracy and the proportion of sequences in the test dataset that are 90% similar to those in the training dataset. The left subplot accounts for similar sequences with concordant labels, e.g., positive in both training and test data, while the right subplot accounts for sequences with discordant labels, e.g., positive in training but negative in test data. Notably, only in the case of the 'vs 1' scenario in the right subplot the correlation reaches the significance threshold ($p<0.05$). Relates to Figure 2B and referenced in Supp. Text 1.

Supplementary Figure 3



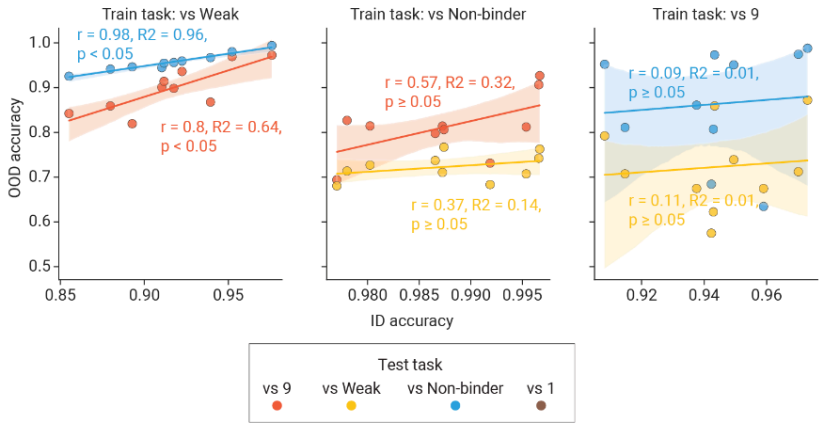
Supp. Figure 3. **Impact of negative data on the performance of machine learning models.** (**a**) Pearson correlation between in-distribution prediction accuracy and sequence similarity calculated using Jensen-Shannon distance and presented per antigen (see Methods). Relates to Figure 2C, which covers all antigens. (**b**) Pearson correlation between the change in Jensen-Shannon distance (JSD) values and the corresponding prediction accuracy for in-distribution and out-of-distribution tasks, separated by positive class antigen. Relates to Figure 2E, which covers all antigens. (**c**) Difference in in-distribution performance between prediction accuracy of logistic regression and SN10. Train and test tasks were the same and are indicated on the x-axis. (**d**) Aggregated *OOD* performance across antigens by training task (indicated by hue) and test task (represented as categories on the x-axis) for SN10 and (e) Logistic regression model. The plots include tables

that display the mean prediction accuracies. Error bars represent one standard deviation from the mean, capturing 68% of the variation in *OOD* performances.

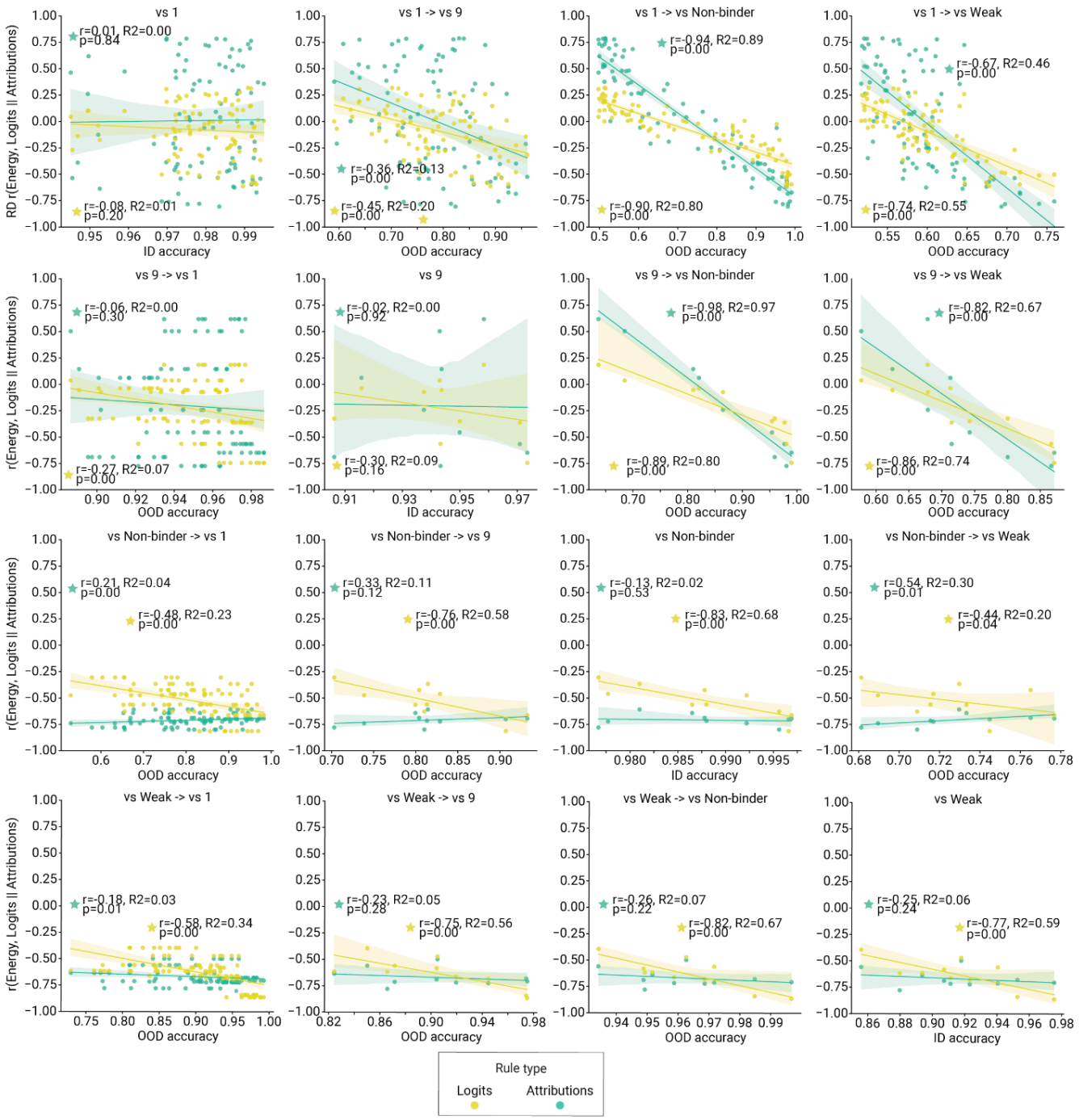
Supplementary Figure 4

a

ID and OOD performance correlation by prediction tasks

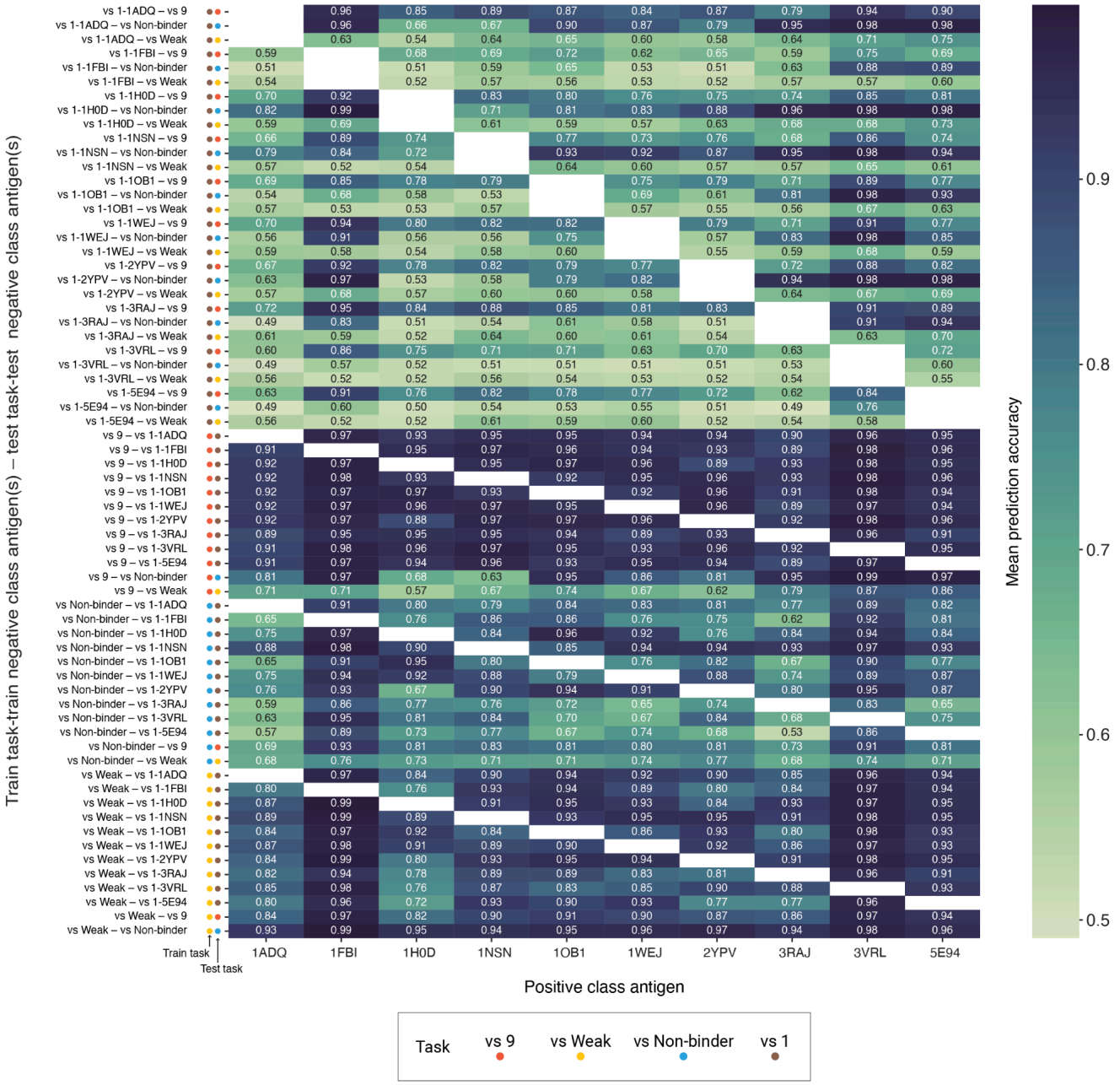


b Rule discovery potential and correlation between ID or OOD performance of SN10 model across tasks.



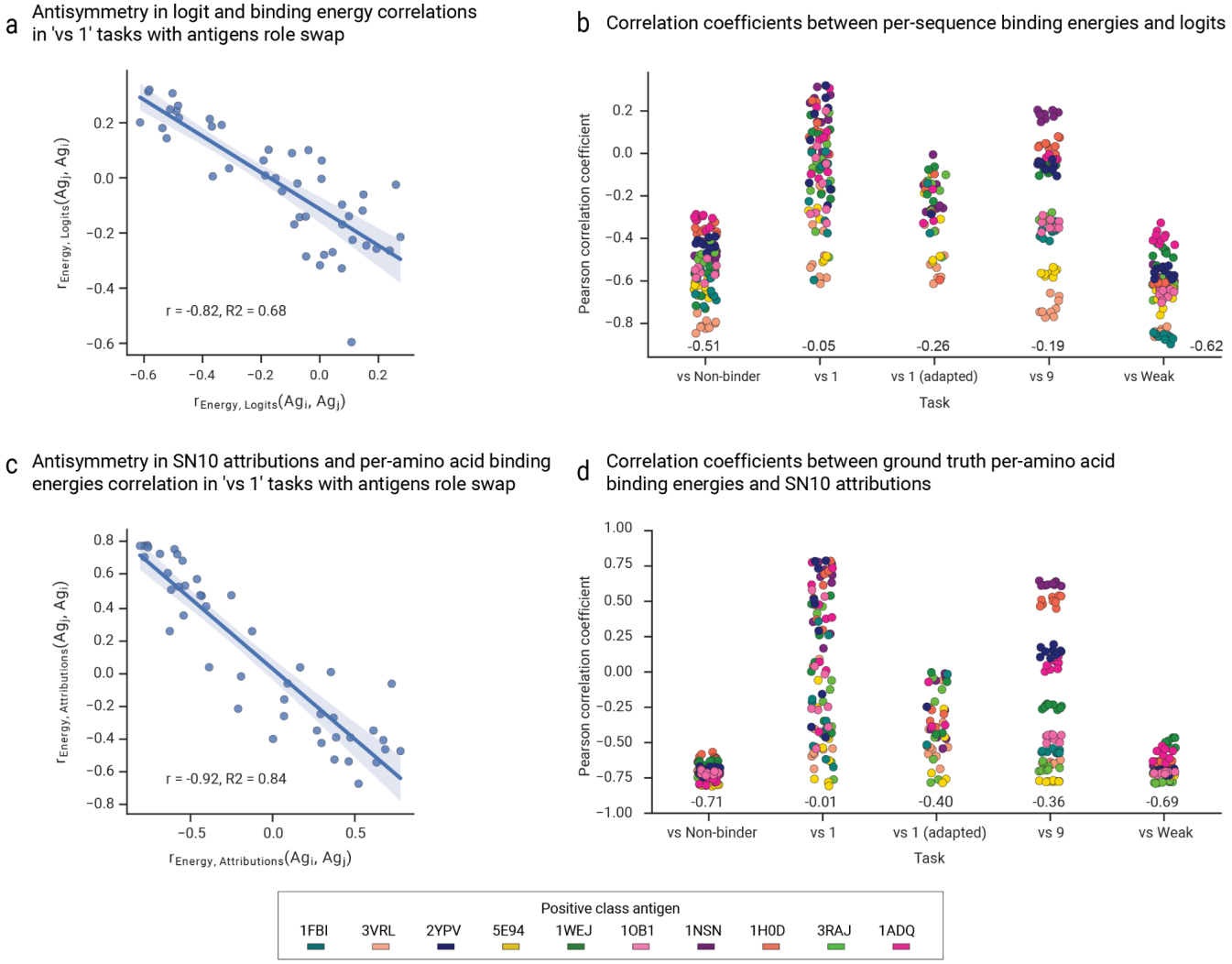
Supp. Figure 4. **Correlation between *ID* and *OOD* prediction accuracy and between *ID*, *OOD* and rule discovery.** (**a**) Scatterplots with Pearson correlation between *ID* and *OOD* prediction accuracy. Each subplot presents data corresponding to one training task (e.g., vs Weak) and the hue indicates the out-of-distribution testing task (e.g., vs Non-binder). For each training task, we present the performance on the 2 other out-of-distribution tasks, with each dot representing the mean over the replicates. *ID-OOD* correlation p-values < 0.05 for model trained on 'vs Weak' and p-value > 0.05 for – 'vs Non-binder' and 'vs 9'. (**b**) Scatterplots with Pearson correlation between rule discovery and either in-distribution prediction accuracy (diagonal) or out-of-distribution prediction accuracy (off-diagonal). Each subplot presents either data corresponding to an in-distribution scenario (task being indicated by the subplot title) or data corresponding to an out-of-distribution scenario (task used for training and task used for testing depicted in the subplot title). The hue indicates the type of rule discovery, i.e., correlations between binding energies and either SN10 logits (yellow) or attributions (green). Data across replicates was aggregated as median values.

Supplementary Figure 5



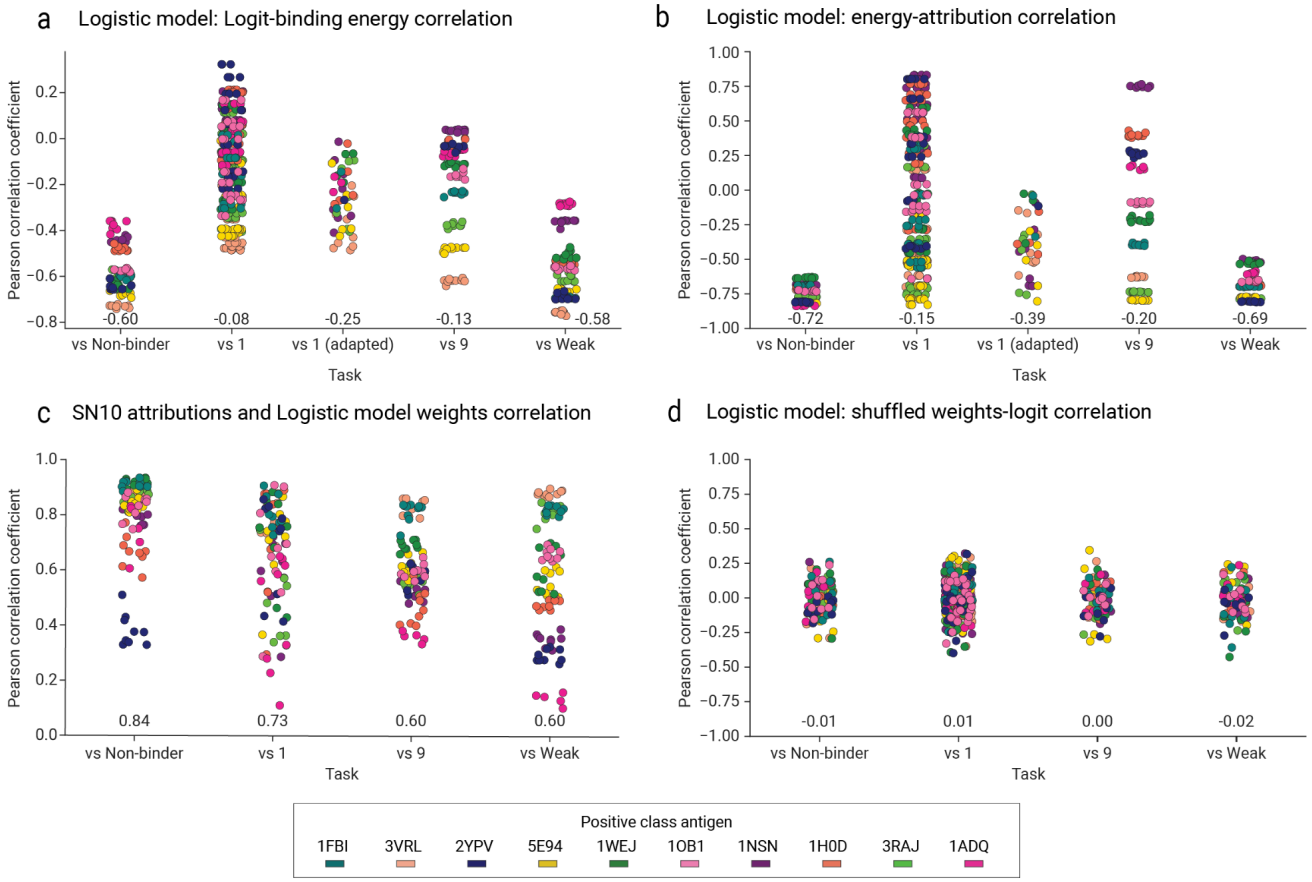
Supp. Figure 5. **Complete out-of-distribution test performance of SN10 model.** Each column of the heatmap represents a positive class antigen. Row labels are named by sequentially aggregating information about the train and test tasks and the target antigens of the negative class antibodies. For example, row 1 column 2 corresponds to 1FBI vs 1ADQ "vs 1" (train task) tested on 1FBI "vs 9" (test task). For convenience, colored circles represent train and test tasks respectively. The values shown in the heatmap are the prediction accuracy means over train-test data replicates. Relates to Figure 2D.

Supplementary Figure 6



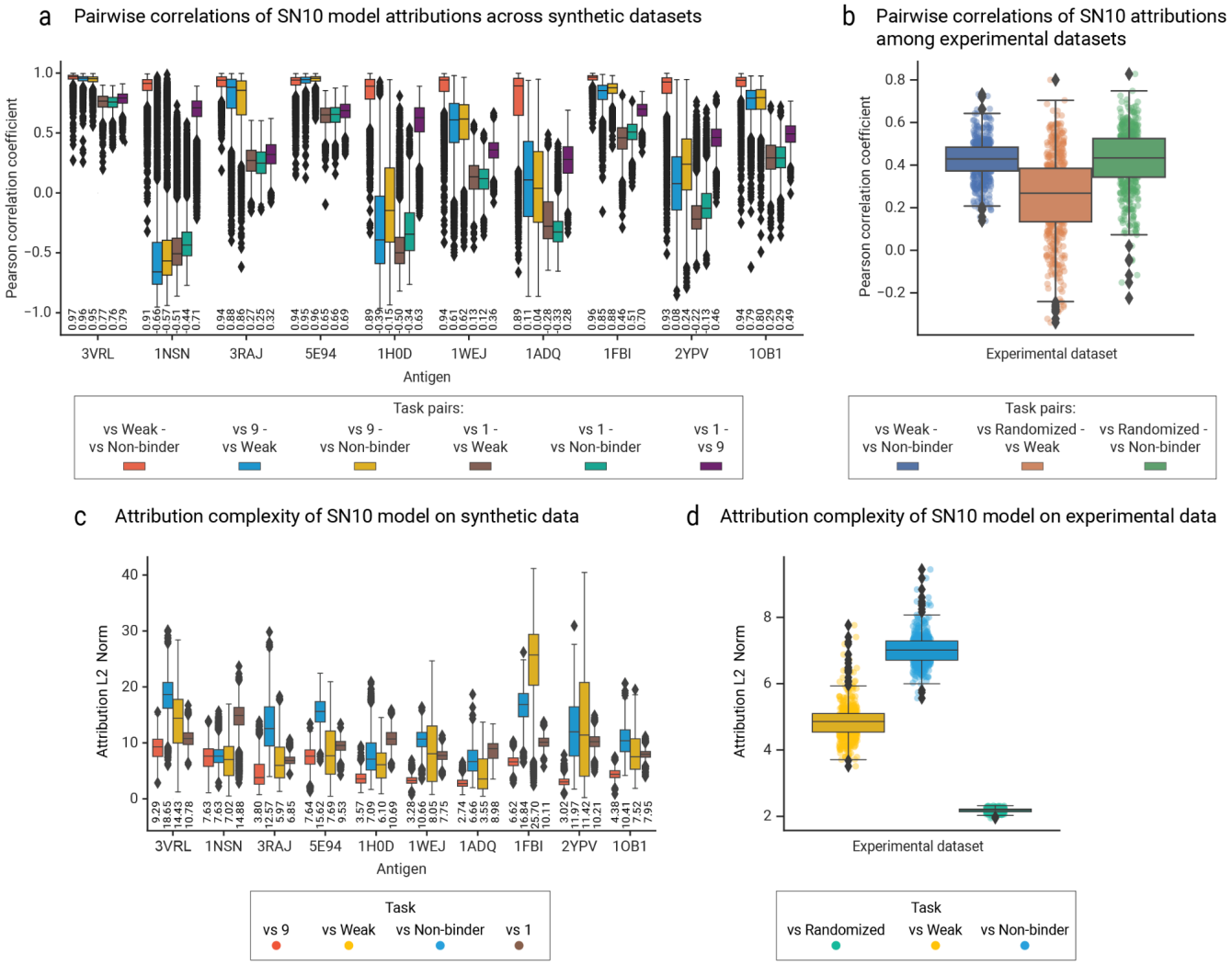
Supp. Figure 6**. Learning energy in "*vs 1*" tasks is asymmetrical and depends on the ordering of the antigen.** (**a**) Pearson correlation between the correlations of logits and per-sequence energies upon switching the role of positive and negative dataset antigen, i.e. computed in antigen pairs sorted alphanumerically first in ascending order (x-axis) and then in descending order (y-axis), e.g., a point corresponds to the correlation obtained in 3VRL vs 5E94 on the x-axis and 5E94 vs 3VRL on the y-axis. (**b**) Pearson correlation coefficients between logits and total binding energy in the positive class examples across different dataset types and antigens. This is an extended plot from Figure 3B with an extra column "vs 1 (adapted)", corresponding to keeping only the "vs 1" cases with the strongest inverse correlation between logits and per-sequence energy. (**c**) Identical to subplot A, except for illustrating correlations between per-residue energy and attribution. (**d**) Pearson correlation coefficients between energy contribution per CDRH3 amino acid and the SN10 amino acid attribution in the positive class samples across all dataset types and antigens. This is an extended plot from Figure 3D with an extra column "vs 1 (adapted)", corresponding to keeping only the "vs 1" cases with the strongest inverse correlation between attributions and per-residue energy. *As always in this work, we compute the correlations on the positive dataset.

Supplementary Figure 7



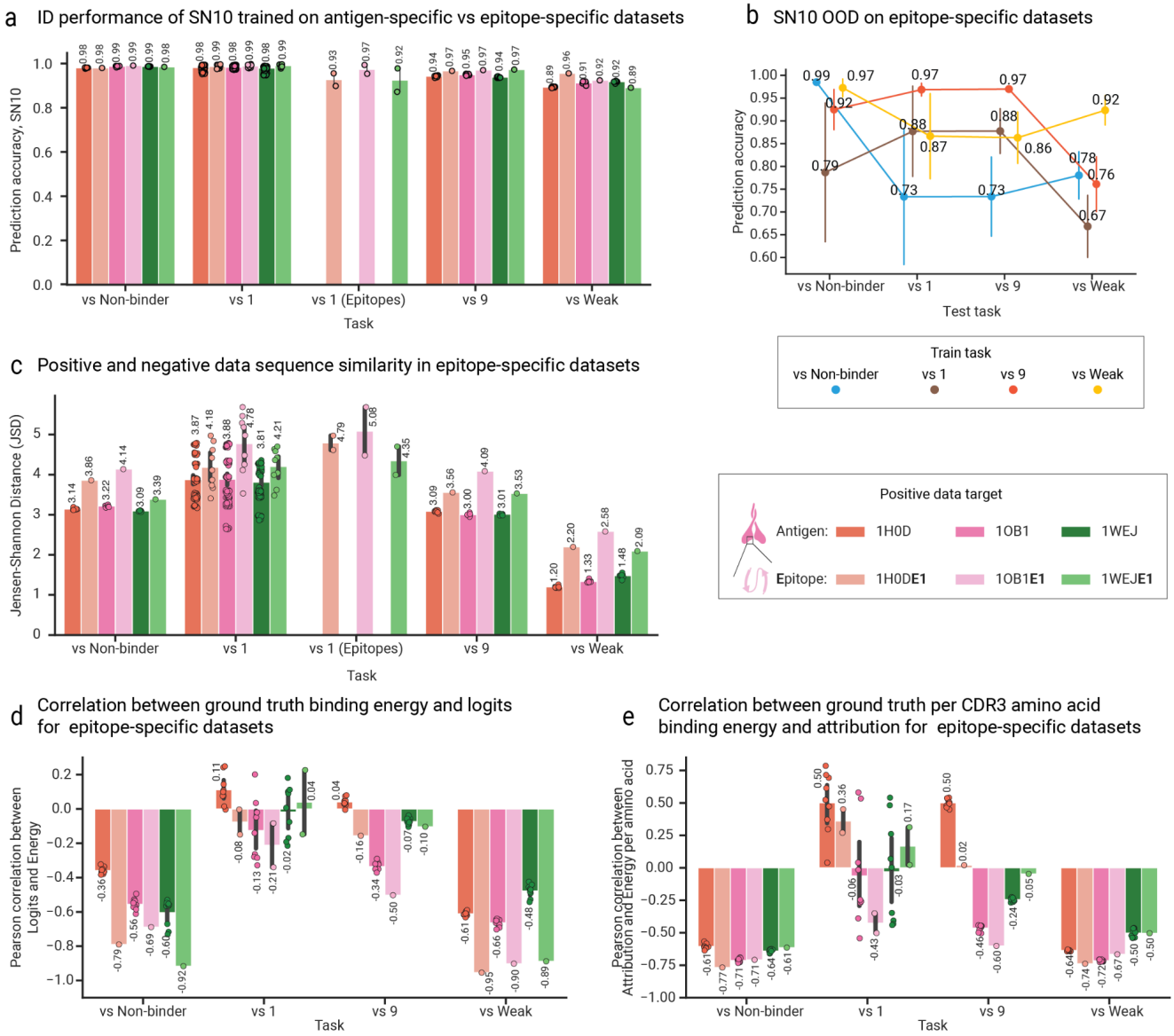
Supp. Figure 7. **Comparison of rule discovery (correlation of binding energy–logits and binding-energy-attributions) as a function of negative data types for SN10 and a logistic regression model.** (**a**) Pearson correlation coefficients between logits and total binding energy in the positive class examples across different dataset types and antigens for Logistic Regression (similar to Figure 3B, but with logistic regression). The numbers at the bottom of the scatter plots indicate the medians of Pearson correlation coefficients within the tasks. Each data point in the scatter plots represents a correlation coefficient calculated for the relationship between logits and total binding energy values corresponding to a model-antigen pair. (**b**) Pearson correlation coefficients between energy contribution per CDRH3 amino acid and the SN10 amino acid attribution in the positive class samples across all dataset types and antigens for Logistic Regression (similar to Figure 3D, but with logistic regression). The median values are presented at the bottom. (**c**) Pearson correlation coefficients between the attributions of SN10 and logistic regression weights, which correspond to contributions to prediction, in the positive class samples across all dataset types and antigens. (d) Pearson correlation coefficients between logits and total binding energy in the positive class examples across different dataset types and antigens for a negative control logistic regression in which the weights were shuffled (negative control corresponding to subplot A), showing no significant correlation, as expected. *As always in this work, we compute the correlations on the positive dataset. Relates to Figure 3B,D.

Supplementary Figure 8



Supp. Figure 8. **Comparison of attributions and attribution magnitudes between negative dataset types in synthetic (Absolut!) and experimental datasets.** (**a**) Pairwise Pearson correlation coefficients between the attributions in the positive class samples of SN10 trained on different negative datasets from Absolut!. Antigen groups are represented on the x-axis, correlation coefficients on the y-axis and the hue indicates the pair of task types, e.g. correlations between attributions of "vs 1" and "vs 9". (**b**) Pearson correlation coefficients between the attributions in the positive class samples of SN10 trained on different negative datasets from the experimental dataset (subplot A for experimental data). (**c**) Magnitudes of attribution vectors (L2 norms) across antigens and dataset types from Absolut!. (**d**) Boxplot of magnitudes of attribution vectors (L2 norms) across antigens and dataset types for the experimental dataset. Relates to Figure 3D.

Supplementary Figure 9



Supp. Figure 9. **Comparison of antigen-based and epitope-based approaches for model performance and rule discovery using different types of negative data (SN10 model).** (**a**) The in-distribution (*ID*) prediction accuracy of the SN10 model is grouped by negative dataset type and colored by positive class antigen. Antigen-based analyses are indicated by the antigen identifier (1H0D, 1OB1, 1WEJ). The corresponding epitope-based analyses are indicated by lighter colors and an adjusted identifier, in which "E1" is appended (1H0DE1, 1OB1E1, 1WEJE1). In the epitope-based analysis, the positive datasets are composed of sequences that bind to the major epitope of the corresponding antigen. Numbers displayed indicate median prediction accuracies between replicates. "vs 1 (Epitopes)" indicates a new category of tasks, in which the negative dataset is a set of sequences that are high binders w.r.t. a single epitope of a different antigen, e.g., 1WEJE1 vs 1H0DE1. (**b**) Aggregated *OOD* performance across antigens for the epitope-based analysis by training task (indicated by hue) and test task (represented as categories on the x-axis) for SN10. The plot includes tables that display the precise mean prediction accuracies. Error bars represent one standard deviation from the mean, capturing 68% of the variation in out-of-distribution performances. Relates to Supp. Figure 3D, which showcases the similar plot for the main, antigen-based analysis. (**c**) Overview of the dissimilarity of the positive class and negative class sets of CDRH3 sequences (JSD) across the antigen-based and epitope-based datasets, as measured by the Jensen-Shannon Distance (JSD). "vs 1 (Epitopes)" represents the same category as in (a). (**d**) Pearson correlation coefficients between logits and total binding energy in the positive class examples across different dataset types and antigens. Tasks, antigens and epitopes are represented as in (a) and (b). We restricted the computation to the positive class examples from a dataset, since in practical scenarios, such as those corresponding to the "vs 1" and "vs 9" tasks, data regarding the binding of negative class sequences to the antigen of the positive class will be unavailable. The numbers indicate the medians of Pearson correlation coefficients within the tasks. Each data point in the bar plots represents a correlation coefficient calculated for the relationship between logits and total binding energy values corresponding to a model-antigen pair. Figure 3B, which presents the results for all antigens from the main antigen-based analysis. (**e**) Pearson correlation

coefficients between energy contribution per CDRH3 amino acid and the SN10 amino acid attribution in the positive class samples across all dataset types and antigens. The median values are presented. Similar plot to (d), except presenting the correlations per amino acid, rather than per sequence. Relates to and referenced in Supp. Text 2.

Supplementary Table 1

Supplementary Table 1. SN10 model prediction accuracy and Jensen-Shannon Distance for negative controls (shuffled data) in the case of in-distribution testing. Controls relate to Figure 2 (b, c).

Task type	Accuracy average	Accuracy standard deviation	JSD average	JSD standard deviation	Pearson correlation and p-value between accuracy and JSD	Number of replicates
vs Non-binder	0.499	0.005	0.398	0.025	r=-0.01, p=0.99	10
vs 1	0.5	0.005	0.403	0.023	r=0.15, p=0.17	90
vs 9	0.497	0.004	0.417	0.022	r=0.38, p=0.28	10
vs Weak	0.501	0.004	0.405	0.018	r=0.40, p=0.25	10

Supplementary Table 2

Supplementary Table 2. SN10 model prediction accuracy for negative controls (shuffled class labels) in the case of out-of-distribution testing. Controls relate to Figure 2 (d, e).

Task type used for training the model	Task type used for testing the model	Accuracy average	Accuracy standard deviation	Number of replicates	Pearson correlation and p-value between delta accuracy and delta JSD
vs 1	vs 9	0.5	0.005	900	r=-0.0, p=0.89, R2=0.0
vs 1	vs Non-binder	0.5	0.005	900	r=0.08, p=0.01, R2=0.01
vs 1	vs Weak	0.5	0.005	900	r=0.16, p=0.0, R2=0.03
vs 9	vs 1	0.5	0.005	900	r=0.17, p=0.0, R2=0.03
vs 9	vs Non-binder	0.5	0.005	100	r=0.1, p=0.31, R2=0.01
vs 9	vs Weak	0.501	0.005	100	r=0.14, p=0.17, R2=0.02
vs Non-binder	vs 1	0.5	0.005	900	r=0.05, p=0.16, R2=0.0
vs Non-binder	vs 9	0.5	0.005	100	r=-0.06, p=0.56, R2=0.0
vs Non-binder	vs Weak	0.499	0.005	100	r=-0.03, p=0.74, R2=0.0
vs Weak	vs 1	0.501	0.005	900	r=0.1, p=0.0, R2=0.01
vs Weak	vs 9	0.5	0.005	100	r=0.22, p=0.03, R2=0.05
vs Weak	vs Non-binder	0.5	0.004	100	r=0.1, p=0.31, R2=0.01

Supplementary Table 3

Supplementary Table 3. Correlations between predicted binding energy and logits / model attributions for the negative controls (shuffled class labels). Controls relate to Figure 3 (b, d).

Task type	**Correlation coefficient between logits and predicted binding energy per CDRH3**				**Correlation coefficient between per-residue model attribution and predicted binding energy per residue of CDRH3**			
	r min	**r median**	**r max**	**N**	**r min**	**r median**	**r max**	**N**
vs 1	-0.05	0.02	0.07	90	-0.11	0.04	0.32	90
vs 9	-0.03	0.01	0.04	10	-0.07	0.01	0.08	10
vs Non-binder	0	0.02	0.08	10	-0.05	0	0.02	10
vs Weak	-0.02	0.03	0.06	10	-0.09	0.06	0.12	10

Supplementary Text 1

Evaluation of the impact of sequence and label similarity between train and test on prediction accuracy

To quantify to which extent sequence similarity between train and test sets relates to in-distribution (*ID*) performance, we determined the frequency of sequences with more than 90% identity, i.e., exactly one-position difference in 11-mer CDRH3 sequences, in train and test dataset across the four negative training dataset definitions. The highest proportion of such sequences was observed in the 'vs 1' datasets (37–50%), followed by the 'vs Weak' dataset (30–44%), 'vs 9' (25–32%), and then 'vs Non-binder' (22–31%) (Supp. Figure 2A). Similarly, in the experimental dataset, a greater number of similar sequences between test and train were found in the 'vs Weak' data (65%) compared to the 'vs Non-binder' (44%) (Supp. Figure 2A). Notably, sequences with 90% similarity may belong to the same or different binding classes (positive and negative), which could result in similar or dissimilar binding energies (Supp. Figure 2B). Focusing on sequences that, in addition to having 90% sequence identity, i.e. having a one point mutation difference, also shared a similar binding class, we observed an insignificant (p-value $\geq$ 0.05) positive correlation with *ID* prediction accuracy: 0.38 for 'vs 9', 0.36 for 'vs Weak', 0.21 for 'vs Non-binder', and 0.20 for 'vs 1' (Supp. Figure 2C). When considering sequences with 90% identity but discordant labels (i.e., positive class in training and negative in testing, or vice versa), we observed a decrease in correlation value for 'vs Non-binder' (r=0.05, p-value>0.05), a change in sign for 'vs 9' (r=-0.27, p-value>0.05), and for 'vs 1' (r=-0.47, p-value<0.05); or, as in the case of the 'vs Weak' task, an increase in correlation value (r=0.54, p-value>0.05). Overall, all the correlations were insignificant, with the exception of the correlation between the proportion of discordant labels and *ID* accuracy in "vs 1".

This investigation emphasizes the increased proportion of discordant sequences between train and test datasets in the 'vs 9' and 'vs Weak' tasks (Supp. Figure 2A). These discordant sequences can potentially "highlight" key amino acids that determine binding. Given that the correlation between the number of similar sequences and prediction accuracy did not reach significance in most cases, we cannot conclude that the presence of similar sequences between the training and test sets contributes to improving *ID* performance. We hold that data leakage from the training to the test dataset due to sharing similar sequences with identical labels was unlikely across all training tasks.

Supplementary Text 2

Antigen versus epitope-based analysis: ID, OOD, rule discovery

We conducted supplementary analyses to further assess the validity of our findings and to examine their sensitivity to the restriction of datasets to a single epitope per antigen. We assessed the accuracy of in-distribution (*ID*) and out-of-distribution (*OOD*) predictions, the Jensen-Shannon distances (JSD), and the logit- and attribution-based rule discovery within epitope-based datasets, contrasting them with antigen-based datasets. Specifically, we compared datasets containing only sequences binding to a single epitope, i.e., epitope-based, with previously utilized datasets containing sequences binding to multiple antigen epitopes, i.e., antigen-based (Supp. Figure 9). To maintain maximal dataset sizes for epitope-based datasets and retain the same training set size as in antigen-based datasets, we had to restrict the analysis to three antigens: 1H0D, 1OB1, 1WEJ (see also Methods).

Epitope-based datasets yielded *ID* prediction accuracies as high as those for antigen-based datasets, with some marginal improvements in "vs 9" and "vs Weak" (Supp. Figure 9A). These marginal enhancements are accompanied by increases in Jensen-Shannon Divergence (JSD) (Supp. Figure 9C), consistent with previous findings indicating that increased JSD is associated with improved performance (Figure 2C). The increased JSD for epitope-based datasets could be attributed to less heterogeneous position-weight matrices for the positive datasets within the epitope-based datasets. Regarding *OOD* predictions, epitope-based datasets demonstrated similar performance and trends across negative class types as the antigen-based datasets (Supp. Figure 9B, Supp. Figure 3D).

Epitope-based datasets exhibited significant improvements over the antigen-based tasks in logit-based rule discovery across "vs Non-binder", "vs Weak", and "vs 9" for the examined antigens (Supp. Figure 9D). This could be intuitively explained by both a more consistent rule discovery due to less heterogeneity in the positive class and a simpler positive class for rule evaluation. However, no consistent improvement was observed in attribution-based rule discovery (Supp. Figure 9E). As noted previously, the discovery of per-residue binding energies appears to be more complex and challenging to achieve compared to per-sequence binding energy discovery.