# SPATIALLY-INFORMED SAMPLING ENABLES ACCURATE PREDICTION OF LARGE-SCALE MUTATIONAL EFFECTS

**Maxime Basse**[1,2], **Dianzhuo Wang**[1,3], **Eugene I. Shakhnovich**[1,†]
[1]Department of Chemistry and Chemical Biology, Harvard University
[2]Ecole Polytechnique, Institut Polytechnique de Paris
[3]John A. Paulson School of Engineering and Applied Sciences, Harvard University
† shakhnovich@chemistry.harvard.edu

## ABSTRACT

Predicting protein binding affinities across large combinatorial mutation spaces remains a critical challenge in molecular biology, particularly for understanding viral evolution and antibody interactions. While combinatorial mutagenesis experiments provide valuable data for training predictive models, they are typically limited due to experimental constraints. This creates a significant gap in our ability to predict the effects of more extensive mutation combinations, such as those observed in emerging SARS-CoV-2 variants. We present PROXICLUST, which strategically combines smaller combinatorial mutagenesis experiments to enable accurate predictions across larger combinatorial spaces. Our approach leverages the spatial proximity of amino acid residues to identify potential epistatic interactions, using these relationships to optimize the design of manageable-sized combinatorial experiments. By combining just two small combinatorial datasets, we achieve accurate binding affinity predictions across substantially larger mutation spaces ($R^2 \approx 0.8$), with performance strongly correlated with capture of high-order epistatic effects. We validated our method in five different protein-protein interaction datasets, including binding of SARS-CoV-2 receptor binding domain (RBD) to various antibodies and cellular receptors, as well as influenza RBD-antibody interactions. This work provides a practical framework for extending the predictive power of combinatorial mutagenesis beyond current experimental constraints, offering applications in viral surveillance and antibody engineering.

## 1 INTRODUCTION

Predicting the dissociation constant ($K_D$) between viral proteins and monoclonal antibodies, as well as the cellular receptor Angiotensin-converting enzyme 2 (ACE2), is critical for assessing viral infectivity for viruses like SARS-CoV-2, as demonstrated in these works Wang et al. (2024); Cia et al. (2022); Hie et al. (2021); Huot et al. (2025). These models, whether based on biophysics or machine learning, utilize the $K_D$ of various antibodies and the receptor-binding domain (RBD) to effectively predict viral fitness.

High-throughput experimental methods provide a means to acquire $K_D$ data across extensive sequence spaces. A prominent technique is Combinatorial Mutagenesis Moulana et al. (2023; 2022). In these experiments, researchers select $L$ specific loci within the Receptor Binding Domain (RBD), testing every possible combination of amino acid mutations at each locus. This method generates a comprehensive dataset comprising $2^L$ combinations, capturing a broad spectrum of mutational impacts. This methodology is particularly crucial in the context of viral evolution, where early in a pandemic, variants of concern (VoCs) with specific mutations are observed Deng et al. (2021); Leung et al. (2021). It is vital to predict the phenotypic or fitness outcomes of variants combining these mutations, as illustrated in Figure 1. However, currently, the capabilities of these methods are restricted to experiments where $L \leq 16$, due to the exponential growth in the number of combinations with larger values of $L$. VoCs such as BA.2 BA.3 and XBB variants Tamura et al. (2024), often

exhibit more than 16 mutations Carabelli et al. (2023), underscoring a gap between current experimental capabilities and the need to effectively predict infectivity of VOCs with more mutations.

Figure 1 illustrates how a combinatorial set of Variants of Concern (VOCs) can be derived from mutations observed in circulating viral strains. Accurate predictions across this combinatorial space are essential for reconstructing viral fitness landscapes, as these sets encompass all possible intermediary variants between wildtype and observed mutants. The mutations included in these sets are particularly significant because they originate from viruses that have demonstrated successful transmission and survival in natural populations. Consequently, new variants combining these individually beneficial mutations have an increased likelihood of being both infectious and viable. While the combinatorial space of $2^L$ variants represents only a small fraction of the total possible sequence space ($20^K$, where $K$ is the total number of residues in the viral protein), it is enriched for potentially concerning variants due to this selective sampling of successful mutations. The ability to accurately predict $K_D$ and fitness across this space is therefore crucial for proactive vaccine development, therapeutic adaptation, and evidence-based public health policy decisions.

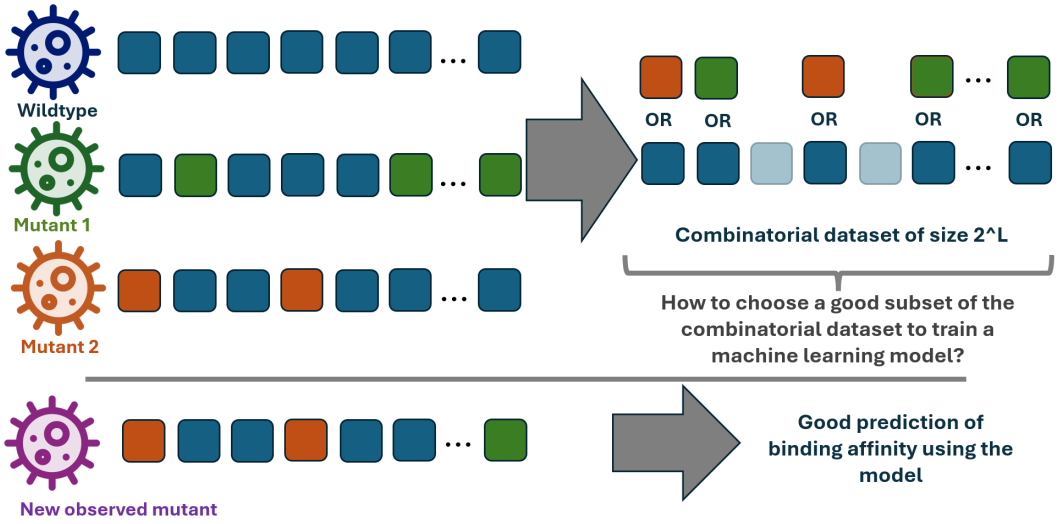



Figure 1: A more infectious Variant of Concern (VoC) may incorporate mutations from existing VoCs. Consequently, it is crucial to predict the biophysical properties of the entire combination of existing VoCs to better understand and anticipate their evolutionary trajectories.

During the early stages of a pandemic, when experimental capacity is often limited relative to the urgent need for data, traditional experimental methods alone may not be sufficient to characterize all emerging viral variants. In these scenarios, machine learning approaches trained on limited available experimental data have proven to be valuable tools for estimating $K_D$ of novel variantsWang et al. (2024); Han et al. (2023); Loux et al. (2024). This computational approach helps bridge the gap between experimental capabilities and the pressing need for rapid viral surveillance.

Significant advances have been made in computational methods for predicting $K_D$ to address limitations in experimental throughput. These approaches span a broad methodological spectrum, from transformer-based architectures Wang et al. (2023) and physics-informed deep learning models Chen et al. (2021) to free energy perturbation calculations Sergeeva et al. (2023). Complementing these modeling advances, active learning (AL) strategies have emerged as powerful tools for optimizing experimental data collection. AL enable strategic sampling from the vast space of possible variants, prioritizing measurements that maximize information gain for model improvement. This could significantly reduce the amount of training data required while maintaining high model accuracy Hie et al. (2020). However, a key limitation of active learning is its inherently sequential nature - experiments must be conducted one after another based on model feedback. This sequential constraint makes AL incompatible with high-throughput experimental methods that require upfront design of

mutation libraries. The benchmark of AL and other low-throughput sampling methods could be-found in SI.

To overcome this fundamental limitation, we investigate optimal methods for combining smaller, experimentally feasible combinatorial datasets to enable accurate predictions across larger combinatorial spaces. To this end, we introduce "PROXICLUST", which leverages structural information from protein-protein binding poses to systematically determine which mutations should be grouped together in combinatorial experiments.

## 2 METHODS

### PROXICLUST

We introduce PROXICLUST, a method based on the principle that combinatorial experiments reveal high-order epistatic interactions among mutations. Our key insight is that mutations with potential interactions should be tested together in experimental designs. These interacting mutations typically form "epistasis clusters" - groups of residues that functionally influence each other. While precise epistatic maps of viruses are rarely available, we leverage the observation that spatial proximity between amino acids serves as a reliable predictor of epistatic interactions Wang et al. (2024). PROXICLUST requires only structural information from the antibody-antigen binding pose, typically available through Protein Data bank Format files of wildtype variants, and proceeds in two steps:

Our method proceeds in two steps. First, we identify interface residues between the antibody-antigen complex by analyzing spatial coordinates in the binding pose. These interface residues can be determined either computationally using protein structure prediction methods like AlphaFold Multimer Jumper et al. (2021), or experimentally through cryo-electron microscopy. Following standard practice in protein-protein interaction studies, we define interface residues as those within a 10Å distance threshold of the binding partner.

Second, we apply a two-stage clustering approach to these interface residues (Figure 2b). We first employ M-means clustering to establish well-distributed centroids across the binding interface, followed by an L-Nearest Neighbors algorithm initiated from these centroids. This hierarchical approach, implemented using scikit-learn's KMeans and KNeighborsClassifier Pedregosa et al. (2011), creates partially overlapping clusters that capture local structural relationships while maintaining diversity across the interface. The overlap between clusters is carefully controlled through centroid initialization to maximize coverage of potential epistatic interactions while minimizing redundant measurements, as validated by our epistasis analysis (Figure 4).

The algorithm ultimately organizes interface residues into M groups of L residues each, where L represents the maximum number of mutations that can be included in a single combinatorial experiment (determined by experimental constraints), and M is the desired number of parallel experiments. Throughout this study, we demonstrate results for $M = 2$, showing that just two well-designed combinatorial experiments can effectively capture the key epistatic interactions in the system.

### DATASETS

In this study, we analyzed five datasets, including the combinatorial datasets of SARS-CoV2 RBD binding with ACE2 Moulana et al. (2022), LY-CoV016, LY-CoV555, REGN10987, and S309 Moulana et al. (2023), as well as the anti-influenza receptor binding site (RBS), CH65 binding to H1 Phillips et al. (2023).

The SARS-CoV-2 datasets Moulana et al. (2022; 2023) explore the combinations of 15 mutations on the RBD corresponding to the 15 mutations observed on Omicron BA.1 variant compared to the Wuhan wildtype: G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H. The influenza dataset Phillips et al. (2023) contains all possible evolutionary intermediates between the unmutated common ancestor and the mature somatic sequence of the CH65 antibody, which binds to diverse H1 strains. The dataset contains all combinations of 16 mutations located on both the heavy and light chain of the CH65 antibody.

These datasets are comprehensive combinatorial mutagenesis scans that measure binding affinity values for all possible combinations of selected mutations. Each mutant variant is represented us-

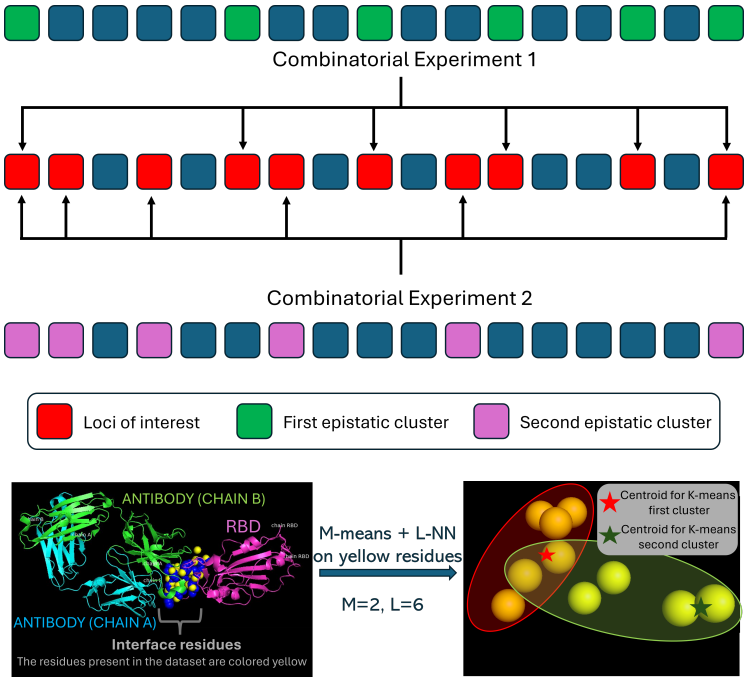



Figure 2: (a) Illustration of using two smaller combinatorial experiments to predict the outcomes of a larger experiment that is out of current capabilities (b) PROXICLUST employs M-means clustering and L-nearest neighbors algorithms to determine which amino acids should be grouped together in the same experiment.

ing a binary one-hot encoding scheme, where '1' indicates the presence of a mutation at a specific position and '0' represents the wild-type residue, as illustrated in Table 1. For a set of L selected mutation sites, this generates a complete combinatorial space of $2^L$ variants, systematically exploring all possible mutational combinations.

## 3 RESULTS

In this section, we address the challenge of assembling multiple smaller combinatorial scans into a training set to predict a larger combinatorial dataset. We benchmark PROXICLUST against random assembly strategies of the same dimensionality. Here, a random assembly strategy refers to selecting mutations for each scan independently as a random subset from the pool of possible mutations.

Figure 3 compares the performance of a random forest machine learning model using two distinct strategies for assembling combinatorial datasets:

- The boxplots on the left display the distribution of $R^2$ scores for strategies that assemble two combinatorial scans, each targeting mutations specifically at the interface of the antigen-antibody binding pose.
- The boxplots on the right show the distribution of $R^2$ scores for strategies that assemble two combinatorial scans, but with no constraints on the mutations included.

Comparing the medians in Figure 3, indicated by the orange horizontal lines, we observe that strategies targeting the interface consistently achieve higher scores. This finding aligns with our current understanding of protein-protein interactions - residues at binding interfaces are known to make the most significant contributions to binding affinity through direct physical contacts Moulana et al. (2022); Vangone & Bonvin (2015). The superior performance of interface-focused strategies can thus be attributed to their ability to capture these critical first-order effects in the training data. By

concentrating on interface residues, these approaches naturally prioritize the most functionally relevant mutations, leading to more accurate predictions of binding affinity changes.

Our analysis reveals that PROXICLUST consistently outperforms both the median combinatorial scan assembly and the median interface assembly strategy (Figure 3). A detailed performance breakdown across all five tested datasets is presented in SI Table 2, where quantile rankings demonstrate how our method compares against a comprehensive pool of randomized residue selection strategies. Further detailed comparisons within interface assembly strategies are visualized in SI Figure 6. The performance advantage of PROXICLUST varies notably across different systems. For REGN10987 and LY-CoV555 antibodies, where the $R^2$ score distributions peak around 0.8, the improvement over median performance is modest ($\leq 0.3$). However, for the Influenza virus and S309 antibody systems, PROXICLUST demonstrates a more substantial advantage, achieving $R^2$ scores of approximately 0.8 compared to random strategies' scores of around 0.4. This variable performance across different systems suggests that the effectiveness of our method may depend on the underlying structural and biochemical properties of the protein-protein interaction being studied.

Note that the assembling strategies being compared are always of the same "dimension" , meaning they consist of an equal number of combinatorial scans, each of the same size. This uniformity is crucial as, aside from the assembly strategy, the size of the training set is a key determinant in the performance of the predictive model. Therefore, it is essential to compare strategies that provide training sets of comparable sizes to ensure valid conclusions about their efficacy.

It is important to note that while strategies of the same dimension incur identical experimental costs (requiring the same number and type of laboratory experiments), they can yield different numbers of unique binding affinity measurements. To illustrate this concept, consider two contrasting scenarios: In the first case, a strategy using two scans with completely distinct sets of 5 mutations each generates $2^5 + 2^5 = 64$ unique measurements. In contrast, a strategy of identical dimension but with 4 overlapping mutations between scans yields fewer unique measurements - $2^5$ measurements from the first scan plus only $2^5 - 2^4 = 16$ additional unique measurements from the second scan, totaling 48 unique mutants in the training set. This difference in measurement efficiency arises from the redundancy introduced by overlapping mutations, despite equivalent experimental effort.

### EPISTASIS ANALYSIS

We have developed a method to quantify the epistasis captured by a training set derived from any given assembly strategy. This method leverages the Walsh-Hadamard transform.

Walsh-Hadamard transform Poelwijk et al. (2016) is a method adapted from signal processing to study high-order epistasis(non-linearity) Faure et al. (2024). This approach is particularly advantageous when working with complete combinatorial datasets, as it provides exhaustive information about the interactions between residues Weinreich et al. (2013). Given that our binding affinity datasets are combinatorial in nature (of size $2^L$), we can derive the effect of any order of epistasis by applying the Walsh-Hadamard transform to the vector of experimentally measured binding affinities.

In the matrix form, this could be written as:

$$\frac{1}{2^L}\boldsymbol{H}\boldsymbol{E} = \boldsymbol{W}$$

where H is the Hadamard matrix with size $2^L \times 2^L$, and E is the Kd vector with size $2^L \times 1$. W is the Walsh coefficients vector with size $2^L \times 1$, where each component $W_g$ represents the importance (or weight) of a specific genotype $g$Weinreich et al. (2013).

A data acquisition strategy composed of multiple combinatorial scans can be presented as $[l_1, \ldots, l_M]$, representing the assembly of $M$ individual combinatorial scans. Each $l_k$ is a subset of $\{1, \ldots, L\}$ that specifies the positions (or loci) in a genotype where mutations (represented by 1s) can occur. A genotype $g$ in $\{0, 1\}^L$ is a sequence of binary values, where a 1 indicates the presence of a mutation at a given position, and a 0 indicates no mutation.

We define $G_k$ as the set of all genotypes $g$ in the combinatorial mutagenesis scan on the loci of $l_k$. Formally:

$$G_k \implies \left\{g \in \{0,1\}^L \mid \forall i,\, g_i = 1 \implies i \in l_k\right\}$$

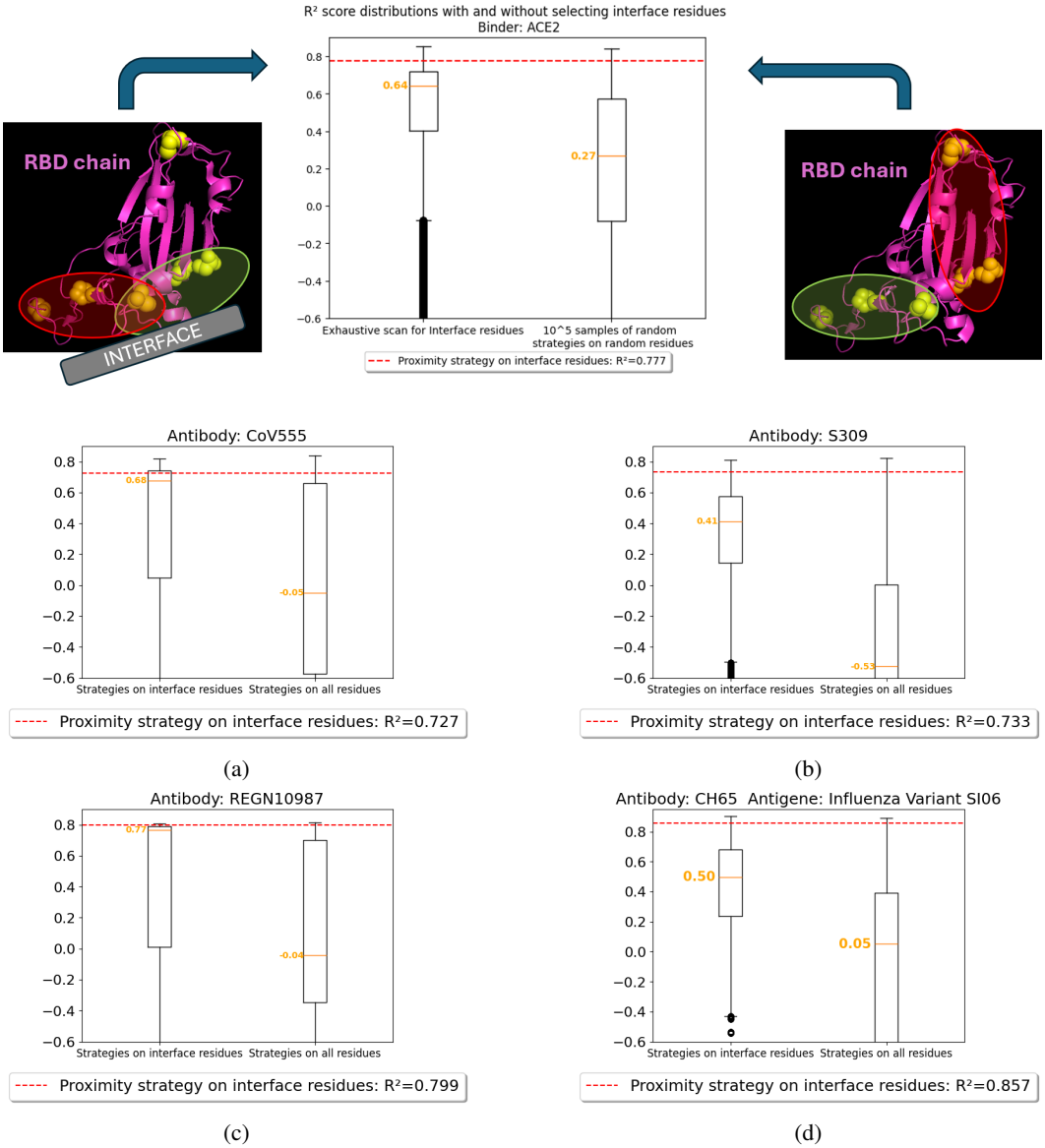



Figure 3: Comparison of $R^2$ scores of models trained using data acquisition strategies focused on the antibody-antigen interface (left boxplots) versus unconstrained strategies (right boxplots). Red line shows the $R^2$ score achieved by PROXICLUST. (a) Comparison for RBD and ACE2 binding. The accompanying protein illustrations represent possible clustering of the dataset's residues shown with yellow spheres. On the left, the clustering is done among the yellow spheres close to the interface whereas on the right, all residues can be included in the clusters. (b)-(d) Similar analyses were conducted for RBD binding with antibodies LY-CoV555, S309, and REGN10987, respectively. (e) Same experiment run on another combinatorial dataset for binding affinity predictions between Influenza variant SI06 and variants of antibody CH65

**Example**: Suppose $L = 5$ and $l_k = \{2, 3\}$. Here, $G_k$ would include genotypes $g$ in $\{0, 1\}^5$ with the following conditions:

- $g_2 = 1$ if and only if $2 \in l_k$
- $g_3 = 1$ if and only if $3 \in l_k$
- For any other position $i$, $g_i = 0$ if $i \notin l_k$

Thus, the valid genotypes in $G_k$ would be $\{01000, 00100, 01100, 00000\}$.

We can then define $G$ as the union of all such sets:

$$G = \bigcup_{k=1}^{M} G_k$$

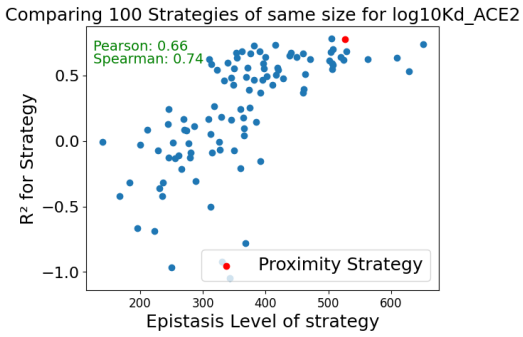
The *epistatic score* for the strategy is given by:

$$\text{epistatic_score} = \sum_{g \in G} |W_g|$$

We demonstrate the *epistatic score* of the strategy correlates with its predictive power. This epistatic score quantifies the extent to which large interaction effects between residues are represented in the dataset produced by a specific strategy. PROXICLUST, designed to identify and utilize epistasis clusters within the protein's binding interface, is expected to perform favorably in terms of epistatic score. Figure 4a presents the epistatic scores for 100 randomly selected assembly strategies, demonstrating a positive correlation between this metric and the $R^2$ score of the machine learning models trained using these strategies, with a Spearman correlation coefficient of 0.74. Notably, PROXICLUST achieves superior performance in both $R^2$ and epistatic scores compared to the random benchmarks. Results for other datatsets are in SI Figure 7.

Further analysis shown in Figure 4b extends these findings in all five data sets, with PROXICLUST consistently ranking within the top 11% of strategies in terms of epistatic score. The relationship between $R^2$ and epistatic score remains robust across these datasets, evidenced by an average Spearman correlation of 0.822 (refer to SI Figure 7).

This comprehensive analysis demonstrates that PROXICLUST's superior performance arises directly from its systematic ability to identify and capture functionally significant epistatic interactions at the protein binding interface, validating our hypothesis that spatial proximity serves as an effective proxy for mutational interdependence.




(a)

| Antibody | Percentile Rank (%) |
|---|---|
| ACE2 | 96.4 |
| S309 | 94.9 |
| REGN10987 | 89.96 |
| CoV555 | 96.34 |
| SI06 (Flu) | 98.58 |

(b)

Figure 4: (a) Correlation between model performance ($R^2$) and epistasis score across 100 random strategies, with PROXICLUST highlighted in red (Spearman: 0.74). (b) PROXICLUST consistently ranks in the top percentiles for epistasis capture across all tested strategies.

## 4 DISCUSSION

This study has demonstrated the potential of using smaller, targeted combinatorial datasets to predict larger, more complex combinatorial datasets with high accuracy at the beginning of a pandemic. PROXICLUST utilizes spatial proximity to infer epistatic interactions, has shown promising results and consistently outperformed random strategies across various datasets.This research demonstrates that just two well-designed combinatorial scans can effectively approximate the entire binding affinity landscape. By strategically leveraging spatial proximity between residues, we focus on mutations with the strongest contributions to binding affinity while avoiding the collection of less significant epistatic effects between residues that are distant in space.

A key implication of our work is its potential application in pandemic preparedness. Early in a pandemic, when new viral variants emerge, our method could guide the strategic design of combinatorial libraries to maximize predictive power. By strategically selecting smaller, experimentally feasible combinatorial sets, researchers could more efficiently predict the effects of larger mutation combinations that might appear in future variants. While existing approaches like the BO-EVO algorithm by Hu et al. (2023) use active learning to guide limited experimental efforts during pandemic response, our method is unique in leveraging a priori structural and biophysical intuition to optimize experimental design before any measurements are taken.

Our findings also raise interesting insights about the fundamental nature of protein-protein interactions. The success of proximity-based clustering in capturing epistatic effects suggests that local structural environment, especially on the binding surface, plays a dominant role in determining binding affinity. This is also demonstrated in several recent publications using local environment to predict protein binding, including the HERMES method which showed that holographic encoding of local atomic environments within 10Å of a focal residue can effectively predict binding affinity and stability effects Visani et al. (2024).

The correlation between epistatic scores and predictive performance provides valuable insights into the mechanistic basis of our method's success. This relationship suggests that spatial proximity and epistasis are a reliable proxy for functional interactions between residues. While epistatic scores cannot be calculated a priori without experimental data, the consistent correlation between spatial clustering and epistatic effects suggests that our proximity-based approach effectively captures functionally important interactions. Furthermore, this correlation offers a potential metric for evaluating future dataset assembly strategies without requiring extensive experimental validation.

However, several limitations and opportunities for future research should be noted. First, while our method has been successfully validated on datasets containing up to 16 mutations - currently the largest available combinatorial datasets - many real-world scenarios involve substantially larger mutation spaces.Kaku et al. (2024) The correlation between spatial proximity and epistatic effects can become more complex with increasing numbers of mutations, and optimal clustering parameters may need to be adjusted. Nevertheless, the consistent success across our diverse test cases and the strong theoretical foundation linking spatial proximity to epistatic interactions suggest that our approach provides a promising framework for tackling larger mutation spaces.

Second, our method currently relies on structural information about the binding interface. While such information is increasingly available through methods like AlphaFold2 Abramson et al. (2024); Bryant et al. (2022), there are still cases where accurate protein complex structures are unavailable. In particular, antibody-antigen complexes and other adaptive immune system interactions remain challenging to AlphaFold modelYin et al. (2022). Developing alternative data assembling strategies for cases without reliable structural information, perhaps based on sequence conservation or other features, could expand the method's applicability.

## MEANINGFULNESS STATEMENT

A meaningful representation of life should capture the complex interdependencies and multi-scale relationships that govern biological systems. In protein-protein interactions, this means understanding how spatial relationships between amino acids translate into mutational and functional effects of proteins. Our work demonstrates that structural information can guide the assembly of smaller experimental datasets to accurately predict broader protein-protein binding effects, bridging the gap between local molecular structure and system-level protein behavior. By connecting physical structure to functional outcomes, our method helps build predictive models that capture biologically meaningful relationships across different scales of protein function.

## CODE AVAILABILITY

The code used in this study is made publicly available at `https://github.com/mbasse0/ProxiClust`

## REFERENCES


Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using alphafold2. *Nature communications*, 13(1):1265, 2022.

Alessandro M Carabelli, Thomas P Peacock, Lucy G Thorne, William T Harvey, Joseph Hughes, Sharon J Peacock, Wendy S Barclay, Thushan I De Silva, Greg J Towers, and David L Robertson. Sars-cov-2 variant biology: immune escape, transmission and fitness. *Nature Reviews Microbiology*, 21(3):162–177, 2023.

Chen Chen, Veda Sheersh Boorla, Deepro Banerjee, Ratul Chowdhury, Victoria S Cavener, Ruth H Nissly, Abhinay Gontu, Nina R Boyle, Kurt Vandegrift, Meera Surendran Nair, et al. Computational prediction of the effect of amino acid changes on the binding affinity between sars-cov-2 spike rbd and human ace2. *Proceedings of the National Academy of Sciences*, 118(42): e2106480118, 2021.

Gabriel Cia, Jean Marc Kwasigroch, Marianne Rooman, and Fabrizio Pucci. Spikepro: a webserver to predict the fitness of sars-cov-2 variants. *Bioinformatics*, 38(18):4418–4419, 2022.

Xianding Deng, Miguel A. Garcia-Knight, Mir M. Khalid, Venice Servellita, Candace Wang, Mary Kate Morris, Alicia Sotomayor-González, Dustin R. Glasner, Kevin R. Reyes, Amelia S. Gliwa, Nikitha P. Reddy, Claudia Sanchez San Martin, Scot Federman, Jing Cheng, Joanna Balcerek, Jordan Taylor, Jessica A. Streithorst, Steve Miller, Bharath Sreekumar, Pei-Yi Chen, Ursula Schulze-Gahmen, Taha Y. Taha, Jennifer M. Hayashi, Camille R. Simoneau, G. Renuka Kumar, Sarah McMahon, Peter V. Lidsky, Yinghong Xiao, Peera Hemarajata, Nicole M. Green, Alex Espinosa, Chantha Kath, Monica Haw, John Bell, Jill K. Hacker, Carl Hanson, Debra A. Wadford, Carlos Anaya, Donna Ferguson, Phillip A. Frankino, Haridha Shivram, Liana F. Lareau, Stacia K. Wyman, Melanie Ott, Raul Andino, and Charles Y. Chiu. Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cell*, 184(13):3426–3437.e8, June 2021. ISSN 00928674. doi: 10.1016/j.cell.2021.04.025. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867421005055.

Andre J Faure, Ben Lehner, Verónica Miró Pina, Claudia Serrano Colome, and Donate Weghorn. An extension of the walsh-hadamard transform to calculate and model epistasis in genetic landscapes of arbitrary shape and complexity. *PLOS Computational Biology*, 20(5):e1012132, 2024.

Wenkai Han, Ningning Chen, Xinzhou Xu, Adil Sahil, Juexiao Zhou, Zhongxiao Li, Huawen Zhong, Elva Gao, Ruochi Zhang, Yu Wang, et al. Predicting the antigenic evolution of sars-cov-2 with deep learning. *Nature Communications*, 14(1):3478, 2023.

Brian Hie, Bryan D Bryson, and Bonnie Berger. Leveraging uncertainty in machine learning accelerates biological discovery and design. *Cell systems*, 11(5):461–477, 2020.

Brian Hie, Ellen D Zhong, Bonnie Berger, and Bryan Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, 2021.

Ruyun Hu, Lihao Fu, Yongcan Chen, Junyu Chen, Yu Qiao, and Tong Si. Protein engineering via bayesian optimization-guided evolutionary algorithm and robotic experiments. *Briefings in Bioinformatics*, 24(1):bbac570, 2023.

Marian Huot, Dianzhuo Wang, Jiacheng Liu, and Eugene Shakhnovich. Few-shot viral variant detection via bayesian active learning and biophysics. *bioRxiv*, 2025. doi: 10.1101/2025.03.12.642881. URL https://www.biorxiv.org/content/early/2025/03/13/2025.03.12.642881.

Bryan E Jones, Patricia L Brown-Augsburger, Kizzmekia S Corbett, Kathryn Westendorf, Julian Davies, Thomas P Cujec, Christopher M Wiethoff, Jamie L Blackbourne, Beverly A Heinz, Denisa Foster, et al. The neutralizing antibody, ly-cov555, protects against sars-cov-2 infection in nonhuman primates. *Science translational medicine*, 13(593):eabf1906, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Yu Kaku, Kaho Okumura, Miguel Padilla-Blanco, Yusuke Kosugi, Keiya Uriu, Alfredo A Hinay, Luo Chen, Arnon Plianchaisuk, Kouji Kobiyama, Ken J Ishii, et al. Virological characteristics of the sars-cov-2 jn. 1 variant. *The Lancet Infectious Diseases*, 24(2):e82, 2024.

Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, et al. Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *nature*, 581(7807):215–220, 2020.

Kathy Leung, Marcus HH Shum, Gabriel M Leung, Tommy TY Lam, and Joseph T Wu. Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. *Eurosurveillance*, 26(1), January 2021. ISSN 1560-7917. doi: 10.2807/1560-7917.ES.2020.26.1.2002106. URL https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.26.1.2002106.

Thomas Loux, Dianzhuo Wang, and Eugene Shakhnovich. More structures, less accuracy: Esm3's binding prediction paradox. *bioRxiv*, pp. 2024–12, 2024.

Alief Moulana, Thomas Dupic, Angela M Phillips, Jeffrey Chang, Serafina Nieves, Anne A Roffler, Allison J Greaney, Tyler N Starr, Jesse D Bloom, and Michael M Desai. Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 omicron BA.1. *Nat. Commun.*, 13(1):7011, November 2022.

Alief Moulana, Thomas Dupic, Angela M Phillips, Jeffrey Chang, Anne A Roffler, Allison J Greaney, Tyler N Starr, Jesse D Bloom, and Michael M Desai. The landscape of antibody binding affinity in SARS-CoV-2 omicron BA.1 evolution. *Elife*, 12, February 2023.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Angela M Phillips, Daniel P Maurer, Caelan Brooks, Thomas Dupic, Aaron G Schmidt, and Michael M Desai. Hierarchical sequence-affinity landscapes shape the evolution of breadth in an anti-influenza receptor binding site antibody. *Elife*, 12, January 2023.

Frank J Poelwijk, Vinod Krishna, and Rama Ranganathan. The context-dependence of mutations: a linkage of formalisms. *PLoS computational biology*, 12(6):e1004771, 2016.

Md Shafiqur Rahman, Min Jung Han, Sang Won Kim, Seong Mu Kang, Bo Ri Kim, Heesun Kim, Chang Jun Lee, Jung Eun Noh, Hanseong Kim, Jie-Oh Lee, et al. Structure-guided development of bivalent aptamers blocking sars-cov-2 infection. *Molecules*, 28(12):4645, 2023.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019. doi: 10.1101/622803.

Alina P Sergeeva, Phinikoula S Katsamba, Junzhuo Liao, Jared M Sampson, Fabiana Bahna, Seetha Mannepalli, Nicholas C Morano, Lawrence Shapiro, Richard A Friesner, and Barry Honig. Free energy perturbation calculations of mutation effects on sars-cov-2 rbd:: Ace2 binding affinity. *Journal of Molecular Biology*, 435(15):168187, 2023.

Rui Shi, Chao Shan, Xiaomin Duan, Zhihai Chen, Peipei Liu, Jinwen Song, Tao Song, Xiaoshan Bi, Chao Han, Lianao Wu, et al. A human neutralizing antibody targets the receptor-binding site of sars-cov-2. *Nature*, 584(7819):120–124, 2020.

Tomokazu Tamura, Takashi Irie, Sayaka Deguchi, Hisano Yajima, Masumi Tsuda, Hesham Nasser, Keita Mizuma, Arnon Plianchaisuk, Saori Suzuki, Keiya Uriu, et al. Virological characteristics of the sars-cov-2 omicron xbb. 1.5 variant. *Nature Communications*, 15(1):1176, 2024.

Anna Vangone and Alexandre MJJ Bonvin. Contacts-based prediction of binding affinity in protein–protein complexes. *elife*, 4:e07454, 2015.

Gian Marco Visani, Michael N Pun, William Galvin, Eric Daniel, Kevin Borisiak, Utheri Wagura, and Armita Nourmohammad. Hermes: Holographic equivariant neural network model for mutational effect and stability prediction. *ArXiv*, pp. arXiv–2407, 2024.

Dianzhuo Wang, Marian Huot, Vaibhav Mohanty, and Eugene I Shakhnovich. Biophysical principles predict fitness of sars-cov-2 variants. *Proceedings of the National Academy of Sciences*, 121 (23):e2314518121, 2024.

Guangyu Wang, Xiaohong Liu, Kai Wang, Yuanxu Gao, Gen Li, Daniel T Baptista-Hon, Xiaohong Helena Yang, Kanmin Xue, Wa Hou Tai, Zeyu Jiang, et al. Deep-learning-enabled protein–protein interaction analysis for prediction of sars-cov-2 infectivity and variant evolution. *Nature Medicine*, 29(8):2007–2018, 2023.

Daniel M Weinreich, Yinghong Lan, C Scott Wylie, and Robert B Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Current opinion in genetics & development*, 23 (6):700–707, 2013.

James RR Whittle, Ruijun Zhang, Surender Khurana, Lisa R King, Jody Manischewitz, Hana Golding, Philip R Dormitzer, Barton F Haynes, Emmanuel B Walter, M Anthony Moody, et al. Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. *Proceedings of the National Academy of Sciences*, 108(34):14216–14221, 2011.

Rui Yin, Brandon Y Feng, Amitabh Varshney, and Brian G Pierce. Benchmarking alphafold for protein complex modeling reveals accuracy determinants. *Protein Science*, 31(8):e4379, 2022.

Zhennan Zhao, Jingya Zhou, Mingxiong Tian, Min Huang, Sheng Liu, Yufeng Xie, Pu Han, Chongzhi Bai, Pengcheng Han, Anqi Zheng, et al. Omicron sars-cov-2 mutations stabilize spike up-rbd conformation and lead to a non-rbm-binding monoclonal antibody escape. *Nature communications*, 13(1):4958, 2022.

# A APPENDIX

### REPRESENTATION OF MUTANTS WITH ONEHOT FORMAT

| **Position** | 339 | 371 | 373 | 375 | 417 | 440 | 446 | 477 | 478 | 484 | 493 | 496 | 498 | 501 | 505 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Wildtype** | G | S | S | S | K | N | G | S | T | E | Q | G | Q | N | Y |
| **Mutant** | G | S | P | F | K | K | S | N | K | E | Q | G | R | Y | H |
| **Onehot representation** | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |

Table 1: Example of onehot representation of one mutant among the $2^{15}$ in the dataset.

## LOW-THROUGHPUT ACQUISITION METHODS

### RANDOM ACQUISITION

Random acquisition is the default data acquisition method for most machine learning pipelines Hie et al. (2020), where the training set is a subset of the whole dataset, with an equal probability of selecting any element of the set.

### DIVERSE ACQUISITION ON ONE-HOT SEQUENCE SPACE

Given a set of labelled mutants $S$, the diverse acquisition scheme samples the unlabelled mutant the furthest from the labelled set at each step, to obtain a new labelled set $S'$ with one additional mutant. More precisely, we studied three possible implementations of that idea:

**Min-linkage** $S' = S \cup \{x\}$ where $x = \arg\max_{y \notin S} \min_{z \in S} H(y, z)$

**Max-linkage** $S' = S \cup \{x\}$ where $x = \arg\max_{y \notin S} \max_{z \in S} H(y, z)$

**Mean-linkage** $S' = S \cup \{x\}$ where $x = \arg\max_{y \notin S} \frac{1}{|S|} \sum_{z \in S} H(y, z)$

Where $H$ is the Hamming distance, $H(x, y) = \sum_{i=1}^{n} \mathbf{1}_{x_i \neq y_i}$ , $x$ and $y$ are two strings of equal length, and $x_i$ and $y_i$ are the $i$-th elements of $x$ and $y$, respectively.

### DIVERSE ACQUISITION ON PROTEIN LANGUAGE MODEL EMBEDDINGS

The *Diverse acquisition on embeddings* sampling method is the same as above but the distance $H$ is defined as the L2 norm on the embedding space of the protein language model ESM-1v Rives et al. (2019), instead of the sequence space. Such a concept of distance in embedding space is similar to ”semanticity” of the protein by Hie et al. in Hie et al. (2021), which serves as a indicator of changes in biophysical properties such as $K_D$.

### SPARSE ACQUISITION ON ESM EMBEDDINGS

The *Sparse Acquisition* sampling consists in selecting variants in the less dense areas of the ESM embeddings space. A kernel density estimator Pedregosa et al. (2011) is trained on the labeled mutants to obtain a density map of the already explored parts of the 1280-dimensional embedding space. Each new variant added to the training set is chosen to have the embedding with the smallest Kernel Density Estimation value.

### COMPARING DIVERSE ACQUISITION WITH RANDOM ACQUISITION

Figure 5b is obtained from plot 8. This plot is created by measuring the $R^2$ prediction performance of random forest models trained on a series of 60 training sets of sizes ranging from 10 to 500 samples. The measurements are repeated 112 times for different initializations of the 10 initial samples in the training set, chosen at random. The transparent areas surrounding the $R^2$ curves represent the standard deviation of the scores, computed over the 112 repetitions. To generate Figure 5b, an

equally spaced range of 50 $R^2$ values are chosen and the corresponding size of training set are obtained on plot 8 by joining the successive points with straight lines junctions. Linear regression analysis, using the scipy linregress package, was performed to calculate the linear fit with a slope of 1.31.

### UNCONSTRAINED DATA ACQUISITION VIA LOW-THROUGHPUT ACQUISITION METHODS

Computing the distribution of $R^2$ scores for interface and non-interface strategies

Inter-residue distances are calculated using PDB files that provide the precise three-dimensional positions of every atom in the antibody-antigen binding poses for our datasets. The specific files used are 6M0J Lan et al. (2020) for RBD-ACE2 binding, 8J26 Rahman et al. (2023) for RBD-REGN10987 binding, 7XCK Zhao et al. (2022) for RBD-S309 binding, 7KMG Jones et al. (2021) for RBD-CoV555 binding, 7C01 Shi et al. (2020) for RBD-CB6 binding, 5UGY Whittle et al. (2011) for SI06-CH65. Positions of the residues are then approximated by the coordinates of each $C_\alpha$ atom.

The boxplots on the left of the figures in 3 are made from an exhaustive scan of all the assembly strategies of interface residues into two combinatorial scans. The number of loci contained in the combinatorial scans was set to $L = 6$ for the Covid datasets and $L = 8$ for the Influenza dataset to account for higher level epistasis effects. The size of the interface is set to 10 for covid datasets and 12 for the Influenza dataset. Thus, the boxplots on the left in subfigures of Figure 3 summarize the distribution of $\binom{\binom{10}{6}}{2} = 21945$ assembly strategies. Note that PROXICLUST is one of those 21945 strategies since it also uses interface residues. Similarly, the left boxplot on 3d summarizes the distribution of $\binom{\binom{12}{8}}{2} = 122265$ assembly strategies.

Boxplots on the right of subplots in Figure 3 result from a sampling of a part of all assembly strategies. Indeed, results for the exhaustive scan of either $\binom{\binom{15}{6}}{2}$ or $\binom{\binom{16}{8}}{2}$ assembly strategies was too computationally heavy. $10^5$ strategies are sampled which account for $0.8\%$ of the total number of assembly strategies for the SARS-CoV-2 datasets and $0.12\%$ of the total number of strategies for the Influenza dataset.

## RESULTS OF LOW-THROUGHPUT ACQUISITION

### UNCONSTRAINED DATA ACQUISITION VIA LOW-THROUGHPUT ACQUISITION METHODS

We aimed to establish how effectively we could predict larger datasets using any smaller subsets and quantify the accuracy of these predictions in terms of the $R^2$ score. Therefore we evaluated various sampling methods to determine the most effective strategy for constructing the training set. In Figure 5a, we see that only the Diverse data acquisition (see methods) consistently outperforms Random Acquisition across all tested training set sizes. Figure 5b shows Diverse Acquisition uses 31% less training samples compared to the Random Acquisition to achieve same $R^2$ scores.

Finally, we determined the number of data points required to achieve a satisfactory $R^2$ score using this strategy. Figure 5d shows the training set sizes necessary to attain high $R^2$ values for combinatorial datasets of various sizes. For a dataset with L loci, which contains $2^L$ samples, the number of points needed to effectively predict the entire dataset grows exponentially. This exponential increase highlights a critical point in protein function prediction: the trade-off between the size of the training set and the model's performance, compounded by the high costs of acquiring experimental data. Notably, while these results are illustrated using RBD-ACE2 data, this trend holds consistently across all datasets we evaluated.

Figure 5d demonstrates that employing unconstrained data acquisition methods, such as Random Acquisition and Diverse Acquisition, is feasible only for combinatorial datasets with a limited number of loci. Given that VoCs sometimes exhibits a lot more mutations than this, and conducting combinatorial experiments on $2^{20}$ combinations or generating random mutations for model training is not experimentally viable. Consequently, to predict much larger combinatorial datasets, we must

construct training sets using high-throughput experimental methods, which allow highly scalable ways to generate measurements.

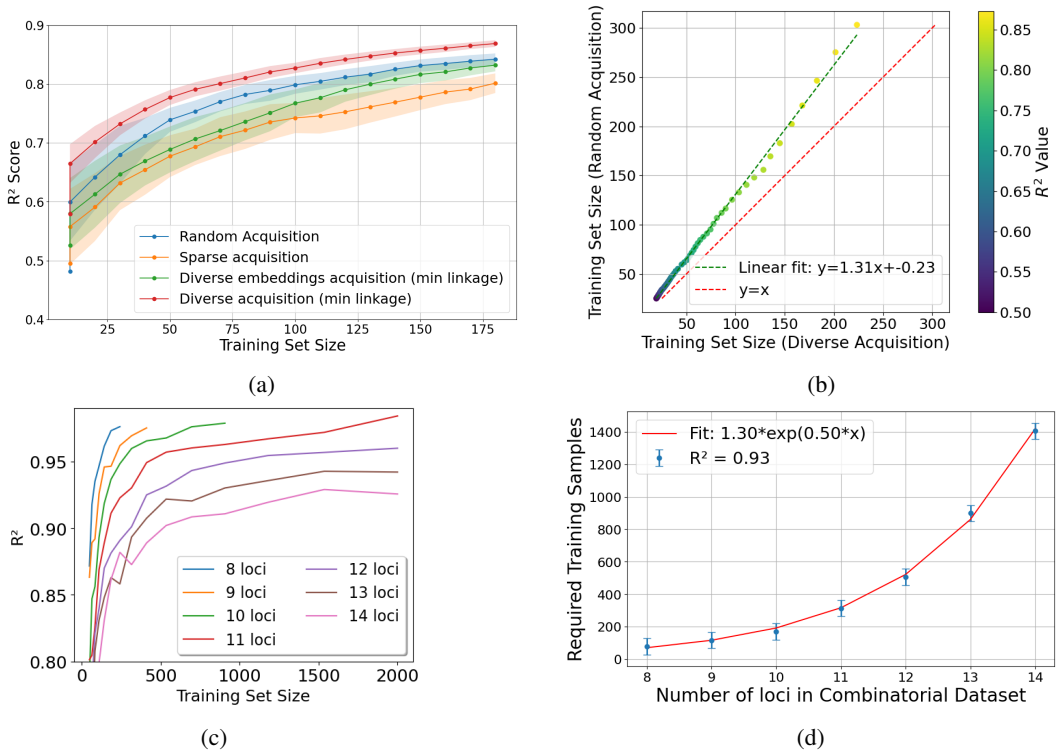



Figure 5: Data acquisition methods comparison (RBD-ACE2 binding affinity prediction). (a) $R^2$ score for varying training set size, compared between four acquisition methods. Only Diverse beats Random. Results are averaged across 112 runs where the initial 10 samples in the training set are initialized differently for each run. (b) Training set sizes required to achieve specific $R^2$ values for Random versus Diverse acquisition method Diverse acquisition saves around 31% of data to achieve similar performance than Random acquisition. (c) $R^2$ scores achieved with random forest models across different training set sizes for datasets of varying complexity (number of loci). (d) Exponential relationship between dataset complexity and required training set size, demonstrating that achieving predictive performance requires exponentially more training data as the number of loci increases.

## ADDITIONAL TABLES AND FIGURES

| Dataset | Median $R^2$ of all strategies scan | Median $R^2$ of interface strategies scan | $R^2$ score of PROXICLUST | Quantile of PROXICLUST's $R^2$ score |
|---|---|---|---|---|
| ACE2 | 0.27 | 0.64 | 0.78 | 0.995 |
| REGN10987 | -0.04 | 0.77 | 0.80 | 0.994 |
| CoV555 | -0.05 | 0.68 | 0.73 | 0.897 |
| S309 | -0.53 | 0.41 | 0.73 | 0.986 |
| SI06 | 0.05 | 0.5 | 0.86 | 0.9997 |

Table 2: Results of PROXICLUST for 5 different antigen-antibody complexes

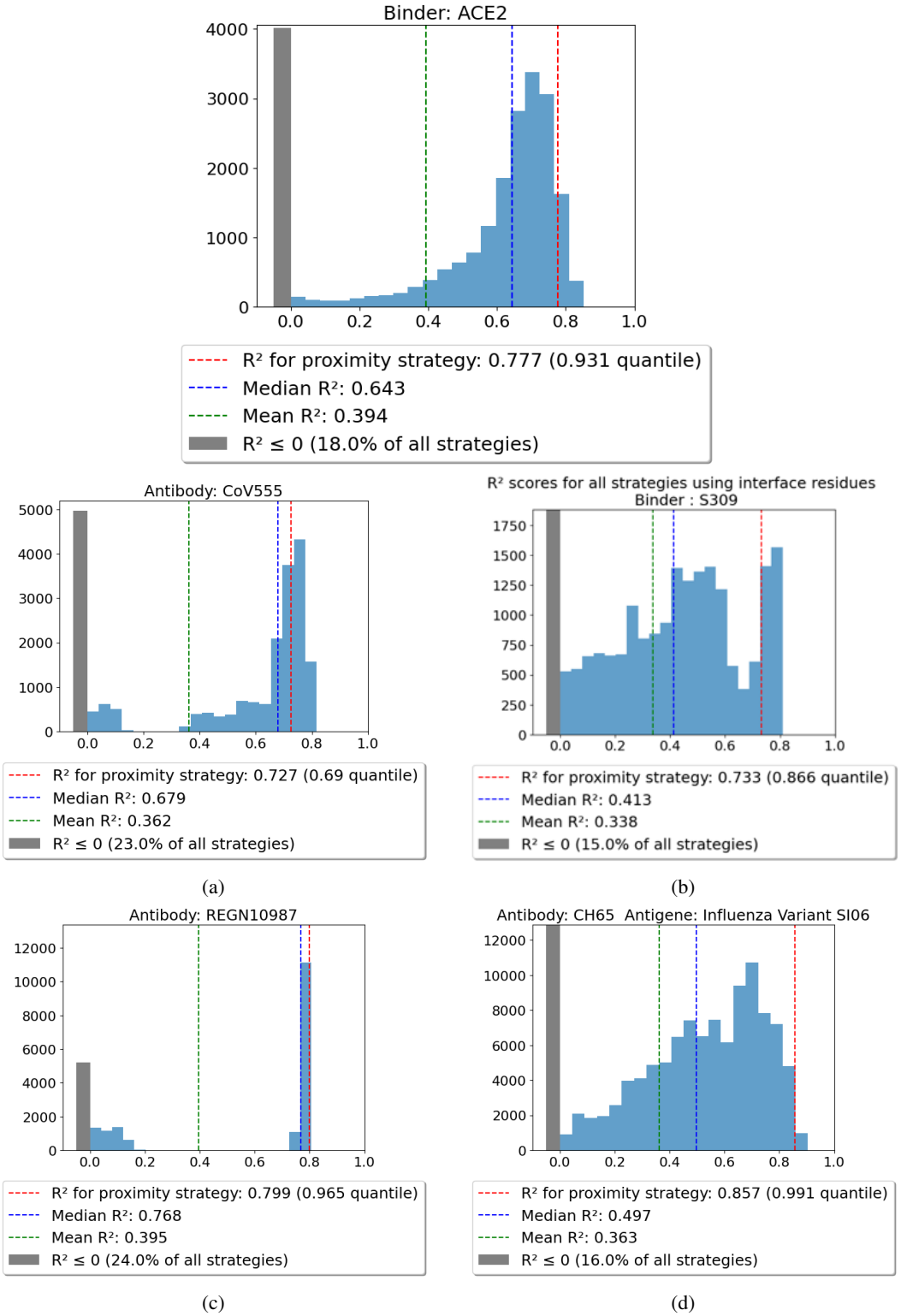



Figure 6: Histograms of $R^2$ scores for models trained from strategies restrained to the antibody-antigen interface. Red line shows that the $R^2$ score achieved by PROXICLUST beats both mean and median values of $R^2$ for strategies sampled at the interface.

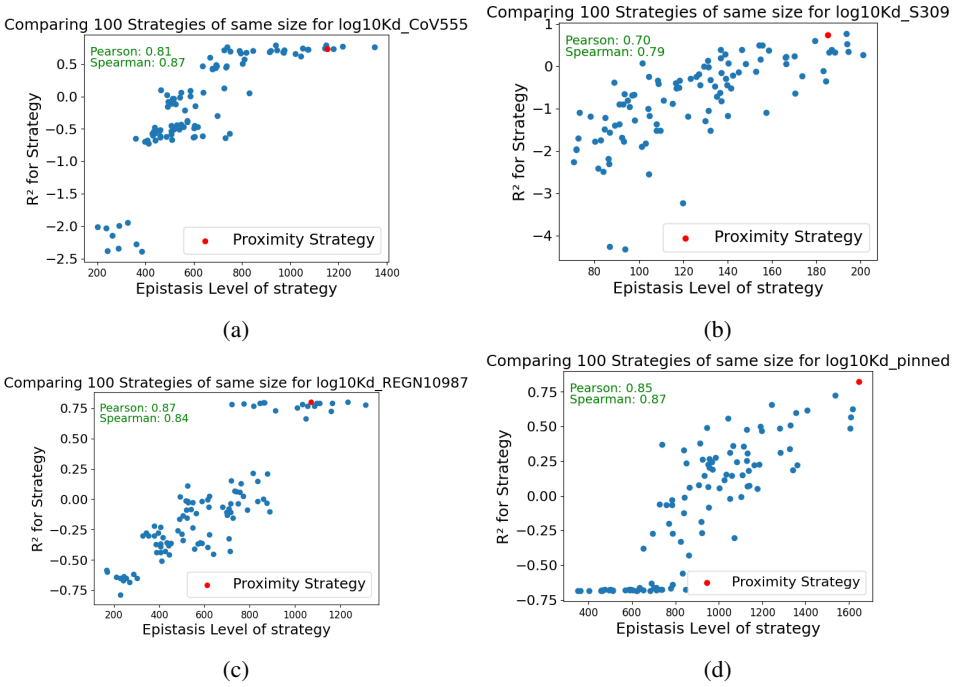



Figure 7: Correlation between model performance and epistasis score across different antibody-antigen binding systems. Each panel shows 100 random mutation selection strategies (blue dots) compared to the ProxiClust strategy (red dot) for: (a) RBD-CoV555, (b) RBD-S309, (c) RBD-REGN10987, and (d) RBD-ACE2 binding. The consistently high Pearson and Spearman correlation coefficients (0.70-0.87) demonstrate that strategies capturing more significant epistatic interactions yield better predictive performance. ProxiClust consistently ranks in the top percentiles for both metrics across all tested systems, validating that spatially-informed clustering effectively identifies functionally important mutational interactions

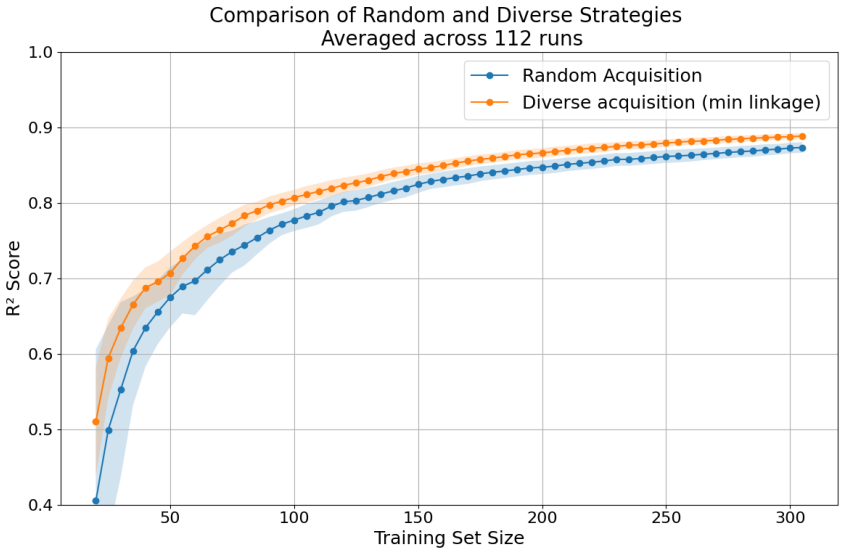



Figure 8: Performance comparison between Random Acquisition and Diverse Acquisition (min linkage) data sampling strategies for RBD-ACE2 binding affinity prediction.