PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

The physical and evolutionary energy landscapes of devolved protein sequences corresponding to pseudogenes

Hana Jaafari[a,b,c], Carlos Bueno[a], Nicholas P. Schafer[a], Jonathan Martin[d], Faruck Morcos[d,e,f], and Peter G. Wolynes[a,c,g,h,1]

Contributed by Peter G. Wolynes; received December 20, 2023; accepted March 26, 2024; reviewed by Robert B. Best and Martin Weigt

Protein evolution is guided by structural, functional, and dynamical constraints ensuring organismal viability. Pseudogenes are genomic sequences identified in many eukaryotes that lack translational activity due to sequence degradation and thus over time have undergone "devolution." Previously pseudogenized genes sometimes regain their protein-coding function, suggesting they may still encode robust folding energy landscapes despite multiple mutations. We study both the physical folding landscapes of protein sequences corresponding to human pseudogenes using the Associative Memory, Water Mediated, Structure and Energy Model, and the evolutionary energy landscapes obtained using direct coupling analysis (DCA) on their parent protein families. We found that generally mutations that have occurred in pseudogene sequences have disrupted their native global network of stabilizing residue interactions, making it harder for them to fold if they were translated. In some cases, however, energetic frustration has apparently decreased when the functional constraints were removed. We analyzed this unexpected situation for Cyclophilin A, Profilin-1, and Small Ubiquitin-like Modifier 2 Protein. Our analysis reveals that when such mutations in the pseudogene ultimately stabilize folding, at the same time, they likely alter the pseudogenes' former biological activity, as estimated by DCA. We localize most of these stabilizing mutations generally to normally frustrated regions required for binding to other partners.

pseudogenes | energy landscapes theory | information theory | protein evolution | resurrected genes

Significance

Pseudogenes are DNA sequences that previously encoded protein sequences but are no longer translated due to sequence degradation. As pseudogenes no longer experience selection pressure to fold into functional proteins stably, they serve as unique examples of naturally occurring protein devolution. We surveyed pseudogenes belonging to protein families of varying length, architecture, and biological function using coevolutionary and optimized physical models. We found the mutations found in pseudogenes typically destabilize their former protein structure. Some mutations that would stabilize pseudogenes in their former structure are found to inhibit or alter their previous biological function.

Natural protein sequences commonly fold into energetically stable, organized three-dimensional protein structures. In order to quickly and robustly fold into such a native state, a protein sequence must encode an energy landscape that is only minimally frustrated, i.e., the interactions between encoded residues are energetically compatible and do not lead to frustrating choices between nearly isoenergetic misfolded intermediates during the folding process (1). Residues and contact pairs encoding such favorable interactions will be conserved by evolutionary pressure to retain the three-dimensional structure, and typically pairs of amino acids will have coevolved so as to fold to similar protein architectures (2).

While arising evolutionarily from homologous protein-coding genes, called parent genes, pseudogenes have relatively altered or diminished transcriptional activity and have lost their translational activity, as a result of random mutations throughout a sequence of retrotransposition events. Such events impair upstream regulatory regions, introduce premature stop codons, or insert novel sequences, thereby inhibiting proper folding or function (3). Since this devolution occurred without the necessity to fold, we might expect pseudogenized genes to encode poorly folding proteins. Nevertheless, some pseudogenized genes occasionally regain their function (4, 5), suggesting they sometimes could yield robust protein folding landscapes despite their multiple sequence alterations. Noncoding genomic regions could then serve as reservoirs of protein diversity. It is unclear how frequently gene resurrection occurs over time, and whether it is a universal phenomenon across many proteins. Pseudogenes, former protein-coding genes identified in numerous protein families, serve as natural candidates to study to what extent protein sequences lose their funneled energy landscapes following reduced selective pressure to function as a foldable protein.

Approximately 14,000 human pseudogenes with identifiable parent genes, belonging to about 2,000 protein families, have been identified using BLAST and manually annotated in the GENCODE database (6, 7). Human protein families enriched with the most pseudogenes include G-protein coupled receptors, RNA-recognition motifs, and immunoglobulins. The parent genes of the identified pseudogenes encode an array of different protein folds, enabling a systematic study of their energy landscapes. Worm, fly, and zebrafish pseudogenes have also been similarly curated; however, the total number

Author contributions: H.J., C.B., N.P.S., J.M., F.M., and P.G.W. designed research; H.J. performed research; H.J., C.B., N.P.S., J.M., F.M., and P.G.W. analyzed data; and H.J., F.M., and P.G.W. wrote the paper.

Reviewers: R.B.B., NIH; and M.W., Sorbonne Universite.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: pwolynes@rice.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2322428121/-/DCSupplemental.

Published May 13, 2024.

of identified pseudogenes for each of these organisms is less than 1,000 (7). Since mouse, macaque, and human pseudogenes are similarly enriched across protein families, we believe the human dataset will be representative of most mammals (7, 8). Unlike in eukaryotic evolution, pseudogenes are typically short-lived in bacterial evolution (9); pseudogenes are however particularly enriched in host-dependent bacteria (10, 11), but their low abundance would limit the findings of any systemic study.

If the sequence statistics of evolution reflect primarily the necessity to fold, the probability that a given sequence in a protein family successfully yields its folded state having an energy E should follow a Boltzmann distribution at an apparent selection temperature T_{sel} (12, 13):

$$P(S) = \frac{e^{-\beta E(S)}}{Z}. \quad [1]$$

In Eq. **1**, the coefficient $\beta = (k_B T_{sel})^{-1}$, contains the selection temperature T_{sel}. E(S) is the energy of a folded structure with a specific sequence S which is measured relative to the average energy of the protein's molten globule states. The molten globule ensemble's energies should be statistically equivalent to the energies of scrambled sequences threaded through the native structure. The normalization Z then becomes a canonical partition function of a sequence-based Potts model. The set of energies for a protein family's native sequences can be said to comprise the evolutionary energy landscape. In this framework, random sequences are more frustrated in an evolutionary sense than native sequences, lacking in both energetically favorable two-body couplings and in conserved single residues relevant to function. The notion of frustration arose in simple models of magnets where the separation of the energy function into local parts (ferromagnetic and antiferromagnetic pair interactions) seems quite natural. In proteins and other systems, in either a physical or evolutionary sense, frustration cannot be so easily read off the Hamiltonian because it is not clear exactly how to separate an energy function uniquely into component parts. Rather, frustration is quantified by comparing the energy of a particular structure to the energies of other possibilities that are found by perturbing the protein structure and the sequence locally in space. For this reason, the local frustration of the evolutionary energy landscape has contributions that come both from the conservation of sequence (the one-body fields term) and the coevolution of pairs (the two-body interaction coupling term). In this sense, one does not separately consider frustration from pairs of sites or on-site fields as might be appropriate in magnets.

The selection temperature T_{sel} in this picture corresponds to the selection pressure to achieve a sufficiently minimally frustrated folding landscape for the protein once it is translated. Given the need for accessible physical stability of a proper structure so as to achieve proper biological function in most proteins, evolution primarily selects for energetically minimized structures, but of course, there are additional selection pressures for specific functional motions as well. As T_{sel} decreases, the probability of finding a sequence being able to fold efficiently decreases, and only the most stabilized sequences are likely to belong to the protein family. T_{sel} values for eight protein families with distinct tertiary structures were found to range between 50 and 130 K, averaging ~110 K (12).

Conversely, every protein sequence has a physical energy landscape, composed of the energies associated with each possible folded configuration. Since natural protein sequences are minimally frustrated, their physical energy landscapes are funneled with the correctly folded state serving as the energetic basin. Given the reduced evolutionary pressure to maintain function once a sequence has been pseudogenized, the physical energy landscapes corresponding to pseudogene sequences should be allowed to yield more rugged landscapes than their parent protein, since it would be irrelevant to avoid kinetic traps that might prevent the sequence from reaching its native folded state, if the protein were ever actually made. In other words, at selection temperature T_{sel}, a pseudogene will be less likely than its parent protein to encode an accessible native protein fold (Eq. **1**). Under the assumption that, for a given sequence, the physical folding landscape and the evolutionary landscape have similar statistics, T_{sel}, the folding temperature T_f, and the glass temperature T_g of its molten globule state, turn out to satisfy a simple relation (13):

$$\frac{2}{T_f T_{sel}} = \frac{1}{T_g^2} + \frac{1}{T_f^2}. \quad [2]$$

The protein's folding temperature, T_f, depends on the energy gap between the proteins selected at T_{sel} and the molten globule, and the glass temperature in the random energy model for the misfolded states depends on the variance in the globule's landscape (1, 14). The protein's glass temperature T_g indicates the temperature at which the protein would typically become trapped in a nonnative energetic well, with an extensive energy cost to refold. In minimally frustrated proteins, T_g is lower than T_f so that this trapping does not become an issue; as T_f/T_g increases from one, the physical energy landscape becomes progressively more funneled. Eq. **2** indicates that, for T_f/T_g to be larger than 1, T_{sel}/T_g must be smaller than 1. Morcos et al. (12) found $T_{sel}/T_g < 1$ for an array of several protein families. Their results quantitatively confirmed that protein sequence evolution has been shaped by the need of selecting for a funneled protein folding landscape (12).

Modern sequencing efforts have now made available the sequences of many members of nearly every protein family. Inverse statistical methods can then exploit this abundance of homologous protein sequence information available from multiple sequence alignments to elucidate correlations in coevolving residue pairs. One such method, direct coupling analysis (DCA), uses a maximum entropy approach to quantify single site and pairwise correlations, which subsequently parameterize a Potts Model Hamiltonian in the sequence space. DCA has efficiently and reliably identified coevolutionary patterns in multiple protein systems, and has proven a key technology in modern protein structure prediction (2, 15–20) as well as for modeling the time course of sequence evolution (21). Using a protein family's coevolutionary information, DCA provides a probability P(S) or evolutionary energy log(P(S)) to an arbitrary sequence to belong to a protein family through the corresponding Potts model Hamiltonian. Meanwhile, physical models centered on using the principle of minimal frustration to learn the folding landscape (14) have been shown to well characterize protein physical free energy landscapes that encode the free energies of structures in sampled folding pathways. The associative memory water-mediated structure and energy model (AWSEM), a coarse-grained free energy Hamiltonian, is a computationally efficient and accurate tool for de novo protein structure and assembly prediction (22, 23), whose parameters have been learned using the principle of minimal frustration as a "loss function" (14).

By taking the ratio of the probabilities of being in the folded and molten globular states (using Eq. **1**) and again assuming that the protein's evolutionary and folding energies are statistically equivalent in their distributions, the selection temperature T_{sel} of a protein family in physical units can be obtained:

$$T_{sel} = \frac{H_{AWSEM,Native} - H_{AWSEM,Globule}}{k_B(H_{DCA,Globule} - H_{DCA,Native})}. \quad [3]$$

Note that in Eq. **3**, DCA energy values are unitless. T_{sel} can then be calculated using a linear least squares fit of the DCA and AWSEM energies of a domain's natural and randomized (globule) sequences.

By comparing the coevolutionary and physical landscapes for the sequences encoded by pseudogenes, we can therefore quantify the origin of evolutionary pressures. We can also specifically compare the energies corresponding to a pseudogene sequence and its parent using the DCA and AWSEM energy functions. Our results, described below, show that all pseudogenes would lead to proteins that would have become relatively more frustrated in an evolutionary sense and that this frustration increases as their sequence similarities to their parent protein decrease, as quantified by the DCA energy function. We also found that for nearly all cases, the purely physical folding energy landscapes corresponding to the hypothetical pseudogene protein, when compared to their parent proteins' energy landscapes, also became more frustrated. It turns out, however, that in some cases there seem to have been some mutations that would increase the pseudogene protein's physical stability, but that would however alter or inhibit their parent protein's function, as revealed by DCA. Careful scrutiny of the localization of the stabilizing mutations in the structure suggests how these mutations lead to disfunction.

Results

We studied the physical and evolutionary energy landscapes of 24 protein families that are listed in Table 1. Given our interest in measuring changes in pseudogenes' physical stability, we did not study disordered proteins and only selected protein families known to encode unique and stable structures. Fourteen of these protein families demonstrate catalytic activity. Three of the families are ribosomal proteins (RP). The lengths of the parent protein sequences ranged from 77 to 415 residues.

The average T_{sel} value of these families is quite low, approximately 24 K, indicating the proteins in these families are quite strongly funneled to their native structure. For illustration, H_{AWSEM} vs H_{DCA} scatter plots of protein families Cyclophilin A, RP S8, and RP L7Ae/L30e/S12e are shown in Fig. 1. The H_{AWSEM} vs H_{DCA} scatter plots of the remaining protein families included in our study are shown in *SI Appendix*, Figs. S1–S6. We calculated T_f/T_g and T_{sel}/T_g ratios for the parent proteins using available, experimentally determined T_f values in the ProTherm database (24). We found that all the evaluated proteins indeed had highly funneled physical energy landscapes ($T_f/T_g \sim 5$) due to selection pressure to fold (Fig. 2).

Pseudogenes Are More Frustrated Than Their Parent Proteins. Several families, including those of Cyclophilin A, Small Ubiquitin-like Modifier-2 (SUMO-2), RP S8, and RP L7Ae/L30e/S12e, turn out to be particularly enriched with pseudogenes (Table 1). The substitution percent for pseudogenes in our study ranged from only 1% to 55%, averaging out to 17%. We characterized the relative differences in the physical and evolutionary energies between the parent proteins and the proteins that would be encoded by their pseudogenes if they were translated. Pseudogene sequence devolution is measured by calculating Δ_{rel}AWSEM and Δ_{rel}DCA values for every pseudogene-parent protein pair. We expect that, as a sequence accumulates random mutations, the network of energetically minimized global contacts will be increasingly perturbed so we expect to have both $\Delta_{rel}\text{AWSEM} < 0$ and $\Delta_{rel}\text{DCA} < 0$. Mutations that inhibit function may also have been introduced into physically frustrated regions that are nevertheless conserved. In this case, we would find $\Delta_{rel}\text{DCA} < 0$, even if the mutations are physically stabilizing. Since DCA learns about all evolutionary correlations that arise from both the need for physical stability and for function, Δ_{rel}DCA is expected to be more sensitive to the mutations occurring after the pseudogene stops being translated. If there were a constant mutation rate, Δ_{rel}AWSEM and Δ_{rel}DCA should become more negative as the evolutionary time since the pseudogene ceased to be translated increases. Details of the way the changes, Δ_{rel}AWSEM and Δ_{rel}DCA, are computed and given in *Materials and Methods*.

All pseudogenes turn out to be more frustrated when evaluated through the evolutionary energy landscape in comparison with their parent protein's fold ($\Delta_{rel}\text{DCA} < 0$) (Fig. 3*A*). We also find the relative difference in evolutionary energy Δ_{rel}DCA to be directly proportional to the apparent number of substitutions from the parent sequence, linearly decreasing (Pearson Coefficient $|R| = 0.9$) over evolutionary time. We find that changes in the coupling terms primarily contributed to Δ_{rel}DCA changes in

Table 1. Parent protein sequence lengths, as well as parent protein and pseudogene counts, associated with all of the studied protein families are shown above listed in descending order of family's pseudogene count

Protein family	PDB length	Parent count	Pseudogene count	M	L
Cyclophilin A	163*	2	20	12,533	155
RP S8	129	1	9	5,965	129
RP L30e/S12e	124⋆	2	9	4,887	95
SUMO-2	77	1	8	1,016	72
Profilin-1	138	1	4	971	121
E2	170†	2	3	7,440	140
AAA domain	173	1	2	2,366	129
ATG8	120‡	2	2	925	104
eIF4E	177	1	2	1,371	165
eRF1 Domain 2	412	1	2	1,052	133
Cofilin	166	1	2	2,009	127
TCTP	172	1	2	2,552	164
Helicase C	379	1	2	78,902	78
HIT domain	110	1	1	7,644	98
Calponin/Actinin	190	1	1	10,193	108
RP S7p/S5e	191	1	1	7,713	148
BART	120	1	1	1,313	116
GLTP	203	1	1	3,760	147
FKBP PPIase	109	1	1	15,112	94
HAD Hydrolase	250	1	1	31,214	176
UCHL-1	223	1	1	978	214
Glutaredoxin	105	1	1	10,830	60
PGK	415	1	1	5,328	384
LMWPP	154	1	1	7,408	138

M denotes the number of aligned sequences featured in the PFAM alignment, while **L** denotes the PFAM alignment length.
Abbreviations: LMWPP: Low molecular weight phosphotyrosine protein phosphatase; TCTP: Translationally controlled tumor protein; E2: Ubiquitin-conjugating enzyme; Helicase C Terminus; GLTP: Glycolipid Transfer Protein.
(⋆) Sequence length of PDB 4KZZ (Chain M), parent protein of six RP L30e/S12e pseudogenes. The other parent protein PDB 3VI6 (Chain A) is 97 residues long.
*Sequence length of PDB 6U5E (Chain A), parent protein of 19 Cyclophilin A pseudogenes. The other parent protein PDB 1QOI (Chain A) is 177 residues long.
†Sequence length of PDB 2HLW (Chain A), parent protein of two E2 pseudogenes. The other parent protein PDB 4IP3 (Chain B) is 151 residues long.
‡Sequence lengths of parent protein PDBs 5GMV (Chain A) and 4CO7 (Chain B) are 120 and 118 residues, respectively.

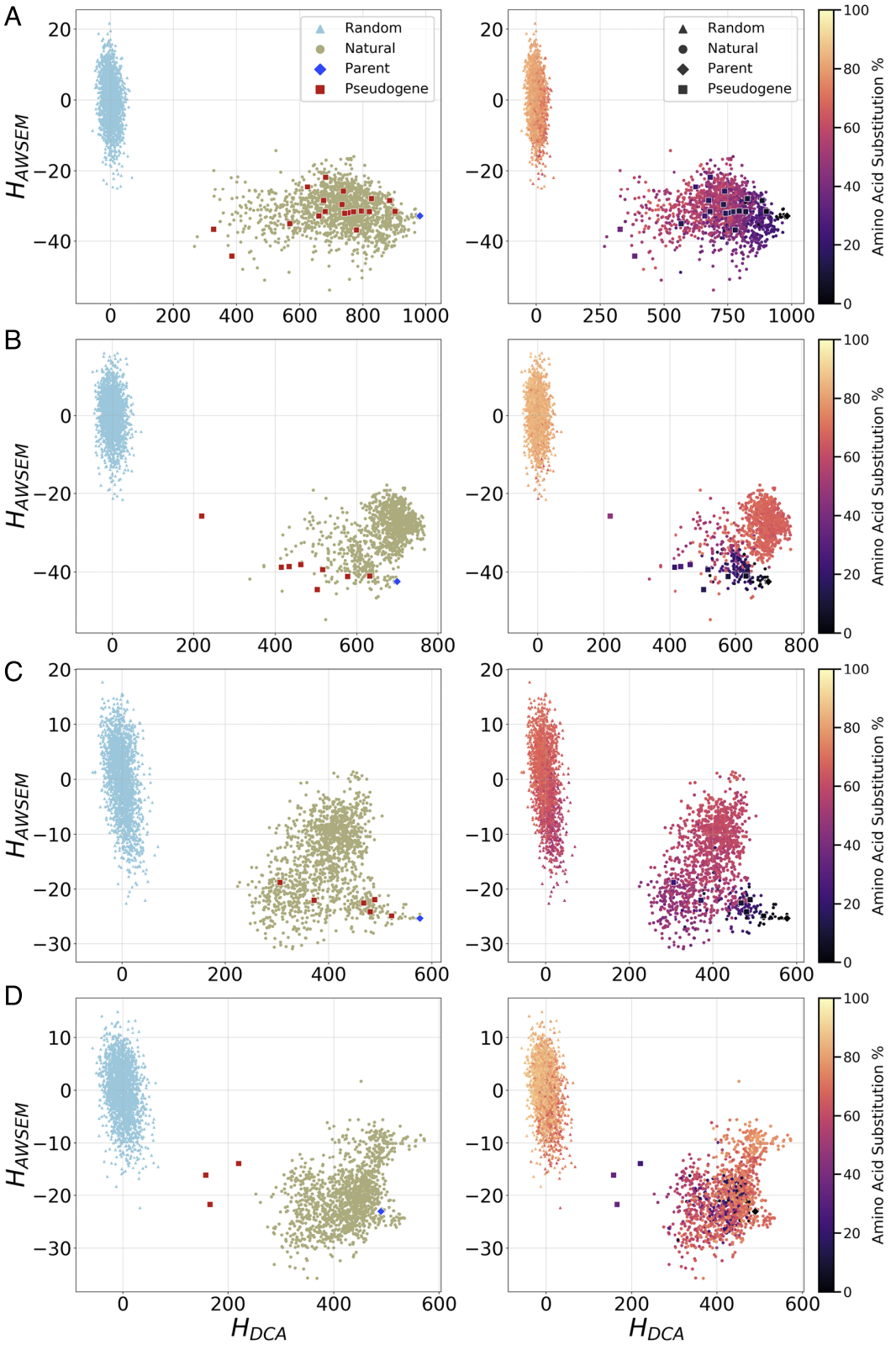

Fig. 1. Scatterplots of AWSEM vs. DCA energies for several families: (*A*) Cyclophilin A (PDB 6U5E, Chain A), (*B*) RP S8 (PDB 4KZZ, Chain W), (*C*) RP S12e (PDB 4KZZ, Chain M), and (*D*) RP L30e (PDB 3VI6, Chain A). For subplots in the right column, data are colored based on sequence divergence relative to the parent protein.

select protein families (*SI Appendix*, Fig. S7). Of course, we can only estimate when the pseudogenization event occurred for each instance. Conceivably a protein product may have been encoded by a mutated sequence that had arisen from gene duplication. Furthermore, for 70% of the pseudogenes, the physical folding energy landscapes corresponding to the hypothetical translated sequences also appear to be comparatively more frustrated than their parent protein's (Δ_{rel}AWSEM < 0). We also compared

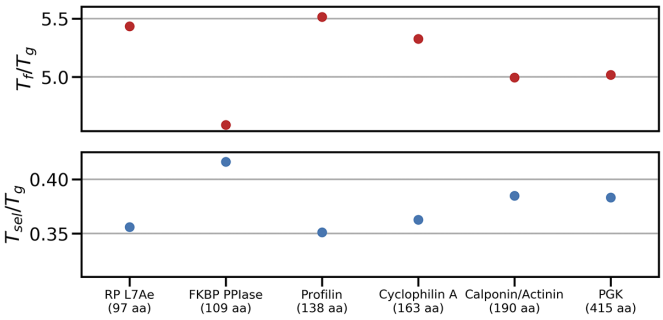

Fig. 2. The protein families are ordered in ascending sequence length. All proteins have funneled landscapes ($T_f/T_g > 1$) as a result of natural selection ($T_{sel}/T_g < 1$).

Δ_{rel}DCA and Δ_{rel}AWSEM values of pseudogenes to those for natural sequences having equivalent degrees of sequence divergence from the parent proteins in several protein families (*SI Appendix*, Fig. S8). Most pseudogene proteins, resulting from random evolution without folding constraints, are more evolutionarily and physically frustrated than their parent protein when compared to natural sequences of equal divergence, which had been, of course, maintained under folding selection pressure during their evolution. These findings are consistent with experimental work showing naturally occurring polymorphisms in *Escherichia coli* protein-coding genes to be notably more neutral than mutations occurring under low selection pressure or high mutation rates (25, 26).

There are however some pseudogenes which, if translated, would encode proteins having more energetically stabilizing interactions than their parent protein (Δ_{rel}AWSEM > 0) has. We will see these sequences would have likely lost functional capabilities if translated (27–29). Many of these pseudogenes belong to the Profilin-1, Cyclophilin A, and SUMO-2 protein families (*SI Appendix*, Fig. S9). We examine the mutational effects on biological activity in these families' pseudogenes using binding affinity and AWSEM frustration measurements in later sections.

Randomly Evolving Sequences Become More Energetically Frustrated Over Time. If pseudogenes are assumed to have evolved randomly over time without functional selection pressure since they stopped encoding a protein, we can simulate the evolutionary and physical energies of an ensemble of putative pseudogene sequences that have progressively undergone sequence devolution without any need to function. We generated such randomly devolved sequences by first selecting a random subset of the parent protein's aligned residues and then mutating each residue into another of the 20 amino acids with equal probability. Putative pseudogenes' substitution rates varied from 1% to >75%. At each substitution count, we generated 5,000 sequences; while the sequences generated in this way will not explore all of sequence space, we expect they should sufficiently capture the population statistics due to the law of large numbers.

After generating these partially de-evolved sequences, we measured the relative changes in evolutionary and physical energies, Δ_{rel}DCA and Δ_{rel}AWSEM, of these randomly evolved sequences for multiple parent proteins. As discussed above, we expect the changes for these nonselected sequences, Δ_{rel}AWSEM and Δ_{rel}DCA, will become more negative as more mutations are introduced. Indeed, the longer the time over which the sequences are taken to have randomly evolved, the more physically and functionally unfavorable residues were introduced relative to the parent protein (Fig. 3 *B*–*D*). As previously observed with naturally occurring pseudogenes (Fig. 3*A*), one finds subpopulations of sequences that have some physically stabilizing mutations.

We also explored another model of random evolution by generating variants of Cyclophilin A, RP S8, and Profilin A parent proteins by using random nucleotide substitutions, rather than by making amino acid substitutions directly. Given the propensity of the genetic code to maintain a mutated residue's physicochemical characteristics, one might expect a slower change of Δ_{rel}DCA and Δ_{rel}AWSEM values of sequences when making random nucleotide substitutions than seen for sequences made directly with random amino acid substitutions. In fact, we do see that Δ_{rel}DCA and Δ_{rel}AWSEM values change somewhat more slowly over apparent evolutionary time, as measured by the number of resulting amino acid substitutions, when the sequences are obtained using random nucleotide substitutions than when amino acid substitutions are directly made, as shown in *SI Appendix*, Figs. S10 and S11. Furthermore, we note that, at a given degree of sequence divergence, the range of variation of Δ_{rel}AWSEM is less for the sequences made using random nucleotide substitutions, than when amino acids are directly randomized.

Mutations Inhibit Cyclophilin A Pseudogenes' Function. While folding is generally a key part of a protein's being able to function, there are other functional constraints too. As an example, we looked at the pseudogenes that are associated with the parent protein Cyclophilin A (PDB ID 6U5E), an enzyme that facilitates the folding of other proteins via cis-trans isomerization of proline residues. The mutations in the pseudogene sequence disrupt the global network of energetically stabilizing interactions found in Cyclophilin A and its homologues in 80% of the Cyclophilin A pseudogenes (Fig. 4*A*). On the other hand, the Cyclophilin A binding interface tends to become more enriched with mutations that would encode those minimally frustrated pseudogenes (Δ_{rel}AWSEM > 0) (*SI Appendix*, Fig. S12). While such mutations would lead to good folding, they likely would diminish the binding affinity or catalytic activity of the hypothetical translated protein, as we quantify below.

Cyclophilin A modulates HIV-1 replication by binding to the capsid protein of HIV (30, 31). To investigate the mutational effects of the sequence change in these pseudogenes on binding affinity to the HIV-1 capsid protein N terminus (PDB ID 1M9C), we compared Δ_{rel}AWSEM values for both the bound and unbound configurations (Fig. 4*B*; data for all Cyclophilin A pseudogenes in *SI Appendix*, Fig. S13). If a pseudogene's Δ_{rel}AWSEM value decreases when bound, its binding affinity possibly will have decreased due to these substitutions thus leading to unfavorable binding interactions. When unbound, the protein that would be encoded by pseudogene ENST00000450588 has a binding interface that would be more minimally frustrated than the actual protein Cyclophilin A, but the opposite is the case for the bound configuration. Thus we see that this type of devolution would decrease the protein's ability to bind to partners at that location. Contacts near and within functional regions including binding pockets tend to be frustrated when in their native, unoccupied configuration, only becoming minimally frustrated once they bind to their partners (32).

We next examined whether the mutations in the pseudogene sequence would affect the catalytic function or the targetability of the protein for posttranslational modifications (*SI Appendix*, Fig. S14). Residues R55 and F60, which are integral to the protein's catalytic activity (33), are modified in pseudogenes

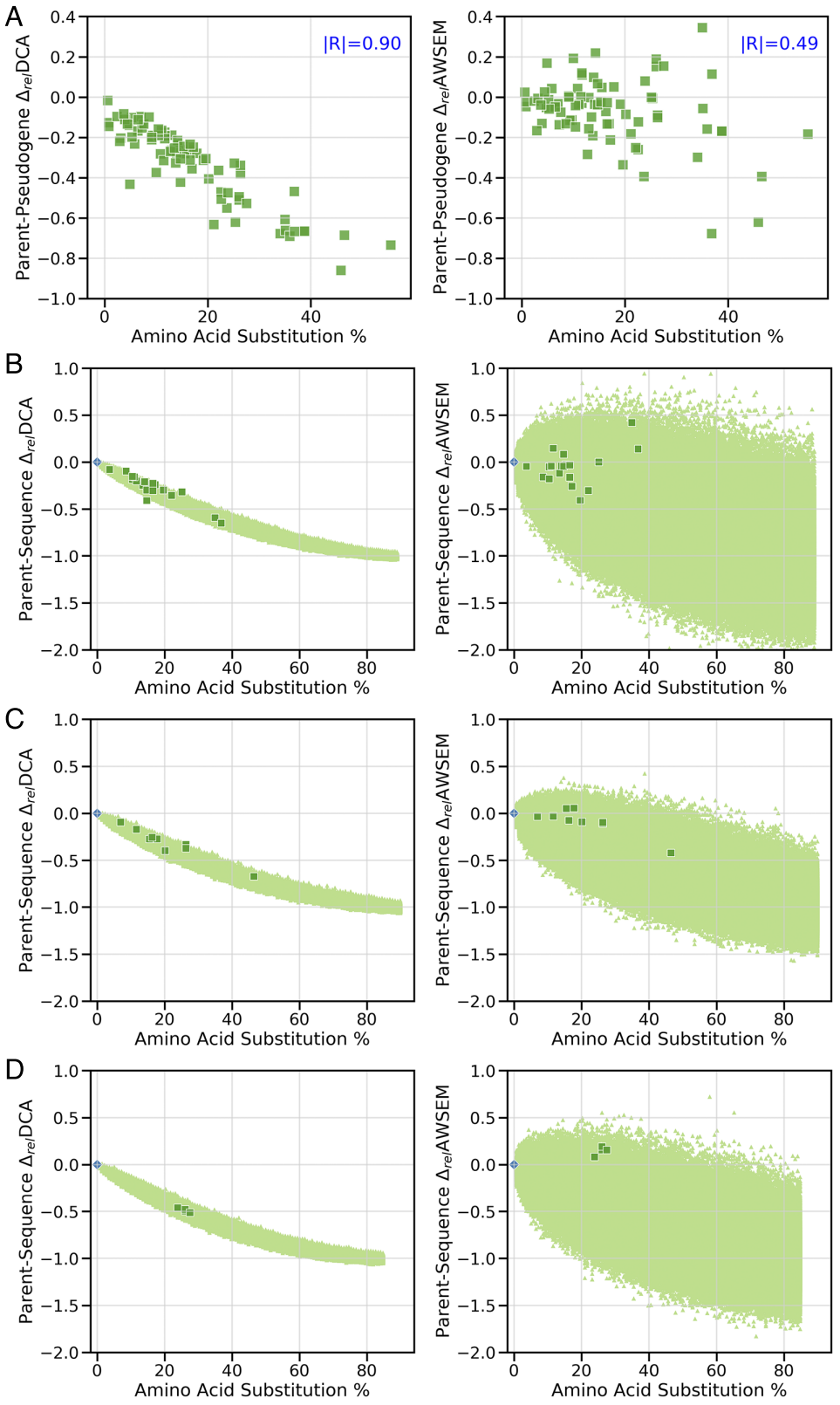

Fig. 3. (*A*) The changes Δ_{rel}DCA and Δ_{rel}AWSEM for all the pseudogenes studied are plotted, as a function of their percentage of substitutions. Δ_{rel}DCA values decrease linearly with apparent evolutionary time. In order to energetically compare pseudogenes and their parent proteins over a larger range of substitutions, we also generated randomly evolving sequences for Cyclophilin A (*B*), RP S8 (*C*), and Profilin A (*D*). The marker shapes and colors correspond with sequence types: pseudogene sequences are denoted by dark green squares, randomly evolving sequences are denoted by light green triangles, and parent protein sequences are denoted by blue diamonds at the origin.

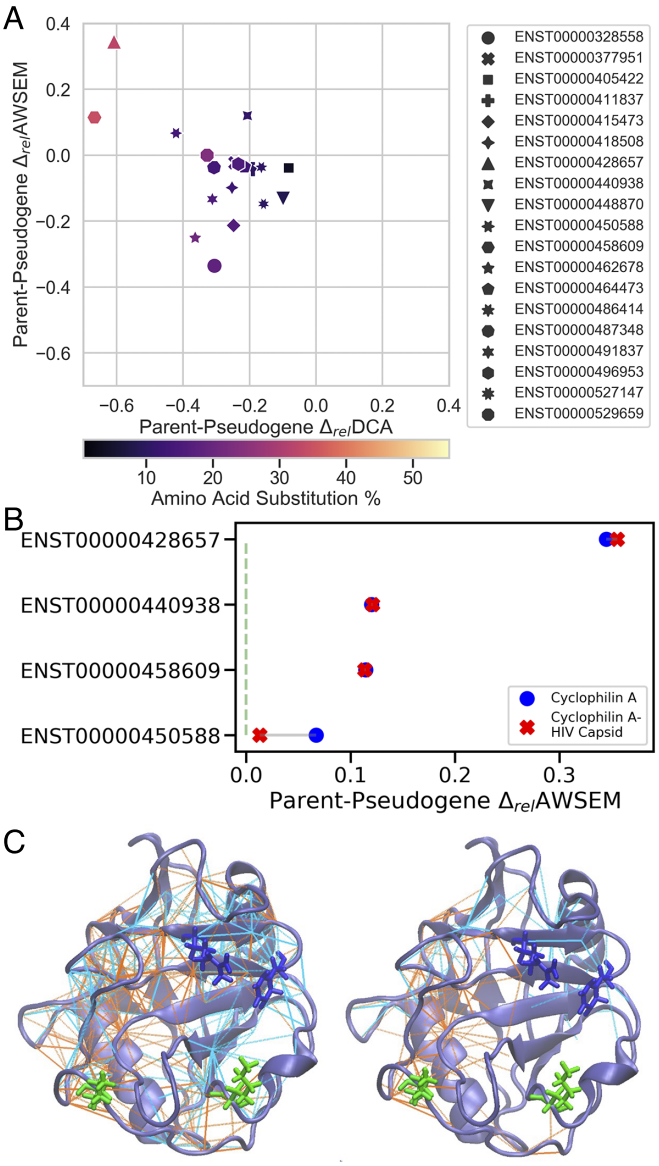

Fig. 4. (*A*) Δ_{rel}AWSEM and Δ_{rel}DCA values for proteins that would be encoded by the Cyclophilin A pseudogenes, along with their sequence divergence from their parent protein, are plotted. (*B*) The binding affinity changes corresponding with all Cyclophilin A pseudogenes are determined by comparing the Δ_{rel}AWSEM values in bound and unbound states. The protein corresponding to pseudogene ENST00000450588 is less likely to bind with the HIV-1 capsid protein N terminus compared to Cyclophilin A itself. (*C*) Differences in AWSEM mutational frustration indices between Cyclophilin A and the protein that would be encoded by pseudogene ENST00000428657 are indicated. Cyclophilin A residues critical for catalytic activity (colored dark blue) and posttranslational modifications (colored neon green) are drawn in the licorice representation. Cyan lines indicate that the parent's contact would be more minimally frustrated than the corresponding contact in the pseudogene encoded sequence, while orange lines indicate the inverse. In the right-hand image, only contacts with frustration indices differing between the ENST00000428657 and parent by one SD or more are shown.

ENST00000428657 and ENST00000458609. These changes therefore would render a protein encoded by the pseudogene sequences catalytically inactive. Acetylation inhibits the protein's enzymatic activity and alters its localization following oxidative stress (34, 35). Residues K82 and K125, which are targets for acetylation, were also found to be mutated in pseudogenes ENST00000428657 (K82S, K125R) and ENST00000458609 (K82S). Significant changes are seen in ENST00000428657 frustration including at the acetylation target site K82, which becomes noticeably more minimally frustrated upon mutation (Fig. 4*C*).

Mutations in the Profilin Pseudogenes Would Lead to Increased Interactions with Other Proteins. Profilin-1 (PDB ID 6NBW, Chain P) nucleates monomeric actin by binding to the barbed end of filamentous actin (F-actin). It serves as a key regulator of actin fiber polymerization. In order to polymerize F-actin efficiently, Profilin-1 forms a complex with proteins containing the polyproline motif, such as the VASP protein, and Actin (PDB ID 3CHW) (36, 37). All four of the pseudogenes associated with Profilin-1 would translate into proteins that would have relatively more minimally frustrated physical energy landscapes (Δ_{rel}AWSEM > 0) than their parent protein. At the same time, the DCA suggests they would have decreased functional fitness (Fig. 5*A*).

Using AWSEM, we can compare the binding affinities of Profilin and the pseudogene proteins to Actin and VASP proteins. One finds the pseudogene-encoded proteins would have more stabilizing binding interactions with both Actin and VASP if the pseudogene substitutions were made (Fig. 5*B*). Despite improved binding affinity, Profilin-1 residue C71 has been linked to the development of familial Amyotrophic lateral sclerosis (ALS). This residue is mutated in all the pseudogenes (*SI Appendix*, Fig. S15). Profilin-1 C71 variants have been shown to maintain native-like F-actin polymerization activity but demonstrates increased aggregation propensity (38, 39). Comparing the parent protein and pseudogenes AWSEM mutational frustration patterns shows that, when residue 71 is mutated, interactions with neighboring contacts become more frustrated, increasing the propensity for aberrant pseudogene-protein interactions (Fig. 5*C*).

Pseudogene-Encoded Mutations Would Modify SUMO-2 Acetylation Sites. SUMO-2 proteins (PDB ID 6JXX) serve as important cellular posttranslational modifiers (40). Monomeric or polymerized SUMO-2 proteins covalently bind to target proteins, modulating their expression and cellular localization. SUMO-2 helps regulate DNA repair mechanisms by catalyzing the disassociation of DNA and thymine-DNA glycosylase, an enzyme that recognizes and removes mismatched bases, by inducing a conformational change. Nearly half of SUMO-2 pseudogenes would yield proteins having more minimally frustrated physical energy landscapes than their parent protein (Fig. 6*A*). At the same time, when bound to thymine-DNA glycosylase (PDB ID 2D07), all SUMO-2 pseudogenes, except ENST00000504193 and ENST00000511179, had decreased binding affinity relative to their parent protein as a result of their sequence degradation (Fig. 6*B*).

Alterations to the ENST00000504193 C terminus appear to amplify its capacity to polymerize by increasing the region's frustration (Fig. 6*C*) (41). SUMO-2 protein chain architecture can be regulated through lysine acetylation (42). Pseudogene ENST00000504193 mutations introduce a lysine into the sequence (R56K) (*SI Appendix*, Fig. S16). In comparison to the parent protein, contacts involving this mutated residue are more frustrated, suggesting this mutated site would be targeted for posttranslational control.

Discussion

We studied the consequences of removing evolutionary pressure to function on several families of proteins by studying the landscapes of pseudogenes derived from these families, using both

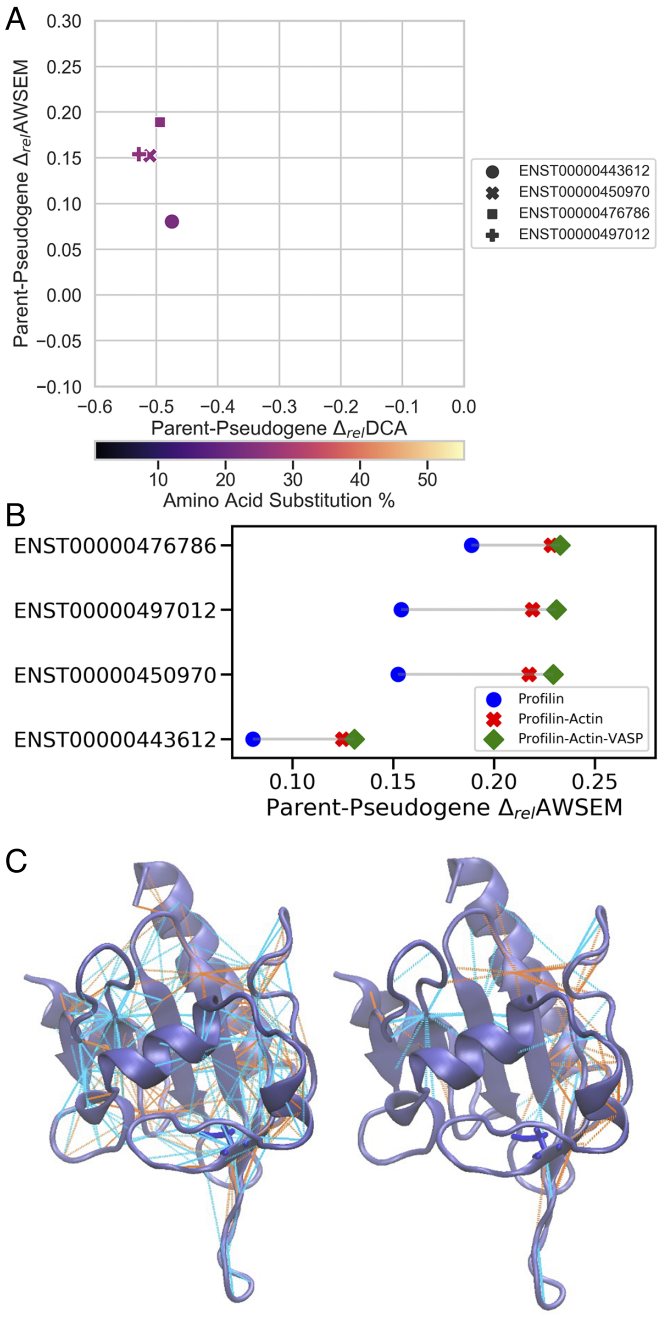

Fig. 5. (*A*) Δ_{rel}AWSEM and Δ_{rel}DCA values for proteins that would be encoded by the Profilin-1 pseudogenes, along with their sequence divergence from their parent protein, are plotted. (*B*) The binding affinity changes corresponding with all Profilin-1 pseudogenes are determined by comparing the Δ_{rel}AWSEM values in bound and unbound states. The proteins that would correspond with all Profilin-1 pseudogenes would bind relatively more favorably to Actin and VASP proteins. (*C*) Differences in AWSEM mutational frustration indices between Profilin-1 and the protein that would be encoded by pseudogene ENST00000497012 are indicated. Residue C71, drawn in the licorice representation in blue, is mutated in ENST00000497012; the region's binding affinity increases as a result of the mutation. In the right-hand image, only contacts with frustration indices differing between the ENST00000497012 and parent by one SD or more are shown.

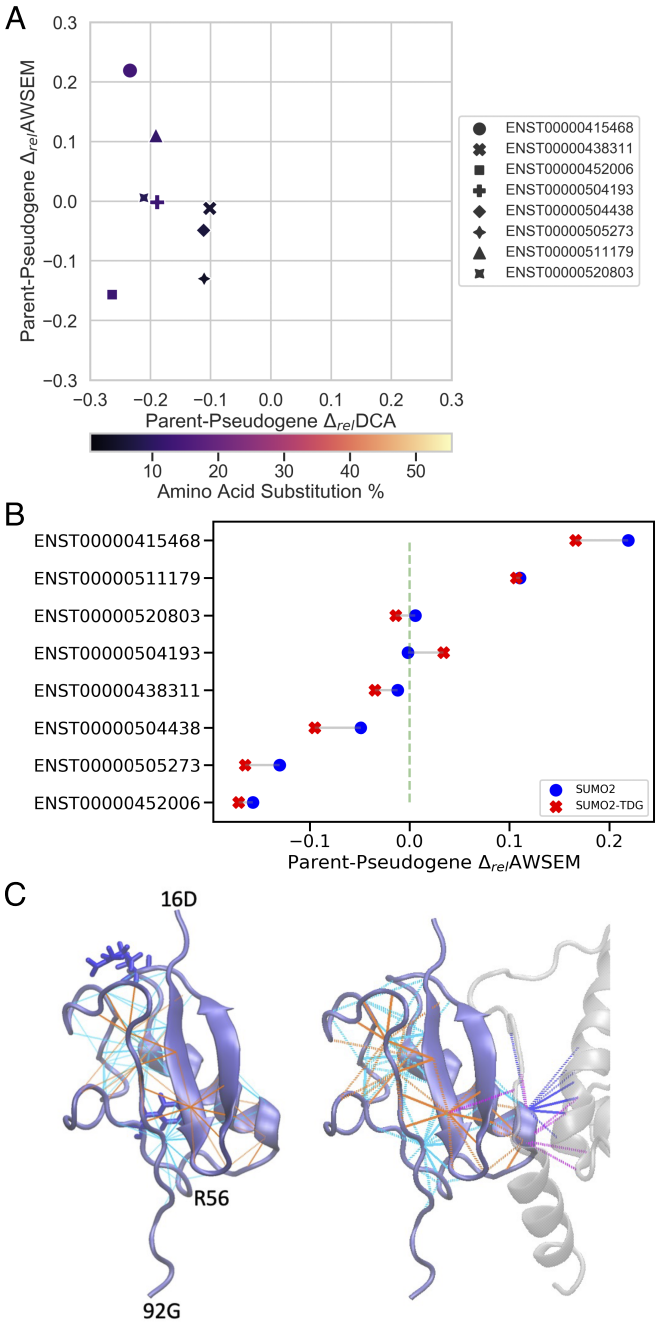

Fig. 6. (*A*) Δ_{rel}AWSEM and Δ_{rel}DCA values for proteins that would be encoded by the SUMO-2 pseudogenes, along with their sequence divergence from their parent protein, are plotted. (*B*) The binding affinity changes corresponding with all SUMO-2 pseudogenes are determined by comparing the Δ_{rel}AWSEM values in bound and unbound states. (*C*) Differences in AWSEM mutational frustration indices between the SUMO-2 and the protein that would be encoded by pseudogene ENST00000504193 in the unbound (*Left*) and bound state (*Right*) are indicated. Dark blue lines indicate an interprotein contact between the parent protein and binding partner is more minimally frustrated than the pseudogene's interprotein contact, while magenta lines indicate the inverse.

a coevolutionary model and an energy landscape optimized for physical folding. As was previously observed by Morcos et al. for a different array of model proteins (12), the energy landscapes for the folds were found to be highly funneled ($T_f/T_g > 1$) apparently due to natural selection ($T_{sel}/T_g < 1$). Employing both models allows us to quantify a protein's degree of minimal frustration. As we refined the experimental single mutation $\Delta\Delta G$ training data previously used to convert AWSEM energy units to laboratory units (*Materials and Methods*), we also recalculated previously published T_f/T_g and T_{sel}/T_g values (12). We found the proteins to be more funneled than previously thought, owing to a change in the weighting of the AWSEM energy terms from that used

previously. Nevertheless, one finds similar T_f/T_g values (ranging between 3.5 and 6) for the parent proteins analyzed in this study, in comparison to the earlier work. Our measured T_f/T_g values fall within the range of lower-limit values of $T_f/T_g \sim 2.5$ (43) and $T_f/T_g \sim 6$ (44), and substantially larger than the values suggested earlier on by the comparison to lattice-model proteins (45). Furthermore, our T_f/T_g values overlap with T_f/T_g values measured for protein families not included in our study, predicted using population genetics and evolutionary energies calculated using DCA (46).

Our analysis of the energy landscapes of proteins that would be encoded by pseudogenes indicates that, if translated, pseudogene coding products would have impaired biological functions in comparison with their parent protein. While most often we find the folding landscape to be impaired, we also found some pseudogene sequence variations that would enhance global interactions favorable for encoding a folding funnel but that would, in contrast, interfere with native binding interactions or posttranslational modifications. We analyzed SUMO-2, Profilin-1, and Cyclophilin A pseudogenes in some detail in this regard. Notably, despite the loss of selection pressure to fold, pseudogenes still structurally and functionally resemble the other natural sequences in their protein family (Fig. 1). If these pseudogenes were to be resurrected and somehow regain their translational activity over evolutionary time, they would be annotated as members of the protein family. The tolerance of pseudogenes to survive sequence variation as to their folding ability is a consequence of their parent proteins having funneled physical energy landscapes.

In this work, we systemically surveyed both biological function and structural stability of proteins that would be encoded by pseudogenes if they were translated across multiple protein families. Since our analysis required the parent proteins to have independent experimentally determined crystal structures and upward of 80% pseudogene homology to their protein domain, we examined a subset of annotated pseudogenes in the GENCODE database which are likely closer to their parents than many actual pseudogenes that are harder to recognize. Nevertheless, since our study includes parent proteins diverse in sequence length, architecture, and biological function, we argue our conclusions are representative for pseudogenes of minimally frustrated parent proteins.

Materials and Methods

Pseudogene Selection Criteria. Parent proteins, and their associated pseudogenes, with a complete experimentally determined NMR or crystal structure were featured in this work. Parent protein structures were identified by searching all available PDB sequences using PSI-BLAST (47). Pseudogene cDNA sequences, available in the Ensembl database (6), were translated with the EMBOSS Transeq algorithm with the default reading frame (48), following the removal of any stop codons. Pseudogenes were included in the analysis if sharing 80% or more sequence similarity with the protein family domain. Natural protein sequences homologous to a parent protein were identified using the associated protein family's multiple sequence alignment (MSA) from the PFAM database (49).

DCA. DCA is an inverse statistical mechanics algorithm commonly used to identify coevolutionary patterns of related protein sequences. Aligned protein sequence sets are used to fit a probabilistic model of the statistical variation between pairs of positions in the aligned sequences, yielding a global model which assigns probabilities to all combinations of amino acids of a defined length. In practice this probability is represented with an energy function, the sequence Hamiltonian, which summarizes the strength of coupling between all residue pairs and the single site frequencies of a particular sequence. Residues pairs with high direct information (DI) values, which are measures of total amino acid coupling strength at two positions, are associated with structural contacts and active sites maintained by natural selection (2). The DCA Hamiltonian H_{DCA} for a sequence of length L is defined as

$$H_{\mathrm{DCA}} = \sum_{i}^{L} h_i(A_i) + \sum_{j}^{L} \sum_{i \neq j} e_{ij}(A_i, A_j). \quad [4]$$

The fields term h_i and couplings term e_{ij} represent the evolutionary energies of residue i and residue pair i, j respectively, and A_i represents the amino acid identity at position i. The h_i and e_{ij} parameters are estimated using the mean-field DCA algorithm with protein family MSAs downloaded from the PFAM 27 database as input, in a procedure similar to previous coevolutionary analysis (2, 49). MSAs of the TCTP (PFAM ID PF00838), BART (PFAM ID PF11527), and GLTP (PFAM ID PF08718) protein families were downloaded from the PFAM 35 database, due to low sequence counts in the families' PFAM 27 MSAs. In order to prevent artifacts in the DCA couplings terms originating from missing data in the alignment, protein sequences featuring consecutive gaps that comprise more than 20% of the sequence were filtered out. MSAs must include approximately more than 1,000 nonredundant, homologous sequences to ensure accurate DCA structural predictions (2). Additionally, noise from long-range contacts was minimized by imposing a distance threshold to contact pairs using a Heaviside step function $\Theta(|r_i - r_j| - r_c)$ (12):

$$H_{\mathrm{DCA}} = \sum_{i}^{L} h_i + \sum_{j}^{L} \sum_{i \neq j} (\Theta(|r_i - r_j| - r_c)) e_{ij}. \quad [5]$$

The C_β-C_β contact maps of the parent protein's crystal structure, with a distance threshold $r_c = 16$ Å identified in Morcos et al. (12), are used in this analysis. Residue pairs involving gaps are ignored, as gaps cannot be mapped to the crystal structure contact map. In a few cases, including Cyclophilin A and RP S12e in Fig. 1, we observed that the parent protein has the highest DCA energy relative to other natural sequences. We found this phenomenon to hold true even if the structure-based distance filter on the couplings term is removed and/or couplings terms related to gaps are included in DCA energy calculations (*SI Appendix*, Figs. S17 and S18). Based on our findings, we believe that the high DCA energies of parent proteins in some protein families result from high sequence conservation or an overrepresentation of homologs of the parent protein in available sequence databases.

"Random" sequences are generated by shuffling aligned residues of natural sequences, ensuring that the sequences' amino acid propensities are preserved. Parent protein and pseudogene protein sequences were aligned to their protein family's hidden Markov model (HMM) profile using the HMMER software package (50). The relative difference in the DCA energies, Δ_{rel}DCA, of each parent protein–pseudogene pair is defined as

$$\Delta_{rel}\mathrm{DCA} = \frac{H_{\mathrm{DCA,Pseudogene}} - H_{\mathrm{DCA,Parent}}}{H_{\mathrm{DCA,Parent}}}. \quad [6]$$

Negative Δ_{rel}DCA values indicate that the parent protein's evolutionary energy landscape is comparatively more funneled than that of the pseudogene.

AWSEM. The AWSEM (22) is a coarse-grained, transferable force field developed for protein structure prediction. AWSEM defines every residue's position by representing its C_α, C_β, and O atoms as individual particles. The AWSEM Hamiltonian H_{AWSEM} is defined as follows:

$$H_{\mathrm{AWSEM}} = H_{\mathrm{Backbone}} + H_{\mathrm{Contact}} + H_{\mathrm{Burial}} + H_{\mathrm{Pap}} + H_{\beta} + H_{\mathrm{Helical}} + H_{\mathrm{AM}} \quad [7]$$

AWSEM energies were calculated using OpenAWSEM, an OpenMM implementation of the AWSEM energy function (23).

Secondary structure information, defined by an associative memory term H_{AM}, improves molecular dynamics simulations' structure prediction accuracy by assisting in the formation of native-like contacts in globular states. In order to

calculate the energetic difference between random globular and native states, H_{AM} is not included in H_{AWSEM} calculations. H_{AWSEM} energies are calculated by threading sequences through the parent protein structure, resulting in fluctuations in the sequence-dependent terms H_{Burial} and $H_{Contact}$.

H_{AWSEM} energies were converted to laboratory units by fitting predicted and experimental mutational energy changes. ΔH_{AWSEM} ($H_{AWSEM,WT} - H_{AWSEM,Mutated}$) values are weighted using multiple linear regression to experimental $\Delta\Delta G$ (kcal/mol) values from the ProTherm database (24) (*SI Appendix*, Fig. S19). Molten globule state are assumed to not energetically fluctuate upon mutation. Single mutation ProTherm entries with the following conditions were used: 1) belonged to protein monomers that fold according to a two-state model, 2) measured at physiological conditions ($20\,°C \leq T \leq 40\,°C$, $6 \leq pH \leq 8$) (entries closest to pH 7 were used for multiple measurements associated with a mutation), 3) affected buried or partially exposed (ASA = <0.4), and 4) introduced a mutation with a change in polarity type (i.e., hydrophilic to hydrophobic or hydrophobic to hydrophilic). Errors in the ProTherm database were corrected by manual curation. A total of 222 ProTherm entries, associated with 12 proteins, met these criteria. Predicted and experimental mutational stability changes were well linearly correlated (Pearson Coefficient $|R| = 0.53$).

The relative difference in the AWSEM energies, Δ_{rel}AWSEM, of each parent protein-pseudogene pair when threaded through the parent protein's unbound structure is defined as

$$\Delta_{rel}\text{AWSEM} = \frac{H_{AWSEM,Pseudogene} - H_{AWSEM,Parent}}{H_{AWSEM,Parent}}. \quad [8]$$

Negative Δ_{rel}AWSEM values indicate that the parent protein's physical energy landscape is comparatively more funneled than the pseudogene's.

AWSEM Protein Frustration. While natural protein sequences are globally minimally frustrated, local structural regions may be frustrated (32). Protein regions associated with biological function, such as binding interfaces and catalytic sites, tend to be frustrated; hence, identifying frustrated regions serves as useful starting points for elucidating a protein's functional sites and mechanisms. A protein's contact pairs' frustration can be characterized by measuring each pairs' mutational and configurational frustration indices. These indices are Z-score values that compare the pairs' stabilization energies with those of other candidate pairs', generated by perturbing only the pair's residue identities or both the residue identities and relative orientations (local density and pairwise distances), respectively. When calculating either mutational or configurational frustration values for a protein's contact pairs, a pair is labeled "minimally frustrated" if its index value is greater than 0.78, "frustrated" if the value is less than −1, and considered "neutral" otherwise.

While a contact pair's (i, j) mutational and configurational frustration indices F_{ij} are similarly defined, the generated decoy states (i', j'), denoted by the superscript U, differ in their physical parameters. Using n decoys, a site's F_{ij} value is defined as

$$F_{ij} = \frac{H^N_{i,j} - \langle H^U_{i',j'} \rangle}{\sqrt{\frac{1}{N}\sum_{k=1}^{n}(H^U_{i',j'} - \langle H^U_{i',j'} \rangle)^2}}. \quad [9]$$

The AWSEM-MD frustratometer algorithm was employed to calculate differences in select parent protein's and pseudogene's mutational and configurational frustration patterns near substitution sites (51). The AWSEM-MD frustratometer indicates frustrated contacts with red lines, minimally frustrated contacts with green lines, and neutral contacts in gray. Water-mediated interactions are drawn with dashed lines.

We measured differences in parent protein and pseudogene AWSEM mutational frustration indices at substitution sites involving direct and water-mediated contacts, $\Delta F_{ij} = F_{pseudogene,ij} - F_{parent,ij}$. Negative ΔF_{ij} values indicate that the pseudogene mutation is comparatively less locally stabilizing than the native residue. *Intraprotein* contacts with negative ΔF_{ij} values are connected with cyan lines, and positive ΔF_{ij} values with orange lines. Furthermore, *interprotein* contacts with negative ΔF_{ij} values are connected with dark blue lines, and positive ΔF_{ij} values with magenta lines.

Data, Materials, and Software Availability. The raw data and analysis scripts used in this study have been deposited in Zenodo (52).

ACKNOWLEDGMENTS. J.M. and F.M. acknowledge support from the NIH (NIH R35GM133631). F.M. acknowledges support from the NSF CAREER award (MCB-1943442). H.J., C.B., and P.G.W. were supported by the Center for Theoretical Biological Physics, sponsored by NSF Grant No. PHY-2019745. Additionally, we wish to recognize the D.R. Bullard Welch Chair at Rice University, Grant No. C-0016 (to P.G.W.).

Author affiliations: [a]Center for Theoretical Biophysics, Rice University, Houston, TX 77005; [b]Applied Physics Graduate Program, Smalley-Curl Institute, Rice University, Houston, TX 77005; [c]Department of Chemistry, Rice University, Houston, TX 77005; [d]Department of Biological Sciences, University of Texas at Dallas, Richardson, TX 75080; [e]Department of Bioengineering, University of Texas at Dallas, Richardson, TX 75080; [f]Center for Systems Biology, University of Texas at Dallas, Richardson, TX 75080; [g]Department of Physics and Astronomy, Rice University, Houston, TX 77005; and [h]Department of Biochemistry and Cell Biology, Rice University, Houston, TX 77005

1. J. D. Bryngelson, P. G. Wolynes, Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7524–7528 (1987).
2. F. Morcos *et al.*, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).
3. L. Poliseno, Pseudogenes, newly discovered players in human cancer. *Sci. Signal.* **5**, re5 (2012).
4. C. Bekpen *et al.*, Death and resurrection of the human IRGM gene. *PLoS Genet.* **5**, e1000403 (2009).
5. K. Esfeld *et al.*, Pseudogenization and resurrection of a speciation gene. *Curr. Biol.* **28**, 3776–3786 (2018).
6. B. Pei *et al.*, The GENCODE pseudogene resource. *Genom. Biol.* **13**, R51 (2012).
7. C. Sisu *et al.*, Comparative analysis of pseudogenes across three phyla. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 13361–13366 (2014).
8. C. Sisu *et al.*, Transcriptional activity and strain-specific history of mouse pseudogenes. *Nat. Commun.* **11**, 3695 (2020).
9. E. Lerat, H. Ochman, Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res.* **33**, 3125–3132 (2005).
10. K. E. Holt *et al.*, Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars Paratyphi A and Typhi. *BMC Genom.* **10**, 3125–3132 (2009).
11. Y. Feng *et al.*, "Pseudo-pseudogenes" in bacterial genomes: Proteogenomics reveals a wide but low protein expression of pseudogenes in *Salmonella enterica*. *Nucleic Acids Res.* **50**, 3125–3132 (2022).
12. F. Morcos, N. P. Schafer, R. R. Cheng, J. N. Onuchic, P. G. Wolynes, Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 12408–12413 (2014).
13. V. S. Pande, A. Y. Grosberg, T. Tanaka, Statistical mechanics of simple models of protein folding and design. *Biophys. J.* **73**, 3192–3210 (1997).
14. R. Goldstein, Z. A. Luthey-Schulten, P. G. Wolynes, Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 4918–4922 (1992).
15. J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, J. N. Onuchic, Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10340–10345 (2012).
16. B. Jana, F. Morcos, J. N. Onuchic, From structure to function: The convergence of structure based models and co-evolutionary information. *Phys. Chem. Chem. Phys.* **16**, 6496–6507 (2014).
17. A. E. Dago *et al.*, Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E1733–E1742 (2012).
18. R. R. Cheng, F. Morcos, H. Levine, J. N. Onuchic, Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E563–E571 (2014).
19. B. J. Sirovetz, N. P. Schafer, P. G. Wolynes, Protein structure prediction: Making AWSEM AWSEM-ER by adding evolutionary restraints. *Proteins* **85**, 2127–2142 (2017).
20. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009).
21. J. A. de la Paz, C. M. Nartey, M. Yuvaraj, F. Morcos, Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 5873–5882 (2020).
22. A. Davtyan *et al.*, AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B* **116**, 8494–8503 (2012).
23. W. Lu *et al.*, OpenAWSEM with Open3SPN2: A fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations. *PLoS Comput. Biol.* **17**, e1008308 (2021).
24. K. A. Bava, M. M. Gromiha, H. Uedaira, K. Kitajima, A. Sarai, ProTherm, version 4.0: Thermodynamic database for proteins and mutants. *Nucleic Acids Res.* **32**, D120–D121 (2004).
25. A. Couce *et al.*, Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E9026–E9035 (2017).
26. L. Vigué *et al.*, Deciphering polymorphism in 61,157 *Escherichia coli* genomes via epistatic sequence landscapes. *Nat. Commun.* **13**, 4030 (2022).

27. G. Schreiber, A. M. Buckle, A. R. Fersht, Stability and function: Two constraints in the evolution of Barstar and other proteins. *Structure* **2**, 945–951 (1994).
28. B. K. Shoichet, W. A. Baase, R. Kuroki, B. W. Matthews, A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 452–456 (1995).
29. B. M. Beadle, B. K. Shoichet, Structural bases of stability-function tradeoffs in enzymes. *J. Mol. Biol.* **321**, 285–296 (2002).
30. Q. Huai *et al*., Crystal structure of calcineurin–cyclophilin–cyclosporin shows common but distinct recognition of immunophilin-drug complexes. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12037–12042 (2002).
31. A. Selyutina *et al*., Cyclophilin A prevents HIV-1 restriction in lymphocytes by blocking human TRIM5 binding to the viral core. *Cell Rep.* **30**, 3766–3777.e6 (2020).
32. D. U. Ferreiro, J. A. Hegler, E. A. Komives, P. G. Wolynes, Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19819–19824 (2007).
33. L. D. Zydowsky *et al*., Active site mutants of human cyclophilin A separate peptidyl-prolyl isomerase activity from cyclosporin A binding and calcineurin inhibition. *Protein Sci.* **1**, 1092–1099 (1992).
34. M. Lammers, H. Neumann, J. W. Chin, L. C. James, Acetylation regulates Cyclophilin A catalysis, immunosuppression and HIV isomerization. *Nat. Chem. Biol.* **6**, 331–337 (2010).
35. N. N. Soe *et al*., Acetylation of cyclophilin A is required for its secretion and vascular cell activation. *Cardiovasc. Res.* **101**, 444–453 (2014).
36. F. Ferron, G. Rebowski, S. H. Lee, R. Dominguez, Structural basis for the recruitment of profilin-actin complexes during filament elongation by Ena/VASP. *EMBO J.* **26**, 4597–4606 (2007).
37. K. Krishnan, P. D. J. Moens, Structure and functions of profilins. *Biophys. Rev.* **1**, 71–81 (2009).
38. E. J. Schmidt *et al*., ALS-linked PFN1 variants exhibit loss and gain of functions in the context of formin-induced actin polymerization. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2024605118 (2021).
39. E. D. Poggetto, F. Bemporad, F. Tatini, F. Chiti, Mutations of profilin-1 associated with amyotrophic lateral sclerosis promote aggregation due to structural changes of its native state. *ACS Chem. Biol.* **10**, 2553–2563 (2015).
40. R. Geiss-Friedlander, F. Melchior, Concepts in sumoylation: A decade on. *Nat. Rev. Mol. Cell Biol.* **8**, 947–956 (2007).
41. Y. Xu *et al*., Structural insight into SUMO chain recognition and manipulation by the ubiquitin ligase RNF4. *Nat. Commun.* **5**, 4217 (2014).
42. A. Gärtner *et al*., Acetylation of SUMO2 at lysine 11 favors the formation of non-canonical SUMO chains. *EMBO Rep.* **19**, e46117 (2018).
43. C. Clementi, S. S. Plotkin, The effects of nonnative interactions on protein folding rates: Theory and simulation. *Protein Sci.* **13**, 1750–1766 (2004).
44. H. Kaya, H. S. Chan, The effects of nonnative interactions on protein folding rates: Theory and simulation. *Proteins* **40**, 637–661 (2000).
45. J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, N. D. Socci, Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3626–3630 (1995).
46. S. Miyazawa, Selection originating from protein stability/foldability: Relationships between protein folding free energy, sequence ensemble, and fitness. *J. Theor. Biol.* **433**, 21–38 (2017).
47. S. F. Altschul *et al*., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
48. F. Madeira *et al*., Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279 (2022).
49. J. Mistry *et al*., Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
50. S. R. Eddy, Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
51. R. G. Parra *et al*., Protein Frustratometer 2: A tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Res.* **44**, W356–W360 (2016).
52. H. Jaafari *et al*., Data and Software: "The Physical and Evolutionary Energy Landscapes of Devolved Protein Sequences Corresponding to Pseudogenes". Zenodo. https://doi.org/10.5281/zenodo.10783361. Deposited 5 March 2024.