



OPEN ACCESS

EDITED BY

Jefferson Russo Victor,
School of Medicine - University of São Paulo
(FM-USP), Brazil

REVIEWED BY

Ashish Kumar Agrahari,
Clemson University, United States
Jerome Anthony E. Alvarez,
George Mason University, United States

*CORRESPONDENCE

Vincent J. Hilser
✉ Hilser@jhu.edu

†PRESENT ADDRESS

Antonieta van den Berg Monsalve,
Department of Biology, Otterbein University,
Westerville OH, United States

RECEIVED 18 November 2025

REVISED 14 January 2026

ACCEPTED 15 January 2026

PUBLISHED 10 February 2026

CITATION

Wrabl JO, Beale J, Fortunato G,
Monsalve AvdB and Hilser VJ (2026)
Ensemble molecular mimicry correlates
with antibody cross-reactivity in
proteome-wide studies.
Front. Immunol. 17:1749369.
doi: 10.3389/fimmu.2026.1749369

COPYRIGHT

© 2026 Wrabl, Beale, Fortunato, Monsalve and Hilser. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Ensemble molecular mimicry correlates with antibody cross-reactivity in proteome-wide studies

James O. Wrabl¹, Josh Beale^{1,2}, Gabriel Fortunato^{1,3,4},
Antonieta van den Berg Monsalve^{1†} and Vincent J. Hilser^{1,3*}

¹Department of Biology, Johns Hopkins University, Baltimore MD, United States, ²Cell, Molecular, Developmental Biology, and Biophysics Graduate Program, Johns Hopkins University, Baltimore MD, United States, ³T. C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore MD, United States, ⁴Program in Molecular Biophysics Graduate Program, Johns Hopkins University, Baltimore MD, United States

Energetics of protein–protein binding necessarily include contributions both from conformational equilibria and from interfacial interactions. In the particular case of an antibody binding to a protein epitope, the conformational contribution is typically neglected as the antibody-bound and free forms of the protein are usually highly similar, leading to the reasonable conclusion that binding affinity in most cases can be reconciled in the context of observed interfacial interactions. However, the phenomenon of molecular mimicry has also been widely observed, wherein antibodies raised against one sequence/structure are able to recognize a completely different sequence/structure. This observation suggests that, in some cases, the conformational contribution could play a significant role in facilitating this cross-reactivity. Here, this conjecture is supported, utilizing a recent discovery that permits evaluation of the thermodynamic compatibility of any sequence for the conformational ensemble of any other protein—in effect providing direct access to the conformational contribution to binding. The importance of the contribution could then be assessed on a proteome-wide scale, in the context of the unexpected cross-reactivity observed when the human proteome is challenged with antibodies raised against a set of virus protein antigens. Because the virus protein antigens and the cross-reactive human proteins share substantial similarity when modeled as thermodynamic ensembles, despite the absence of detectable sequence or structural similarity, we hypothesize that these cross-reactive epitopes share a novel kind of immunological molecular mimicry, termed “ensemble molecular mimicry” (EMM). To investigate potential mechanisms, a sequence-based algorithm was developed to probe for the relationship between high scoring sequence segments and cross-reacting epitopes, and it was discovered that 9 of 11 medically relevant cross-reactive epitopes taken from the literature exhibited

higher-than-expected local EMM values. Taken together, the results suggest that conformational equilibrium can affect affinity and that it is hypothetically possible for cross-reactive epitopes to share a pairwise thermodynamic signature, even in the absence of sequence or structural similarity.

KEYWORDS

autoimmunity, binding energetics, conformational equilibrium, polyclonal, protein ensemble

1 Introduction

Thousands of high-resolution structures of antibody–antigen complexes exist in the Protein Data Bank (1, 2). This wealth of data has been transformative for the understanding of antibody specificity and computation of the intermolecular energetics underlying affinity. However, routine estimation of accurate binding energies from structure remains elusive and often requires resource-intensive simulations. As such, significant efforts have been directed towards force-field development and artificial intelligence-driven molecular modeling (3–7), but despite these efforts, reliable prediction of which antibodies will bind to which antigens, and how tightly, is not currently achievable (8).

While the binding energetics at the antibody–antigen interface are undoubtedly important and most accessible for study, less attention has been paid to the idea that every protein–protein binding reaction also includes a contribution from the ensemble conformational equilibrium. There is always a separate energy cost for the antigen (as well as the antibody) to adopt the specific conformations necessary to form the complex, i.e., a free energy of stability upon which the population of binding-competent antigen, as dictated by statistical thermodynamics, depends. In the case of high-affinity, monoclonal antibody complexes, this contribution might be safely ignored, as strong interfacial atomic interactions would be expected to dominate overall binding affinity. For weaker polyclonal responses, however, ensemble conformational equilibrium can be the decisive factor, as demonstrated by the elegant pioneering experiments of Anfinsen and colleagues (9, 10).

If ensemble conformational equilibrium is relevant to the polyclonal response to antigen, how can it be detected? In this work, we take a biophysical approach to the question by deploying state-of-the-art proteome chip technology, which enables large-scale measurements of relative binding affinities for a panel of polyclonal antibodies against nearly the entire human proteome ($N \sim 20,000$ proteins). These relative affinities are then compared to high-throughput computational estimates of ensemble conformational equilibrium developed in this laboratory, which are calculated in pairwise fashion between highest-affinity proteins from the chip and the antigens against which the polyclonal antibodies were raised.

In these experiments we test the null hypothesis that affinity and conformational equilibrium are unrelated, and based on modest but significant correlation between empirical affinities and computations, we conclude that ensemble conformational equilibrium should not be neglected as part of the polyclonal response. Furthermore, because the antigens are derived from a virus and the binding proteins come from the human proteome, the results report unexpected cross-reactivity and may help to explain documented cases of molecular mimicry in varied auto-immune diseases characterized by cross-reactive epitopes that do not exhibit obvious sequence or structure similarity (11–14). Instead, we hypothesize a new type of epitope based on “ensemble molecular mimicry” (EMM), which could be revealed through detailed thermodynamic analysis of local conformational equilibria between pairs of proteins. This ensemble information may be useful for the future prediction of auto-antigenic proteins and epitopes.

2 Results

2.1 Fluorescence as a proxy for total binding free energy

We consider any macromolecular binding reaction, such as that between an antibody (Ab) and a specific antigen, as a series of two coupled equilibria (Figure 1A) (9). In this scheme, binding free energy (ΔG_{total}) is influenced by at least two energetic contributions: An *intrinsic* contribution (ΔG_{int}) resulting from the interactions of particular chemical groups between the antibody and the antigen (such as hydrogen bonding, charge–charge interactions, and hydrophobic packing), and a conformational contribution (ΔG_{conf}) reflecting the free energy required for the epitope to adopt a binding-competent geometry. We expect that the particular sequence of the amino acids at the binding interface is primarily accounted for by the ΔG_{int} term, whereas contributions associated with redistributing the conformational ensemble to those conformations capable of binding the antibody, e.g., protein folding, comprise the ΔG_{conf} term.

The classic view of antibody binding is one of rigid-body association, which assumes that the ΔG_{conf} term is relatively small

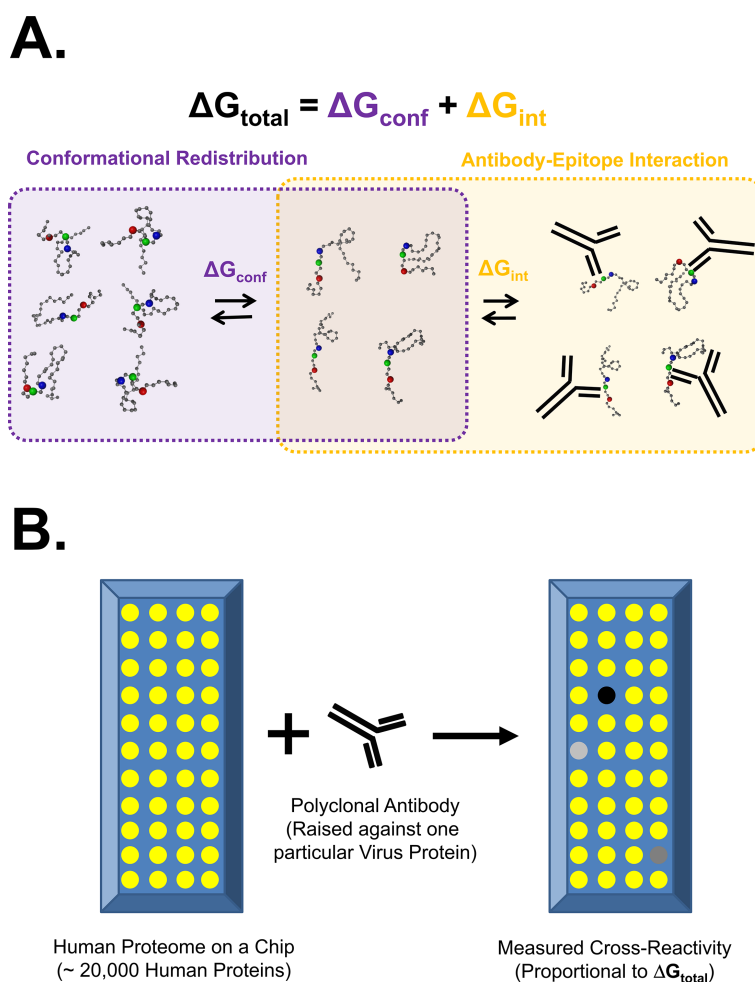


FIGURE 1

Conformational equilibrium contributes to binding affinity and affinity can be measured by proteome-on-a-chip technology. **(A)** Conformational equilibrium contributes to binding affinity. In general, coupled equilibria exist describing the binding between an antibody (black cartoon Y) and an antigen (colored beads-on-string cartoon). The free energy of binding (ΔG_{total}) has two major contributions: a free energy of the antigen epitope interacting with the antibody paratope (ΔG_{int}), and a free energy of the antigen epitope adopting the correct conformation to permit interaction (ΔG_{conf}). For binding to occur (orange box, right side), the antigen must adopt a binding-competent conformation (purple box, right side) as opposed to a binding-incompetent conformation (purple box, left side). Conformational equilibrium of the epitope between binding-incompetent and binding-competent shapes is suggested by the purple box. The hypothesis tested in this work is that while epitope sequence conservation is reflected by ΔG_{int} , thermodynamic molecular mimicry in the absence of sequence identity can be facilitated by a favorable ΔG_{conf} . **(B)** Affinity can be measured by proteome-on-a-chip technology. Almost the entire human proteome can be expressed, purified, and spotted on a nitrocellulose/glass chip (left side) using technology developed by CDI Laboratories, Inc. (Baltimore, MD). Each yellow spot represents many copies of a localized particular human protein. An antibody or serum sample (black cartoon Y) can be passed over the chip, probing all of the proteome simultaneously with a primary antibody. Binding of primary antibody to localized proteins on the chip can be quantified with a fluorescent secondary antibody (not shown), akin to a Western blot. Schematized locations of binding are shown by dots with shades of gray. It is assumed that degree of fluorescence is proportional to relative affinity.

and that tight binding originates from high structural compatibility between the Fab region of the antibody and the antigen that it binds. Indeed, high similarity is often observed when the structures of unbound and antibody-bound antigens are compared, supporting the notion that the conformational free energy difference between unbound and bound states is small, at least in those cases ($\Delta G_{\text{conf}} \sim 0$) (15, 16). However, it is also well known that the antibody maturation process involves an initial polyclonal response, which consists of numerous sequence-distinct antibodies whose individual binding affinities are significantly lower than the mature (monoclonal) antibody, but whose collective binding affinity is relatively high. Moreover, the two contributors to overall binding

energy may compensate for each other. For example, if an antigen readily populates a large sub-ensemble of conformational states, each of which can only provide sub-optimal contacts at the binding interface(s), then this “penalty” of a relatively high ΔG_{int} can be countered by a relatively low ΔG_{conf} (Figure 1A). In other words, the bound complex can be conformationally heterogeneous, as schematically depicted in Figure 1A.

Although the gold-standard measurement of binding affinity (ΔG_{total}) requires bulk-solution biophysical methods that monitor direct binding, such as isothermal titration calorimetry, fluorescence anisotropy, or nuclear magnetic resonance, these low-throughput methods are not amenable to proteome-scale

measurements of antibody–target complexes. To estimate ΔG_{total} on a large scale, we instead measure the secondary fluorescence signal generated when a primary antibody binds to a target protein localized on a nitrocellulose/glass chip, as implemented on *HuProt* chips by CDI Labs, Inc. (Figure 1B, Section 4.1). One *HuProt* chip can localize large numbers of folded proteins, including essentially the entire human proteome (Figure 1B), and the relative binding of an antibody sample probing the entire proteome can be quantified.

2.2 Thermodynamic similarity (eTFR) as a proxy for free energy of conformational equilibrium

The conformational equilibrium shown in Figure 1A is akin to a free energy of folding, which is notoriously difficult to measure by high-throughput methods [although progress is being made (17, 18)]. For this work, we continue the development of a computational model of the energetics of the protein ensemble (COREX), which recently has been shown to measure the relative energetic distance (ΔG_{conf}) between two protein conformations (19–21). In direct support of the current work, we demonstrated in those studies that protein sequences sharing no similarity can nonetheless adopt the same conformation, and that this propensity is predictable based on our characterization of ensembles from numerous proteins. Here, we go further and ask, if a sequence has a propensity to adopt an alternative fold, can that protein also be recognized by a ligand (or antibody) that binds that alternative fold? To address this question, we modified our previously presented methodology, as described briefly below.

Our approach is termed *eScape* Thermodynamic Fold Recognition (eTFR). eTFR merges two previously published computational resources. The *eScape* (energy landScape) algorithm is a sequence-based predictor of the local folding stability (ΔG) of a protein, along with the enthalpic (apolar and polar ΔH) and entropic ($T\Delta S$ of apolar and polar solvation, as well as conformational $T\Delta S$) components of that stability (22, 23). This algorithm generates the energy landscape of the protein ensemble from information provided by *thermodynamics*, not amino acid sequence or structure (24–26). Specifically, the energy landscape identifies which regions of the protein are more or less locally stable (i.e., more or less likely to be populated in a locally structured conformation).

Thermodynamic Fold Recognition uses “thermodynamic environments” (27–29), defined by the clustering of a large database of vectors $\{\Delta G, \Delta H, T\Delta S\}$, to create a protein’s “thermodynamic profile” (30, 31), which is used as a query to search a database of arbitrary amino acid sequences. Each pairwise alignment in the search is scored by dynamic programming subject to a thermodynamic substitution matrix, and the significance of each score is computed against a calibrated null model (19, 32). In short, eTFR uses a thermodynamic profile derived from *eScape* as a query to search a sequence database with a *thermodynamic* substitution matrix instead of an *amino acid* substitution matrix.

Importantly, this procedure essentially quantifies the compatibility of a protein thermodynamic ensemble with an amino acid sequence, providing a potentially powerful alternative to traditional sequence–sequence, sequence–structure, or structure–structure algorithms, especially when the information output of the traditional methods is undetectably low.

Highest-significance eTFR matches are reliably obtained when the sequence of an extant protein is queried against its native thermodynamic profile, while next-highest matches often occur when *homologous sequences* are queried against the same profile (32). It has also been demonstrated that thermodynamic similarity can exist even below the “twilight zone” of sequence identity (31). It is thus likely that significant eTFR matches between two sequence-dissimilar proteins could predict a substantially favorable value of ΔG_{conf} indicating that the conformational ensembles of the two proteins overlap to some degree and that each protein (or parts thereof) may transiently populate similar local conformations to the other. If true, then antibodies raised against one protein (e.g., protein A) may bind to a second protein (e.g., protein B). This could happen because, as shown previously (19–21), high sequence compatibility corresponds to a low ΔG_{conf} for the sequence of protein B to adopt the structure of protein A. Importantly, these potential cross-reactivities could occur even in the absence of sequence or structural similarity.

We have demonstrated predictable antibody cross-reactivity on a small scale in at least three prospective cases of sequence-dissimilar pairs of proteins—one a member of the SARS-CoV-2 proteome and the other identified from the human proteome—which exhibit a high eTFR significance between them (Supplementary Figure S1A) and bind to the same antibody on Western blot (Supplementary Figure S1B). Of potential medical significance, these unexpected observations of cross-reactivity correspond to three of seven human proteins for which autoantibodies have been previously documented in the blood sera of acute COVID-19 patients (12, 13). We hypothesize that eTFR can report on the likelihood that two proteins, even ostensibly unrelated ones, share conformational ensemble similarity. In as much as the computed similarity has been shown to allow one protein to even adopt the fold of the second protein (19), without explicit consideration of structural complementarity, we were motivated to determine if those same principles applied to the thermodynamic signatures of epitopes, thus motivating the current proteome-wide study.

The workflow for eTFR matching is shown in Figure 2. Briefly, a thermodynamic profile of the virus antigen corresponding to a cognate polyclonal antibody (Figure 2, Box 1, left) was computed with *eScape* from its amino acid sequence. As alluded to above, *eScape* is a machine-learning algorithm that in the simplest case generates a four-dimensional vector of thermodynamic descriptors $\{\Delta G, \Delta H_{\text{ap}}, \Delta H_{\text{pol}}, T\Delta S_{\text{conf}}\}$ from each amino acid in a sequence, corresponding to the stability of the native (folded) state of the protein. Thermodynamic descriptors are mapped to thermodynamic environments (i.e., the closest cluster center of descriptors) observed in a large database of diverse proteins (Figure 2 and Box 2). This reduces the four-dimensional

thermodynamic descriptor space to a one-dimensional thermodynamic profile (Figure 2 and Box 3, left, colored rectangles), for computational efficiency. Using standard dynamic programming, the thermodynamic profile can be optimally aligned to an arbitrary amino acid sequence (Figure 2 and Box 3, left) and a significance can then be computed (Figure 2 and Box 3, right). As mentioned above, in this work, the significance is considered a proxy for the ΔG_{conf} between the virus and human proteins used in the calculation (Figure 2, bottom).

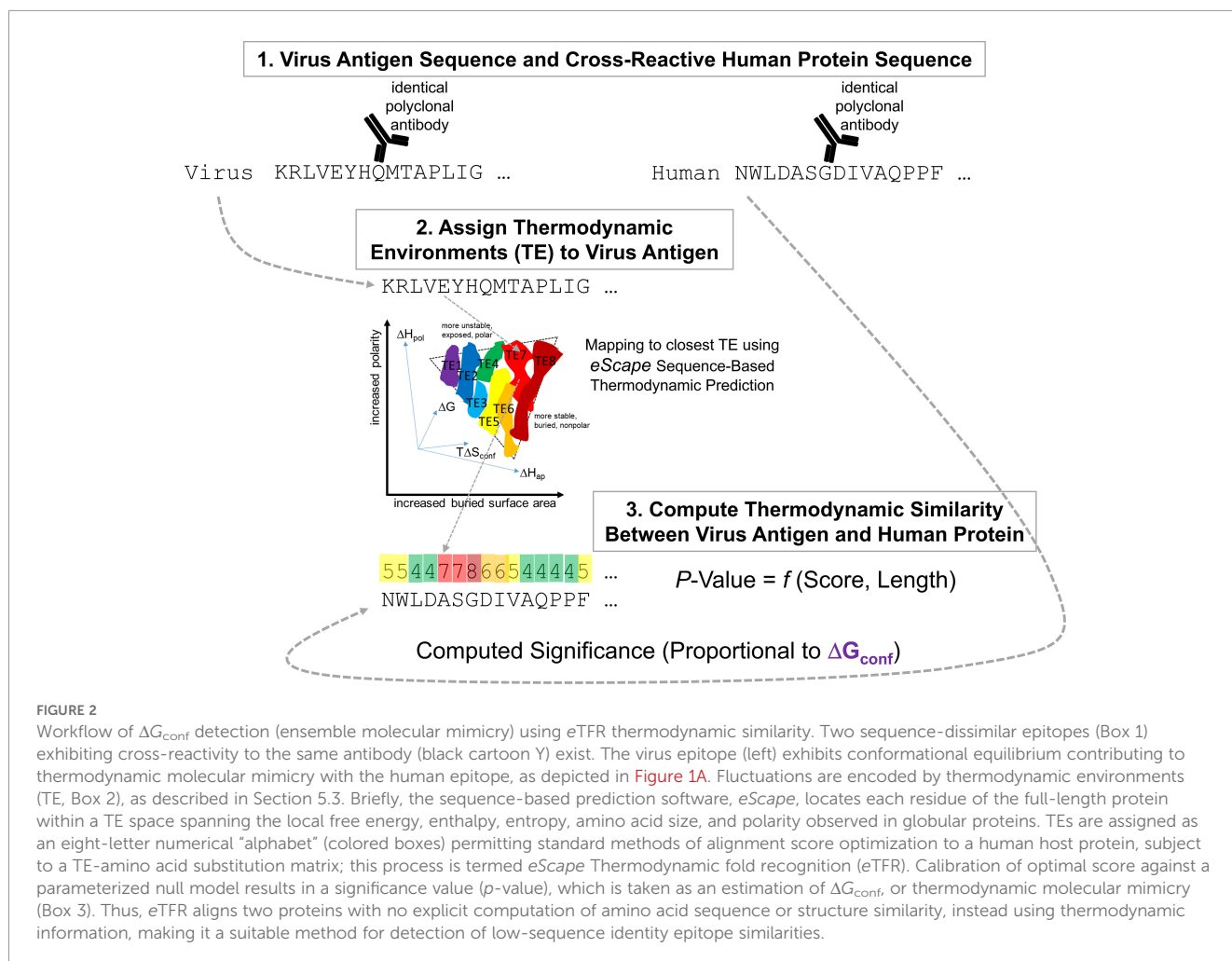
2.3 Correlation between eTFR significance and chip fluorescence suggests association between ΔG_{conf} and ΔG_{total}

A panel of seven commercial polyclonal primary antibodies, raised against seven SARS-CoV-2 full-length protein antigens, were passed over separate human proteome chips, under both native and denaturing conditions (Sections 4.1 and 4.2). These antibodies were chosen from a preliminary scan of the thermodynamic profiles of the SARS-CoV-2 proteome against the amino acid sequences of all proteins in the Protein Data Bank (2) (Section 4.4, Supplementary Figure S1A). For each human protein spot on the chip, secondary

antibody fluorescence proportional to primary antibody affinity was reported as a Z-score relative to all proteins on the chip, and the sum of native and denatured Z-scores for each human protein was calculated (sums were taken because fluorescence signals were found to be relatively independent of native or denatured conditions, cf. Figure 3).

The 10 strongest-binding human proteins against each of the seven anti-virus antibodies of the panel were prioritized as part of CDI Labs' analysis report, so that $7 \times 10 = 70$ protein pairs of virus antigen and human protein were reported as cross-reactive in the company's analysis (Supplementary Table S1). The eTFR significances computed for each virus antigen's thermodynamic profile aligned to its paired human amino acid sequence was compared to the experimental binding affinity for the same pair. To explore the relationship between eTFR significance and binding affinity, the binding data from all seven chips were pooled, and the linear correlation between affinity and significance was directly calculated.

It might be expected that no relationship would exist between binding Z-score and eTFR significance—after all, the protein ensemble model COREX/eScape and the TFR/eTFR methods were not parameterized with any energetic information from protein-protein binding, nor were they trained with the structures of any



antibody–antigen complexes—but indeed, a significant relationship ($p = 5 \times 10^{-3}$) was observed (Figure 4A). Insofar as Z-score and eTFR significance are related to ΔG_{total} and ΔG_{conf} respectively, this finding suggests that (1) ΔG_{conf} contributes to ΔG_{total} , and (2) the magnitude of the contribution (i.e., the R^2 value) is modest, approximately 10% for these cross-reactive protein pairs.

To rule out amino acid sequence identity as the driver for the correlation in Figure 4A, sequence similarity was computed for the 70 virus–human protein pairs by optimally aligning their full-length sequences using FASTA36 (33) (Section 5.3). This local alignment algorithm is a stringent test for possible conserved sequence motifs as only the shortest, best match of highest sequence identity is reported, and random alignments are not considered. As no correlation was observed between Z-score and sequence similarity (Figure 4B), the possibility that amino acid identity represents the source of the correlation is not supported by these studies, which further suggests that the compatibility of a human protein sequence with a viral protein fold (i.e., a viral protein ensemble) is contributing to the overall binding affinity, ΔG_{total} .

3 Discussion

3.1 Thermodynamic similarity between full-length proteins and relation to epitopes

The importance of the finding that conformational compatibility between full-length sequences can contribute to antibody cross-reactivity cannot be overstated. Yet, identifying the underlying determinants in each case and establishing causality is

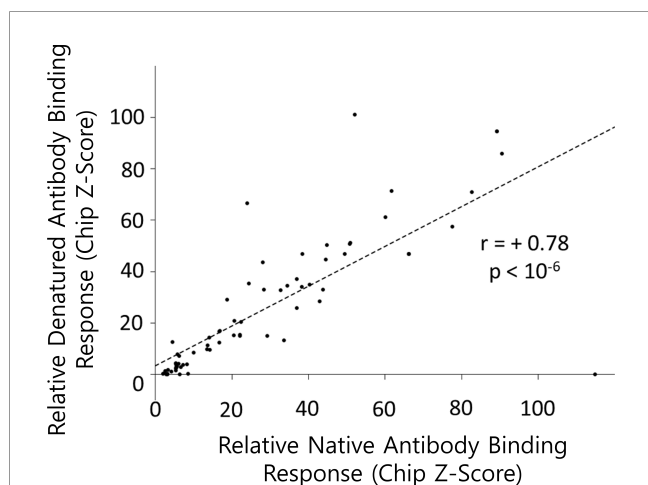


FIGURE 3

Cross-reactive epitopes are likely to be unstructured and continuous. Each point in the plot represents the fluorescence signal of one anti-virus antibody binding to the same human protein, measured under native conditions (x-axis) or denatured (8 M urea) conditions (y-axis). The strong correlation suggests that structured epitopes were not a general feature of any of the human proteins studied. There is one exception, namely, human NMD3 protein binding anti-SARS-Cov-2 nsp13 polyclonal, which suggests a signal for structured epitope(s). Data for these figures, and the proteins involved, are listed in Supplementary Table S1.

challenging due to the nature of the information involved. To highlight these challenges, we first make a distinction between our previous findings and what we report here. In earlier work, we demonstrated that proteins, rather than being classified in static structural terms, could be represented in thermodynamic terms, which reflect the relative stabilities of the different parts of the protein in the functional ensemble (19). Importantly, we showed that globular proteins appear to share common patterns of amino acid composition within this thermodynamic framework, providing the basis for a statistical free energy calculation. This allowed us to evaluate the compatibility of any sequence with the thermodynamic

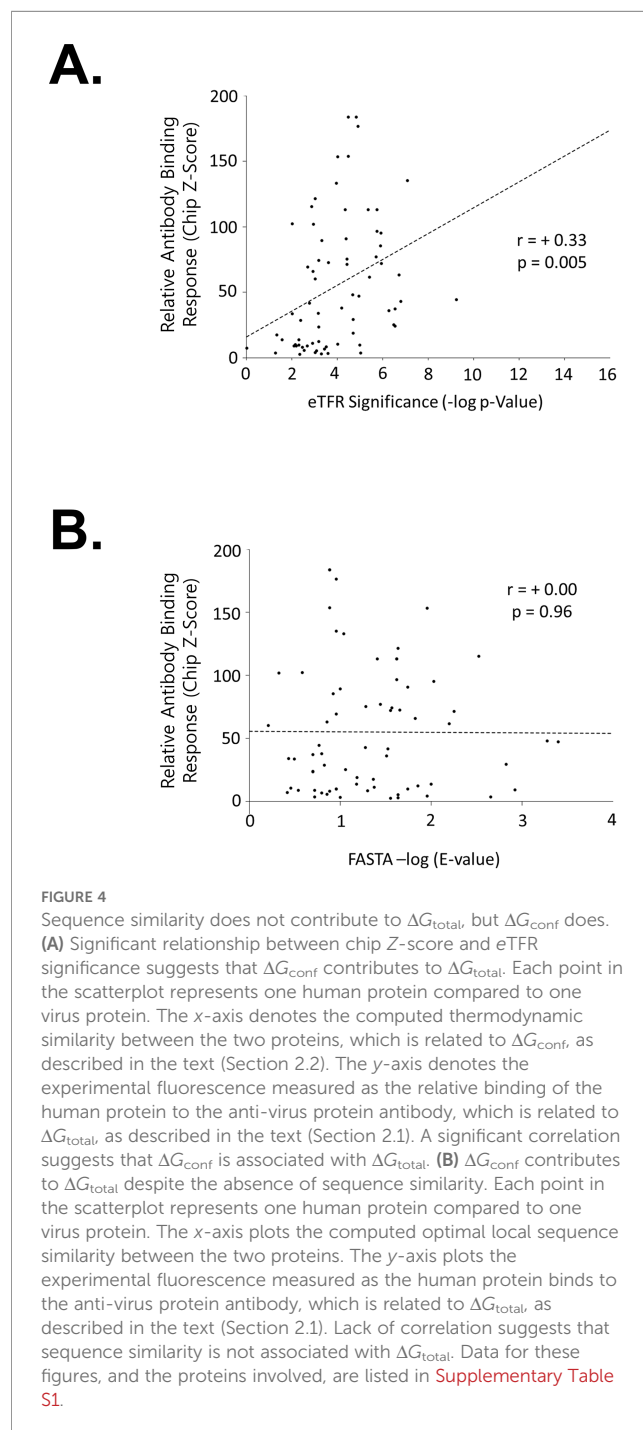


FIGURE 4

Sequence similarity does not contribute to ΔG_{total} , but ΔG_{conf} does.

(A) Significant relationship between chip Z-score and eTFR significance suggests that ΔG_{conf} contributes to ΔG_{total} . Each point in the scatterplot represents one human protein compared to one virus protein. The x-axis denotes the computed thermodynamic similarity between the two proteins, which is related to ΔG_{conf} , as described in the text (Section 2.2). The y-axis denotes the experimental fluorescence measured as the relative binding of the human protein to the anti-virus protein antibody, which is related to ΔG_{total} , as described in the text (Section 2.1). A significant correlation suggests that ΔG_{conf} is associated with ΔG_{total} . (B) ΔG_{conf} contributes to ΔG_{total} despite the absence of sequence similarity. Each point in the scatterplot represents one human protein compared to one virus protein. The x-axis plots the computed optimal local sequence similarity between the two proteins. The y-axis plots the experimental fluorescence measured as the human protein binds to the anti-virus protein antibody, which is related to ΔG_{total} , as described in the text (Section 2.1). Lack of correlation suggests that sequence similarity is not associated with ΔG_{total} . Data for these figures, and the proteins involved, are listed in Supplementary Table S1.

environments of any other protein structure, regardless of sequence or structural similarity. Furthermore, the fact that (1) the score of a sequence for a fold was correlated with the empirical stability of that sequence in that fold (19), and (2) sequence mutations that increased the score for a different fold caused the sequence to switch folds (if the score were high enough) (19), directly demonstrates that the previously reported TFR scores provide a measure of thermodynamic compatibility (*i.e.*, ΔG_{conf}) of the two ensembles—in effect validating the notion that a new type of EMM is real and, to at least some degree, predictable in the context of full-length protein sequences.

The work presented here, on the other hand, investigates whether that same EMM plays a role in the well-known phenomenon of antibody cross-reactivity. We reasoned that for binding reactions wherein the ΔG_{int} term for antibody binding (cf. Figure 1A) is expected to be weak (as would be the case for the binding of proteins to antibodies raised against an entirely different protein), tight binding, if it were to be found at all, would be more likely to rely on a more favorable ΔG_{conf} term. Thus, challenging the human proteome with antibodies raised against seven viral proteins, which have no human homologs, provides a means for both identifying cases where cross-reactivity occurs, and determining the thermodynamic compatibility of the human sequence for the viral protein ensemble.

Unfortunately, the total binding energy (*i.e.*, ΔG_{total}) for the cross-reacting antibodies consists of both ΔG_{conf} , which is computed, and the individual ΔG_{int} terms, which will likely vary on a case-by-case basis and may involve other mechanisms. Thus, although the absence of sequence and structural similarity precludes other types of reported mimicry mechanisms playing a role [see (34) and discussion below], we cannot conclude that all of the cross-reactivity is due to the ensemble similarity. However, the fact that a modest but statistically significant ($p = 0.005$) correlation is observed (Figure 4A), even in the absence of recognizable sequence identity (Figure 4B), strongly supports the hypothesis that conformational free energy (ΔG_{conf}) plays a role in antibody cross-reactivity across the proteome. The fact that the approach further reveals high thermodynamic scores for known cross-reactive epitopes (cf. Figure 5 and discussion below), even though none of the proteins involved (or any antibodies) were included in the algorithmic development, demonstrates this point and highlights what we can conclude from these results and what we cannot. Succinctly put, the approach identifies what other protein sequences can sample the conformational ensemble of the antigenic protein, but it does not propose a specific mechanism for antibody cross-reactivity for each case. Instead, the results demonstrate that the full-length protein ensemble compatibility (*i.e.*, mimicry) reported previously (19) and embodied in the ΔG_{conf} term in Figure 1A appears to contribute meaningfully to ΔG_{total} , when viewed across the proteome, being especially evident when probing specifically for cross-reactivity. Perhaps the most important result is that this approach can provide investigators with the ability to determine which protein sequences have a significant compatibility with the ensemble of another protein (and thus can potentially cross-react with antibodies raised

against that protein). This capability could be especially useful in detecting thermodynamic compatibility where no recognizable sequence or structural similarity exists.

Given the correlation between binding and conformational contribution reported here, and the caveats noted above, the questions, nonetheless, rightly turn to identifying the range of biophysical mechanisms for the unexpected cross-reactivity, and the location of the epitope(s). Is there a correlation between the epitope and the calculated high-scoring residues? Should there be? We note that unlike comparing experimental and calculated energies arising from the binding interface (35, 36), where the contribution of individual positions is explicitly computed, there is no obligatory position-specific correspondence; sequence determinants that result in high agreement of one sequence for an ensemble need not correspond to the amino acids that serve as the epitope for the actual binding part of the reaction (cf. Figure 1A). This point is made clear by considering that amino acids in stable, internal environments will contribute positively to high TFR scores, but epitopes are usually located primarily in less stable, surface-exposed sites (1, 15, 16). Consistent with this reasoning, it is noteworthy that our computational methods consider the full-length sequence of the antigen and cross-reacting proteins, and make no predictions about where the binding interfaces between antibody and antigen might be, as they are in principle unrelated. The situation is complicated further by the use of polyclonal antibodies, as, for example, the following two alternatives cannot be differentiated given an empirical binding affinity: (a) That many antibodies bind weakly to many epitopes on a given human protein, and the experimental signal is an integration of this heterogeneous response by the protein averaged over the ensemble, or (b) that few antibodies bind strongly to a single epitope on a given human protein, leading to the same experimental signal as (a).

The observed correlation between conformational equilibrium and relative antibody binding is also formally consistent with alternative models. For example, two or more distinct antibody populations within the polyclonal response could each bind unrelated epitopes on the virus and human proteins. In such a scenario, the association with ensemble similarity would still hold, but there would be no correspondence between the epitope and the scoring. These possibilities notwithstanding, identifying the location(s) of epitopes would help identify the range of scenarios and would further our understanding of this phenomenon. Distinguishing between these several possibilities positions techniques like epitope-mapping as the most straightforward way to address these questions, and studies to resolve these issues on a case-by-case basis are currently underway.

Given the challenges noted above, several lines of evidence support the notion that the computational methods may, nonetheless, provide clues about the nature and location of the epitopes in some cases. For example, fluorescence Z-scores obtained under denaturing vs. native conditions (Figure 3) suggest that the binding affinity is essentially independent of whether the human protein binder is folded or unfolded, implying that most epitopes are likely to be continuous. We further reasoned that modifying the

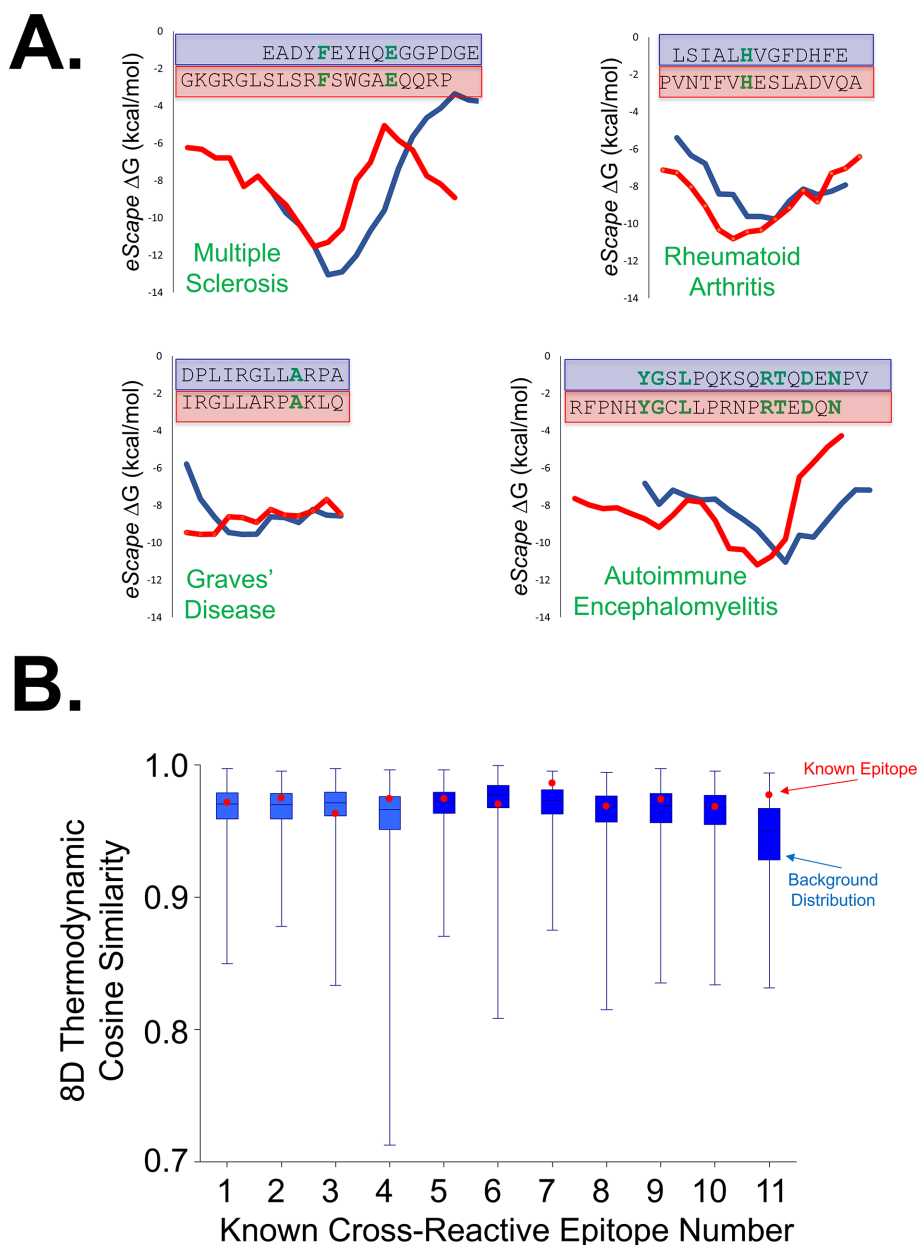


FIGURE 5

Known examples of medically relevant cross-reactive epitopes anecdotally share thermodynamic similarity in the absence of sequence similarity. **(A)** Four cases of known cross-reactive epitopes implicated in auto-immune disease molecular mimicry. For each example, an intriguing correspondence between the predicted eScape local stabilities of the published aligned epitopes are observed. Amino acid identities are highlighted in bold green font, emphasizing the lack of homology between the epitopes as the identities range from only 8% to 33%. Multiple sclerosis (38) (upper left): virus protein Epstein–Barr Virus Nuclear Antigen-1, residues 411–426 (blue), and human Myelin Basic Protein, residues 205–224 (red); Graves' disease (37) (lower left): human thyroid peroxidase, residues 536–547 (blue), and human thyroid peroxidase, residues 539–550 (red); auto-immune encephalomyelitis (11) (lower right): human Myelin Basic Protein, residues 68–95 (blue), and Chlamydia cysteine-rich outer membrane protein peptide (CRP, red); rheumatoid arthritis (11) (upper right): Glucose-6-phosphate isomerase, residues 282–294 (blue), and Bovine RNase, residues 42–46 (red). **(B)** FVC program indicates that known cross-reactive epitopes exhibit higher-than-average thermodynamic similarity. Box-and-whiskers plots were made using cosine similarity output from FVC, as described in Section 4.6. Error bars encompass the maximum and minimum cosine similarity scores of the background, the blue box encloses one-half of the background data, and the horizontal line within denotes the average. Cosine similarities for known epitopes (red points) are consistently higher than expected by the background distributions for the full-length proteins from which they are derived (blue). The significance of 9 out of 11 epitopes scoring higher than average is estimated to be $p < 0.05$. Numbers on the x-axis correspond to the following epitopes: (1) Multiple Sclerosis PLP and MHV proteins (11), (2) Multiple Sclerosis PLP and *H. influenzae* proteins (11), (3) Multiple Sclerosis MHV and *H. influenzae* proteins (11), (4) Encephalomyelitis MBP and *C. pneumoniae* proteins (11), (5) Myocarditis Myosin and CRP proteins (11), (6) Irritable Bowel Syndrome Mouse Hsp60 and Mycobacterium GroEL (11), (7) Rheumatoid Arthritis GPI and RNase (11), (8) Multiple Sclerosis Human MBP and Epstein–Barr Virus protein (11), (9) Liver autoimmunity Mouse FAH and MHV protein (46), (10) Multi-inflammatory syndrome in children Human SNX8 and SARS-CoV-2 orf9 (47), and (11) Multiple Sclerosis EBNA-1 and Human MBP (38). Numbers 4, 7, and 11 are also explicitly shown in Figure (A).

full-length thermodynamic matching to a more localized thermodynamic matching (e.g., global or local sequence alignments in *FASTA* or *BLAST*) may provide candidates for continuous epitopes. Support for this hypothesis was found by anecdotal inspection of *eEscape* ΔG values for several published sequence-dissimilar cross-reactive auto-immune epitopes (Figure 5A), which suggested that the local thermodynamic stabilities over 10- to 15-residue (i.e., epitope-sized) sequence fragments could exhibit similarity to the naked eye (although this similarity by itself does not rise to a level of statistical significance).

Armed with these observations, we developed an algorithm to extract optimal local matches from two *eEscape* thermodynamic profiles, using only sequence information as input. The resultant *Fragment Vector Comparison* (FVC) algorithm calculates the $8 \times L$ -dimensional cosine similarity between all pairwise sets of thermodynamic profiles of a specified window size, L , between two proteins (Section 4.6). Thus, a perfect correlation between the thermodynamics of two epitopes would yield an FVC score of 1.0, with lesser values indicating poorer matches. The resulting cosine similarities are then sorted, and the best local matches are retrieved.

(The FVC package, including *eEscape*, is released along with this manuscript, and is freely available at the following site: <https://github.com/jBeale23/FragmentVectorComparison>).

Of note, when this approach was implemented, it was discovered that a majority of published cross-reactive epitopes from medical literature exhibited higher-than-average thermodynamic similarities (Figure 5B), a statistically significant result ($p < 0.05$). Furthermore, weighting each ϵ TFR significance in Figure 4A by the fraction of the sequence that scores greater than 0.95 on FVC when matched with the viral protein measurably improved the correlation (Figure 6), as well as its significance ($p = 2 \times 10^{-3}$). This suggests that optimal local (epitope-sized) thermodynamic similarities, which may be subsumed within the optimal full-length similarities, may play an important role in the relationship between ΔG_{conf} and ΔG_{total} , perhaps suggestive of a connection between ΔG_{conf} and ΔG_{int} . Of course, the ultimate validity of this approach awaits additional experiments to determine whether FVC can actually predict the locations of the continuous cross-reactive epitopes responsible for Figures 3, 4A. To support such experiments, a list of candidate FVC matches for each protein pair in Figure 4 has been tabulated (Supplementary Table S3). It would truly be noteworthy if the determinants of the compatibility of a sequence for an ensemble overlapped with the epitope, but this is an open question, and one of the many possibilities requiring future studies.

3.2 Thermodynamic molecular mimicry: source of cross-reactivity?

One common property shared among all documented cross-reactive epitopes in Figure 5 is that these medically relevant sequence fragments have undetectable pairwise identity with each other. How could one antibody bind to two epitopes in the absence of sequence similarity? Does a single subpopulation of the polyclonal antibody bind both the virus and human protein, or are there differences in the subpopulations that bind each protein? To explain this conundrum, the well-known phenomenon of

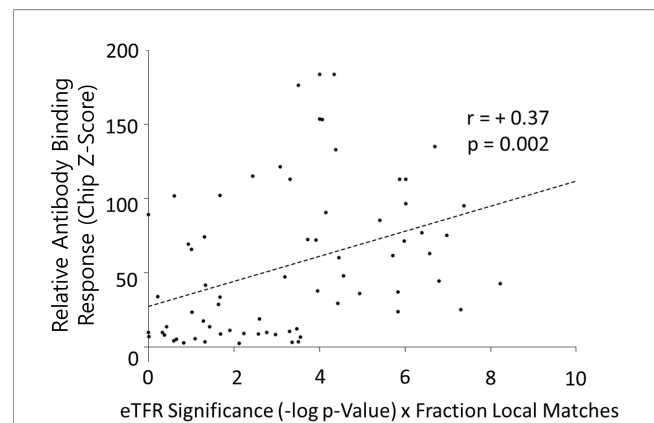
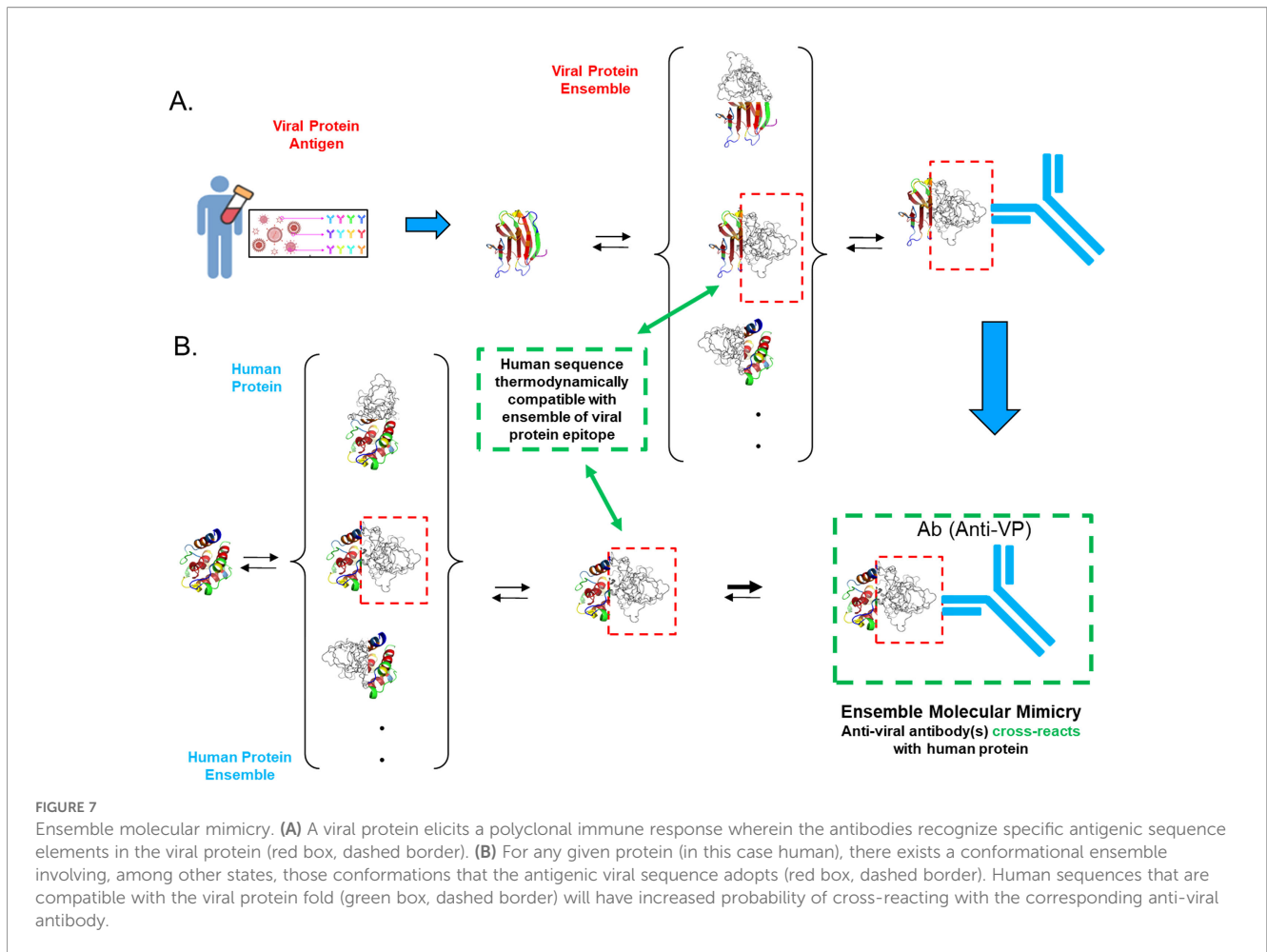


FIGURE 6

Simple weighting of ϵ TFR significance by fraction of optimal local similarity improves correlation between ΔG_{total} and ΔG_{conf} . Each point in the scatterplot represents one human protein compared to one virus protein. The x-axis plots the computed full-length thermodynamic similarity between the two proteins, which is related to ΔG_{conf} , as described in the text (Section 2.2), weighted by the fraction of FVC matches greater than 0.99 contained within the full-length similarity. The y-axis plots the experimental fluorescence measured as the human protein binds to the anti-virus protein antibody, which is related to ΔG_{total} , as described in the text (Section 2.1). A significant correlation suggests that ΔG_{conf} is associated with ΔG_{total} . Because this correlation coefficient is improved over Figure 4A, this suggests that epitope-sized regions of strongest thermodynamic similarity are important contributors to the correlation (compare with Figure 4A).

molecular mimicry (34) has been proposed. Many examples of immunologic molecular mimicry have been documented (11, 37, 38), with Rojas et al. succinctly listing four distinct types spanning a continuum (34), ranging from complete structural similarity and sequence identity (Type 1) to structural similarity in the absence of sequence identity (Type 4). The type of molecular mimicry observed here is difficult to classify within this framework, because the common thermodynamic property is neither sequence nor structural in nature. These results suggest that the previously developed TFR scores in conjunction with the newly developed FVC matches provide a useful means of investigating the relationship between a more local ΔG_{conf} and the ΔG_{int} associated with the cross-reactive epitopes. Progress on these questions await epitope mapping studies on numerous systems.

The most general physical interpretation of EMM is that antibodies do not necessarily bind to a unique state, but instead to a subset of thermodynamically similar conformational states within an ensemble (Figure 1A). Pairs of sequences can, to different extents, populate states that are thermodynamically similar to each other; thus, our hypothesis is that in such cases, antibodies raised to only one sequence can nonetheless bind both, as shown in the schematic model of EMM depicted in Figure 7. Importantly, even in the absence of detectable structural similarity in the aligned regions, chemical groups such as hydrophobic side chains or hydrogen bond donors/acceptors would be geometrically placed to permit the same constellation of antibody(ies) to bind both antigens (37). This is the same principle that allows us to evaluate the preference of an amino acid sequence to adopt one conformation over another simply based on the scoring of amino acids in the corresponding



environments of the new conformation without explicit consideration of structural context or interactions (19).

Finally, with regard to medical relevance, the phenomenon of EMM identified here may play a role in unexplained aspects of auto-immunity, such as promiscuous antibodies (34, 39, 40) or the breaking of immune tolerance in the absence of sequence similarity between self and non-self (41, 42). However, almost all extant cases of mimicry have been discovered individually in resource-intensive experiments (11, 37, 38, 43), motivating the need for cross-reactivity-prediction tools that are less reliant on amino acid sequence identity. It is our hope that the observations reported here, as well as the tools accompanying this manuscript, will provide a new avenue for addressing this need and identifying possible molecular mimicry within a proteome.

4 Methods

4.1 Human proteome-on-a-chip experiments

Rabbit polyclonal IgG primary antibodies raised against purified complete SARS-CoV-2 proteins were purchased off-the-shelf from commercial sources, as described below. Primary

antibodies listed below, and in [Supplementary Table S1](#), were sent to CDI Laboratories (Baltimore, MD and Mayaguez, PR) and passed over separate *HuProt* chips containing essentially complete representatives of each protein in the human proteome (*HuProt v4.0*, 21,215 distinct amino acid sequences). Binding of antibody to individual proteins was assessed by fluorescence (Alexa 555) of secondary anti-rabbit IgG according to the established procedure of CDI Laboratories. Binding was quantified for each human protein with a Z-score, representing the number of standard deviations above or below the average fluorescence signal over the entire chip. Objective analysis reports were received from CDI Laboratories listing the top 10 highest binding human proteins on each chip ([Supplementary Table S1](#)).

With the exception of the anti-SARS-CoV2-orf10 antibody, two chips were run for each antibody: a “native” chip, where the proteins on the chip were never exposed to denaturant, and a “denatured” chip, where the proteins on the chip were exposed to 8 M urea before incubation with the primary antibody. Native and denatured chips were compared to assess the possibility of structured epitopes, and it was found that there was generally a high correlation between native and denatured binding measurements ([Figure 3](#)). This suggested the lack of widespread structured epitopes, indicating largely redundant information between native and denatured ([Supplementary Table S1](#)). A

technical limitation resulted in the unavailability of denatured binding data for the orf10 experiment. Because of the high correlation between native and denatured observed for the other samples, orf10 native state values were substituted where necessary, without loss of information. However, it is emphasized that this substitution introduces a small but non-negligible source of uncertainty.

Although *HuProt* assays are reported to be reasonably reproducible, we explicitly tested the reproducibility of a separate antibody sample on chips from different batches, 1 year apart. This was polyclonal anti-human CD53 (Bioss Antibodies bs-13625R). The results indicated high correlation between the binding affinities of the two chips under native conditions (Supplementary Figure S2). Therefore, the measurements and conclusions are believed to be robust.

4.2 Polyclonal antibodies used

All primary antibodies used here were polyclonal Rabbit IgG from commercial sources, at stock concentrations of 0.1–1.0 mg/mL and generally raised against the full-length protein. Manufacturers included anti-SARS-CoV-2-orf9 (Novus Biologicals NBP3-00510), anti-SARS-CoV-2-nsp13 (Life Technologies/Invitrogen PA5120711), anti-SARS-CoV-2-nsp16 (Life Technologies/Invitrogen PA5120701), anti-SARS-CoV-2-Spike (Life Technologies/Invitrogen PA5116916), anti-SARS-CoV-2-orf6 (Life Technologies/Invitrogen PA5120715), anti-SARS-CoV-2-orf8 (Antibodies-Online ABIN7383791), and anti-SARS-CoV-2-orf10 (Antibodies-Online ABIN6952939). These antibodies were chosen based on a preliminary *eTFR* scan of the SARS-CoV-2 proteome against the amino acid sequences of the Protein Data Bank (2) described below and in Supplementary Figure S1A.

4.3 Computational methods: *eTFR* and FASTA alignments

The *eTFR* thermodynamic alignment algorithm was executed as described in previous publications, without modification. In detail, the process was as follows: One full-length virus protein sequence was fed into the *eScape* algorithm as input (included within the *FVC* code released with this manuscript, web app at best.bio.jhu.edu/eScape) (22, 23). The output, two sets of four-dimensional vectors $\{\Delta G, \Delta H_{ap}, \Delta H_{pol}, T\Delta S_{conf}\}$ corresponding to the native state and denatured state thermodynamic descriptors for each residue of the protein, were mapped to the eight native state and eight denatured state thermodynamic environments, as previously described (44). This information constitutes the complete query thermodynamic profile, and the process was repeated for each virus protein (i.e., the antigen against which each polyclonal was raised).

The workflow proceeded as individual pairwise gapless alignments of each query with the 10 amino acid sequences prioritized by CDI Labs, comparing the shorter of the pair in all possible registers with the longer sequence (32). Each register was

scored by consulting a thermodynamic substitution matrix for each aligned amino acid–environment pair and summing the total over the entire alignment (32). The total score was converted into a significance (*p*-value) according to the null model and length-dependent equation, as described previously (32). Separate scores and significances were computed for the native state and the denatured state, and the summed negative logs of the native and denatured significances were taken as the total score for that register (19). Finally, the maximum over all registers was taken as the optimal alignment and significance for that thermodynamic query–amino acid sequence pair. (These values are plotted on the *x*-axis of Figure 4 and are listed in Supplementary Table S1.) Thus, for the seven proteins in the SARS-CoV-2 proteome considered, the *eTFR* procedure resulted in two outputs (1): a pairwise alignment between the virus amino acid sequence and the human sequence (Supplementary Table S2), and (2) a significance estimation for that alignment (Supplementary Table S1).

Amino acid sequence alignments were performed using the FASTA36 package (33) and BLOSUM62 substitution matrix (45), with all default settings. The most significant *E*-value was always taken as the single optimal result between two pairs of proteins; these values are plotted on the *x*-axis of Figure 4B.

4.4 Preliminary *eTFR* scan of the Protein Data Bank

Thermodynamic profiles of the Wuhan-Hu-1 SARS-CoV-2 proteome were computed using *eScape* from the 28 amino acid sequences translated from the NCBI Genomes accession NC_045512.2 (ncbi.nlm.nih.gov). Each of the 28 proteins in the SARS-CoV-2 proteome was used separately as query to exhaustively match all 594,420 amino acid sequences in the Protein Data Bank [download version date 06/14/21, rcsb.org (2)]. In detail, the process was as follows: One virus protein was fed into the *eScape* algorithm as input (included with the *FVC* code released with this manuscript, or web app at best.bio.jhu.edu/eScape). The output, two sets of four-dimensional vectors $\{\Delta G, \Delta H_{ap}, \Delta H_{pol}, T\Delta S_{conf}\}$ corresponding to the native state and denatured state thermodynamic descriptors for each residue of the protein, were mapped to the eight native state and eight denatured state thermodynamic environments, as previously described (44). This information constitutes the complete query thermodynamic profile, and the process was repeated for each virus protein.

The query was then used to exhaustively search a database of all amino acid sequences of known structure, as contained in the Protein Data Bank. The search proceeded as individual pairwise gapless alignments of the query with one amino acid sequence, comparing the shorter of the pair in all possible registers with the longer one (32). Each register was scored by consulting a thermodynamic substitution matrix for each aligned amino acid–environment pair and summing the total over the entire alignment (32). The total score was converted into a significance (*p*-value) according to the null model and length-dependent equation, as described previously (32). Separate scores and significances were

computed for the native state and the denatured state, and the summed negative logs of the native and denatured-state significances were taken as the total score for that register (19). Finally, the maximum over all registers was taken as the optimal alignment and significance for that thermodynamic query–amino acid sequence pair. These significance values are plotted on the *y*-axis of **Supplementary Figure S1A**.

The highest-scoring human proteins against each viral protein thermodynamic profile were tabulated by manual inspection (these proteins are indicated by red bars in **Supplementary Figure S1A**). Seven virus–human matches were subjectively chosen on the basis of commercial availability of polyclonal antibody raised against the full-length viral protein and commercial availability of purified full-length human protein for Western blot analysis, described in the following. These seven matches chosen for testing are circled and numbered in **Supplementary Figure S1A**. The seven polyclonal antibodies comprised the panel of reagents used for the proteome-on-a-chip experiments shown in **Figure 4** and **Supplementary Table S1**.

4.5 Western blot experiments confirming cross-reactivity

Purified protein was obtained from commercial sources, as described below, at concentrations of approximately 0.1–1.0 mg/mL. Generally, 10–20 μ L of this protein was loaded into 12%–20% SDS-polyacrylamide Laemmli reducing slab gels to completion at ambient temperature (1 h, 200 V). Electrophoretic transfer to nitrocellulose membrane from the gel was performed under Tris–glycine–methanol pH 8.3 buffer conditions to completion at 4 °C on ice (2 h, 400 A constant current). The membrane was blocked with Blocking Buffer (137 mM NaCl, 20 mM Tris, 0.1% Tween-20, and 5% dry milk, pH 7.6) for 24–48 h at 4°C. Then, the membrane was incubated with the primary antibody from the list above for at least 1 h at room temperature, washed 3 \times with TBST (137 mM NaCl, 20 mM Tris, and 0.1% Tween-20, pH 7.6), incubated at least 30 min at room temperature with anti-rabbit IgG secondary antibody linked to horseradish peroxidase, and washed 3 \times with TBST. Chemoluminescence was generated with no more than 24-h-old SignalFire reagents (Cell Signaling Technology, Danvers, MA), according to the manufacturer's directions. Imaging was performed with CL-XPosure film (ThermoFisher Scientific, Baltimore, MD) at a short exposure of 30 s, and again at a long exposure of 2 h. Imaging was also performed in separate experiments with fluorescence scanning after incubation with DyLite 800 PEG 4 \times conjugate secondary antibody, according to the manufacturer's directions (Cell Signaling Technology, Danvers, MA). Images are displayed in **Supplementary Figure S1B**. All cross-reactivity results shown in **Supplementary Figure S1B** were reproducible at least three times.

Primary antibody solutions were generally diluted 1:1,000 in Blocking Buffer, assuming an antibody concentration of approximately 0.1 mg/mL. Secondary antibodies were generally diluted 1:1,000 in Blocking Buffer, assuming an antibody

concentration of approximately 0.1 mg/mL. Primary and secondary antibodies were reused for periods of up to 2 weeks, with storage at 4 °C and added 0.1% final concentration sodium azide during the intervals.

All proteins used were commercially available at stock concentrations of 0.1–1.0 mg/mL and were full length (or the longest length available, subject to coverage of the region of computational prediction). We purchased expected antigens as well as predicted cross-reactive proteins to test the efficacy of the polyclonal antibodies: all polyclonals were judged effective by Western blot analysis (data not shown). Manufacturers from which antigens and proteins were obtained were as follows: SARS-CoV-2-orf9 (R&D Systems/Bio-technie 11033-CV-100), SARS-CoV-2-nsp13 (ProSci 10-427), SARS-CoV-2-nsp16 (ProSci 20-243, N-term uncleaved GST fusion), SARS-CoV-2-Spike (Life Technologies/Invitrogen RP87671, aa16-1213), SARS-CoV-2-orf6 (ProSci 20-193, N-term uncleaved MBP), SARS-CoV-2-orf8 (Life Technologies/Invitrogen RP87666, aa16-121), SARS-CoV-2-orf10 (ProSci 20-189, N-term uncleaved MBP), human-IL11 (Life Technologies/Gibco PHC0115), human-IGLL5 (MyBioSource MBS1029820), human-IL6 (Life Technologies/Invitrogen RP8619), human-SCN4B (OriGene TP323951), human-KDR/VEGFR2 (OriGene TP710248, aa20-764), human-CXCR4 (Abnova H00007852-G01), and human-CD53 (Abnova H00000963-G01).

4.6 Fragment vector comparison

The python implementation of the *FVC* program was used with all default settings, except that similarity cutoff was set to zero to record all local pairwise comparisons of fragment length *L* between two amino acid sequences. *L* was set to values between 9 and 25 residues long, depending on the length of the published alignment between one pair of cross-reactive epitopes. For each pair of cross-reactive epitopes, all cosine similarity scores computed for length *L* were used as the background distribution to make the box-and-whiskers plots in **Figure 5B**. The single score corresponding to the fragment of length *L* in the register of the published alignment (i.e., only the documented epitope) was used to plot the red dots in **Figure 5B**. The source code of *FVC*, which includes the *eScape* algorithm, is released as part of this manuscript, including a C implementation of the cosine similarity calculation that allows for higher throughput on supported device architectures. The code is freely and publicly available at the following site: <https://github.com/jBeale23/FragmentVectorComparison>.

For **Figure 6**, *FVC* was run separately on each of the 70 virus antigen–human protein pairs with window size set to 20 and similarity cutoff set to zero. Then, for each virus–human pair, only *FVC* scores corresponding to the overlapping windows of the register found in the full-length thermodynamic *eTFR* alignment were retained. Of these *N* scores, a weight *f* was computed from $f = M/N$, where *M* was the number of scores with cosine similarity greater than 0.95. The *x*-axis of **Figure 6** was thus

computed from $S \times f$, where S is the e TFR significance ($-\log p$ -value) from Figure 4A. Values for f , M , and N obtained are listed in Supplementary Table S1. This procedure merely weighted the full-length significance by the fraction of “very similar” local thermodynamic matches within the full-length alignment.

To facilitate future experiments, a list of the three highest matches of 20 residues found by FVC in the context of the full-length e TFR alignment between each pair of proteins in Figure 4 is given in Supplementary Table S3. Note that these data may also be used to independently check the proper working of a locally installed version of FVC.

Data availability statement

The datasets presented in this study can be found in the Supplementary Material.

Author contributions

JW: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. VH: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing. JB: Investigation, Methodology, Software, Validation, Writing – review & editing. GF: Data curation, Formal analysis, Methodology, Software, Writing – review & editing. AM: Conceptualization, Data curation, Formal analysis, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. Funding was provided by NIH: R01-GM63747 and R01-GM121567. JB was supported by NIH T32-GM007231 training grant and GF was supported by NIH T32-GM008403 training grant.

References

- Madsen AV, Mejias-Gomez O, Pedersen LE, Preben Morth J, Kristensen P, Jenkins TP, et al. Structural trends in antibody-antigen binding interfaces: a computational analysis of 1833 experimentally determined 3D structures. *Comput Struct Biotechnol J*. (2024) 23:199–211. doi: 10.1016/j.csbj.2023.11.056
- Burley SK, Berman HM, Duarte JM, Feng Z, Flatt JW, Hudson BP, et al. Protein data bank: A comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students. *Biomolecules*. (2022) 12. doi: 10.3390/biom12101425
- Childers MC, Daggett V. Molecular dynamics methods for antibody design. *Methods Mol Biol*. (2023) 2552:109–24. doi: 10.1007/978-1-0716-2609-2_5
- Jeliazkov JR, Frick R, Zhou J, Gray JJ. Robustification of rosettaAntibody and rosetta snugDock. *PLoS One*. (2021) 16:e0234282. doi: 10.1371/journal.pone.0234282
- Kenlay H, Dreyer FA, Kovaltsuk A, Miketa D, Pires D, Deane CM. Large scale paired antibody language models. *PLoS Comput Biol*. (2024) 20:e1012646. doi: 10.1371/journal.pcbi.1012646
- Joubbi S, Micheli A, Milazzo P, Maccari G, Ciano G, Cardamone D, et al. Antibody design using deep learning: from sequence and structure design to affinity maturation. *Brief Bioinform*. (2024) 25. doi: 10.1093/bib/bbae307
- Bennett NR, Watson JL, Ragotte RJ, Borst AJ, See DL, Weidle C, et al. Atomically accurate de novo design of antibodies with RFdiffusion. *Nature*. (2025) 649:183–93. doi: 10.1101/2024.03.14.585103
- Hummer AM, Schneider C, Chinery L, Deane CM. Investigating the volume and diversity of data needed for generalizable antibody-antigen DeltaDeltaG prediction. *Nat Comput Sci*. (2025) 5:635–47. doi: 10.1038/s43588-025-00823-8
- Sachs DH, Schechter AN, Eastlake A, Anfinsen CB. An immunologic approach to the conformational equilibria of polypeptides. *Proc Natl Acad Sci U S A*. (1972) 69:3790–4. doi: 10.1073/pnas.69.12.3790
- Furie B, Schechter AN, Sachs DH, Anfinsen CB. An immunological approach to the conformational equilibrium of staphylococcal nuclease. *J Mol Biol*. (1975) 92:497–506. doi: 10.1016/0022-2836(75)90305-8

Acknowledgments

We thank Sara Rahman, Aria Heidsek and Buffy Wrabl for assistance and discussions.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fimmu.2026.1749369/full#supplementary-material>

11. Kohm AP, Fuller KG, Miller SD. Mimicking the way to autoimmunity: an evolving theory of sequence and structural homology. *Trends Microbiol.* (2003) 11:101–5. doi: 10.1016/S0966-842X(03)00006-4
12. Wang EY, Mao T, Klein J, Dai Y, Huck JD, Jaycox JR, et al. Diverse functional autoantibodies in patients with COVID-19. *Nature.* (2021) 595:283–8. doi: 10.1038/s41586-021-03631-y
13. Wang EY, Dai Y, Rosen CE, Schmitt MM, Dong MX, Ferre EMN, et al. High-throughput identification of autoantibodies that target the human exoproteome. *Cell Rep Methods.* (2022) 2. doi: 10.1016/j.crmeth.2022.100172
14. Song Y, Li J, Wu Y. Evolving understanding of autoimmune mechanisms and new therapeutic strategies of autoimmune disorders. *Signal Transduct Target Ther.* (2024) 9:263. doi: 10.1038/s41392-024-01952-8
15. Laver WG, Air GM, Webster RG, Smith-Gill SJ. Epitopes on protein antigens: misconceptions and realities. *Cell.* (1990) 61:553–6. doi: 10.1016/0092-8674(90)90464-P
16. Benjamin DC, Berzofsky JA, East IJ, Gurd FR, Hannum C, Leach SJ, et al. The antigenic structure of proteins: a reappraisal. *Annu Rev Immunol.* (1984) 2:67–101. doi: 10.1146/annurev.iy.02.040184.000435
17. Tsuboyama K, Dauparas J, Chen J, Laine E, Mohseni Behbahani Y, Weinstein JJ, et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature.* (2023) 620:434–44. doi: 10.1038/s41586-023-06328-6
18. Escobedo A, Voigt G, Faure AJ, Lehner B. Genetics, energetics, and allostery in proteins with randomized cores and surfaces. *Science.* (2025) 389:eadq3948. doi: 10.1126/science.adq3948
19. Voortman-Sheetz K, Wrabl JO, Hilser VJ. Impact of local unfolding fluctuations on the evolution of regional sequence preferences in proteins. *Protein Sci.* (2025) 34:e70015. doi: 10.1002/pro.70015
20. Millard CEF, Wrabl JO, Brantley SJ, Grasso E, Schmitz A, White JT, et al. The ensemble basis of allostery and function: insights from models of local unfolding. *J Mol Biol.* (2025) 437:169287. doi: 10.1016/j.jmb.2025.169287
21. Hilser VJ, Wrabl JO, Millard CEF, Schmitz A, Brantley SJ, Pearce M, et al. Statistical thermodynamics of the protein ensemble: mediating function and evolution. *Annu Rev Biophys.* (2025) 54:227–47. doi: 10.1146/annurev-biophys-061824-104900
22. Gu J, Hilser VJ. Predicting the energetics of conformational fluctuations in proteins from sequence: a strategy for profiling the proteome. *Structure.* (2008) 16:1627–37. doi: 10.1016/j.str.2008.08.016
23. Gu J, Hilser VJ. Sequence-based analysis of protein energy landscapes reveals nonuniform thermal adaptation within the proteome. *Mol Biol Evol.* (2009) 26:2217–27. doi: 10.1093/molbev/msp140
24. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem.* (1997) 48:545–600. doi: 10.1146/annurev.physchem.48.1.545
25. Hilser VJ, Garcia-Moreno EB, Oas TG, Kapp G, Whitten ST. A statistical thermodynamic model of the protein ensemble. *Chem Rev.* (2006) 106:1545–58. doi: 10.1021/cr040423+
26. Wand AJ. Deep mining of the protein energy landscape. *Struct Dyn.* (2023) 10:020901. doi: 10.1063/4.0000180
27. Wrabl JO, Larson SA, Hilser VJ. Thermodynamic environments in proteins: fundamental determinants of fold specificity. *Protein Sci.* (2002) 11:1945–57. doi: 10.1110/ps.0203202
28. Larson SA, Hilser VJ. Analysis of the “thermodynamic information content” of a Homo sapiens structural database reveals hierarchical thermodynamic organization. *Protein Sci.* (2004) 13:1787–801. doi: 10.1110/ps.04706204
29. Wang S, Gu J, Larson SA, Whitten ST, Hilser VJ. Denatured-state energy landscapes of a protein structural database reveal the energetic determinants of a framework model for folding. *J Mol Biol.* (2008) 381:1184–201. doi: 10.1016/j.jmb.2008.06.046
30. Vertrees J, Wrabl JO, Hilser VJ. Energetic profiling of protein folds. *Methods Enzymol.* (2009) 455:299–327. doi: 10.1016/S0076-6879(08)04211-0
31. Wrabl JO, Hilser VJ. Investigating homology between proteins using energetic profiles. *PLoS Comput Biol.* (2010) 6:e1000722. doi: 10.1371/journal.pcbi.1000722
32. Hoffmann J, Wrabl JO, Hilser VJ. The role of negative selection in protein evolution revealed through the energetics of the native state ensemble. *Proteins.* (2016) 84:435–47. doi: 10.1002/prot.24989
33. Pearson WR. Finding protein and nucleotide similarities with FASTA. *Curr Protoc Bioinf.* (2016) 53:3 9 1–3 9 25.
34. Rojas M, Herran M, Ramirez-Santana C, Leung PSC, Anaya JM, Ridgway WM, et al. Molecular mimicry and autoimmunity in the time of COVID-19. *J Autoimmun.* (2023) 139:103070. doi: 10.1016/j.jaut.2023.103070
35. Lee J, Seok C, Ham S, Chong SH. Atomic-level thermodynamics analysis of the binding free energy of SARS-CoV-2 neutralizing antibodies. *Proteins.* (2023) 91:694–704. doi: 10.1002/prot.26458
36. Clark AJ, Gindin T, Zhang B, Wang L, Abel R, Murrett CS, et al. Free energy perturbation calculation of relative binding free energy between broadly neutralizing antibodies and the gp120 glycoprotein of HIV-1. *J Mol Biol.* (2017) 429:930–47. doi: 10.1016/j.jmb.2016.11.021
37. Quaratino S, Thorpe CJ, Travers PJ, Londei M. Similar antigenic surfaces, rather than sequence homology, dictate T-cell epitope molecular mimicry. *Proc Natl Acad Sci U S A.* (1995) 92:10398–402. doi: 10.1073/pnas.92.22.10398
38. Jog NR, McClain MT, Heinlen LD, Gross T, Towner R, Guthridge JM, et al. Epstein Barr virus nuclear antigen 1 (EBNA-1) peptides recognized by adult multiple sclerosis patient sera induce neurologic symptoms in a murine model. *J Autoimmun.* (2020) 106:102332. doi: 10.1016/j.jaut.2019.102332
39. James LC, Roversi P, Tawfik DS. Antibody multispecificity mediated by conformational diversity. *Science.* (2003) 299:1362–7. doi: 10.1126/science.1079731
40. Van Regenmortel MH. Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition. *J Mol Recognit.* (2014) 27:627–39. doi: 10.1002/jmr.2394
41. Koncz B, Balogh GM, Manczinger M. A journey to your self: the vague definition of immune self and its practical implications. *Proc Natl Acad Sci U S A.* (2024) 121:e2309674121. doi: 10.1073/pnas.2309674121
42. Medzhitov R, Iwasaki A. Exploring new perspectives in immunology. *Cell.* (2024) 187:2079–94. doi: 10.1016/j.cell.2024.03.038
43. Robinson WH, Steinman L. Epstein-Barr virus and multiple sclerosis. *Science.* (2022) 375:264–5. doi: 10.1126/science.abm7930
44. Chin AF, Wrabl JO, Hilser VJ. A thermodynamic atlas of proteomes reveals energetic innovation across the tree of life. *Mol Biol Evol.* (2022) 39. doi: 10.1093/molbev/msac010
45. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A.* (1992) 89:10915–9. doi: 10.1073/pnas.89.22.10915
46. Mathieu PA, Gomez KA, Coutelier JP, Retegui LA. Sequence similarity and structural homologies are involved in the autoimmune response elicited by mouse hepatitis virus A59. *J Autoimmun.* (2004) 23:117–26. doi: 10.1016/j.jaut.2004.05.006
47. Bodansky A, Mettelman RC, Sabatino JJJr., Vazquez SE, Chou J, Novak T, et al. Molecular mimicry in multisystem inflammatory syndrome in children. *Nature.* (2024) 632:622–9. doi: 10.1038/s41586-024-07722-4