Decoupled Information Theoretic Feature Selection for Rapid Protein Key Tuning Residue Identification

Haris Saeed[1]*, Aidong Yang[1] and Wei E. Huang[1]*

[1] Department of Engineering Science, University of Oxford, Oxford, OX1 3PJ, United Kingdom

*Corresponding author:

Haris Saeed haris.saeed@stcatz.ox.ac.uk

Wei E. Huang wei.huang@eng.ox.ac.uk

Abstract

Identifying key residues that govern protein function, stability, and interactions remains a significant challenge in protein engineering. We present an information-theoretic approach, inspired by signal processing methods to detect such 'key tuning' residues from protein sequences. Our approach integrates both unsupervised and supervised methods using Shannon entropy and mutual information as initial priors, respectively. A key aspect of our method is the use of Cramer's V and Thiel's U to decouple coevolved residues. The supervised method combining mutual information and evolutionary decoupling, significantly outperforms an existing state of the art, DeepSequence, ($pvalue = 0.0149$). The unsupervised approach that requires only a BLAST derived multiple sequence alignment (MSA) offers a simple yet competitive alternative ($pvalue = 0.14$). We validated our methods by applying them to three distinct protein datasets: Green Fluorescent Proteins (GFPs), rhodopsins, and alkanal monooxygenases, demonstrating a high predictive efficacy compared to established tools. When applied to well-characterised nanobody-antigen interactions, our methods not only effectively identified the binding residues but also accurately reconstructed a contiguous binding surface, surpassing current state-of-the-art computational tools. This work provides a flexible, interpretable, computationally efficient and transparent strategy for identifying key residues, paving the way for new strategies in rational protein design, enzyme optimization and drug discovery.

Keywords— Protein Engineering, Signal Processing, Semi-rational Design, key residues, fluorescent protein, rhodopsin, alkane monooxygenase, nanobody

Introduction

Proteins, the molecular workhorses of biological systems, contain key tuning residues that regulate complex functions, influencing catalytic efficiency, ligand specificity, and spectral properties [1-3]. These residues are essential for fine-tuning enzyme activity, determining ligand binding affinity, modulating the spectra of fluorescent proteins for bioimaging, enhancing industrial enzyme stability, and improving therapeutic protein efficacy [1-11]. Recent advances in computational methods, particularly those leveraging information-theoretic measures such as sequence entropy and mutual information, have improved the identification of these residues [12 – 15]. Emerging techniques, including machine learning models for allosteric site identification [16] and studies on the evolutionary significance of specific residues [17], emphasize the necessity for continued refinement of computational approaches alongside experimental validation. Our proposed method employs a novel information-theoretic approach to tackle the challenges in identifying key tuning residues, promising enhanced accuracy and efficiency that could significantly influence protein engineering, drug discovery, and our understanding of protein evolution and function.

Traditional methods like site-directed mutagenesis and X-ray crystallography are limited in scope and scalability [12, 13]. High-throughput sequencing and the growth of protein sequence databases present opportunities and challenges, necessitating computational methods to swiftly and accurately identify critical residues across diverse protein families. Existing computational tools have made progress but often require precise structural data or lack explainability, hindering their ability to process large datasets and provide interpretable results. These tools can be broadly categorized into structure-based and sequence-based approaches. Notable tools include EVcouplings, DeepSequence, and Hotspot Wizard. EVcouplings uses evolutionary couplings from multiple sequence alignments (MSAs) to predict residue interactions and functional sites via Direct Coupling Analysis (DCA) but relies on large datasets, limiting accessibility for non-specialists [18, 19]. DeepSequence employs Variational Autoencoders (VAEs) to assess mutation effects on protein function using high-quality MSAs, though its "black box" nature complicates interpretation [20, 23]. Hotspot Wizard integrates structural data with sequence conservation analysis to identify

functional hotspots but requires high-resolution protein structure data and can be complex to customize [21,22]. Overall, while these tools are powerful, they face challenges in data requirements, complexity, and interpretation, indicating the need for more accessible methods that can provide accurate predictions with less complex data, benefiting a broader range of researchers in protein science and engineering.

In this study, we present a novel approach that reframes the identification of key tuning residues as a feature selection problem, using information theory concepts. By treating protein sequences as complex signals and employing measures such as Shannon entropy and mutual information, we quantify residue variability and inter-residue dependencies as shown in Figure 1**a**, **b**. This method allows us to highlight high-entropy regions as potentially functionally significant while accounting for coupled protein networks, providing an explainable 'white box' model for identifying key residues in proteins.

To evaluate our method's efficacy, we first used three diverse datasets: (1) Green Fluorescent Proteins (GFPs) with peak excitation Wavelength, (2) rhodopsins with peak absorbance wavelength, and (3) alkanal monooxygenases with optimal pH (utilizing a pre-existing neural network to estimate optimal pH values). These datasets represent a range of protein families and functional properties, allowing for a comprehensive assessment of our method's performance across different biological contexts. We compared our method with three existing tools: (1) Hotspot Wizard, (2) EVCouplings, and (3) DeepSequence (refer to Figure S1 for pipelines), assessing both agreement with literature-identified key residues and the ability to detect residues with high informativeness based on protein sequences. We further used the case study of nanobody-antigen binding to validate the agreement between identified key residues and residues that are observed to influence binding affinity from docking simulations.

Our results demonstrate significant advancements in identifying key tuning residues, with both supervised and unsupervised approaches based on signal processing outperforming existing specialized methods. The supervised approach, utilizing Cramer's V-adjusted Mutual Information, achieved exceptional performance, with a statistically significant improvement (p-value = 0.0149) over the best existing tool tested, DeepSequence. Concurrently, our unsupervised approach, focused on refined

Shannon Entropy measures, yielded promising results with a p-value of 0.14. These outcomes underscore the effectiveness of our integrated methodology, which combines novel information-theoretic measures with sophisticated decoupling techniques, leading to improved accuracy and reliability in key residue identification compared to current computational methods.

This research not only introduces a powerful new tool for protein analysis but also bridges the gap between machine learning techniques and biological insights. By framing key residue identification within the context of feature selection, we open new avenues for applying established data science methodologies to fundamental questions in protein biology. Our findings have far-reaching implications for protein engineering, drug discovery, and evolutionary biology, offering a robust, generalizable framework for unraveling the complex relationships between protein sequence, structure, and function.

Results

Key Residue Identification as a Feature Selection Problem

Understanding the relationship between specific key residues and their functional properties is crucial in the study of protein optimization and design. This can be effectively framed as a mathematical optimization problem. Let $\mathcal{A}$ be the set of all amino acids, and $S_i \in \mathcal{A}^N$ represent a protein sequence of length N from a family of proteins. A property of interest for these proteins can be described by a function $f: \mathcal{A}^N \to \mathbf{R}$.

The problem of identifying key residues can be formulated as finding the subset of residues that maximizes the impact on the property of interest. Mathematically, we seek to find:

$$\arg \max_{j \in \{1,\ldots,N\}} \left| \frac{\partial f}{\partial s_j} \right| \#(1)$$

where s_j represents the j-th residue in the sequence. This formulation is analogous to the feature selection problem in machine learning [48], where we aim to identify a subset of features $\mathcal{F} \subset \{1, \dots, N\}$ that maximizes an objective function $J(\mathcal{F})$:

$$\arg \max_{\mathcal{F} \subset \{1,\dots,N\}} J(\mathcal{F}) \# (2)$$

In the context of protein sequences, $J(\mathcal{F})$ could represent the impact of the selected residues on the property of interest, such as:

$$J(\mathcal{F}) = \sum_{j \in \mathcal{F}} \left| \frac{\partial f}{\partial s_j} \right| \# (3)$$

This formalism allows us to apply various feature selection techniques to the problem of key residue identification. Here, the labelled data refer to a set of protein sequences with known values for the property of interest (e.g., activity level, binding affinity), allowing models to learn relationships between sequences and these specific properties. Unlabeled data, in contrast, consists of only the sequences themselves, without any associated property information, requiring the models to infer patterns solely from the sequence data.

1. **Supervised approaches:** With labelled data (S_i, y_i), methods like Lasso regression or Random Forest can be used to identify key residues [49].
2. **Unsupervised approaches:** Without labelled data, information-theory-based methods, such as maximizing sequence subset entropy, can be applied [50].
3. **Semi-supervised approaches:** Sequence conservation data (unsupervised) can be combined with limited experimental validation (supervised) to identify key residues [51].

We used two feature selection-based sequence-based information-theoretic methods for identifying key tuning residues:

1. Entropy-Based Methods: These unsupervised methods assess the variability and conservation of residues within a multiple sequence alignment (MSA). Entropy measures uncertainty or disorder, highlighting residues with high variability [50].

Such high-entropy residues are often key tuning residues, as their variability can significantly impact protein properties without compromising core functionality.

2. Mutual Information-Based Methods: These supervised methods examine the mutual information (MI) between a target property (e.g., a specific functional or structural characteristic) and the amino acid at each residue location in the aligned sequences [50]. By evaluating the dependency between residues and the target property, mutual information identifies residues strongly correlated with the target property. High mutual information indicates that changes in the residue are strongly associated with changes in the target property, marking them as key tuning residues.

Both entropy and MI serve as initial estimates of residue importance based on a protein sequence. While MI necessitates a labeled dataset, entropy can be derived from a single sequence without external annotations by running a BLAST of the focus sequence. Both of these metrics are inherently one-dimensional and overlook the global structure and interdependencies that may be present in the sequence. This oversight becomes particularly significant when residues are strongly interdependent or coupled.

Consider a protein sequence represented as

$$S = (s_1, s_2, \dots, s_N)$$

where N is the total number of residues. Let us define a partition

$$\mathcal{P} \subset S$$

with $|\mathcal{P}| = M < N$, representing the top-ranked residues according to an importance score $I(s)$ representing the impact of residue 's' on a given property, where 1 indicates 's' deterministically predicts said property. Specifically, we assume that

$$I(s) < 1 \text{ for } s \in \mathcal{P}, \text{ and } I(s) < min([I(s) \text{ for } s \in \mathcal{P}]), \text{ for } s \notin \mathcal{P}.$$

In a worst-case scenario, all residues within $\mathcal{P}$ are perfectly coupled; that is, the value of any single residue in $\mathcal{P}$ deterministically reveals the values of all other residues in the partition. Under these conditions, the effective dimensionality of $\mathcal{P}$ reduces to one, meaning that only one residue contributes independent information. A naive ranking

approach would require the examination of at least $M + 1$ residues (i.e. $\mathcal{P}$ and the next top ranked residue) to identify the top two independently functionally relevant residues.

As $M \to N \to \infty$, the 2-accuracy-defined as the probability of correctly identifying two independent key residues within the top $\frac{2}{M}$ fraction of the ranking-tends to zero.

To address this challenge, decoupling methods are employed to mitigate the effects of interdependent residues. Decoupling involves evaluating the interactions between residues and adjusting their rankings to account for these dependencies. Decoupling allows us to determine the standalone impact of each residue, improving the clarity of which residues are genuinely critical for the protein's function, rather than being artifacts of correlated variation. Figure 1**a** and **b** illustrate the comprehensive pipeline for each of these sequence-based measures, detailing the steps from sequence acquisition to the final identification of key tuning residues, including decoupling by both Theil's U and Cramer's V (Chi Squared).

Framing key residue identification as a feature selection problem enables us to use a diverse range of algorithms and theoretical frameworks from machine learning and data analysis. This enables the development of more robust and computationally efficient strategies for pinpointing functionally important residues, thereby enhancing our understanding of protein structure-function relationships and facilitating more effective protein engineering.

Removing Evolutionary Coupling Significantly Reduced Fingerprint Noise Regardless of Decoupling Matrix

Generating accurate residue importance rankings using information-theoretic measures begins with an initial ranking based on either mutual information or entropy. This is followed by decoupling highly coupled residues to refine the resulting fingerprint. This approach not only improves interpretability by reducing noise but also serves as a form of dimensional reduction. Figure 2**a** shows the distribution of amino acid probabilities and their corresponding conservation scores across the fluorescent protein dataset (refer to Supplementary Figures S2 & S3 for other protein families). The analysis identifies regions with both highly conserved and highly variable residues, each of which

is crucial for understanding functional significance. By focusing on high-entropy residues, which are inherently less conserved, we eliminate redundant data and reduce noise. This targeted approach streamlines the analysis by concentrating on the most informative residues. Excluding overly conserved or misaligned residues enhances accuracy and emphasizes regions with a higher likelihood of true functional relevance, allowing for more robust identification of residues that are likely to drive protein function.

Figures 2**b** and 2**d** display coupling matrices for fluorescent proteins, calculated using Cramer's V and Theil's U metrics. These matrices indicate that regions with high conservation scores typically show strong internal coupling. These highly internal coupled regions often have limited predictive power for high-entropy residues. According to information theory, and as demonstrated in Figures 2**b** and 2**d**, low-entropy regions cannot achieve perfect symmetric coupling with high-entropy regions due to irreversible information loss, highlighting the concept behind decoupling for protein analyses. Our novel approach specifically focuses on high-entropy regions post-decoupling, enabling the identification of critical residues that are less likely to be alignment artefacts. This targeted strategy improves the reliability of residue importance rankings, providing deeper insights into the functional relevance of these residues and supporting more accurate predictions of protein function for future experimental applications.

Figures 2**c** and 2**e** present information plots before and after decoupling using Theil's U and Cramer's V based on both entropy and mutual information (MI) for initial rankings. These Figures show that decoupling significantly sharpens the results, enhancing clarity and reducing noise in the data. Decoupling works by assessing the interactions between residues and adjusting their rankings to account for these dependencies. This process isolates the specific impact of each residue, ensuring that identified key residues are not artefacts of correlated changes but are genuinely influential on the protein's function. Notably, the effectiveness of decoupling is independent of the specific matrix employed; both Theil's U and Cramer's V effectively isolate distinct peaks and reduce noise, regardless of whether MI or entropy was used for the initial ranking. However, the

structure of the final fingerprint remains sensitive to the choice of the initial information metric, as it fundamentally shapes our understanding of residue importance. This highlights the critical need for selecting an appropriate initial ranking measure, while the choice of decoupling approach has comparatively less influence on the outcome.

The application of decoupling methods across all three protein families significantly improved the precision of key residue identification, sharpening the protein fingerprints in each analyzed dataset. This enhancement indicates that decoupling effectively reduces noise and clarifies the functional significance of residues. To further quantify this improvement, we used autocorrelation as a function of phase shift to measure noise levels. A non-zero-mean white noise process would display a phase-independent autocorrelation of 1, serving as a benchmark for comparison (see methodology for details).

Figure 3**a** shows the autocorrelation patterns of information-theoretic measures for the fluorescent protein dataset (see Supplementary Figure S4 for data on other protein families). Initial metrics, prior to decoupling, exhibited high autocorrelation across the phase sweep across all datasets, indicating high noise. In contrast, decoupling significantly reduced autocorrelation at phase shifts away from 0. This noise reduction is further quantified in Figure 3**c**, which illustrates that decoupling reduced phaseaveraged autocorrelation by over fivefold for both Mutual Information (MI) and entropy, on average, across all datasets (refer to Supplementary Figure S4 for a breakdown by protein family). Decoupled rankings consistently maintained noise factors below 0.08, reflecting uniform noise reduction across different coupling measures and initial rankings.

The effectiveness of decoupling methods is validated by Hedge's G analysis (Figure 3**d**), which consistently shows values below -1 between noise factors for initial and decoupled rankings, indicating significant noise reduction. Hedge's G values for decoupled measures of the same type range from -0.25 to 0.25, suggesting minimal differences between decoupling methods. The average cross-correlation between decoupled measures exceeds 0.8 (Figure 3**b**), further supporting this consistency.

The choice of initial ranking method significantly impacted results, pre-decoupling Mutual Information (MI) and entropy exhibited high autocorrelations (> 0.8) across all datasets (see Supplementary), attributed to their white noise like natures. Post-decoupling, the cross-correlation between different rankings (MI vs. entropy) averaged below 0.5, with datasets, such as Alkane Monoxygenases, showing correlations below 0.25 (Figure S4). These results further indicate that while decoupled measures are stable, they can be sensitive to the initial ranking method, said sensitivity influenced by dataset characteristics. This emphasizes the importance of carefully selecting the initial ranking method for different protein families.

Comparison of Information Theoretic and Existing Approaches for Key Residue Identification

We evaluated the efficacy of algorithms in identifying key tuning residues through a comparative analysis with established tools for key residues from scientific literature. The study focused on three datasets: Alkane Monooxygenases (optimal pH), Rhodopsins (peak absorbance wavelength), and Fluorescent Proteins (peak excitation wavelength). These datasets were chosen for their diverse biochemical properties and documented key tuning residues, serving as a benchmark for computational predictions.

Figure 4**a** shows the performance of existing computational tools across the datasets, with importance plots overlaid by dashed lines representing literature-identified key residues. Significant discrepancies were noted between computational predictions and key residues from the literature, demonstrating the challenges in accurately identifying functionally essential residues in diverse biochemical contexts. Similarly, Figure 4**b** presents importance plots generated by our methods alongside literature-derived ground truths. While our methods demonstrated improved performance, they still showed notable discrepancies when comparing to literature identified key residues, suggesting a common limitation across various algorithmic approaches.

To facilitate effective quantitative analysis of different methods, we define coherence as the relative alignment between the identified key residues and those reported in literature. An analysis of the 12 evaluated methods-comprising 6 existing and 6 newly

developed approaches-yielded promising insights, with average coherence values generally remaining below 0.15 and notable variability across the assessed protein datasets, as illustrated in Figure 4c. Among the established methods, EV and Hotspot Wizard achieved the highest average performances, with coherences of 0.0758 and 0.0694, respectively. In contrast, our newly developed methods significantly exceeded these, with the leading information-theoretic approach (Cramer MI) achieving an average coherence of 0.137 and a maximum accuracy of 0.241 (due to the large variance observed, these differences were not statistically significant).

Figure 4**d** presents the Hedge's G matrix, which quantifies the effect size between different computational tools, looking at the coherence between the developed fingerprint and literature identified key residues. Hedge's G is a statistical measure used to evaluate the magnitude of differences between two groups, typically applied to assess the effectiveness of various methods. Unlike other metrics, it accounts for sample size bias, making it a reliable measure for comparing the performance of different approaches. The results reveal significant differences among the tools, with the maximum Hedge's G value between developed and existing methods generally being positive. Notably, the Hedge's G values for the decoupled developed methods are within ± 0.3, indicating no significant performance differences among decoupled measures, regardless of the initial ranking or decoupling approach used.

While our findings suggest improved performance, the overall coherence remains modest, with significant variability across datasets. This variability may arise from inherent limitations in computational methodologies that impact the prediction of functionally essential residues, or from ground truths derived from mutagenesis studies, which, while informative, might be too specific to the experimentally assessed proteins to generalize reliably to the protein's family. These results show we need improved tools for protein key residue prediction across different families and functions. We also need improved testing metrics and datasets, possibly by combining computational predictions with targeted experiments to make residue identification more reliable by providing more accurate benchmarks.

Information Theoretic Measures Significantly Improve Performance at Identifying Informative Residues

While literature-based measures provide valuable insights into key residues for specific proteins, their reliance on accurate ground truths and limited generalizability across protein families necessitate the exploration of data-driven approaches for comprehensive assessment. We thus developed a novel metric based on the informativeness of predicted key residues, as illustrated in Figure 5**a** and discussed in methods. This method involves sequentially fitting simple linear models to residue subsets, with each subset including the next highest-ranked residue. Each regression model aims to predict the target protein property, for which measurements are available, over a training partition. The resulting curve, from the proposed method, expresses the relationship between R^2 and sequence length (i.e., residues considered) (see methods for more details). An ideal theoretical ranking would reach a maximum R^2 at one residue and remain stationary thereafter, generalizing to realistic cases, an ideal ranking would have a weakly monotonically decreasing gradient. Figures 5**b, d, e, f**. and **g** present characteristic R^2 vs. aligned sequence length curves for top-ranked residues in the fluorescent protein dataset, focusing on peak excitation wavelength using various methods.

Figure 5**h** displays the average performance of each method across all datasets, normalized to an optimal value of 1. Analysis of existing measures reveals substantial performance variability, with an average standard deviation of approximately 0.08 across different datasets. DeepSequence (both absolute and positive) shows the highest performance among existing tools, indicating a strong correlation between key functional residues with high deleteriousness and key tuning residues. Positive EVCoupling ranks next, suggesting residues with high evolutionary coupling are also critical tuning residues. In contrast, our developed methods demonstrate significant improvement in both performance and variability. The average standard deviation in observed performance across all datasets was reduced to less than 0.04, with the highest average performance exceeding 0.97. Across developed methods, decoupling residues enhanced performance, especially using Cramer's V as a coupling measure. Comparison

between the optimal developed measure (Cramer MI) and optimal existing measure (AbsDeep) yields a statistically significant p-value of 0.0149, despite limitations of the dataset. This substantial improvement in accuracy and consistency demonstrates the efficacy of our developed methods in reliably identifying critical tuning residues.

Figure **5** illustrates Hedge's G values comparing different methods, offering detailed insights into their practical significance. Among existing measures, Absolute Deep emerges as the most effective, with Hedge's G exceeding 1 compared to Hotspot Wizard, negative EVCouplings, and Positive Deep. In contrast, Cramer's Adjusted MI performs best among our developed methods. For methods decoupled similarly (None, Cramer's V, Thiel's-U), MI-derived methods show positive Hedge's G, indicating superior performance. Decoupling enhances performance for both MI and Entropy, with the highest improvement observed between MI and Cramer's MI (Hedge's G of 0.67).

These results collectively demonstrate the superior performance of our information theoretic approaches, particularly Cramer's Adjusted MI, in identifying key tuning residues across diverse protein datasets. The statistically significant improvements in accuracy and consistency over existing tools signifies the potential of these methods to advance our understanding of protein structure-function relationships and facilitate more targeted approaches in protein engineering and drug design.

Nanobody-Antigen Binding

The three existing datasets previously assessed-(1) Rhodopsins, (2) Alkanal Monooxygenases, and (3) Fluorescent Proteins-demonstrated that the developed methods effectively identify potential key tuning residues with higher performance than existing methods, using both unlabeled data (entropy) and labeled data (information theoretic) approaches. These datasets are well suited to the developed tools, as key tuning residues are not necessarily directly correlated with binding residues.

To further evaluate our methods, we examined nanobody interaction with specific binding ligand: the heptad repeat 2 (HR2) region of the SARS-CoV-2 spike protein [52]. This case is well suited for three key reasons: (1) the large dataset of experimentally

379 assessed nanobody sequences (3,127) provides a statistically robust benchmarking
380 dataset, (2) structural docking enables direct validation of identified residues through
381 structural analysis, and (3) the biomedical significance of HR2 as a critical binding
382 epitope in SARS-CoV-2, highlighting the importance of accurately mapping nanobody-
383 antigen interfaces for therapeutic nanobody design.

384 In line with our earlier datasets, the performance assessment of the developed tools
385 demonstrates that both mutual information (MI)- and entropy-based approaches
386 effectively extract protein fingerprints. Figures 6**a**. show that both methods produce
387 similar residue importance profiles, capturing clusters of residues at comparable
388 locations and highlighting conserved regions likely to be functionally important.

389 Furthermore, both decoupling approaches yield comparable sharpening effects: metrics
390 such as Theil's U and Cramer's V effectively reduce indirect correlations, resulting in
391 sharper peaks and more precise identification of high-importance residues. In contrast
392 to the previous three datasets, decoupling does not significantly enhance sharpness or
393 reduce autocorrelation in this case. This can be attributed to the low maximum
394 correlation between residue pairs, as shown in the cross-correlation matrices (Figure
395 S5) resulting in similar observed noise proportions between coupled and decoupled
396 variants (Figure S6). Specifically, the maximum pairwise correlation does not exceed
397 0.3, indicating that each residue provides largely independent information-i.e., for any
398 pair of residues A and B, knowledge of A offers minimal additional information about B.
399 When applying our informativeness metric (as defined in Section S1.2), we observe a
400 substantial performance gain over the best existing tool(AbsDeep)-a relative
401 improvement exceeding 70%, with our method achieving a score of 0.7789 compared to
402 0.4502 (Figure 6c).

403 We demonstrate that our developed methods outperformed existing approaches in the
404 context of nanobody-antigen binding for SARS-CoV-2 (Figure 6c). Furthermore, we
405 propose that incorporating structure-based metrics offers an additional dimension for
406 evaluating key residue identification, thereby enhancing the robustness and biomedical
407 relevance of our assessment.

The nature of nanobody-antigen binding posits three expectations for key tuning residues: (1) they should be located on the solvent-accessible surface, (2) they should be spatially clustered, and (3) they should overlap significantly with non-essential binding residues that facilitate-but are not critical for-the binding event. To assess the structural interaction of nanobody-antigen complex, Figure 7 directly compares our developed methods against existing methods (DeepSequence) against the first two expectations. By mapping all solvent-accessible residues onto a unit sphere and coloring them based on their ranking (with red indicating the highest importance and blue the lowest), Figure 7 shows that all our developed methods (both MI and Entropy based with and without decoupling) consistently identified a contiguous set of residues forming a well-defined binding surface.

In contrast, existing methods, although capable of highlighting relevant regions, tend to produce a broader, less focused pattern with a sparse central cluster of important residues, further demonstrating the improved performance of our developed methods. Quantitative F1 score analysis (Figure S7) supports these observations. Our developed methods achieving scores approaching 0.8 for the top 20th percentile of residues-significantly outperforming the best existing tool, scoring only 0.6, showing the increased contiguousness and specificity of the binding surface identified by our approach.

Docking was performed using ClusPro to assess the overlap with binding-relevant residues. As described in the Supplementary Materials, binding analyses identified 50 solvent-accessible residues within 10 □ of atoms in the HR2 spike protein (highlighted in red in Figure 8**a** and 8**b**. Among these, 22 residues were predicted to be involved in antigen binding. When comparing the top 75 residues identified by each method, our best-developed approach detected 33 of the 50 residues within the 10,□ cutoff-a 50% improvement over the 22 residues identified by the best existing method (AbsDeepSequence [20, 23], from assessments against the prior three protein families). Our method also captured 64% of the potentially binding residues, compared to only 36% by the leading method (DeepSequence), thereby nearly doubling performance.

Figure 8**c-f** demonstrate this difference: conventional methods yield sparse interfacial residues (Figure 8**e**), whereas our approach produces a contiguous, coherent binding interface (Figure 8**f**). Notably, our method also identifies allosteric residues outside the immediate contact zone that potentially modulate binding affinity via long-range interactions-an essential consideration for a comprehensive understanding of the binding mechanism. Further validation is provided in Figure 8**i**, which shows a stronger correlation between predicted importance and estimated interfacial bond energies for our best-developed method compared to the best existing method ($R^2 = 0.63$ vs. 0.41), corroborated by the Akaike Information Criterion scores (279 vs. 302). Furthermore, Figure S8 shows the variation in overlap going from most important identified residue to least important, and it is clearly visible that for any subset (i.e. top n residues) the developed methods match or surpass existing methods when identifying binding or interfacial residues.

All developed methods surpassed existing methods when looking at the three structural priors for key tuning residues: (1) high overlap with surface accessible residues, (2) spatial co-location, and (3) significant overlap with nonessential binding residues that facilitate but are not essential for binding. In the case study of nanobody-antigen interactions, our developed methods significantly outperform existing tools, achieving over 50% higher performance. The effectiveness of our approach is verified through three metrics: (1) approximately two-fold increased informativeness compared to existing tools, (2) generation of a contiguous binding surface when examining structural contiguousness, and (3) identification of a greater number of residues within the nanobody binding pocket.

Discussion

This study reconceptualizes key protein residue identification as a feature selection problem, demonstrating how information-theoretic approaches can outperform more complex models while providing greater explainability. The central innovation lies in our decoupling methodology, which effectively isolates individual residue influences by

466 mitigating noise from highly coupled systems-a persistent challenge in biological
467 datasets.

468 Our feature selection framework offers a powerful alternative to opaque, model-driven
469 strategies traditionally employed in residue identification. While deep learning models
470 like DeepSequence have shown effectiveness, they often function as black boxes with
471 limited interpretability. In contrast, our approach provides a simpler yet robust solution
472 through residue interaction decoupling, disentangling correlated effects to yield clearer
473 assessments of each residue's functional contribution. Notably, Cramer's V-based
474 decoupled Mutual Information demonstrated superior performance to the best
475 performing existing method assessed, DeepSequence ($p = 0.0149$). When evaluating
476 identified key residues in nanobody antigen binding, where structural priors exist, our
477 method generated a contiguous binding surface encompassing nearly twice the binding
478 residues, demonstrating superior structural rationality. This highlights that a relatively
479 simple information-theoretic approach can outperform sophisticated deep learning
480 models when noise in coupled systems is properly addressed. The enhanced precision
481 from decoupling minimizes autocorrelation effects and extracts functionally relevant
482 residues with higher consistency across diverse protein families, challenging the
483 assumption that more complex models are inherently superior for residue identification
484 tasks.

485 Furthermore, we compared supervised and unsupervised approaches to residue
486 identification by evaluating mutual information (MI), a supervised method requiring
487 target variable data, against entropy, an unsupervised approach based solely on multiple
488 sequence alignments. While unsupervised methods offer broader applicability without
489 labeled training data, they lack the precision of supervised methods. Our results showed
490 MI outperforming entropy, demonstrating supervised methods' capacity to capture
491 more intricate residue relationships. The entropy approach remains valuable for
492 scenarios lacking labeled data, and while it showed lower average performance, the
493 difference was not statistically significant ($p = 0.4945$). The implications of this work
494 extend beyond residue identification to protein engineering and drug discovery. By
495 providing more accurate and explainable methods for identifying key residues, our
496 approach can enhance protein design efforts through more targeted functional

modifications. In drug discovery, identifying critical binding residues can lead to more effective therapeutics. The method's scalability and broad applicability make it particularly valuable for identifying novel therapeutic targets when experimental data is limited.

Despite its promise, our method has limitations. While focusing on sequence data allows broad applicability, it may not capture all aspects of protein behavior, particularly those related to three-dimensional structure and dynamic interactions. Integrating structural and dynamic information into the feature selection framework represents an important direction for future research. While the decoupling approach significantly improves precision in residue ranking, further refinement would benefit cases where highly coupled residues contribute to complex, multi-residue interactions that remain after simple decoupling.

In conclusion, by reframing residue identification as a feature selection problem with a novel decoupling approach, we introduce a powerful new tool for protein analysis. Our method enhances identification of functionally critical residues while offering a more transparent and generalizable framework for advancements in protein engineering and drug discovery. As we refine these methods and integrate them with structural data, we anticipate further improvements in both precision and applicability.

Methodology

Test Datasets

To ensure the generalisability of the computational tools for identifying key residues across various proteins, we used three datasets from different protein families, summarized in Table 1. These datasets include information on protein sequences and their corresponding target variables, which are crucial for evaluating the accuracy and performance of the tools. The datasets encompass a range of proteins and functions, providing a comprehensive benchmark for comparison.

These datasets were selected to ensure a representative sample of protein types and functions, facilitating a robust comparison of the computational tools.

- Rhodopsins: The dataset consists of 884 bacterial rhodopsins from multiple subfamilies with labelled absorption wavelengths [29].
- GFP and Variants: This dataset includes 611 sequences with labelled quantum yield, excitation wavelengths, and emission wavelengths [31].
- Alkanal Monoxygenases: The dataset contains 1880 sequences where optimal pH was predicted using deep learning [35].

The key residues for each protein family, as detailed in Table 1. provided a basis for evaluating the accuracy of the predictions made by each tool.

Tool Execution

To identify key residues involved in functional shifts, we employed three advanced tools: HotSpot Wizard, EVcouplings, and DeepSequence. These tools integrate structural, functional, and evolutionary data to predict mutational hotspots and identify co-evolved residues that are functionally important (refer to Supplementary Section S1.1 for detailed breakdown).

HotSpot Wizard: We used HotSpot Wizard 3.0 to predict mutational hotspots by integrating structural and evolutionary data. AlphaFold models provided accurate protein structures for input into HotSpot Wizard's analysis, which cross-referenced known protein structures to identify functionally important residues.

EVcouplings: This tool was used to detect evolutionary couplings by analyzing co-evolutionary patterns between residues. Multiple sequence alignments were generated using Kalign, and evolutionary coupling data were used to pinpoint residues that maintain protein structure and function.

DeepSequence: DeepSequence leverages deep generative models to predict mutation effects. This tool exhaustively samples mutations and predicts their impact on protein function using a variational autoencoder to model the sequence-function relationship.

Shannon Entropy and Mutual Information Analysis

In our study, we employed two key information-theoretic measures-Shannon Entropy and Mutual Information (MI) to identify critical residues in protein sequences.

Shannon Entropy: We used Shannon Entropy to assess the variability of residues at each position in a multiple sequence alignment (MSA). Positions with low entropy were considered highly conserved and likely functionally or structurally important, while high-entropy positions showed greater variability, potentially suggesting adaptive significance. Entropy values were scaled for comparison across different datasets, ensuring consistency regardless of protein family size or sequence variability.

Mutual Information: Mutual Information (MI) was used to quantify the relationship between residue positions and functional characteristics. MI values were calculated between residue positions in protein sequences and a quantitative target variable, such as optimal pH. To manage continuous target data, we applied binning strategies based on the Freedman-Diaconis rule. Normalized MI values were used to find residues with the strongest functional influence, aiding in pinpointing critical residues involved in protein function.

For further details on the specific calculations and methodologies, refer to Supplementary Section S1.2.

Removing the Effect of Coupling

The multidimensional nature of amino acid sequences can introduce biases when ranking residues based on Mutual Information (MI) or entropy alone. These rankings may mistakenly identify residues as crucial due to evolutionary coupling between them. To mitigate this issue, we employed additional statistical methods to reduce the effects of these residues and ensure a more accurate identification of key residues.

Thiel's U: Thiel's U is a measure of forecast accuracy, comparing the performance of a predictive model to that of a naive using information theory. The formula for Thiel's U is:

$$U(X \mid Y) = \frac{H(X) - H(X \mid Y)}{H(X)} \#(4)$$

Where:

$$H(X) = -\sum_{i} P(x_i)\log\ P(x_i)$$

is the entropy of variable X, and:

$$H(X \mid Y) = \sum_{j} P(y_j)H(X \mid Y = y_j)$$

is the conditional entropy of X given Y. A U value close to than 1 indicates that the model provides better predictions than the naive approach.

Chi-squared (χ^2): The Chi-squared test evaluates the association between two categorical variables. It is computed as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \#(5)$$

where O_i are the observed frequencies and E_i are the expected frequencies. Large χ^2 values indicate strong associations between the variables.

To remove the effects of coupling and sharpen residue rankings, we implemented an algorithm to adjust the entropy and MI values by factoring in evolutionary coupling between residues. This decoupling algorithm effectively reduces biases caused by indirect correlations, allowing for more accurate identification of key residues involved in protein function and stability. For an Algorithmic breakdown refer to Supplementary section S1.3.

Incorporating these decoupling methods refines residue ranking by ensuring that the detected interactions are statistically significant, and not artifacts of coupled residue behavior. The resulting refined analysis improves the reliability of functional predictions and enhances the robustness of evolutionary studies. This approach is implemented in tools such as EVcouplings, which improves predictions of protein residue interactions and helps deepen our understanding of protein function and evolution [18, 19].

Quantifying Performances of Different Methods

To assess the alignment and consistency of rankings from different methods, we used two statistical techniques: Fourier Transform-based Autocorrelation and Cross-correlation.

Fourier transforms were employed to examine the periodicity in ranking sequences. The autocorrelation function was computed in the frequency domain, using the inverse Fourier transform:

$$\text{Autocorrelation } (R, \tau) = \mathcal{F}^{-1}\{\mathcal{F}(R) \cdot \overline{\mathcal{F}(R)}\}(\tau) \#(6)$$

The normalized autocorrelation and the average autocorrelation were also computed to evaluate the consistency and potential periodicity of the rankings.

Cross-correlation was used to measure the similarity between two ranking sequences:

$$\text{CrossCorrelation}(R_1, R_2) = \frac{\sum_{i=1}^{n} \left(r_{1,i} - r_1\right)\left(r_{2,i} - r_2\right)}{\sqrt{\sum_{i=1}^{n} \left(r_{1,i} - r_1\right)^2 \sum_{i=1}^{n} \left(r_{2,i} - r_2\right)^2}} \#(7)$$

These methods provided insights into the alignment, cyclic patterns, and consistency across rankings, enabling a more thorough evaluation of the computational methods' results.

Assessing Performance

We evaluated the computational tools using two measures: Literature-based Coherence and Machine Learning-based Performance.

Literature-based Coherence: This metric quantifies the overlap and alignment of the model's identified key residues with those from the literature, based on a swap distance metric and normalized importance values.

Machine Learning-based Performance: This approach assesses the predictive power of a model using a Bayesian Ridge Regressor. It evaluates performance through a one-hot encoding process and the R^2 score, providing a comprehensive metric for model efficacy.

For a complete mathematical breakdown refer to Supplementary Section S1.5

Comparing Measures Hedge's G

To compare the metrics obtained from different methods, we employed Hedge's G, a statistical method commonly used to assess the effect size between two groups or conditions. Here, we adapted Hedge's G to evaluate the differences in performance metrics across various computational methods.

For each pair of methods A and B, Hedge's G was calculated using the formula:

$$G = \frac{X_A - X_B}{s_{\text{pooled}}} \cdot \sqrt{\frac{n_A n_B}{n_A + n_B}} \#(8)$$

where:

- X_A and X_B represent the mean values of the selected metrics for methods A and B, respectively.
- s_{pooled} is the pooled standard deviation of the metrics across methods A and B.
- n_A and n_B are the sample sizes (number of observations) for methods A and B, respectively.

A G value close to 0 indicates minimal difference in metrics between methods A and B. Positive G values suggest that the metrics from method A are larger than those from method B, while negative G values indicate the opposite. While Hedge's G provides valuable insights into methodological differences, its interpretation may be influenced by sample size and variance.

Nanobody Binding Validation

To further validate the performance of our developed methods, we opted to use the well-studied problem of Nanobody binding to the SARS-CoV-2 spike protein [52]. This dataset comprised over 3,000 unique nanobody sequences after rigorous filtering to remove duplicate entries and negative controls, ensuring that only sequences with measurable binding were included. The primary objective was to identify key tuning

residues that modulate binding affinity—residues that are hypothesized to drive the interaction between the nanobody and the spike protein.

We combined sequence-based methods (as previously discussed) with structural analyses from docking simulations. We used ClusPro 2.0 [53–58], which employs FFT-based rigid-body docking to sample binding poses and identify probable binding modes. This modelling generated 3D representations of the nanobody-spike complex, mapping key tuning residues onto the binding interface.

To assess agreement between predictions and docking results, we used:

1. Interface Mapping: We identified interface residues by calculating interatomic distances, considering residues within 10 Å of the antigen as interface residues.
2. Overlap Analysis: We compared predicted important residues with interface residues to quantify overlap, with high overlap indicating agreement with structural data.
3. Energy Analysis: We calculated interaction energies for each residue with the most probable docking pose (S1.7) and assessed correlation between binding energy contributions and residue importance.

This approach provides a framework for evaluating our methods. The agreement between results from data-based and docking analyses confirms an algorithm's reliability in identifying critical residues affecting nanobody binding to the SARS-CoV-2 spike protein.

Acknowledgements

W.E.H. thanks EPSRC (EP/M002403/1 and EP/Y014073/1) for financial support.

Author Contributions Statement: H.S conceived the original idea; H.S., A.Y. and W.E.H. designed research; H.S. performed research; H.S. analysed the data; H.S., A.Y. and W.E.H. interpreted the data; H.S. drafted the manuscript, and all authors revised the manuscript.

Competing Interests Statement

The authors declare no competing financial interest.

Availability of Data and Materials

The destination vectors described in this study are available at https://github.com/hs280/NaiveEntropicMeasures.

Reference:

[1] Warshel, A. ” Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites.” Journal of Biological Chemistry, vol. 276, no. 39, 2001, pp. 33235-33239.

[2] Agarwal, P.K. ” Role of protein dynamics in reaction rate enhancement by enzymes.” Journal of the American Chemical Society, vol. 124, no. 11, 2002, pp. 1525-1531.

[3] Ranganathan, R., et al. ” Evolutionary trace method defines binding surfaces common to protein families.” Journal of Molecular Biology, vol. 257, no. 4, 1996, pp. 841-858.

[4] Teichmann, S.A. ” The evolution of protein function.” Current Opinion in Structural Biology, vol. 12, no. 3, 2002, pp. 354-360.

[5] Tsien, R.Y. ” The green fluorescent protein.” Annual Review of Biochemistry, vol. 67, 1998, pp. 509-544.

[6] Chudakov, D.M., et al. ” Fluorescent proteins as a toolkit for in vivo imaging.” Trends in Biotechnology, vol. 23, no. 12, 2005, pp. 605-613.

[7] Arnold, F.H. ” Combinatorial and computational challenges for biocatalyst design.” Nature, vol. 409, no. 6817, 2001, pp. 253-257.

[8] Hult, K., and P. Berglund. ” Enzyme promiscuity: mechanism and applications.” Trends in Biotechnology, vol. 21, no. 1, 2003, pp. 3-9.

[9] Bornscheuer, U.T., and R.J. Kazlauskas. ” Catalytic promiscuity: an expanding phenomenon in synthetic chemistry.” Angewandte Chemie International Edition, vol. 45, no. 18, 2006, pp. 2968-2978.

[10] Leader, B., Q.J. Baca, and D.E. Golan. ” Protein therapeutics: a summary and pharmacological classification.” Nature Reviews Drug Discovery, vol. 7, no. 1, 2008, pp. 21-39.

[11] Strohl, W.R. ” Therapeutic Antibody Engineering: Current and Future Advances Driving the Strongest Growth Area in the Pharmaceutical Industry.” Woodhead Publishing, 2015.

[12] Carter, P. ” Site-directed mutagenesis.” Biochemistry, vol. 30, no. 26, 2001, pp. 6269-6275.

[13] Davies, G.J., and B. Henrissat. ” Structures and mechanisms of glycosyl hydrolases.” Structure, vol. 3, no. 9, 2006, pp. 853-859.

[14] Chakrabarti, S., and A. Panchenko. ” Protein-protein interaction interfaces and functional modules.” Current Opinion in Structural Biology, vol. 20, no. 4, 2012, pp. 484-490.

[15] Lapedes, A.S., et al. ” Correlated mutations in models of protein sequences: phylogenetic and structural effects.” Lecture Notes-Monograph Series, 1999, pp. 236-256.

[16] Verma, S., et al. ” Elucidating functional residues of a protein by integrating evolutionary and physicochemical patterns.” Scientific Reports, vol. 8, 2018, article 33984.

[17] Starr, T.N., and J.W. Thornton. ” Epistasis in protein evolution.” Proceedings of the National Academy of Sciences, vol. 111, no. 22, 2014, pp. E2301-E2309.

[18] Marks, D.S., et al. ” Protein 3D structure computed from evolutionary sequence variation.” PLoS One, vol. 6, no. 12, 2011, e28766.

[19] Hopf, T.A., et al. ” Mutation effects predicted from sequence co-variation.” Nature Biotechnology, vol. 35, no. 2, 2017, pp. 128-135.

[20] Riesselman, A.J., et al. ” Deep generative models of genetic variation capture the effects of mutations.” Nature Methods, vol. 15, no. 10, 2018, pp. 816-822.

[21] Pavelka, A., et al. ” HotSpot Wizard: a web server for identification of hot spots in protein engineering.” Nucleic Acids Research, vol. 37, Web Server issue, 2009, pp. W376-W383.

[22] Sumbalova, L., Stourac, J., Martinek, T., Bednar, D., & Damborsky, J. (2018). HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input

information. Nucleic Acids Research, 46(W1), W356-W362. doi:10.1093/nar/gky500.

[23] Kingma, D.P., & Welling, M. ” Auto-Encoding Variational Bayes.” arXiv, 2014.

[24] Shannon, C. E. ” A mathematical theory of communication.” Bell System Technical Journal, vol. 27, 1948, pp. 379-423.

[25] Cover, T. M., & Thomas, J. A. ” Elements of Information Theory.” Wiley-Interscience, 2006.

[26] G¨obel, U., Sander, C., Schneider, R., & Valencia, A. ” Correlated mutations and residue contacts in proteins.” Proteins: Structure, Function, and Bioinformatics, vol. 18, no. 4, 1994, pp. 309-317.

[27] Dorsam, R. T., & Gutkind, J. S. ” G-protein-coupled receptors and cancer.” Nature Reviews Cancer, vol. 7, no. 2, 2010, pp. 79-94.

[28] UniProt. https://www.uniprot.org/

[29] Karasuyama, M., Inoue, K., Nakamura, R., Kandori, H., & Takeuchi, I. Understanding colour tuning rules and predicting absorption wavelengths of microbial rhodopsins by data-driven machine learning approach. Scientific Reports 8, 16334 (2018).

[30] Tsujimura, M., & Ishikita, H. Insights into the Protein Functions and Absorption Wavelengths of Microbial Rhodopsins. Journal of Physical Chemistry B 2020, 124, 12956-12965. https: //pubs.acs.org/doi/10.1021/acs.jpcb.0c08910

[31] FPbase. https://www.fpbase.org/

[32] Heim, R., Prasher, D. C., & Tsien, R. Y. Wavelength mutations and posttranslational autoxidation of green fluorescent protein. Proceedings of the National Academy of Sciences 91, 12501-12504 (1994).

[33] Orm¨o, M., Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y., & Remington, S. J. Crystal structure of the Aequorea victoria green fluorescent protein. Science 273, 1392-1395 (1996).

[34] NCBI. https://www.ncbi.nlm.nih.gov/

[35] Deep learning prediction of enzyme optimum pH. bioRxiv (2021). https://www.biorxiv. org/content/10.1101/2021.06.25.449983v1

[36] Huang, S., et al. Identification and Characterization of a Catalytic Base in Bacterial Luciferase by Chemical Rescue of a Dark Mutant. Biochemistry 37, 8614-8614 (1998). DOI:10.1021/bi985039j. PMID:9622513.

[37] Romero, E., et al. Same Substrate, Many Reactions: Oxygen Activation in Flavoenzymes. Chem Rev 118, 1742-1769 (2018). DOI:10.1021/acs.chemrev.7b00650. PMID:29323892.

[38] Campbell, Z.T., et al. Crystal structure of the bacterial luciferase/flavin complex provides insight into the function of the beta subunit. Biochemistry 48, 6085-6094 (2009). DOI:10.1021/bi900003t. PMID:19435287.

[39] Li, C.H., et al. Active Site Hydrophobicity Is Critical to the Bioluminescence Activity of Vibrio harveyi Luciferase. Biochemistry 44, 12970-12977 (2005). DOI:10.1021/bi050935y. PMID:16185065.

[40] Low, J.C., et al. Functional Roles of Conserved Residues in the Unstructured Loop of Vibrio harveyi Bacterial Luciferase. Biochemistry 41, 1724-1731 (2002). DOI:10.1021/bi011958p. PMID:11827516.

[41] Li, H., et al. Effects of Mutations of the His45 Residue of Vibrio harveyi Luciferase on the Yield and Reactivity of the Flavin Peroxide Intermediate. Biochemistry 38, 4409-4415 (1999). DOI:10.1021/bi982396l.

[42] Eckstein, J.W., et al. Mechanism of bacterial bioluminescence: 4a,5-Dihydroflavin analogs as models for luciferase hydroperoxide intermediates and the effect of substituents at the 8-position of flavin on luciferase kinetics. Biochemistry 32, 404-411 (1993). DOI:10.1021/bi00053a004. PMID:8422349.

[43] Tinikul, R., et al. Structure, Mechanism, and Mutation of Bacterial Luciferase. Adv Biochem Eng Biotechnol 154, 47-74 (2016). DOI:10.1007/102014281. PMID:25487767.

[44] Jumper, J., et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589 (2021). https://www.nature.com/articles/s41586-021-03819-2

[45] Lassmann, Timo. ” Kalign 3: multiple sequence alignment of large datasets.” Bioinformatics, vol. 36, no. 6, pp. 1928–1929, 2019. https://doi.org/10.1093/bioinformatics/ btz795.

[46] Alexey A. Pakhomov and Vladimir I. Martynov, GFP Family: Structural Insights into Spectral Tuning, Chemistry & Biology, vol. 15, no. 8, pp. 755–764, 2008. https://doi.org/10.1016/j.chembiol.2008.07.009.

[47] E. Capriotti and P. Fariselli,PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants, Nucleic Acids Res.,vol. 45, no. W1, pp. W247-W252,Jul.2017.https://doi. org/10.1093/nar/gkx369.

[48] Chandrashekar, G. Sahin, F. A survey on feature selection methods. Computers Electrical Engineering 40, 16–28 (2014).

[49] Wang, Y. et al. Identification of key biomarkers for STAD using filter feature selection approaches. PeerJ 10, e14408 (2022).

[50] Solorio-Fern´andez, S., Carrasco-Ochoa, J. A. Mart´ınez-Trinidad, J. F. A review of unsupervised feature selection methods. Artificial Intelligence Review 53, 907–948 (2020).

[51] Sheikhpour, R., Sarram, M. A., Gharaghani, S. Chahooki, M. A. Z. A survey on semi-supervised feature selection methods. Pattern Recognition 64, 141–158 (2017).

[52] Engelhart, E., Emerson, R., Shing, L., Lin, Y., Tanno, Y., Sormanni, P., Georges, G. A dataset comprised of binding interactions for 104,972 antibodies against a SARS-CoV-2 peptide. Scientific Data 9, 653 (2022). DOI:10.1038/s41597-022-01779-4.

[53] Jones, G., Jindal, A., Ghani, U., Kotelnikov, S., Egbert, M., Hashemi, N., Vajda, S., Padhorny, D., Kozakov, D. Elucidation of protein function using computational docking and hotspot analysis by ClusPro and FTMap. Acta Crystallogr D Struct Biol 78(Pt 6), 690-697 (2022). DOI:10.1107/S205979832200682X.

[54] Desta, I.T., Porter, K.A., Xia, B., Kozakov, D., Vajda, S. Performance and Its Limits in Rigid Body Protein-Protein Docking. Structure 28(9), 1071-1081 (2020). DOI:10.1016/j.str.2020.06.006.

[55] Vajda, S., Yueh, C., Beglov, D., Bohnuud, T., Mottarella, S.E., Xia, B., Hall, D.R., Kozakov, D. New additions to the ClusPro server motivated by CAPRI. Proteins: Structure, Function, and Bioinformatics 85(3), 435-444 (2017). DOI:10.1002/prot.25219.

[56] Kozakov, D., Hall, D.R., Xia, B., Porter, K.A., Padhorny, D., Yueh, C., Beglov, D., Vajda, S. The ClusPro web server for protein-protein docking. Nature Protocols 12(2), 255-278 (2017). DOI:10.1038/nprot.2016.169.

[57] Kozakov, D., Beglov, D., Bohnuud, T., Mottarella, S., Xia, B., Hall, D.R., Vajda, S. How good is automated protein docking? Proteins: Structure, Function, and Bioinformatics 81(12), 2159-2166 (2013). DOI:10.1002/prot.24403.

[58] Brenke, R., Hall, D.R., Chuang, G.-Y., Comeau, S.R., Bohnuud, T., Beglov, D., Schueler-Furman, O., Vajda, S., Kozakov, D. Application of asymmetric statistical

840 potentials to antibody-protein docking. Bioinformatics 28(20), 2608-2614 (2012).
841 DOI:10.1093/bioinformatics/bts493.

842

843 **Tables and Figures:**

844

845 **Table 1**: Summary of Protein Families and Target Variables with Key Residues, For
846 Rhodopsins the Focus Sequence was GR, for GFP the Focus sequence was aacuGFP1,
847 and for Alkanal Monoxygenases the Focus Sequence Was P07739.2. Datasets contained:
848 884, 611, and 1880 distinct proteins respectively

Protein Family	Target Variable	Key Residues	Citations	Size	Cramer MI
Rhodopsins	Absorption wavelength	M52, V80, A84, Y119, D121, W122, T125, V126, L129, M158, I159, G162, G178, S181, T182, F185, W222, Y225, P226, Y229, Y249, D253, A256, K257	[28–30]	884	F38, N43, W95, L101, S110, K112, D121, T125, L129, E143, S168, T172, T173, L200, V201, R202, S237, V243, V245, D253, A256, L269, D275, H282
GFP and Variants	Fluorescence, Excitation, Quantum yield	Q68, Y69, G70, R97, W145, K147, E150, H199, I201, Q219	[5, 28, 31-33, 46]	611	Q68, N71, R72, L127, V163, E165, H199, G207, M214, P227
Alkanal Monoxygenases	Optimum pH	H44, H45, C106, N113, S227	[28, 34-43]	1880	L34, P120, Q124, E248, K263

849

850

851

852

853

854

855

856

857

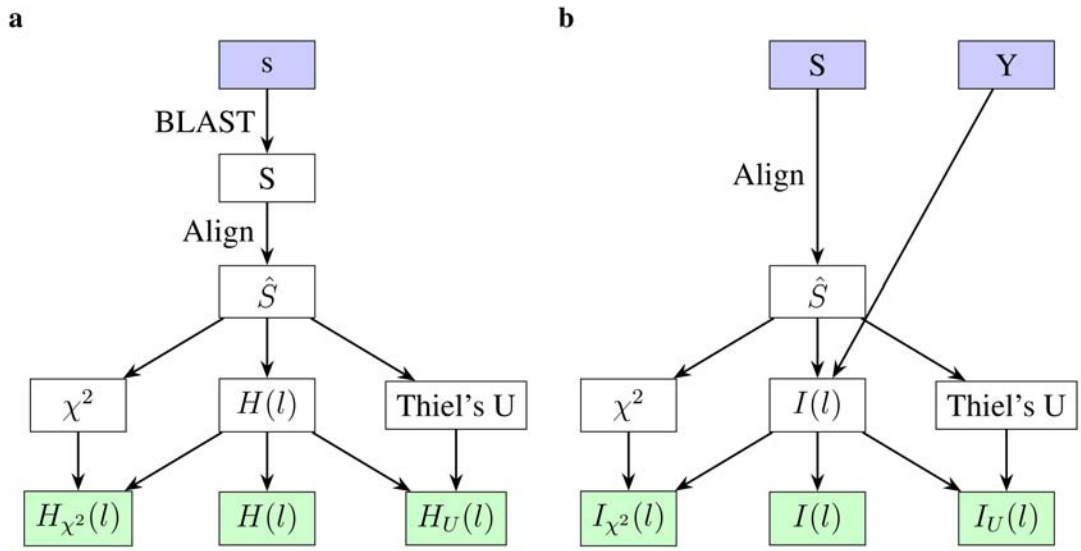

Figure 1: a Flow chart showing the process of calculating and modifying entropies using BLAST, alignment, and subsequent calculations of $H(l)$, χ^2, and Thiel's U. The sole required input is a single amino acid sequence (s). b Flow chart showing the process of calculating and modifying information measures using BLAST, alignment, and subsequent calculations of $I(l)$, χ^2, and Thiel's U. The required inputs are a database of amino acid sequences (S), and their corresponding target parameter e.g. wavelength (Y).

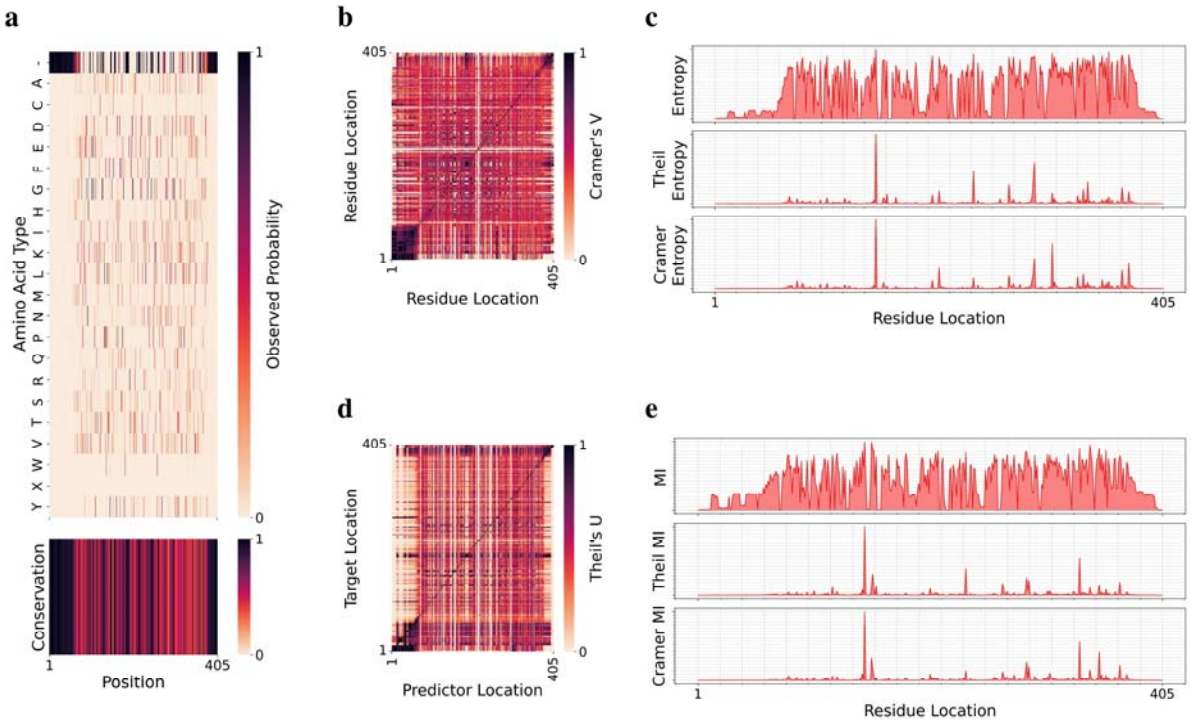

Figure 2: Detailed analysis of residue decoupling for the Fluorescent Protein dataset, focusing on peak excitation wavelength as the target variable. Panel a characterises the dataset of fluorescent proteins, Top: Probability distribution of residues based on observed occurrences in a dataset of 611 Fluorescent Protein variants after Multiple Sequence Alignment (MSA), highlighting residue frequency across sequence positions. Bottom: Visualization of entropy-based residue conservation along the sequence, together these figures underscore the strong correlation between gap frequency and conservation. Panels b and d show coupling matrices obtained using Cramer's V (symmetric measure) and Theil's U (non-symmetric measure) across the aligned sequences, revealing residue as- sociations across the sequence length, with Cramer's V being a measure of association based on Chi squared, and Theil's U being a measure of predictive power. Panels c and e show Residue importance plots demonstrating the enhancement of fingerprint precision using Theil's U or Cramer's V as coupling metrics. In these two panels the upper plot shows the raw importance scores, while the two lower plots show Theil's-U and Cramer's-V decoupled importance scores, respectively. These plots illustrate how these measures sharpen the protein fingerprint, giving a more confident prediction of key residues crucial for peak excitation wavelength in fluorescent proteins.

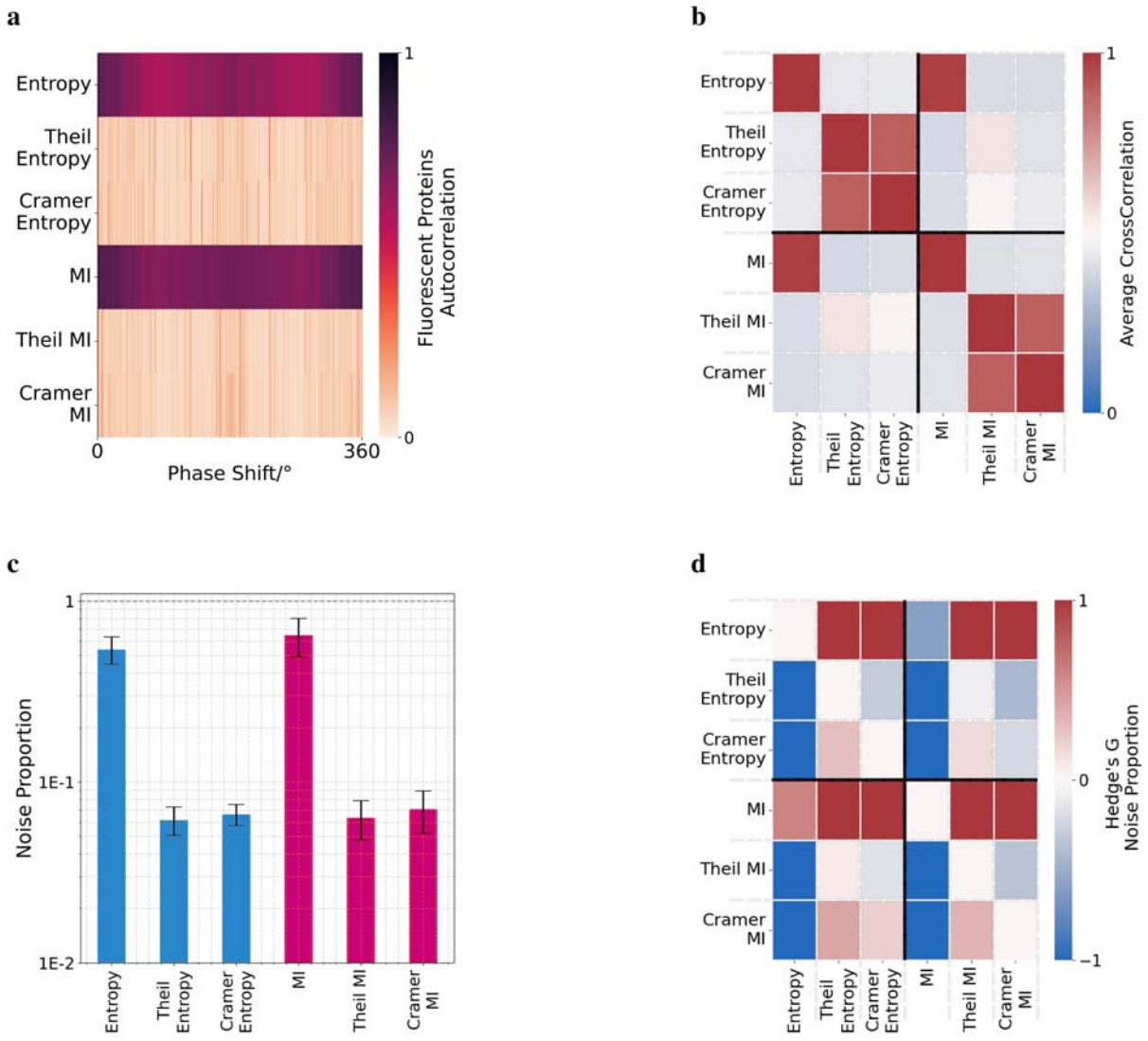

Figure 3: Autocorrelation and ranking consistency analyses across datasets. a shows the autocorrelation profile for Fluorescent Proteins, indicating the effect of decoupling on noise reduction. b presents average cross-correlation among ranking datasets, where values close to 1 denote strong correlations. c illustrates the average noise propor- tion across methods, with 1 representing a white noise process, (refer to Supplementary Equation S11). d displays the Hedge's G Matrix for all measures, highlighting correlations between ranking metrics.

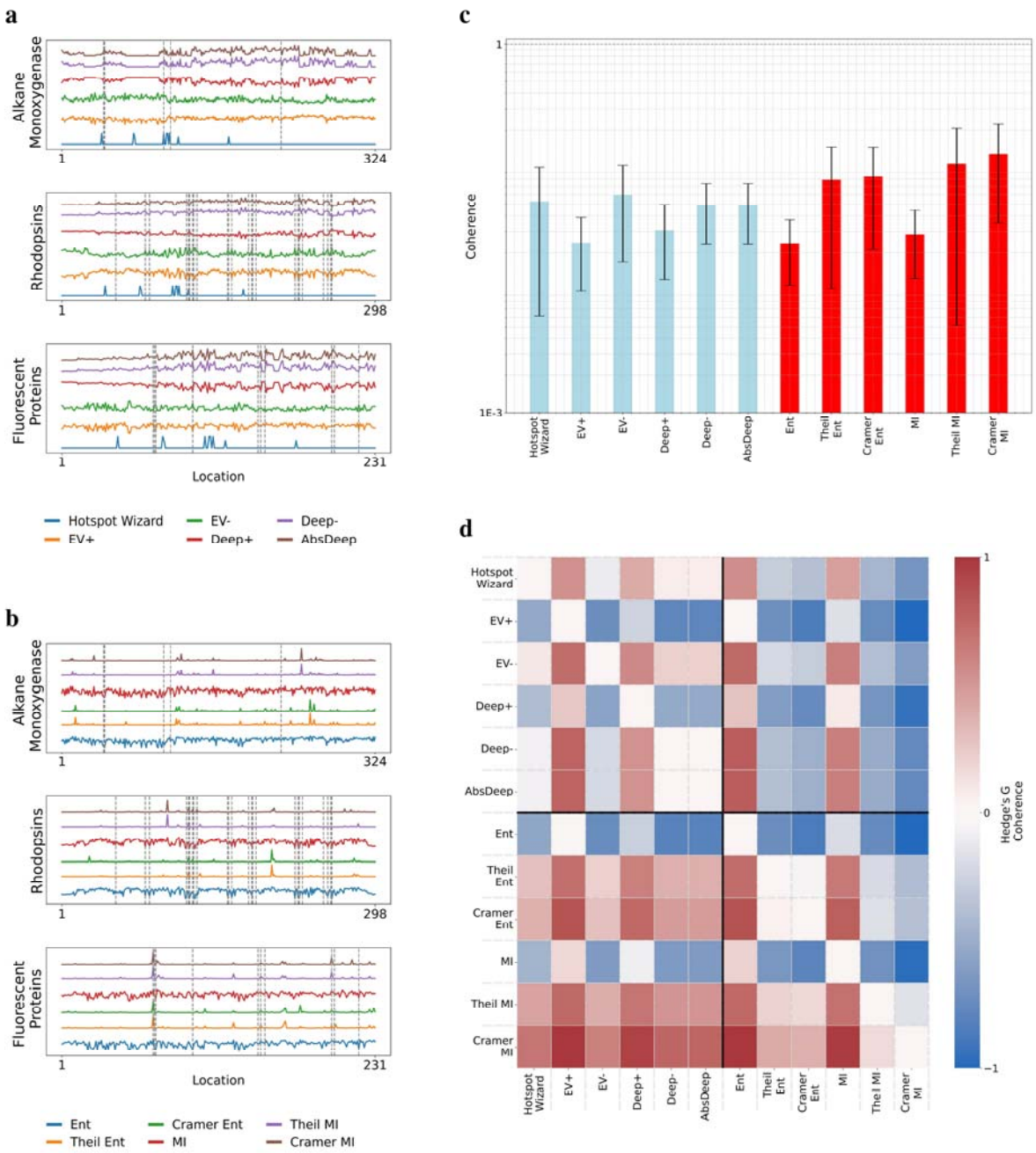

Figure 4: a and **b** show the ranking fingerprints for the 6 existing and 6 developed methods, respectively, for the unaligned primary sequence. Gray dashed lines indicate key residues identified in the literature. Fingerprints are presented for (top) alkanal monooxygenases, (middle) Rhodopsins, and (bottom) Fluorescent Proteins excitation wavelength. **c** shows the estimated coherence of each of the 12 measures assessed, averaged across each dataset. **d** presents the Hedge's G matrix comparing coherence among different measures.

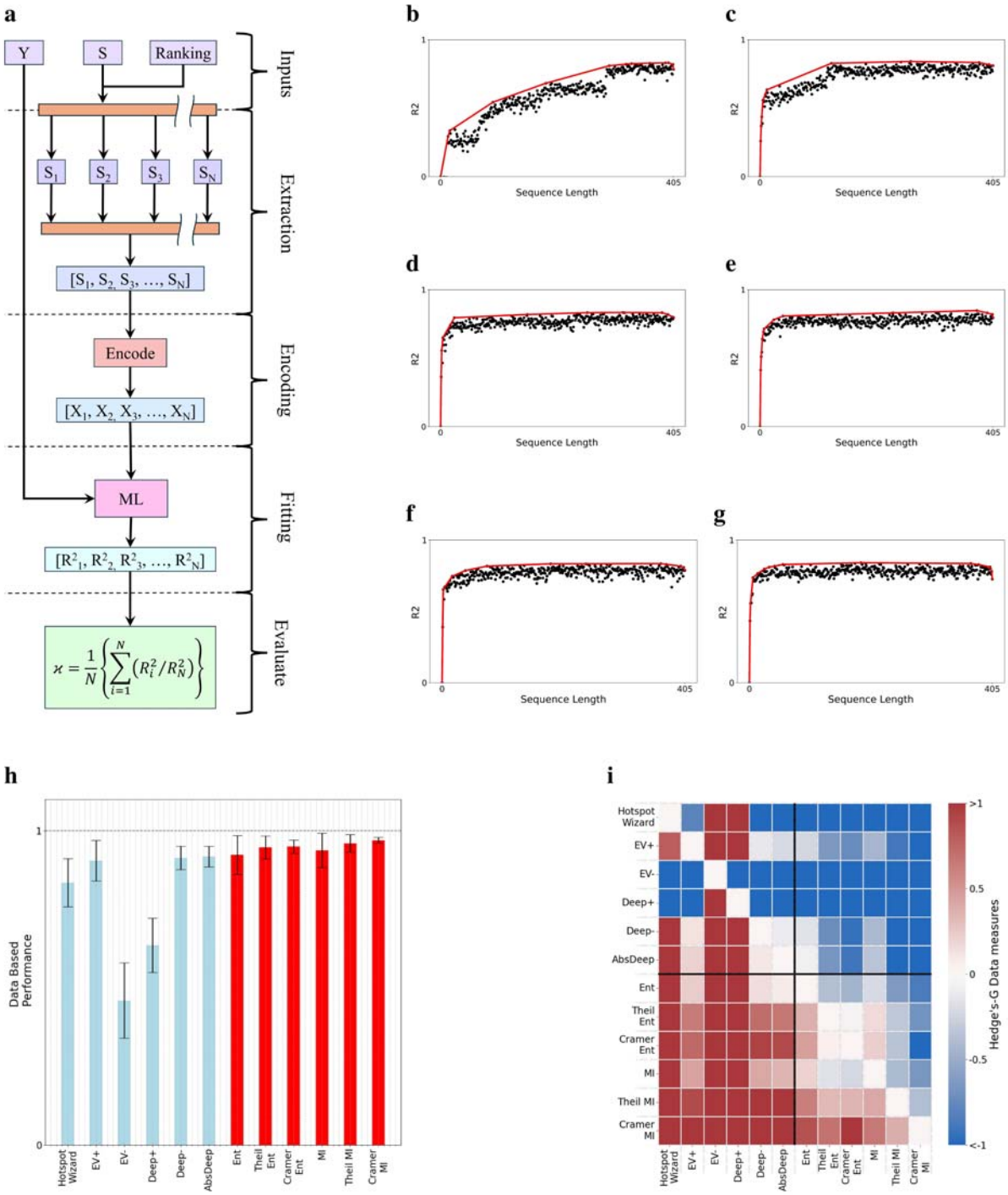

Figure 5: a Design of the overall pipeline for data-based ranking assessment. This method requires labelled data (Y) for a protein sequence set (S) and an initial ranking developed by a residue ranking tool. The ranking is obtained from a method's generated fingerprint and is simply the ordered list of indices sorted by fingerprint intensity at a

given location. The process has four key steps: extraction, encoding, fitting and evaluation. **b, c, d, e, f,** and **g** show characteristic curves of R^2 vs. aligned sequence length for top-ranked residues for the fluorescent protein dataset with peak excitation wavelength using: Deep Sequence+, Deep Sequence-, Entropy, Cramer Entropy, MI and Cramer MI respectively. h shows the average data-based score across all datasets for each of the existing and naive measures, where a performance of one indicates ideal performance. **i** presents the Hedge's G matrix when comparing the data-based performance of the 12 different measures assessed, with large positive values indicating that row entries outperform column entries, and large negative values indicating the inverse.

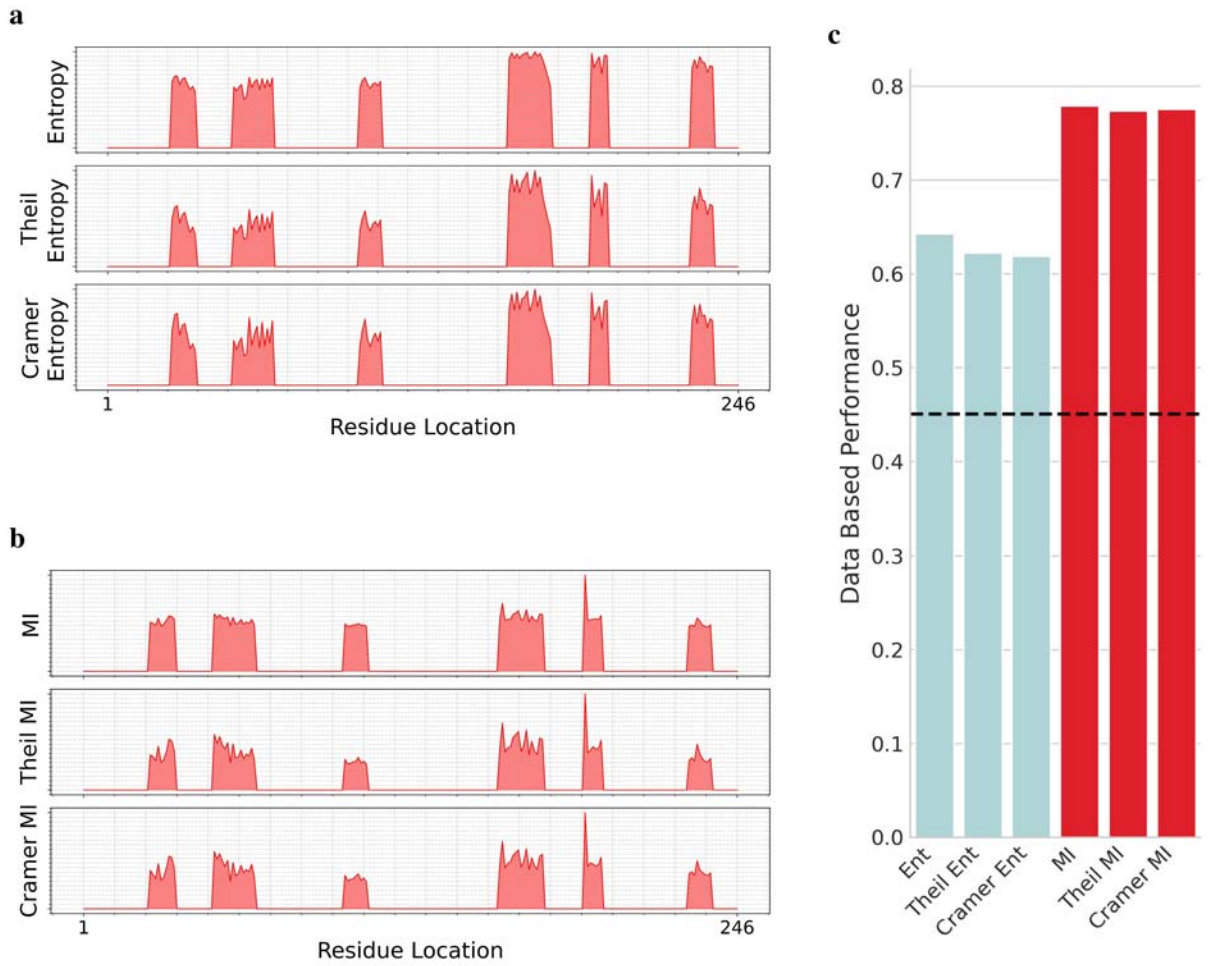

Figure 6: a and **b** present residue importance plots that demonstrate enhanced fingerprint precision via coupling metrics, Theil's U and Cramer's V. In each panel, the upper sub-panel displays the raw importance scores, whereas the lower sub-panels show the decoupled scores after applying Theil's U and Cramer's V, respectively. These results indicate that decoupling sharpens the protein fingerprint, yielding more confident predictions of key residues linked to peak excitation wavelength in fluorescent proteins. Although the overall fingerprint structures are similar for both entropy- and MI-based approaches, the MI-based method exhibits sharper, more specific peaks. **c** shows the data-based performance of the entropy-based (blue) and MI-based (red) methods, with the black dashed line marking the performance of the best existing method evaluated on the prior three datasets (Alkane Monooxygenases, Fluorescent Proteins, and Rhodopsins). Both MI and entropy-based methods outperform the best existing approach, with MI-based techniques consistently demonstrating enhanced performance over entropy-based approaches.

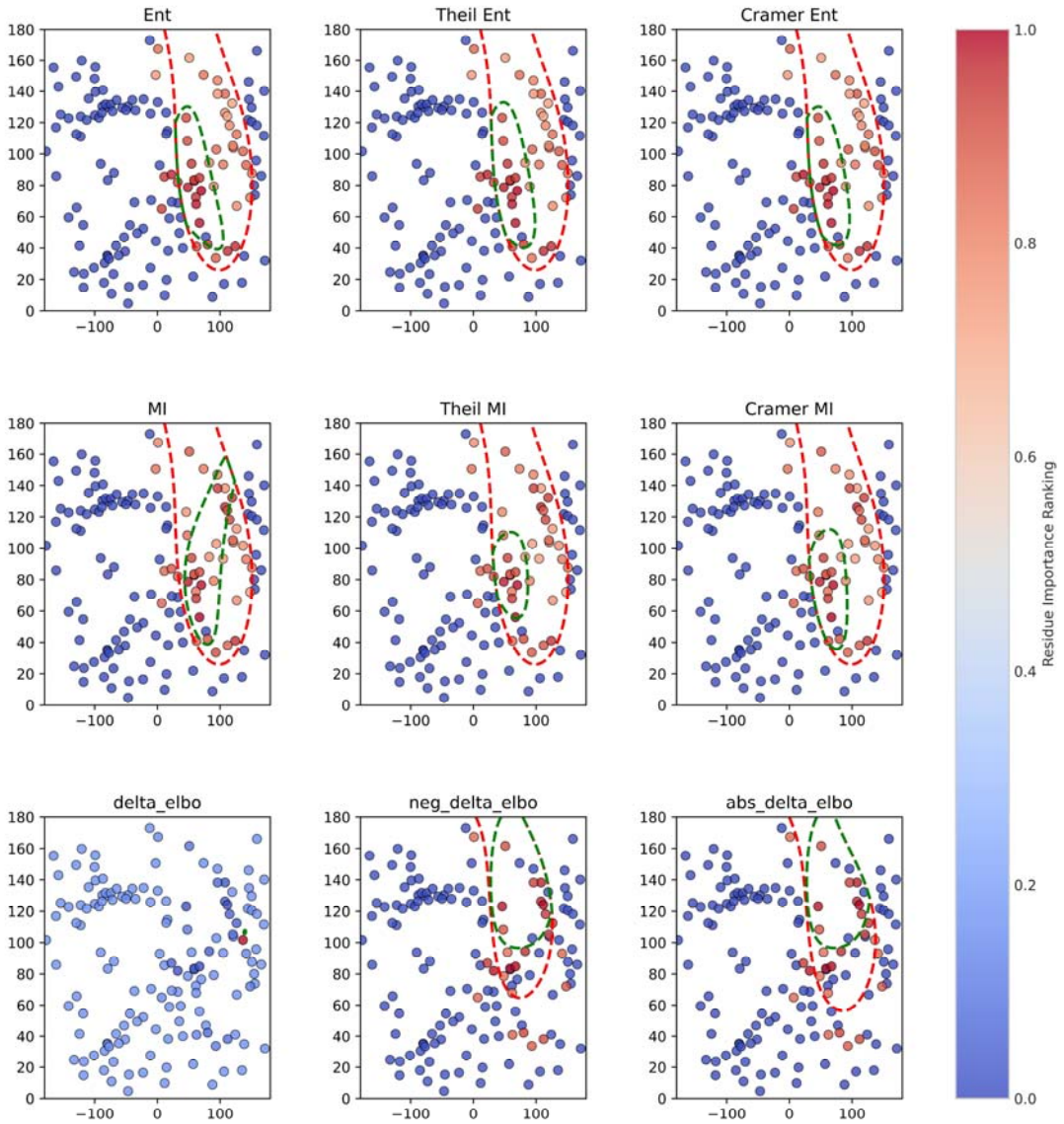

Figure 7: Residue rankings based on importance, for solvent-accessible residues, mapped onto a unit sphere centered at the protein's centroid. The first two rows correspond to the developed methods, while the last row represents existing approaches. Importance rankings are normalized between 0 and 1 for direct comparison. The developed methods distinctly highlight a well-defined surface of key residues, in contrast to existing approaches, which produce broader and less localized importance distributions. This improved selectivity suggests that the developed methods offer enhanced sensitivity in identifying functionally significant regions. The contour in red highlights the top 50% of residues, while the contour in green highlights the top 10% of residues.

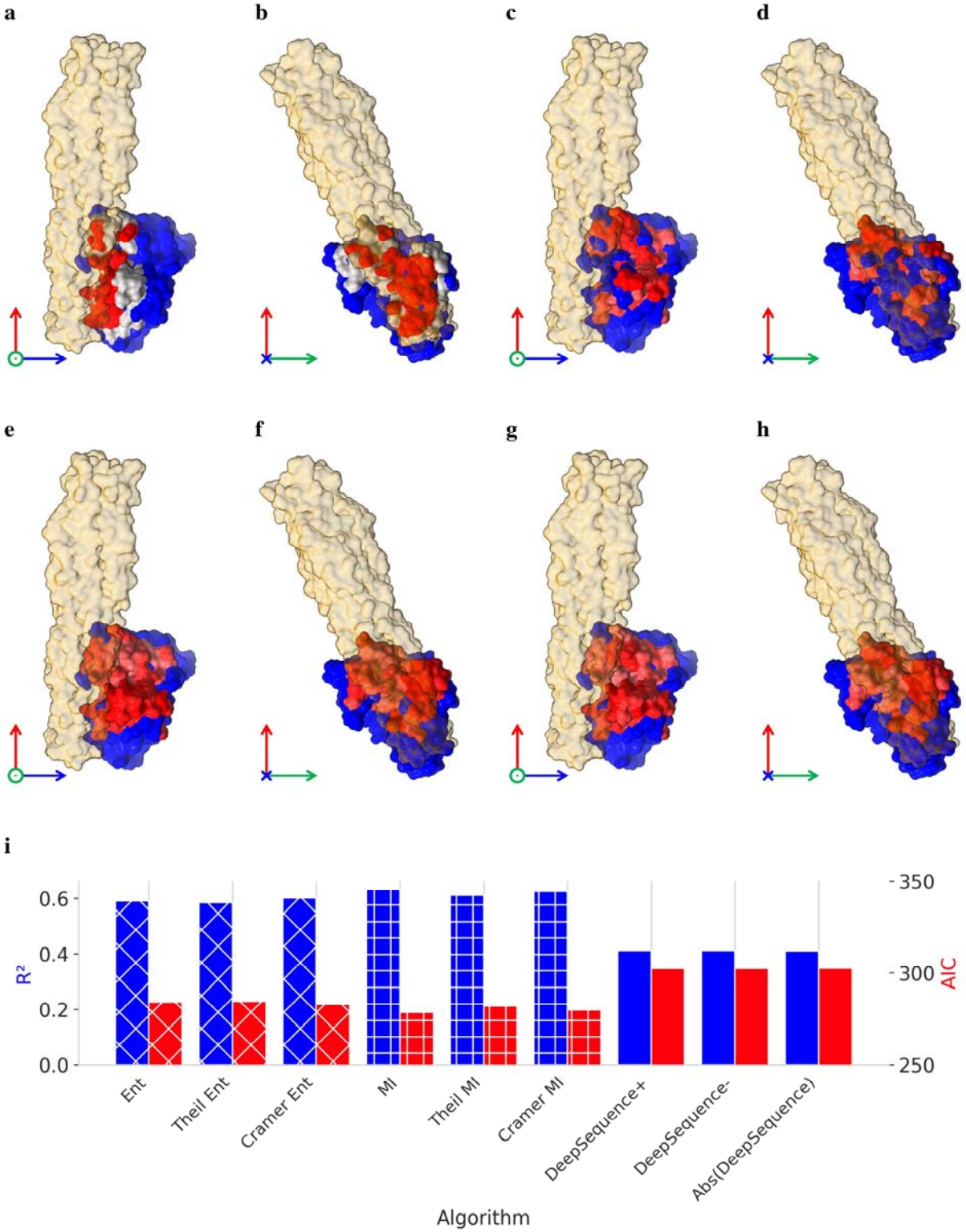
a
b
c
d
e
f
g
h
i
0.6
0.4
0.2
0.0
R²
350
300
250
AIC
Ent
Theil Ent
Cramer Ent
MI
Theil MI
Cramer MI
DeepSequence+
DeepSequence-
Abs(DeepSequence)
Algorithm

Figure 8: Comparison of residue importance scores across computational methods. Side and top views of binding residues (**a,b**), the best existing tool (DeepSequence+) (**c,d**), entropy-based extraction (**e,f**), and MI-based extraction (**g,h**). Top views are rotated 90° around the Z-axis. Panel **i** shows R^2 and AIC values for residue rankings versus binding energies (Methods S1.9).