Formation of native shortcut networks and two-state protein folding

Susan Khor

February 15, 2019

Abstract

A dynamic network-centric approach to study two-state folding from native structure called network dynamics (ND) is introduced. ND applies two fundamental principles of protein folding: hydrophobicity and loop-entropy, on a protein's native residue network (PRN0) to generate its native shortcut network (SCN0). ND generates barrier heights that correlate significantly (-0.7) with folding rate, and positions the transition-state (TS) for 52 proteins within $0.1 \leq Q < 0.5$ of its reaction coordinate, which monitors SCN0 completion. ND trajectories through SCN0 space are reasonable; they support our previous work on identifying initial fold steps. Both relative contact order (RCO) and network clustering coefficient *C*, computed on ND generated SCN0s, correlate significantly with all three folding kinetic variables: folding rate, TS placement and native-state (NS) stability. Some of these correlations are stronger or become significant when computed on partial than on complete SCN0s. In particular, contrary to previous findings, RCO can correlate significantly with NS stability. This revelation is made possible by ND, which provides a computationally light way to explore partially folded proteins via incomplete SCN0s. ND analysis affirms the presence of NS topology in its TS structures, and finds *C* to be a better measure of protein topology than RCO, to shed light on the hypothesized relationship between NS structure and folding rate. Within the ND-TS region, C_SCN0 correlates significantly with folding rate, while SCN0_RCO does not. However, strong TS NS structural correlations, in terms of both SCN0_RCO and C_SCN0, is also producible by a randomized version of ND.

1. Introduction

A protein's native shortcut network (SCN0) is pertinent to protein folding [Khor18]. Single-value descriptors calculated on SCN0, such as relative contact order (RCO) [Plaxco98] and clustering coefficient *C*, correlate significantly with folding rate, transition-state (TS) placement and native-state (NS) stability of proteins that fold in a two-state manner. SCN0s are information rich contact maps from which plausible folding pathways can be identified.

The unique mix of short- and long-range (in terms of sequence separation) contacts that make up a SCN0 is identified by a local search algorithm called Euclidean Distance Search (EDS) [Khor16], which stems from the concept of a small-world in social networks. EDS uses Euclidean distance information from the placement of amino acids (3D coordinates) in a protein's native structure, and the connectivity of the amino acids in a protein residue network (PRN0), to find EDS paths between pairs of amino acids.

Due to its limited view of the underlying PRN0, the choice of next hop in a path is locally optimal but not necessarily globally advisable. Therefore, EDS may need to backtrack when it encounters a dead-end. Shortcut edges are PRN0 edges that help EDS avoid this backtracking cost. A shortcut contact is always part of a PRN0 cycle.

Such dependency of shortcut contacts on the position of multiple amino acids aligns with the importance of many-body interactions to protein folding cooperativity [Chan11, Portman10], but complicates the effort of a direct approach to SCN0 formation. Moreover, it is PRN0 node degree and not SCN0 node degree that is more indicative of residue burial (sections 3.1 & 4.2). Hence, SCN0 formation is entwined with PRN0 formation in the background, and we take this PRN0→SCN0 approach.

PRN0 (SCN0) formation is via a network generation model called Network Dynamics (ND) (section 2.2), which applies directly two fundamental drivers of protein folding: (i) hydrophobicity, and (ii) loop entropy. The PRN0→SCN0 approach, as opposed to an approach based solely on SCN0, affords two advantages: (i) the relationship between network characteristics of PRN0 and SCN0, namely the influence of key folding residues on folding kinetics, can be tested, and (ii) non-native shortcuts, but still native contacts, (these are PRN0 edges identified as shortcuts in partially formed PRN0s, but are not SCN0 edges) can be detected. Inclusion of non-native contacts injects ruggedness into the minimally frustrated landscape of SCN0 formation projected by the ND model. This feature is crucial for calculating the TS energy barrier and for locating the TS within our reaction coordinate (section 2.4).

2. Method

2.1 Preliminaries

PRN0 and SCN0 construction from a protein's NS 3D coordinates, and the network characteristics thereof, were described in [Khor16]. Both PRN0 and SCN0 are simple undirected single component graphs.

Node degree is the number of edges incident on a node. Hub nodes are nodes with high connectivity, i.e. large (conventionally, larger than average) node degree. Node clustering measures the extent to which the direct neighbors of a node form a complete graph. C_SCN0 denotes the clustering coefficient of a SCN0; it is the average node clustering of all nodes in a SCN0 computed on the SCN0. Identification of folding pathways using C_SCN0 was demonstrated in [Khor18]. SCN0_RCO is the contact order (average sequence separation) of SCN0 edges divided by protein chain length, *N*. PRN0_RCO is similarly computed but on a PRN0. SCN0 triangles refer to native shortcut cycles of length three.

2.2 The Network Dynamics (ND) model

Initial condition and node selection. ND begins with a PRN0 devoid of its edges, and each node

endowed with its maximum contact potential (node degree in the complete PRN0). At time each step, nodes with larger potential are selected preferentially to make contacts; nodes lose their potential in proportion to the PRN0 edges they actualize. In this way, the PRN0 hub nodes act as key folding residues (typically hydrophobic), whose influence over ND node selection wanes as they become more buried in a protein's core.

Network growth and contact creation. PRN0 edges of a selected node are percolated with probability p, where p is the natural logarithm of an edge's sequence distance, scaled to [0.0, 1.0] range. EDS is run to find shortcuts after ND has tried to actualize every as-yet latent edge incident on a selected node. The shortcuts identified by EDS which are also shortcuts in the complete PRN0 compose a ND generated SCN0 for the ND generated PRN0. The number of PRN0 edges increases monotonically over simulation time, but this is not necessarily so for SCN0 edges. It is possible for a SCN0 edge that appears at one time step to disappear in a future time step due to changes in PRN0 edges.

Terminating condition. ND terminates upon SCN0 completion. At ND termination, the underlying PRN0 may not be complete (typically it is very nearly complete); as such, the presence of non-native shortcuts is still possible.

Reaction coordinate and ND time. Progress of a ND trajectory is measured by Q, the proportion of SCN0 edges formed. Observations of a ND trajectory are made at regular intervals of 0.1, i.e. Q = 0.1, 0.2, 0.3, ..., 0.9, 1.0. $Q = z$ means when Q equals or first exceeds z, e.g. Q = 0.4 means $0.4 \leq Q < 0.5$. ND time at $Q = z$ is the number of times node selection is performed to advance a ND trajectory from the start of simulation to when $Q = z$.

2.3 Evaluation

We test three variants of the ND model: (i) RNRE: random node selection ($p_n = 1/N$) with random edge creation (p = 0.5), (ii) RNPE: random node selection with probabilistic edge creation ($p = 1/ln(|x\text{-}y|)$) where $|x\text{-}y|$ is the sequence distance between residues x and y, and (iii) DNPE: dynamic degree node selection with probabilistic edge creation. DNPE is the default ND variant (described in section 2.2). RNPE tests the influence of PRN0 hub nodes on SCN0 formation. RNRE is the null model; it assumes random folding, i.e. the order in which native contacts are made is irrelevant.

ND statistics for each test protein are averaged over 100 independent runs. In the case of proteins with multiple NMR models, 100 independent runs per model are made, and the ND statistics are averaged over all those runs. Pearson (linear) correlation is used throughout. Unless stated otherwise, a p-value < 0.05 is required for statistical significance.

At minimum, DNPE should outperform RNRE, and produce the expected significant correlations with a non-trivial set of experimental protein kinetic data [Henry04]. The validity of the ND model is

assessed with the experimental folding rates of 52 two-state proteins (15α, 16αβ, and 21β) [Kaya13], two-state TS placement for 23 proteins (5α, 7αβ and 11β) [Micheletti03], and NS stability measures for 23 proteins (4α, 9αβ and 10β) [Weikl03]. The two last datasets were also used in [Khor18], which examined correlations between protein folding kinetics and structural metrics computed on complete SCN0s only. In contrast, the ND model provides an opportunity to re-examine these relationships with incomplete SCN0s.

At a more detailed level, ND should also support the selection of the initial fold step on C_SCN0 folding pathways identified previously for several small model proteins [Khor18], and provide evidence of two-state folding behavior. The latter is accomplished by analyzing energy profiles of ND generated SCNs (section 2.4).

2.4 Energy of ND generated SCNs and TS placement on the Q reaction coordinate.

A ND generated SCN is a set of edges identified by EDS as shortcuts from a ND generated PRN0. A ND generated SCN differs from a ND generated SCN0 in that the former includes non-native shortcut edges, while the latter excludes them.

Energy of a ND generated SCN E, is calculated using inter-residue potentials from the upper triangle of Table 3 in [Miyazawa96], and are in RT units. For a ND generated SCN with f native shortcuts and g non-native shortcuts, E is the sum of the inter-residue potential of the f native shortcuts minus the sum of the inter-residue potential of the g non-native shortcuts. The presence of non-native shortcuts raises E, while the presence of native shortcuts lowers E.

The energy profile for a ND trajectory is composed of the E values recorded at the 10 Q points of observation. This data, averaged over 100 independent ND trajectories, makes the energy profile for a protein. The energy profiles (of proteins) typically have a single peak. The maximum E value of an energy profile is recognized as the height of the (free) energy barrier separating the denatured and NS structures (assumes energy of denatured structures is 0 since ND starts with an empty SCN). The few downhill (maximum E is at $Q = 0.1$) energy profiles are typically for α-helix proteins.

The Q where E peaks situates the TS for a protein on our reaction coordinate. The SCN0s generated by ND at the Q associated with a protein's barrier height (*bh*) are deemed to capture the essential native contacts[1] present in the protein's transition-state ensemble (TSE).

[1] A preliminary test with reconstruct [Lappe10] finds PRN0/SCN0 contact maps less effective (in terms of minimizing RMSD to native structure) as distance constraints for generating native structures than "cone-peeled" [Lappe09] native contact maps. These two types of contact maps have different objectives, whose simultaneous satisfaction may be a conflict. The cavity preserving feature of PRN0/SCN0 contact maps [Khor16] could be a disadvantage for the structure prediction task. Are the partially generated PRN0s/SCN0s useful as distance constraints to generate TSE structures? What additional support would be helpful? This could be a topic for future work.

3. Results

3.1 PRN0 hub nodes as key folding residues

There is interest in identifying key folding residues from native protein structures through computational means to discern critical factors for productive folding [Shmy05]. Hub nodes are natural candidates in the view of protein folding driven by burial of hydrophobic amino acids. However, this intuition has not been fully realized, and other network metrics such as node betweenness [Vend02] and node closeness [Li08], both of which involves computing all-pairs shortest paths, have been proposed as substitutes instead.

Here we find that, with the PRN0 representation of native structures, this intuitive idea need not be abandoned. First, there is a strong positive correlation (0.84) between PRN0 node degree and burial of residues (Fig. A1). Second, taking a cue from the positive results of rigidity analysis [Thomas07], and the characterization of rigid nodes in a network as those with larger degrees (more globally connected) and smaller clustering coefficients (less locally connected) [Piazza08], we observe that PRN0 hub nodes, i.e. nodes with larger than average node degree, are enriched with key folding residues and their nucleation interactors (Figs. A2 to A12). Third, preferentially selecting nodes with larger PRN0 degree yields a significant positive impact on ND (section 3.2).

3.2 Correlation with folding rate of two-state proteins

DNPE time correlates significantly with experimental folding rate of 52 two-state proteins, and is strongest (-0.76) at Q = 0.4 (Fig. 1 top-left). DNPE produces stronger correlations than RNRE, thereby demonstrating the two principals of protein folding: hydrophobicity and loop-entropy, at work in ND. SCN0 formation is positively affected by preferentially selecting nodes with larger PRN0 degree to form contacts, and preferentially creating local (short-range) over global (long-range) contacts.

When nodes are selected uniformly at random as in RNPE, ND time correlation with folding rate weakens slightly, but significantly (one sided paired t-test p-value = 0.03). Similar weakening is observable also with Spearman and Kendall correlations (Fig. B1). This difference highlights the influence PRN0 node degree wields on SCN0 formation. NS PRN0 node degree and NS SCN0 node degree are moderately, though significantly, correlated (Table C1).

Both DNPE and RNPE generate barrier heights that correlate significantly and strongly (≈ -0.7) with experimental folding rate (Table 1). This result shows that ND is capturing salient aspects of two-state folding kinetics. The contrast with RNRE barrier heights, which barely have a significant correlation with folding rate, demonstrates the importance of loss of chain entropy in determining the free energy barrier to folding.

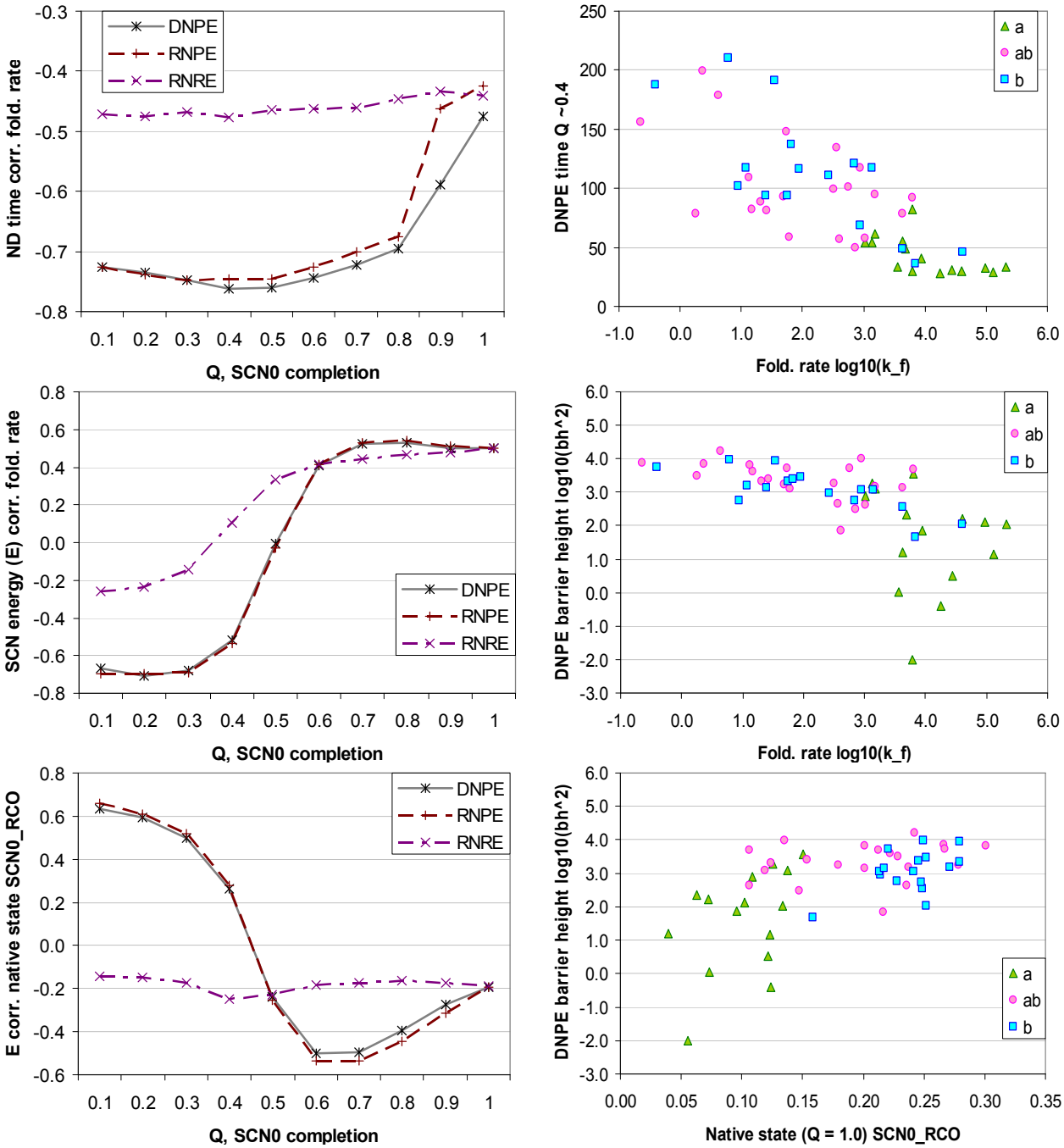


Fig. 1 Top-left: Correlation between ND time (averaged over 100 independent runs per protein) and folding rate ($\log_{10}$) of 52 two-state proteins. All correlations have p-value < 0.002. **Top-right:** Scatter of DNPE time at $Q = 0.4$ against folding rate, by fold type. **Middle-left:** Correlation between energy of ND generated SCNs E (averaged over 100 independent runs per protein) and folding rate ($\log_{10}$) of 52 two-state proteins. Only negative correlations are considered since they are meaningful. DNPE and RNPE correlations have p-value < 0.003 for $Q \leq 0.4$; RNRE correlations are insignificant. **Middle-right:** Dispersion of DNPE barrier heights against folding rate, by fold type. α-helix (a) proteins occupy the widest barrier height range. **Bottom-left:** Correlation between energy of ND generated SCNs E (averaged over 100 independent runs per protein) and NS SCN0_RCO. Only positive correlations are considered since they are meaningful. DNPE and RNPE correlations have p-value < 0.0002 for $Q \leq 0.3$; RNRE correlations are insignificant. **Bottom-right:** Dispersion of DNPE barrier heights against NS SCN0_RCO, by fold type.

Table 1 Barrier height statistics gathered by ND from the 52 two-state proteins.

ND model	(1) log10 kf corr. *bh*	(2) *bh* at *Q*=0.1	(3) log10 kf corr. log10(*bh*^2)	Range of log10(*bh*^2)	Mean *Q*	Range of *Q*
DNPE	-0.700 (4) 0	3α	-0.612 0	6.218	0.244 (5) ± 0.067	0.1 ... 0.4
RNPE	-0.701 0	2α	-0.654 0	5.443	0.258 ± 0.064	0.1 ... 0.4
RNRE	-0.239 0.087	1α 4αβ 6β	-0.283 0.042	1.940	0.179 ± 0.041	0.1 ... 0.2

Table 1 continued

ND model	Native state (*Q* = 1.0) SCN0_RCO corr. log10(bh^2)	log10(*bh*^2) corr. *N*
DNPE	0.579 0	0.575 0
RNPE	0.608 0	0.569 0
RNRE	-0.139 0.331	0.703 0

(1) Correlation between logarithm of folding rate and barrier height *bh*.
(2) Instances where an energy profile peaks at *Q* = 0.1.
(3) The square of *bh* is used to enable taking the logarithm of negative values.
(4) p-value of a (Pearson) correlation is printed below its coefficient.
(5) one standard deviation from the mean.

From the 52 two-state protein study, both DNPE and RNPE locate the TS within $0.1 \leq Q < 0.5$ (Table 1). Prior to $Q = 0.5$, the energy of both DNPE and RNPE generated SCNs correlates significantly and negatively with folding rate (Fig. 1 middle-left); this correlation is strongest around the mean *Q* value (0.25) identified with barrier heights (Table 1). Two-state proteins with lower free energy barriers fold faster; the ND model is able to express this principal via these negative correlations.

The barrier heights of both DNPE and RNPE correlate significantly and positively (≈ 0.57) with *N* (Table 1), in accord with the intuition that longer protein chains fold more slowly. Folding rate of the 52 two-state proteins correlates with *N* with a strength of 0.435 (p-value = 0.001). The correlation between RNRE barrier height with *N* is stronger, but it deviates from the empirical correlation (0.27 = 0.70 - 0.43) two times more than the correlation for DNPE (0.14 = 0.57 - 0.43) does.

Both DNPE and RNPE produce barrier heights that correlate significantly (≈ 0.6) with NS SCN0_RCO (Table 1). Prior to $Q = 0.4$, the DNPE and RNPE generated barrier heights correlate significantly and positively with NS SCN0_RCO, while the RNRE barrier heights do not (Fig. 1 bottom-left). This positive correlation follows since faster folders are associated with smaller NS RCOs. However, comparable positive correlations can also be obtained with RCO computed on RNRE generated SCN0s (section 4.1).

Both DNPE and RNPE produce a wider range of barrier height values than RNRE (Table 1, Fig. B2). Nonetheless, their ranges are only about half the six orders of magnitude spanned by the folding rates of two-state proteins. Increasing the diversity of simulated folding rates [Portman10, Kaya13] is an area for future work. The approach by [Ejtehadi04] in particular, may be relevant since the proportion of SCN0 triangles completed by ND around $Q = 0.4$, averages at 20% (Fig. B3).

Summary. ND has satisfied the minimum requirement for a protein folding model. It produces expected significant correlations with a non-trivial set of experimental folding rate data, and outperformed the null model, RNRE. It also captures a critical aspect of two-state folding: its barrier heights correlate significantly and negatively with experimental folding rate. Next, we delve deeper into the behavior of DNPE (the default ND model) within the ND identified TS region ($0.1 \leq Q < 0.5$).

3.3 ND trajectory supports C_SCN0 initial fold step

[Khor18] used C_SCN0 to identify folding pathways for several model proteins from their native structure. Here, we examine whether the C_SCN0 folding bias exhibited by a native structure, match those expressed by ND generated SCN0s. Such an agreement gives assurance that ND is traversing SCN0 space (where each point is some combination of native shortcuts) in a meaningful way.

Of particular interest is the initial step or first secondary structure element pairing on a C_SCN0 folding pathway, since the presence of these substructures, which are supported empirically, signal nascent productive folding activity. Let P_{init} be the proportion of ND generated SCN0s at each Q where the initial step matches that expected from the native structure (Table C1). P_{init} is inspired by the P_{fold} notion used to identify TS structures, which asserts that a two-state protein has equal chance of folding or unfolding (P_{fold} = 50%) at its TS [Faisca09]. Hence, our focus on Q where P_{init} first equals or exceeds 0.5.

In general, the initial fold step made by ND generated SCN0s agrees with that selected by the native structure. Except for bad models (the initial fold step chosen by their native structure disagrees with the expected), P_{init} increases as SCN0 reaches completion (Fig. C1). For all but five proteins (2PTL, 1MHX, 1MI0, 2CI2, 1QYS), $P_{\text{init}} \geq 0.5$ as Q approaches 0.5 (Table 2), which is within the ND-TS region ($0.1 \leq Q < 0.5$). These observations support the claim that ND makes reasonable trajectories in SCN0 space; DNPE is "edge-percolating" SCN0s in a reasonable order.

P_{init} for 2PTL remains below 0.5 until $Q = 0.9$, but the P_{init} over good 2PTL models start to diverge from the P_{init} over bad 2PTL models starting from $Q = 0.3$ (Fig. C1). 1MHX and 1MI0 are protein G (1GB1) mutated to fold like protein L (2PTL). Interestingly, their P_{init} plots resemble 2PTL more than 1GB1. 2CI2, the quintessential nucleation-condensation (NC) model, achieves $P_{\text{init}} \geq 0.5$, a little later, at $Q = 0.5$. This delay could be due to the more distributed nature of its folding mechanism[2]. 1QYS is a small non-two-state protein; 1QYS achieves $P_{\text{init}} \geq 0.5$ later still, at $Q = 0.6$.

RNRE trajectories also reach $P_{\text{init}} \geq 0.5$ within the ND-TS region for most proteins (Table 2), but they do not generate meaningful barrier heights (section 3.2). The DNPE barrier heights for these 13

[2] We note a similarity, due to the presence of parallel β-strands in CI2, between the contact map of CI2 and the HIFF-II [Khor07] interaction scheme, which because of inter-level conflict, is a challenge to solve with a bottom-up approach, but requires "cooperation" from different levels to avoid local minima.

proteins correlate significantly (0.680, p-value = 0.011) with NS SCN0_RCO; here NS SCN0_RCO is used as a substitute for experimental folding rate. The RNRE Q P_{init} values indicate that a measure like P_{init} is not selective enough to distinguish between SCN0s generated in a non-random (probabilistically biased) versus random (probabilistically unbiased) manner. This finding has negative implications for the application of a measure like P_{init} to identify structures for inclusion in a TSE. Further, P_{init} is unable to situate a protein's TS within the Q reaction coordinate; typically Q where $P_{init} \geq 0.5$ is larger than Q associated with the barrier height for a given protein (Table 2).

Table 2 Critical Q values.

PDB id		1BDD	2ABD	2PTL	1QYS	1MI0	1MHX	1GB1
(1) |SCLE|		6	27	27	55	28	34	32
NS SCN0_RCO		0.084	0.108	0.118	0.142	0.164	0.170	0.201
(2) Q bh	DNPE	0.2	0.3	0.3	0.2	0.3	0.3	0.2
(3) Q P_{init} = 0.5	DNPE	0.39	0.41	0.86	0.61	0.71	0.65	0.44
	RNRE	0.39	0.74	0.57	0.33	0.32	0.64	0.58
	(4) DNPE*	0.39	0.41	0.91	0.68	0.81	0.72	0.44

Table 2 continued

PDB id		2KJW	2KJV	1SHG	1APS	1SRM	2CI2
(1) |SCLE|		67	66	43	79	46	46
NS SCN0_RCO		0.227	0.231	0.247	0.267	0.268	0.278
(2) Q bh	DNPE	0.3	0.3	0.3	0.3	0.2	0.3
(3) Q P_{init} = 0.5	DNPE	0.23	0.30	0.23	0.36	0.43	0.51
	RNRE	0.32	0.41	0.37	0.47	0.54	0.45
	(4) DNPE*	0.24	0.31	0.23	0.38	0.48	0.61

(1) Number of long-range edges in NS SCN0s.
(2) Q associated with DNPE barrier height (bh).
(3) Q associated with P_{init} = 0.5 for DNPE and for RNRE trajectories, found by linear regression (over all good models) from plots in Fig. C1.
(4) DNPE* ignores long-range native shortcuts identified at each Q when computing P_{init} for a Q (section 6).

4. Structural correlations

The relationships between two-state folding kinetic variables and protein structure descriptors: RCO and C computed on ND generated SCN0s, are examined in this section.

4.1 Folding rate

Both RCO and C were shown to correlate significantly with folding rate of two-state proteins when computed on complete (NS) SCN0 [Khor18]. However, when computed on partial (ND generated) SCN0s, RCO and C exhibit *qualitatively* different behaviors. Crucially, in the ND-TS region ($0.1 \leq Q < 0.5$), SCN0_RCO does not correlate significantly with folding rate, while C_SCN0 correlates more

strongly with folding rate than NS C_SCN0 (Fig. 2 left). RCO and *C* also behave differently in the randomized ND variant, RNRE.

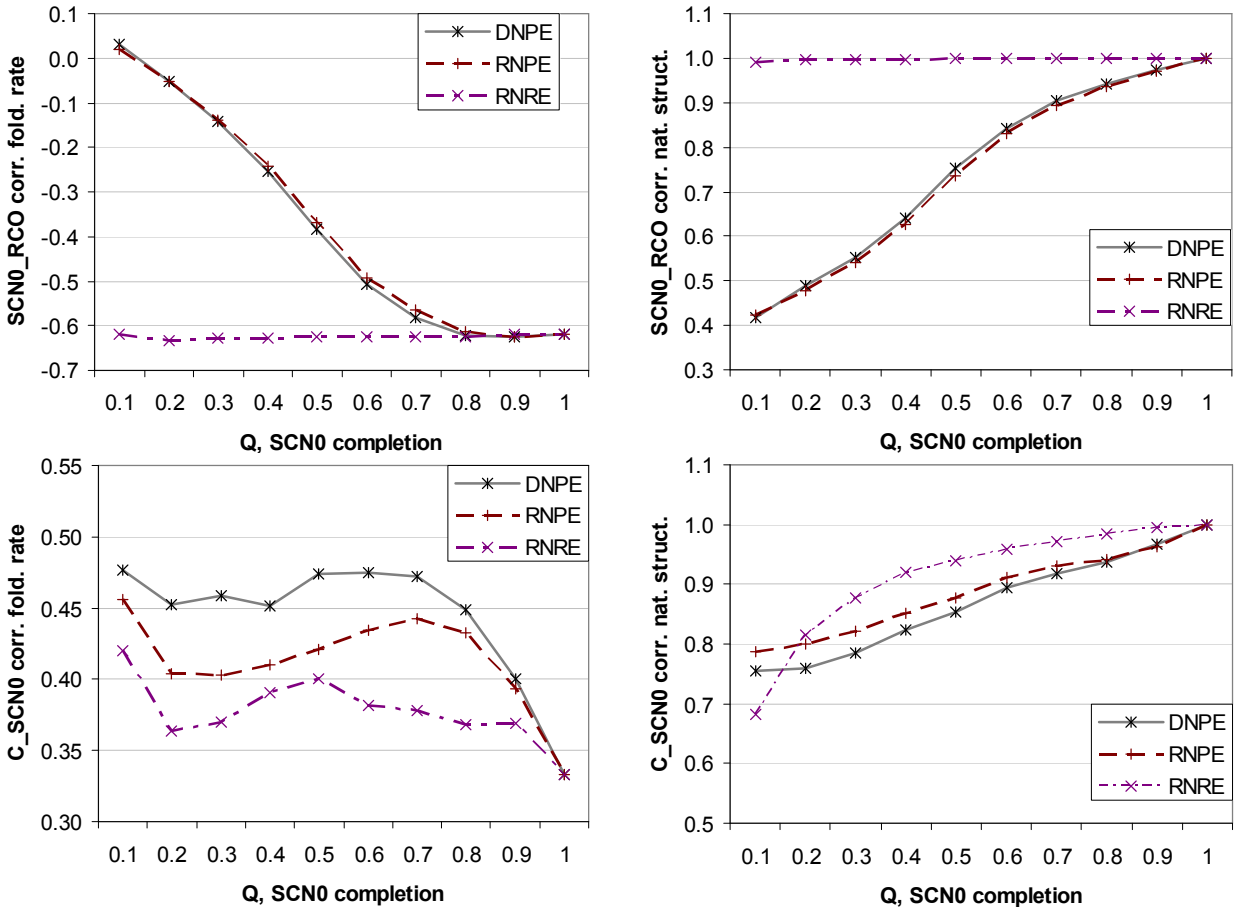


Fig. 2 Top-left: Correlation between RCO of ND generated SCN0s (averaged over 100 independent runs per protein) and folding rate ($\log_{10}$) of 52 two-state proteins. Both DNPE and RNPE correlations have p-value < 0.007 for $Q \geq 0.5$; RNRE correlations have p-value ≈ 0; all other correlations have p-value > 0.05. **Top-right:** Correlation between native-state (NS) RCO of 52 two-state proteins and the RCO of their respective ND generated SCN0s (averaged over 100 independent runs per protein). All correlations have p-value < 0.002. **Bottom-left:** Correlation between *C* of ND generated SCN0s (averaged over 100 independent runs per protein) and folding rate ($\log_{10}$) of 52 two-state proteins. All correlations for $Q < 1.0$ have p-values < 0.008. Correlation at $Q = 1.0$ has p-value = 0.016. **Bottom-right:** Correlation between NS C_SCN0 of 52 two-state proteins and the C_SCN0 of their respective ND generated SCN0s (averaged over 100 independent runs per protein). All correlations have p-value ≈ 0.

RCO for RNRE generated SCN0s are almost perfectly correlated with NS SCN0_RCO (Fig. 2 top-right), and show uniformly strong correlations with folding rate for all *Q* (Fig. 2 top-left). In contrast, the correlation between RCO for DNPE generated SCN0s and NS SCN0_RCO increases in strength with increase in *Q*. This gradual increase in RCO reflects the slow loss in chain-entropy during protein folding [Weikl03], and attests to the non-random formation of DNPE SCN0s. RNRE, where PRN0 edges have equal chance of forming regardless of their length (sequence distance), is suggestive of protein folding being a random affair. The correlation between DNPE SCN0_RCO and folding rate start to be significant only after the ND-TS region, i.e. $Q \geq 0.5$. In short, these observations tell us that unless there is some

random element in the formation of SCN0s, ND-TS SCN0_RCO does not correlate with folding rate even if it correlates significantly with NS SCN0_RCO.

C for RNRE generated SCN0s correlates with NS C_SCN0 more strongly than *C* for DNPE generated SCN0s (Fig. 2 bottom-right). Both DNPE and RNRE C_SCN0 correlate significantly with folding rate for all *Q*, but RNRE C_SCN0 correlates with folding rate significantly less strongly than DNPE C_SCN0 (Fig. 2 bottom-left). In short, these observations tell us that C_SCN0 within the ND-TS region ($0.1 \leq Q < 0.5$) correlates significantly with folding rate (random folding not required and not desirable). Further, ND-TS region C_SCN0 correlates more strongly with folding rate than NS C_SCN0.

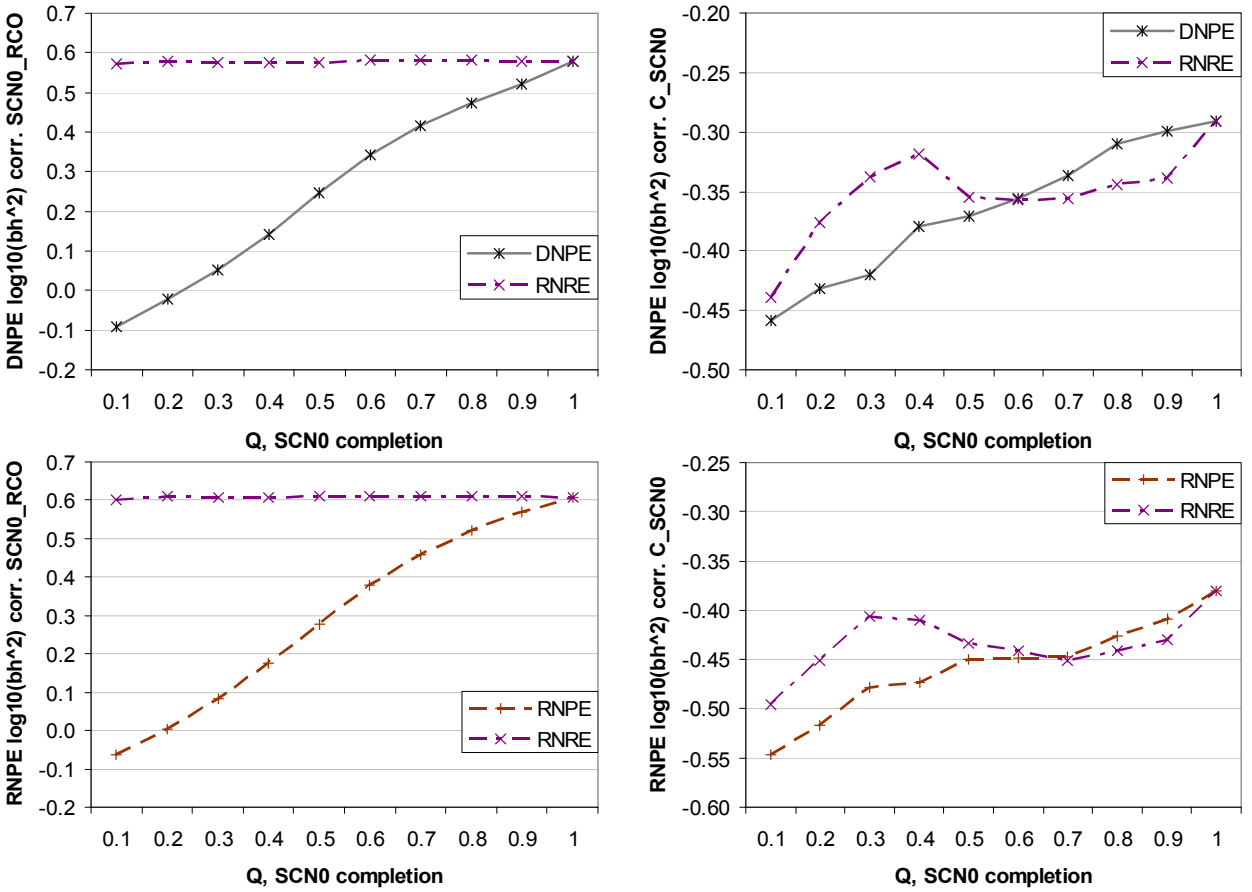


Fig. 3 A log10(bh^2) value (bh=barrier height) is the maximum *E* in a protein's energy profile, squared to permit the logarithm of negative values. **Top-left:** Correlation between DNPE generated barrier heights for 52 two-state proteins, and RCO of DNPE and of RNRE generated SCN0s (averaged over 100 independent runs). DNPE correlations have p-values < 0.013 when $Q \geq 0.6$; all RNRE correlates have p-values ≈ 0; all other correlations have p-values > 0.081. **Top-right:** Correlation between DNPE generated barrier heights for 52 two-state proteins, and *C* of DNPE and of RNRE generated SCN0s (averaged over 100 independent runs). DNPE correlations have p-values < 0.007 when $Q \leq 0.5$, and p-values > 0.010 but < 0.036 when $Q > 0.5$. All RNRE correlations have p-values < 0.036. **Bottom-left:** Correlation between RNPE generated barrier heights for 52 two-state proteins, and RCO of RNPE and of RNRE generated SCN0s (averaged over 100 independent runs). RNPE correlations have p-values < 0.005 when $Q \geq 0.6$; all RNRE correlates have p-values ≈ 0; all other correlations have p-values > 0.047. **Bottom-right:** Correlation between RNPE generated barrier heights for 52 two-state proteins, and *C* of RNPE and of RNRE generated SCN0s (averaged over 100 independent runs). All RNPE and all RNRE correlations have p-values < 0.005.

The abovementioned qualitative differences between RCO and *C* are observable too with both DNPE and RNPE generated barrier height data (a.k.a. simulated folding rate) from section 3.2. DNPE barrier heights correlate significantly and uniformly strongly with RNRE SCN0_RCO for all *Q*, but only begin to correlate significantly with DNPE SCN0_RCO when $Q \geq 0.6$ (Fig. 3 top-left). DNPE barrier heights correlate significantly with both DNPE and RNRE C_SCN0 for all *Q*, but within the ND-TS region, the correlation with DNPE C_SCN0 is stronger than with RNRE C_SCN0 (Fig. 3 top-right).

Analogous observations with RNPE barrier heights (Fig. 3 bottom) show that the qualitative differences between RCO and *C* uncovered here hold even when the role of PRN0 hub nodes as key folding residues hypothesized in section 3.1 is rejected. For the 52 two-state proteins, the RNPE version of ND produces significantly stronger correlations between barrier height and C_SCN0 than DNPE.

4.2 Transition-state placement

Transition-state placement (θ_m) quantifies TS structures in terms of the extent to which they bury solvent-accessible surface areas buried in their respective NS structures [Plaxco98]. TS structures with θ_m values closer to 1.0 are more native-like in their compactness.

There is a negative correlation between θ_m and folding rate (-0.439, p-value= 0.036), which is not, as we stated previously, counter-intuitive. Due to their higher NS-TS similarity, proteins with larger θ_m values would have more long-range contacts in their TS, and be slower folders. This intuition is borne out by viewing the θ_m values by fold type (Fig. 4 top-left); this plot also shows θ_m's positive relationship with RCO, and negative relationship with *C*. The DNPE barrier heights for this dataset ($Q = 0.270 \pm 0.070$) correlate significantly (-0.563, p-value = 0.005) with folding rate, but not with θ_m (0.175).

However, the number of long-range contacts in NS SCN0s (|SCLE|) is not a determinant of θ_m. [Khor18] explains that this is possibly due to size effect; unlike NS SCN0_RCO and NS C_SCN0, |SCLE| is significantly correlated with *N*, whereas TS placement is size independent, but dependent on topological complexity of the NS structure [Plaxco98]. And indeed, the correlation between θ_m and *N* is -0.347 (p-value = 0.105) for our 23 point dataset, and θ_m correlates significantly with both NS SCN0_RCO and NS C_SCN0 [Khor18]. But we find this size-independence explanation for |SCLE| unconvincing since |SCSE|, the number of short-range contacts in NS SCN0s, is strongly correlated (0.758) with *N*, and yet also significantly correlated (-0.526, p-value = 0.010) with θ_m. Nonetheless, the negative sign of this correlation does lend indirect support to a positive relationship between long-range contacts and θ_m.

RCO for DNPE generated SCN0s is more strongly correlated with θ_m than RCO for RNRE generated SCN0s (Fig. 4 bottom-left). This indicates that the order in which SCN0 forms is relevant also for TS placement on a surface area burial scale. The DNPE SCN0_RCO correlation with θ_m is strongest when

$0.6 \leq Q \leq 0.9$. This reflects the presence of more long-range edges in the ND generated SCN0s, and their role in protein packing, i.e. increasing the (global) compactness of structures formed within the ND-TS region.

RCO for ND generated PRN0s exhibit a similar pattern of correlation with θ_m as RCO for ND generated SCN0s, but the PRN0_RCO correlations are stronger (Fig. 4 bottom-right). This follows since θ_m places TS structures on a reaction coordinate that monitors burial of hydrophobic surface area, and PRN0 node degree correlates more strongly with burial of residues than SCN0 node degree (Fig. A1).

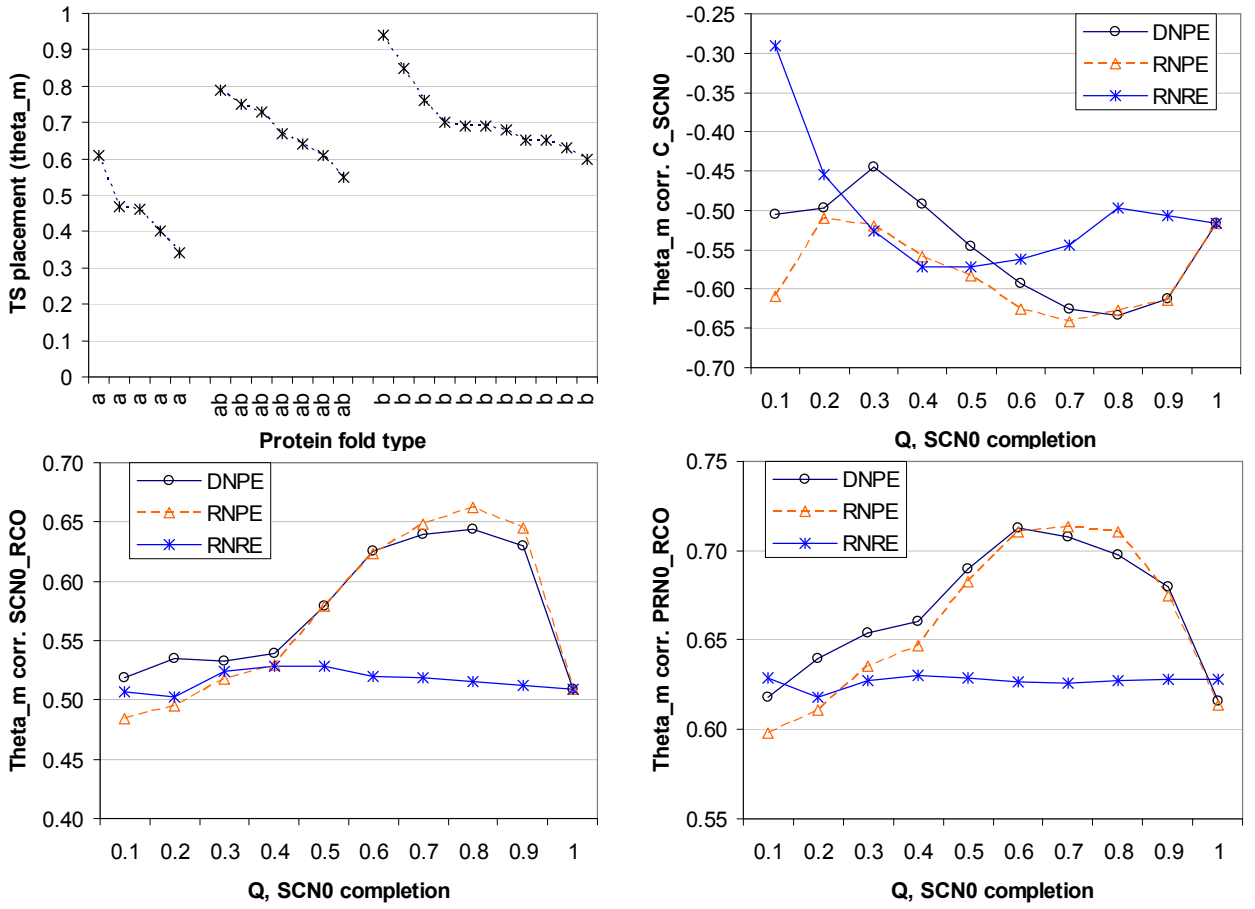


Fig. 4 Top-left: TS placement (θ_m) data for 23 two-state proteins by fold type. **Top-right:** Correlation between *C* of ND generated SCN0s (averaged over 100 independent runs per protein) and θ_m. DNPE, RNPE and RNRE correlations have p-value < 0.007 for $0.5 \leq Q \leq 0.9$, $0.4 \leq Q \leq 0.9$, and $0.4 \leq Q \leq 0.7$, respectively. Except the RNRE correlation at Q=0.1 which has a p-value > 0.1, all other correlations have p-value < 0.04. **Bottom-left:** Correlation between RCO of ND generated SCN0s (averaged over 100 independent runs per protein) and θ_m. DNPE correlations have p-value < 0.009 for $0.2 \leq Q \leq 0.9$; RNPE correlations have p-value < 0.009 for $0.4 \leq Q \leq 0.9$; all other correlations have p-value < 0.02. **Bottom-right:** Correlation between RCO of ND generated PRN0s (averaged over 100 independent runs per protein) and θ_m. All correlations have p-value < 0.002.

Within the abovementioned post ND-TS but pre-NS region ($0.6 \leq Q \leq 0.9$) is also where *C* for DNPE generated SCN0s correlate most strongly with θ_m, and is not outperformed by RNRE C_SCN0 (Fig. 4 top-right). However, unlike SCN0_RCO, DNPE C_SCN0 performs more poorly than RNRE

C_SCN0 in terms of correlation strength with θ_m, within the ND-TS region. RNPE produces the best overall C_SCN0 correlations with θ_m; it skirts the bottoms of both DNPE and RNRE plots.

4.3 Stability

NS stability (ΔG) quantifies how much free energy is required to unfold a NS protein. Although proteins with more stable native structures are expected to fold faster, the linear relationship between ΔG and folding rate is not evident [Plaxco00, Dinner01]. The correlation is 0.185 for our 23 point dataset.

From a static analysis of NS structures, [Khor18] proposed that the positive relationship between ΔG and folding rate can be demonstrated via an intermediary variable, C_SCN0, but not SCN0_RCO. Here, we confirm that C_SCN0 is able to fill this intermediary role. C_SCN0 correlates with both folding rate and NS stability in the ND-TS region, demonstrating that the native contacts which underlie the calculation of C_SCN0, are determinants of both NS stability and the folding barrier. The DNPE barrier heights ($Q = 0.248 \pm 0.059$) for this dataset correlate significantly (-0.783) with folding rate. Our analysis of ND generated SCN0s also reveals that while SCN0_RCO can correlate with NS stability, it does so only within the ND-TS region. But because SCN0_RCO does not correlate significantly with folding rate in the ND-TS region (section 4.1), SCN0_RCO is unable to play this intermediary role.

[Khor18] reported a significant correlation between NS C_SCN0 and stability, but only after excluding two outliers (1lmb and 1urn) from the dataset. Fig. 5 top-left shows how *C* for ND generated SCN0s correlate with stability on this reduced (21 points) dataset. In contrast to the RNRE correlations, both DNPE and RNPE correlations on partial SCN0s show little variance from the NS C_SCN0 correlation with stability data. The DNPE correlation is slightly stronger at $Q = 0.6$ (0.561, p-value = 0.008), than at $Q = 1.0$ (0.513, p-value = 0.017). For all Q, C_SCN0 on ND generated SCN0s did not produce significant correlations with the original (23 points) dataset.

[Khor18] did not find a significant correlation between NS SCN0_RCO and NS stability. However, we find that SCN0_RCO can correlate significantly with NS stability, without need to exclude any point from the dataset, when RCO is computed on SCN0s within the ND-TS region ($0.1 \leq Q < 0.5$) (Fig. 5 bottom-left). These significant DNPE correlations cannot be obtained with RNRE SCN0s; RNRE correlations across all Q are insignificant, again showing that the order in which SCN0 edges are formed matters.

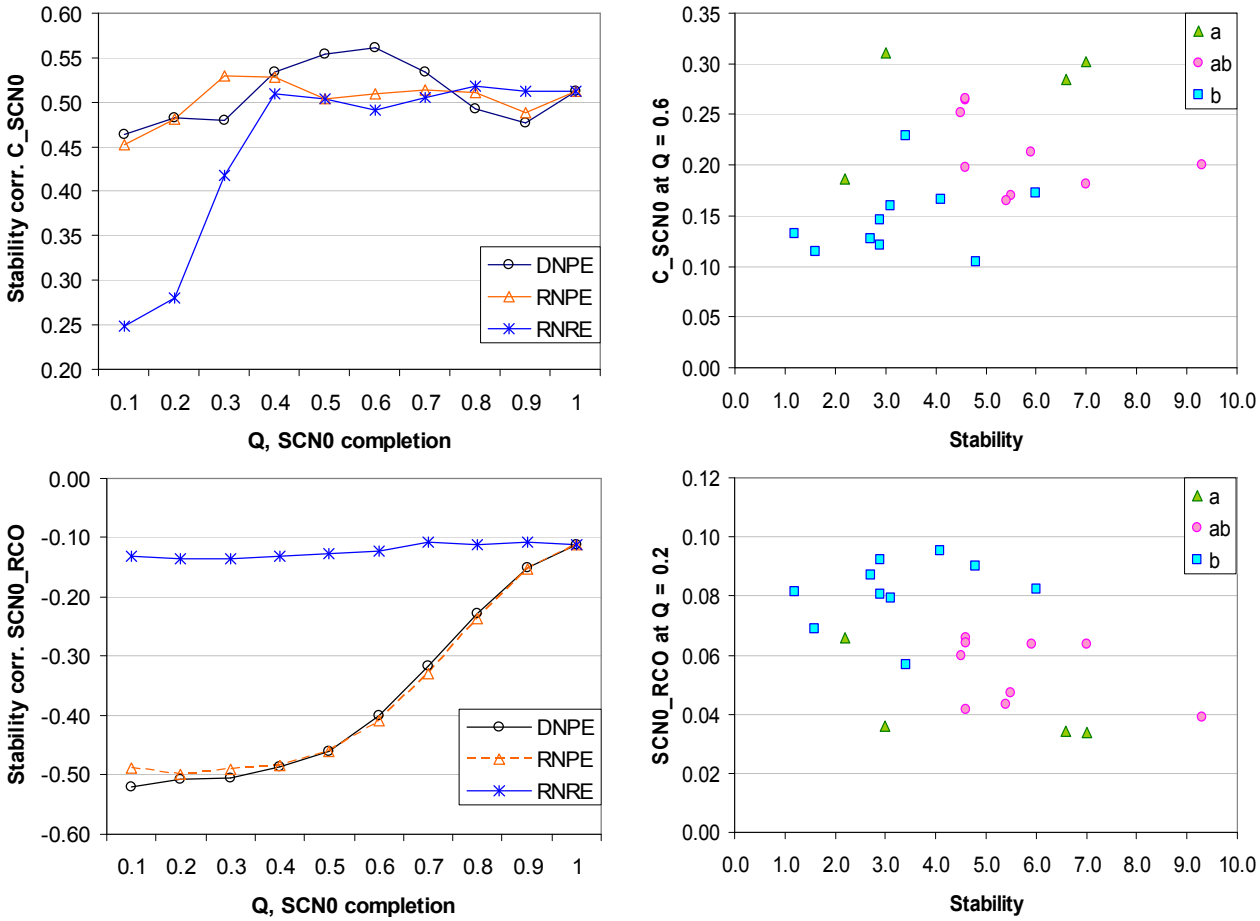


Fig. 5 Top-left: Correlation between *C* of ND generated SCN0s (averaged over 100 independent runs per protein) and NS stability for 21 data points. DNPE and RNPE correlations are significant (p-value < 0.05) for all *Q*; RNRE correlations are significant for $Q \geq 0.4$. **Top-right:** Scatter of C_SCN0 values at $Q = 0.6$ against NS stability, by fold type. All 23 data points are shown. **Bottom-left:** Correlation between RCO of ND generated SCN0s (averaged over 100 independent runs per protein) and NS stability for 23 data points. DNPE and RNPE correlations have p-value < 0.027 for $Q \leq 0.5$; all other correlations are insignificant (p-value > 0.05). **Bottom-right:** Scatter of SCN0_RCO values at $Q = 0.2$ against NS stability, by fold type. All 23 data points are shown.

4.4 Discussion

RCO was proposed as a single-value measure of a protein's overall native topology or structural complexity, for the purpose of studying the influence a protein's native structure exerts over its folding kinetics and possibly mechanism. Despite its consistent and strong performance as a determinant of two-state folding rate, a puzzle remains as to why the significant NS RCO folding rate correlation should exists at all since folding rate is actually a function of the barrier height at the TS. A prevailing explanation is that TS structures already resemble NS structures in some pertinent way [Paci05, Faisca12]. NS RCO captures information about the entropy loss experienced by a protein chain to reach its TS. An underlying expectation/assumption here is that TS RCO correlates significantly, or even more strongly than NS RCO, with folding rate.

However, assuming early folding is not a random affair, our findings with ND hitherto do not support this expectation. In the ND-TS region ($0.1 \leq Q < 0.5$), the RCO of DNPE generated SCN0s do not correlate significantly with folding rate, although they correlate significantly with NS SCN0_RCO (Fig. 2 top). However, the contacts that go into the calculation of these RCOs do contribute to configuration energies that correlate significantly with folding rate. Thus, what is needed here may not be a different explanation, but a different topological complexity measure of protein structure.

We propose that this need can be satisfied by *C*. In the ND-TS region ($0.1 \leq Q < 0.5$), the *C* of DNPE generated SCN0s correlates significantly with both folding rate and NS C_SCN0. The use of *C* as a structural descriptor of proteins for the purpose of studying two-state folding kinetics does not diminish the role chain-entropy, the basis of the RCO metric, plays in folding, as the weaker RNRE C_SCN0 correlation with folding rate show (Fig. 2 bottom-left). Further, RNRE generated barrier heights do not correlate significantly with folding rate (section 3.2). The C_SCN0 results in Fig. 2 (bottom-left) also clearly show the advantage of giving hub nodes preferential treatment in ND; DNPE correlations are stronger than RNPE ones.

A limitation of *C* in its present form is its correlation with folding rate is weaker than that of RCO, and as such, has less mechanistic predictive value. A remedy could be a weighted clustering coefficient [Opsahl09] that considers not only the arrangement of contacts, but also their sequence distance.

5. ND analysis of phi-generated TS structures

In this section, we test our assertions in section 4.4 by examining structures of 10 transition-state ensembles (TSE) generated with phi restraints [Paci02]. We refer to these TSE structures as PTS structures, to distinguish them from the general notion of TS structures.

To keep uniformity of source for native structures, we use PDB coordinates (not the supplied minimized version[3]) to obtain NS PRN0s for the 10 proteins (2α, 4αβ and 4β). ND is performed on these 10 NS PRN0s to obtain partial (ND generated) SCN0s. The PRN0 of a PTS structure comprises all and only the PRN edges of the PTS structure which are also in the native structure. In keeping with our PRN0→SCN0 approach, the SCN0 of a PTS structure is found by EDS on its PRN0, and excludes all non-native shortcuts (these are possible given the partial nature of PRN0s in PTS structures).

[Paci05] reports that despite their heterogeneity, in terms of root-mean-square-deviation (RMSD) to the native structure, the PTS structures resemble their respective native structures in terms of RCO with a correlation of 0.854. Further, RCO of both NS structures (-0.791) and PTS structures (-0.753) correlate strongly and negatively with folding rate for the 10 proteins [Paci05]. These three PTS attributes find counterparts in our methodology. Heterogeneity of PTS structures is exhibited in terms of SCN0, with

[3] PRN0s from the minimized versions are more dense (have more edges).

each structure generating a unique SCN0 (Table E1). The PTS structures resemble their respective native structures in terms of SCN0_RCO, with a correlation of 0.803. SCN0_RCO's correlation with folding rate is -0.716 for NS structures, and -0.802 for PTS structures.

The PTS structures share an average of 30% to 50% of the shortcuts in their respective NS structures (Table E1); this range places them towards the upper end of the ND-TS region identified in section 3.2.

Although the PTS structures have sparser SCN0s than their NS structures, the PTS native shortcuts are arranged such that PTS C_SCN0 is strongly correlated (0.830) with NS C_SCN0 (Fig. E1); same with *C* of ND generated SCN0s on a larger dataset (Fig. D1). Thus, C_SCN0 provides another support for the claim that the overall topology of NS structures is established in TS structures; a claim that is the lynchpin for the observed empirical relationship between folding rates and topological descriptors of NS structures.

However, both RCO and *C* of RNRE generated SCN0s can produce strong, if not stronger, correlations with that of NS SCN0s (Fig. 2 right). Further, stronger TS-NS structural resemblance need not translate to stronger correlation with folding rate. Unlike SCN0_RCO, where a stronger correlation with NS SCN0_RCO is a positive in terms of improving correlation with folding rate, the reverse is true for C_SCN0, because NS C_SCN0 has a weaker relationship with folding rate than non-NS C_SCN0 (section 4.1).

C_SCN0 of the PTS structures does not correlate significantly with folding rate (Table E1). This is not at variance with results in section 4.1, and is explained by the weaker (compared with SCN0_RCO) but still significant C_SCN0 correlation with folding rate combined with sample size difference (52 vs. 10). p-values can be influenced by sample size; smaller samples tend to have larger p-values. At p-value < 0.05, the probability of a significant correlation between folding rate and C_SCN0 for a random set of 10 two-state proteins is about (12/50) 24% (Table E2). This probability is lowered by random elements in the generative process (e.g. unbiased node selection in RNPE, and unbiased node selection and unbiased PRN0 edge creation in RNRE), and decreases as SCN0 nears completion.

However, SCN0_RCO's strong correlation (p-value = 0.005) with folding rate for the PTS structures is inconsistent with findings in section 4.1, if the PTS structures are true TS structures (or at least structures within the ND-TS region). At p-value < 0.01, the probability of a significant correlation between folding rate and SCN0_RCO for a random set of 10 two-state proteins is 0% within the ND-TS region (Table E2). This probability is raised by random elements in the generative process, and increases as SCN0 nears completion.

This inconsistency suggests either the existence of random elements in the formation of PTS SCN0s, or the PTS ensembles are heavily populated with post-TS structures. Both of these suggestions can be checked: the former, by comparing PTS C_SCN0 with DNPE and with RNRE C_SCN0, on the principal

that randomly generated networks have significantly smaller clustering coefficients than comparable less randomly generated networks; and the latter by using PTS SCN0_RCO to situate the PTS structures within DNPE trajectories, products of a non-random generative process. These checks are performed with the aid of linear regression on data in Figs. E4 & E5, and the results are presented in Table 3.

Table 3 Linear regression results.

PDB id	(3) PTS *Q*	C_SCN0			SCN0_RCO	
		PTS	(1) DNPE	(2) RNRE	(5) PTS	(4) DNPE *Q*
1imq α	0.365 ± 0.074	0.100 ± 0.050	0.165	0.102	0.076 ± 0.031	0.874
1lmb4 α	0.484 ± 0.085	0.174 ± 0.069	0.238	0.206	0.034 ± 0.004	0.700
1bf4 αβ	0.531 ± 0.123	0.095 ± 0.050	0.166	0.135	0.105 ±0.016	0.893
2ptl αβ	0.352 ± 0.052	0.109 ± 0.040	0.117	0.075	0.093 ± 0.011	0.722
2ci2 αβ	0.276 ± 0.057	0.023 ± 0.031	0.055	0.042	0.095 ± 0.033	0.489
1aps αβ	0.401 ± 0.048	0.073 ± 0.033	0.115	0.082	0.188 ± 0.025	0.783
1fmk β	0.352 ± 0.043	0.039 ± 0.032	0.060	0.062	0.152 ± 0.030	0.531
1bk2 β	0.356 ± 0.060	0.070 ± 0.040	0.078	0.075	0.171 ± 0.032	0.632
1shf β	0.422 ± 0.042	0.074 ±0.026	0.086	0.064	0.157 ± 0.021	0.606
1ten β	0.378 ± 0.097	0.055 ± 0.025	0.054	0.058	0.170 ± 0.025	0.693
Mean± std. dev.	0.392 ± 0.073	0.081 ± 0.042	0.113 ± 0.060	0.090 ± 0.048	0.124 ± 0.050	0.692 ± 0.134

Extrapolated (linear regression) DNPE (1) and RNRE (2) C_SCN0 values for given PTS *Q* (3) from Fig. E4 plots.
Extrapolated (linear regression) DNPE *Q* (4) for given PTS SCN0_RCO values (5) from Fig. E5 plots.

PTS SCN0s are more similar to RNRE SCN0s than DNPE SCN0s in terms of *C*. For the given PTS *Q* values, PTS C_SCN0 is not significantly different from RNRE C_SCN0, but is significantly (one-sided paired t-test p-value = 0.002) smaller than DNPE C_SCN0 (Table 3). This observation supports the presence of random elements in the formation of PTS SCN0s.

The RCO values of PTS SCN0s place them in the post ND-TS region of DNPE trajectories. Averaged over the 10 proteins, DNPE generated SCN0s produce RCO values comparable to those of PTS SCN0s at $Q \approx 0.7$ (Table 3). This finding supports the notion that the reason why PTS RCO correlates significantly with folding rate is because they are heavily populated with post-TS structures, and hence actually says little about the relationship between TS RCO and folding rate.

TS structures are suppose to be those associated with energy of the folding barrier for two-state proteins, making them most transient, least stable and therefore difficult to capture experimentally. However, the energy of PTS structures (calculated per section 2.4) do not correlate with folding rate. Fig. E3 shows where the PTS structures are situated with respect to their respective ND generated energy profiles; most are situated after the peak of a DNPE energy profile. In contrast, the DNPE barrier heights correlate (-0.865) significantly with folding rate for the 10 proteins (Fig. E2).

Our ND analysis of the phi restraint generated TSEs leads us to conclude that these structures do not lie within the ND-TS region. Since they are not TS structures within the ND model, they have little to say

about the TS, and thus cannot be used to support the prevailing explanation for the relationship between NS RCO and folding rate of two-state proteins. In any case, the ND model does not produce evidence for a relationship between TS RCO and folding rate of two-state proteins (section 4.1).

This unexpected conclusion reaffirms the importance of the presence of long-range contacts in putative TS structures for TS RCO to correlate significantly with folding rate. From their lattice-model simulation, [Faisca12] found that high contact order (CO) structures exhibit a stronger correlation with folding rate than low CO structures, even though the activation energy of folding (barrier height) for both high and low CO structures correlate strongly with logarithmic folding rate. They attribute this difference to the formation of tertiary contacts which are mainly long-range in the TSE of high CO structures.

6. Discussion

DNPE and RNRE differ primarily in the order in which PRN0 edges are generated, which in turn influences which SCN0 edges get identified by EDS. While DNPE abides by the loop-entropy principle, RNRE ignores it. Consequently, RNRE enables long-range PRN0 (and SCN0) edges to form more easily (Fig. F1). This bias, together with the distinct roles short- and long-range contacts play in protein folding, can account for the RCO results in section 4; SCN0_RCO correlation with NS stability is significant for $Q \leq 0.5$, and SCN0_RCO correlation with TS placement is stronger when $Q > 0.5$. RCO correlations are more sensitive to contact sequence distances than C. Like C, our energy formulation of SCNs E, is also independent of contact sequence distance. Table 4 summarizes the ND view of how factors of protein folding relate with each other.

Table 4 Generalization of the relationships between folding factors discussed in this paper.

Folding rate	Barrier height	RCO	C	TS native-like compactness (θ_m)	NS stability	Long-range contacts
Larger Faster	Smaller Lower	Smaller	Larger	Smaller Lower	Larger More stable	Fewer
Smaller Slower	Larger Higher	Larger	Smaller	Larger Higher	Smaller Less stable	More

Herein lies a possible criticism of our ND model: its execution of the loop-entropy principle lacks the element of explicit edge dependency that enables long-range contacts to adjust their contact order as short-range contacts are formed. The use of the logarithmic form instead of raw sequence distance to calculate the probability of making a PRN0 edge was an effort to produce this effect, but it may be inadequate[4]. The presence of long-range native contacts in TS structures is a hallmark of proteins that fold via the NC mechanism, e.g. 2CI2 and 1APS. On the other hand, DNPE does not seem burdened by this

[4] A ND model applying the notion of effective contact order [Weikl03] to create PRN0 edges (DNEE), quickly morphs into RNRE with probability of PRN0 edge creation approaching a uniform 50% (Fig. F1).

lack, and actually performs better for proteins with larger SCN0_RCO (section 3.3); their Q for $P_{init} = 0.5$ is within the ND-TS region (Table 2). Further, the early presence of random long-range native shortcuts, as in RNRE, need not imply a better (within the ND-TS region) Q for $P_{init} = 0.5$ (Table 2).

What is vital for NC folding is the few select long-range native contacts that help stabilize the folding nucleus. For most proteins in Table 2, Q for $P_{init} = 0.5$ increases if long-range native shortcuts (SCLE) discovered at each Q are ignored by C_SCN0 when selecting the initial fold step for a Q (SCLE from previous Qs are included). This adverse effect on DNPE trajectories give evidence that the SCLE identified within DNPE do exert influence over productive folding.

While long-range edges appear in DNPE SCN0s at a much slower rate than in RNRE SCN0s (Fig. F1 left), DNPE generates more meaningful SCLE at the Q where folding energy peaks than RNRE. For the 10 proteins in section 5, at the Q where their respective DNPE barrier height is measured, DNPE SCN0s have fewer long-range edges, but report larger match values than RNRE SCN0s (Table F1). A match value is the proportion of SCLE generated in at least 10% of SCN0s at a Q, that is also present in at least 10% of PTS structures. A closer inspection of these matched SCLE for three proteins: 2PTL, 1APS and 2CI2, finds that they are clustered in secondary structure element (SSE) pairs associated with initial fold steps for the respective proteins (Table F2).

References

[Khor18] Khor S. Folding with a protein's native shortcut network. Proteins: Structure, Function and Bioinformatics 2018; 86(9):924-934.

[Plaxco98] Plaxco KW, Simons KT and Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. 1998; 277:985-994.

[Khor16] Khor S. Protein residue networks from a local search perspective. Journal of Complex Networks 2016; 4(2):245-278.

[Chan11] Chan HS, Zhang Z, Wallin S and Liu Z. Cooperativity, local-nonlocal coupling and nonnative interactions: Principles of protein folding from coarse-grained models. Annu. Rev. Phys. Chem. 2011; 62:301-326.

[Portman10] Portman JJ. Cooperativity and protein folding rates. Current opinion in Structural Biology 2010; 20:11-15.

[Henry04] Henry ER and Eaton WA. Combinatorial modeling of protein folding kinetics: free energy profiles and rates. Chemical Physics 2004; 307:163-185.

[Kaya13] Kaya H, Uzunoglu Z and Chan HS. Spatial ranges of driving forces are a key determinant of protein folding cooperativity and rate diversity. Phys Rev E 2013; 88:044701.

[Micheletti03] Micheletti C. Prediction of folding rates and transition-state placement from native-state geometry. Proteins 2003; 51:74-84.

[Weikl03] Weikl TR and Dill KA. Folding rates and low-entropy-loss routes of two-state proteins. J. Mol. Biol. 2003; 329:585-598.

[Miyazawa96] Miyazawa S and Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J. Mol. Biol. 1996; 256:623-644.

[Lappe10] Duarte JM, Sathyapriya R, Stehr H, Filippis I and Lappe M. Optimal contact definition for reconstruction of contact maps. BMC Bioinformatics 2010; 11:283.

[Lappe09] Sathyapriya R, Duarte JM, Stehr H, Filippis I and Lappe M. Defining an essence of structure

determining residue contacts in proteins. PLoS Comput. Biol. 2009; 5(12):e1000584.

[Shmy05] Shmygelska A. Search for folding nuclei in native protein structures. Bioinformatics 2005; 21 Suppl 1:i394-402.

[Vend02] Vendruscolo M, Dokholyan NV, Paci E and Karplus M. Small-world view of the amino acids that play a key role in protein folding. Physical Review E 2002; 65:061910.

[Li08] Li J, Wang J and Wang W. Identifying folding nucleus based on residue contact networks of proteins. Proteins 2008; 71:1899-1907.

[Thomas07] Thomas S, Tang X, Tapia L and Amato NM. Simulating protein motions with rigidity analysis. J. Comp. Biol. 2007; 14(6):839-855.

[Piazza08] Piazza F and Sanejouand Y-H. Discrete breathers in protein structures. Physical Biology 2008; 5:026001.

[Ejtehadi04] Ejtehadi MR, Avall SP and Plotkin SS. Three-body interactions improve the prediction of rate and mechanism in protein folding models. PNAS 2004; 101(42):15088-15093.

[Faisca09] Faisca PFN. The nucleation mechanism of protein folding: a survey of computer simulation studies. J. Phys.: Condens. Matter 2009; 21:373102.

[Khor07] Khor S. HIFF-II: A hierarchically decomposable problem with inter-level interdependency. IEEE Symposium on Artificial Life 2007; 274-281.

[Plaxco00] Plaxco KW, Simons KT, Ruczinski I, Baker D. Topology, stability, sequence and length: Defining the determinants of two-state protein folding kinetics. Biochemistry 2000; 39(37): 11177-11183.

[Dinner01] Dinner AR and Karplus M. The roles of stability and contact order in determining protein folding rates. Nature Structural Biology 2001; 8(1):21-22.

[Paci05] Paci E, Lindorff-Larsen K, Dobson CM, Karplus M and Vendruscolo M. Transition state contact orders correlate with protein folding rates. J. Mol. Biol. 2005; 352(3):495-500.

[Faisca12] Faisca PFN, Travasso RDM, Parisi A and Rey A. Why do protein folding rates correlate with metrics of native topology? PLoS ONE 2012; 7(4):e35599.

[Opsahl09] Opsahl T and Panzarasa P. Clustering in weighted networks. Social Networks 2009; 31(2):155-163.

[Paci02] Paci E, Vendruscolo M, Dobson CM and Karplus M. Determination of a transition state at atomic resolution from protein engineering data. J. Mol. Biol. 2002; 324:151-163.

Supplementary Information

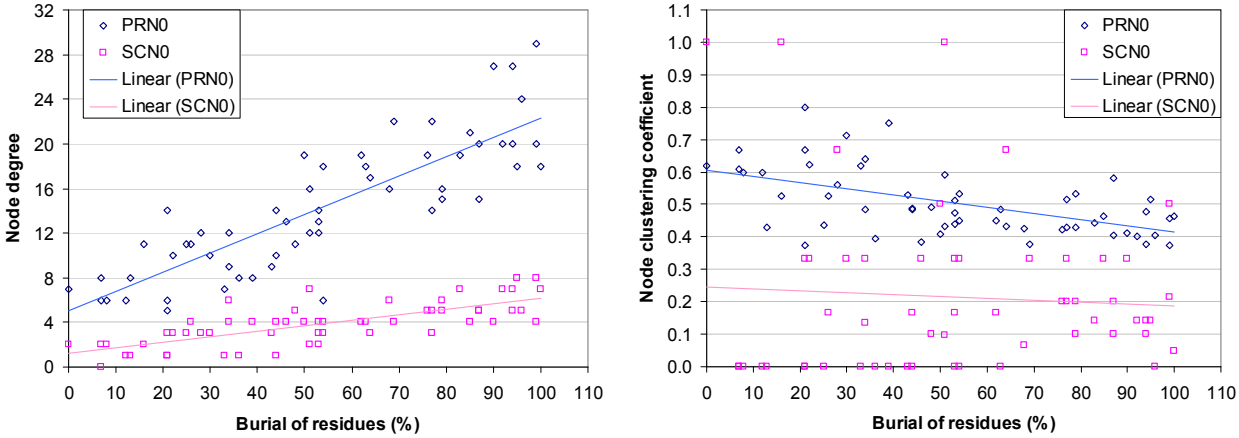


Fig. A1 Burial of src SH3 domain residues in the native state (Burial data from Table 1 of ref. [Riddle99]) against 1SRM's PRN0 and SCN0 node statistics. **Left**: Pearson's correlation coefficient for PRN0 node degree and Burial of residues is 0.8448 (p-val=2.22E-16), and for SCN0 node degree and Burial of residues is 0.7442 (p-val=4.92E-11). **Right**: Pearson's correlation coefficient for PRN0 node clustering and Burial of residues is -0.5500 (p-val=1.13E-05), and for SCN0 node clustering and Burial of residues is -0.0667 (p-val=0.6249).

The key folding residues are sourced from the references listed below. The conserved rigid residues from Suppl. Mat. of ref. [Sacquin15] are referred to as mechanically rigid (M-R) sites. For 2IGD and 1SRM, we use the provided data for 1IGD and 1SRL, respectively. The M-R sites confer structural stability in a protein's native state, but many of these sites also coincide with folding nuclei identified in the literature. The hydrogen-deuterium exchange (H-X) sites (Table 1 in ref. [Li99]) include amide protons (NHs) which are slowest to exchange out, or first to gain protection. The H-X probes do not identify nucleation sites per se, but rather the neighborhood where they might be found. Key folding sites on turns (which are crucial to trigger β-hairpin formation) are not detected. In Figs. A2 to A12, residues closer to the lower right of a plot are deemed more rigid than those closer to the upper left of a plot. PRN0 node degree is normally distributed [Khor16], so characteristic node degree would be about the midpoint of a node degree range.

1BDD (Fig. A2): Residues involved in frequently formed contacts in the transition state structures between the first and second helices (1H-3H) are: F14, L18, F31, I32 and L35; and between the second and third helices (3H-5H) are: L45, F31 and L35 [Kmiecik12]. The H-X sites are Y15...L18, R28, L35, K36, A49, K50 and K51. Except for L18 which is a turn residue immediately after 1H, the key folding residues congregate on the three α-helices. The 3H-5H nucleus residues are more rigid than the 1H-3H nucleus residues, and accordingly folding begins with 3H-5H pairing on the C_{SCN0} folding pathway for 1BDD.

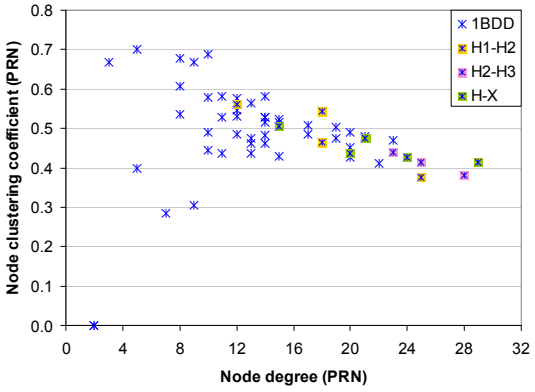


Fig. A2 Key folding residues for 1BDD are highlighted in pink, orange and green. H-X sites which are not also H1-H2 or H2-H3 nucleation sites are highlighted.

2IGD (Fig. A3): The M-R sites are Y8, L10, A31, F35, F57 and V59. These sites are located on the first and forth β-strands (1S and 9S), and the α-helix (5H), and they overlap with 1GB1's key folding residues in a structural alignment (PDBe Fold v2.59). The remaining four most rigid residues belong to 1S, 5H, 6T and 7S. The rigid residue on the third β-strand (7S) is W48, which corresponds to 1GB1's W43 nucleus residue in a structural alignment (PDBe Fold v2.59).

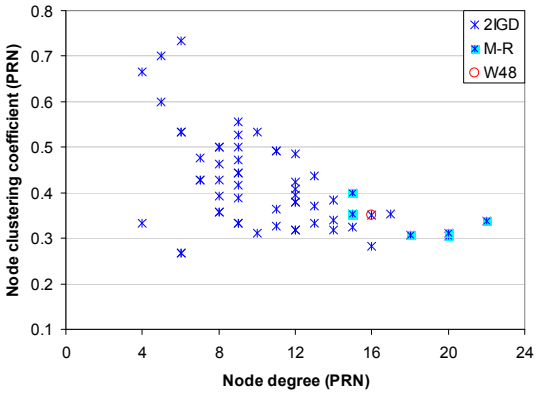


Fig. A3 Key folding residues for 2IGD are highlighted in blue.

1GB1 (Fig. A4): The folding nucleation sites are Y3, L5, F30, W43, Y45 and F52 [Kmiecik08, Hubner04]. These six residues are evolutionarily conserved in protein G-like folds, and make frequent long-range native interactions with K4, I6, L7, A26, T51, T53 and V54 as part of the nucleation growth process in simulations (Table 1 in ref. [Kmiecik08]). The H-X sites are L5, I6, T25, A26, E27, F30, T44, K50, and T51...V54. Except for K50 which is a turn residue immediately before 8S, the key folding residues congregate on the first, third and forth β-strands (0S, 6S and 8S), and the α-helix (4H).

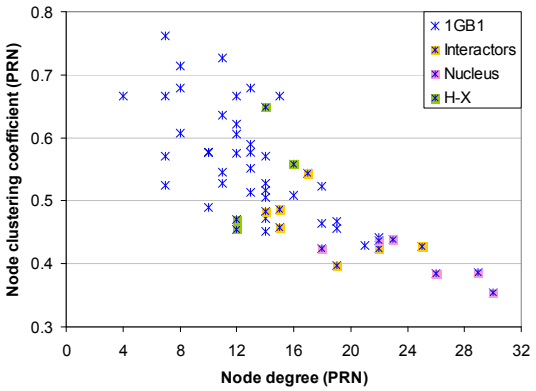


Fig. A4 Key folding residues for 1GB1 are highlighted in pink and green. Interactor residues (yellow) are those that partner with a Nucleus residue in Table 1 of ref [Kmiecik08]. H-X sites which do not also fall in the other two categories are highlighted.

2PTL (Fig. A5): The H-X sites are I20, A22, L24, I25, F36, S45, A47...D52, L72, I74, and K75. The key folding residues congregate on the first, second and forth β-strands (0S, 2S and 8S), and the α-helix (4H). For 2IGD, 1GB1 and 2PTL, key folding residues are absent on one of the β-strands of the β-hairpin that forms later, i.e. 2S for 2IGD and 1GB1, and 6S for 2PTL. This observation is in agreement with the initial folding step for these proteins (Table C1).

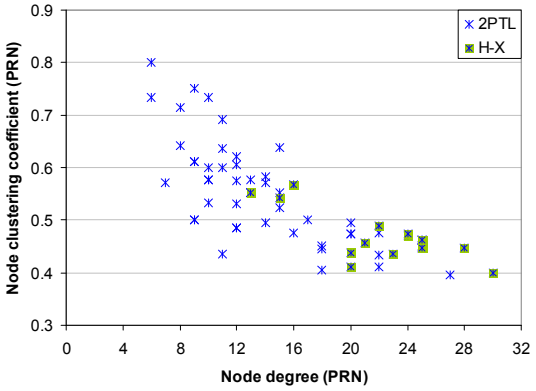


Fig. A5 Key folding residues for 2PTL are highlighted in green.

2CI2 (Fig. A6): The folding nucleation sites for 2CI2 are A35, I39, V66, L68 and I76 [Kmiecik07, Itzhaki95]. Four of these sites (A35, I39, V66, L68), together with R67 and P80 are M-R sites. The H-X sites are K30, I39, L40, I49, L51, V66, L68, F69 and V70. Except for K30, I76, and P80, the key residues congregate on the α-helix (3H), and the second and third β-strands (5S and 7S). K30 is a turn residue immediately before the helix (3H), I76 is a residue in the turn (8T) between the third and forth β-strands (7S and 9S), and P80 is a residue in the forth β-strand (9S). Key folding residues are absent from the first β-strand (1S), which only gets involved at the later stages according to our C_{SCN0} folding pathway for 2CI2.

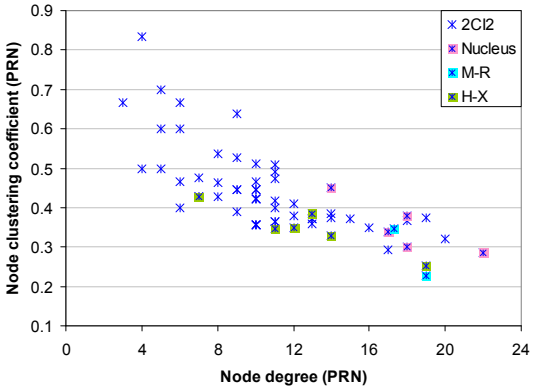


Fig. A6 Key folding residues for 2CI2 are highlighted in pink, blue and green. M-R sites which are not also Nucleus residues are highlighted. H-X sites which are not also Nucleus or M-R residues are highlighted.

1SHG (Fig. A7): The M-R sites are M25, V44 and V53. These three M-R sites, together with V9, A11, V23, L31, L33 and V58, form the hydrophobic core [Ventura02]. The M-R sites are found on the third and forth β-strands (5S, 7S) and towards the end of the RT-loop. 5S and 7S are the earliest folding elements in our C_{SCN0} folding pathway for 1SHG (Table C1). The two other most rigid residues in Fig. A7 are W42 which belongs to 5S, and L31 which belongs to 3S. L31 occupies a protein sequence position that is rigid in all the other proteins studied in the SH-3 domain family [Sacquin15]. The hydrophobic core sites touch all the β-strands.

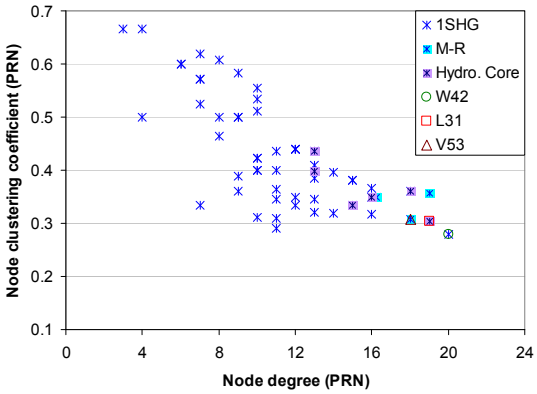


Fig. A7 Key folding residues for 1SHG are highlighted in blue. Hydrophobic core residues which are not also M-R sites are highlighted in purple.

1SRM (Fig. A8): The M-R sites are F26, L32, A45 and I56. These four M-R sites, together with F10, A12, L24, I34, W43 and V61 form the hydrophobic core [Riddle99, Lindorff04]. The hydrophobic core sites touch all the β-strands, and the diverging turn (26...32). I56 is the most rigid residue and it plays a central role in hydrophobic collapse of src SH3 domain circular permutants (the structurally aligned site in α-spectrin SH3 domain is V53) [Grantcharova 01].

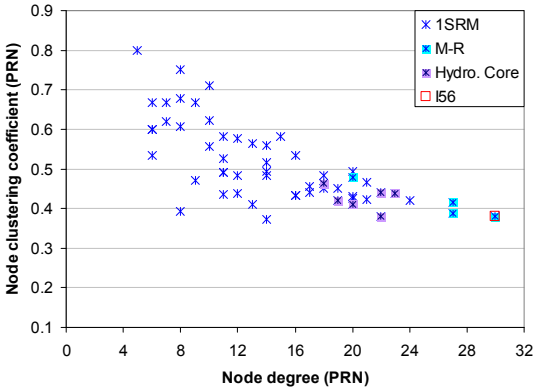


Fig. A8 Key folding residues for 1SRM are highlighted in blue. Hydrophobic core residues which are not also M-R sites are highlighted in purple.

2KJV (Fig. A9) and its p54-55 circular permutant *2KJW* (Fig. A10): The M-R sites for 2KJV are E5, V6, N7, I8, I26, F60, L61, W62, Y63, V65 and L79. Three of these sites (V6, I8 and I26), together with L30 form the hydrophobic core for 2KJV [Lindberg06]. The M-R sites for 2KJW are L8, Y10, L22, E25, R29, V48, I50, L52, L61, E64, I68 and L72. M-R site L61, together with Y63 and V65 form the hydrophobic core for 2KJW [Koga01]. For both 2KJV and 2KJW, the M-R sites congregate on the first and third β-strands (1S and 7S, or β1 and β3 for both 2KJV and 2KJW) and the two α-helices (3H and 9H). β1, α1 and β3 make up one of S6's two folding cores; the other comprises β1, α2 and β4 [Lindberg07].

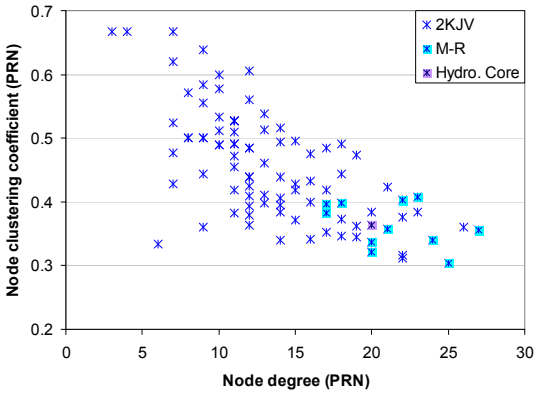


Fig. A9 Key folding residues for 2KJV are highlighted in blue. Hydrophobic core residues which are not also M-R sites are highlighted in purple.

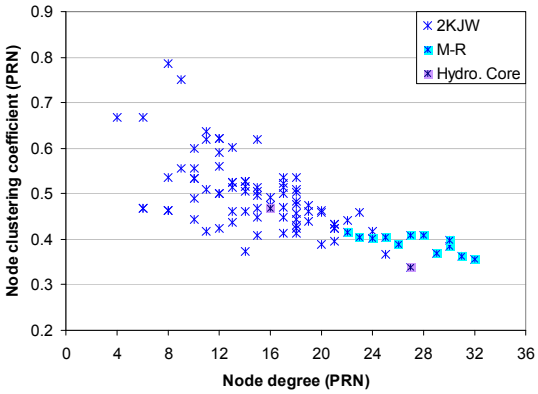


Fig. A10 Key folding residues for 2KJW are highlighted in blue. Hydrophobic core residues which are not also M-R sites are highlighted in purple.

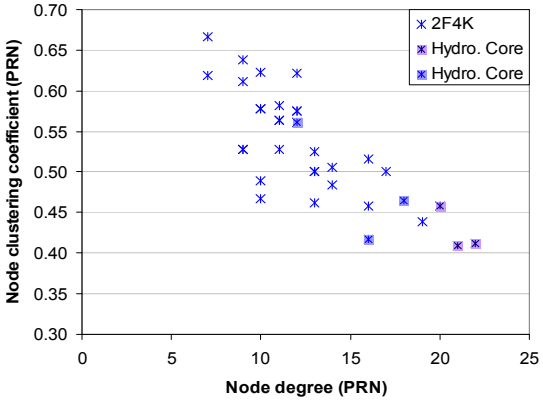


Fig. A11 Phenylalanine (F) residues 47, 51 and 58 in the hydrophobic core of 2F4K are highlighted in purple. The other hydro core residues L42, V50 and L69 mentioned in [Frank02] are highlighted in periwinkle.

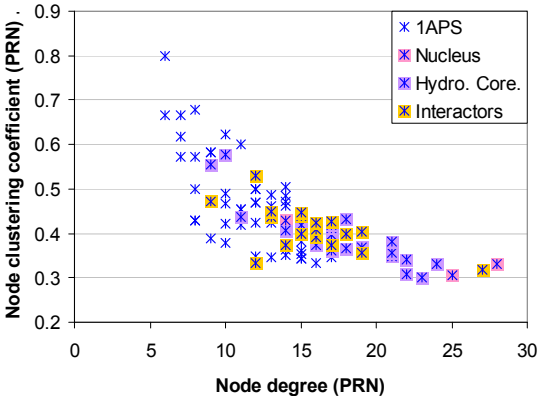


Fig. A12 The three key folding residues (Y11, P54 and F94) [Vend01] for 1APS are highlighted in pink. Hydrophobic core residues [Chiti99] which are not also Nucleus sites are highlighted in purple. Residues making long-range native contacts with the Nucleus sites [Vend01] are highlighted in orange.

References for section A

[Riddle99] Riddle DS, Grantcharova VP, Santiago JV, Alm E, Ruczinski I and Baker D. Experiment and theory highlight role of native state topology in SH3 folding. Nat. Struc. Biol. 1999; 6(11):1016-1024.

[Sacquin15] Sacquin-Mora S. Fold and flexibility: what can proteins' mechanical properties tell us about their folding nucleus? J. R. Soc. Interface 2015; 12:20150876.

[Li99] Li R and Woodward C (1999) The hydrogen exchange core and protein folding. Prot. Sci. 8:1571-1591.

[Khor16] Khor S. Protein residue networks from a local search perspective. Journal of Complex Networks 2016; 4(2):245-278.

[Kmiecik12] Kmiecik S, Gront D, Kouza M and Kolinski A. From coarse-grained to atomic-level characterization of protein dynamics: Transition state for the folding of B domain of protein A. J. Phys. Chem. B 2012;116:7026-7032.

[Kmiecik08] Kmiecik S and Kolinski A. Folding pathway of the B1 domain of protein G explored by multiscale modeling. Biophysical J. 2008; 94:726-736.

[Hubner04] Hubner IA, Shimada J and Shakhnovich EI. Commitment and nucleation in the protein G transition state. J. Mol. Biol. 2004; 336:745-761.

[Kmiecik07] Kmiecik S and Kolinski A. Characterization of protein-folding pathways by reduced-space modeling. PNAS 2007; 104(30):12330-12335.

[Itzhaki95] Itzhaki LS, Otzen DE and Fersht AR. The structure of the transition state for folding of Chymotrypsin Inhibitor 2 analyzed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding. J. Mol. Biol. 1995; 254:260-288.

[Ventura02] Ventura S, Vega MC, Lacroix E, Angrand I, Spagnolo L and Serrano L. Conformational strain in the hydrophobic core and its implications for protein folding and design. Nat. Struc. Biol. 2002; 9(6):485-493.

[Lindorff04] Lindorff-Larsen K, Vendruscolo M, Paci E and Dobson CM. Transition states for protein folding have native topologies despite high structural variability. Nature Structural & Molecular Biology 2004; 11(5):443-449.

[Grantcharova01] Grantcharova VP and Baker D. Circularization changes the folding transition state of the src SH3 domain. J. Mol. Biol. 2001; 306:555-563.

[Lindberg06] Lindberg MO, Haglund E, Hubner IA, Shakhnovich EI and Oliveberg M. Identification of the minimal protein-folding nucleus through loop-entropy perturbations. PNAS 2006; 103(11):4083-4088.

[Koga01] Koga N and Takada S. Roles of native topology and chain-length scaling in protein folding: A simulation study with a Gō-like model. J. Mol. Biol. 2001; 313:171-180.

[Lindberg07] Lindberg MO and Oliveberg M. Malleability of protein folding pathways: a simple reason for complex behavior. Current Opinion in Structural Biology 2007; 17:21-29.

[Frank02] Frank BS, Vardar D, Buckley DA and McKnight CJ. The role of aromatic residues in the hydrophobic core of the villin headpiece subdomain. Protein Science 2002; 11:680-687.

[Vend01] Vendruscolo M, Paci E, Dobson CM and Karplus M. Three key residues form a critical contact network in a protein folding transition state. Nature 2001; 409:641-645.

[Chiti99] Chiti F, Taddei N, White PM, Bucciantini M, Magherini F, Stefani M and Dobson CM. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. Nature Structural Biology 1999; 6(11):1005-1009.

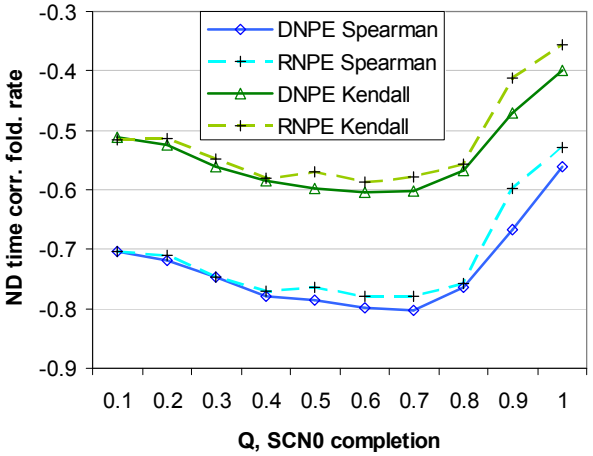


Fig. B1 Spearman and Kendall correlations between ND time (averaged over 100 independent runs per protein) and folding rate of 52 two-state proteins. All correlations have p-value < 0.01. DNPE correlations are significantly stronger than RNPE correlations (one sided paired t-test p-value is 0.009 for Spearman, and 0.003 for Kendall).

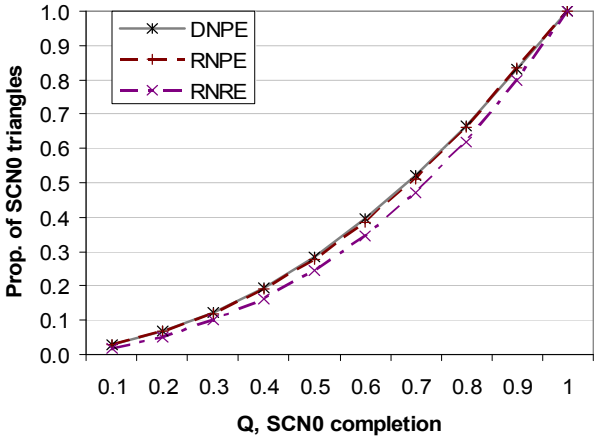


Fig. B3 Proportion of SCN0 triangles completed (averaged over 100 independent runs per protein).

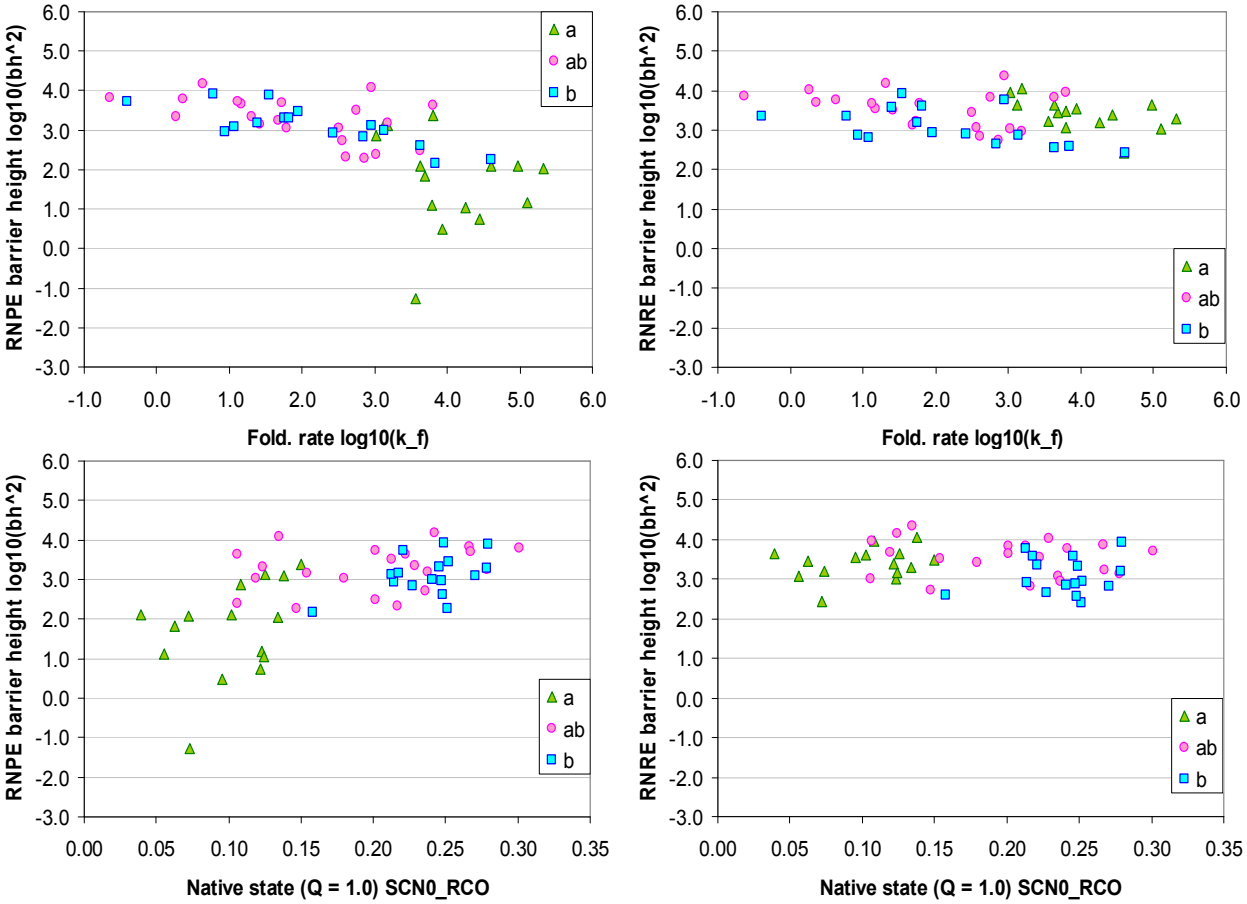


Fig. B2 Dispersion of RNPE and RNRE barrier heights (squared to handle negative values) for 52 two-state proteins against folding rate in a log10-log10 scatter plot (top), and against native state SCN0_RCO (bottom). Analogous plots for DNPE are in Fig. 1 of the main text.

Table C1 C_{SCN0} initial fold step selection by native (PDB) structures

PID	Expected initial fold step (1)	Good Models (2)	Bad models	Total	PRN0 SCN0 node degree corr. (3)
2F4K	(1H, 3H)	1	0	1	0.506
1BDD	(3H, 5H)	1	0	1	0.507
1GB1	(6S, 8S)	49	11	60	0.578
2PTL	(0S, 2S)	6	14	21	0.403
1MHX	(0S, 2S)	1	0	1	0.634
1MI0	(0S, 2S)	1	0	1	0.647
2CI2	(5S, 7S)	1	0	1	0.624
1SHG	(5S, 7S)	1	0	1	0.660
1SRM	(5S, 7S)	13	7	20	0.654
2ABD	(5H, 7H)	20	9	29	0.328
1APS	(5S, 7S)	5	0	5	0.366
2KJV	(5S, 7S)	20	0	20	0.583
2KJW	(5S, 7S)	20	0	20	0.487
1QYS	(1H, 3H)	1	1	1	0.652

(1) naming convention and secondary structure delineation as in [Khor18]

(2) Good models are those native structures where the expected initial fold step appears as the initial reaction on a C_{SCN0} folding pathway; bad models are those where this does not happen. For instance, 6/21 2PTL models are classified "good" because their initial fold step is formation of the N-terminal β hairpin (0S, 2S), while the remaining 14 2PTL models are "bad" because their initial fold step is not (0S, 2S).

(3) Pearson correlation coefficient between PRN0 and SCN0 node degree; all p-values < 0.002.

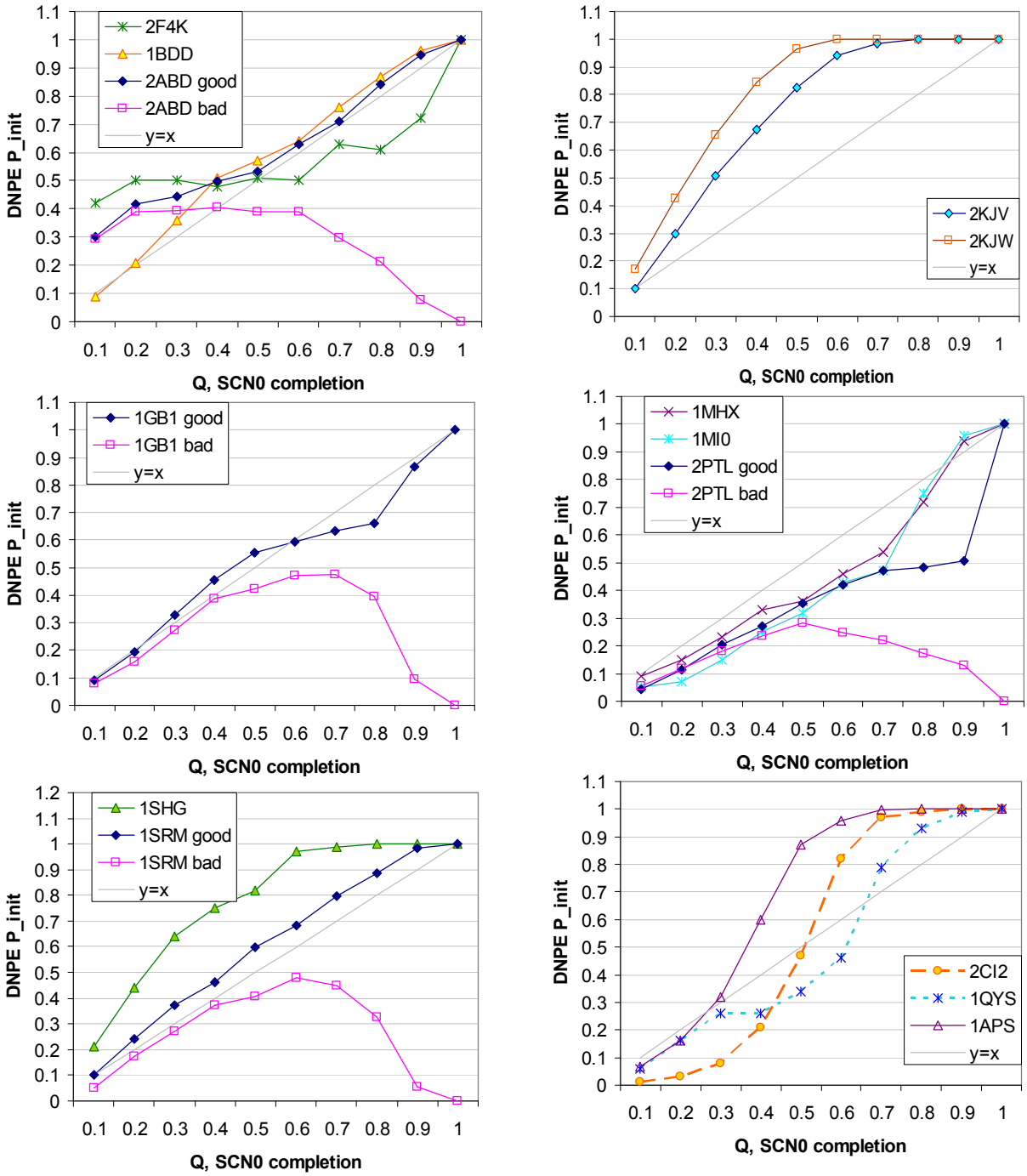


Fig. C1 Proportion of ND (DNPE) generated SCN0s (y-axis) at each *Q* (x-axis) where the initial fold step matches the expected of the native structure. The "good" ("bad") qualifier denotes protein native structures with (without) the expected initial step. For instance, the expected initial fold step for protein G (1GB1) is the pairing of its C-terminal β-strands: (6S, 8S); at $Q = 0.4$, 46% of ND generated SCN0s using the "good" models have (6S, 8S) as their initial fold step, whereas only 39% of ND generated SCN0s using the "bad" models have (6S, 8S) as their initial fold step. The expected initial fold step for each protein is listed in Table C1.

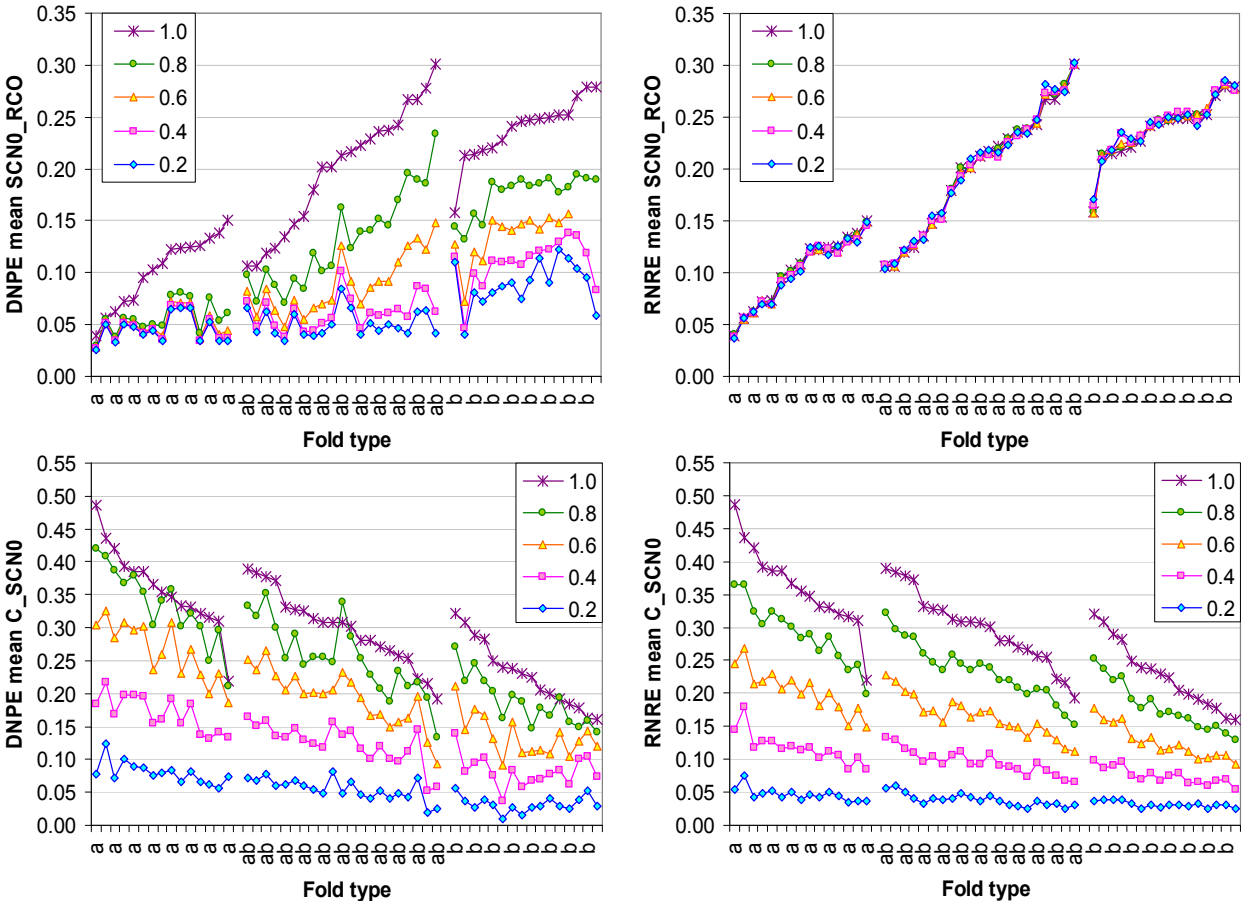


Fig. D1 Top: SCN0_RCO of DNPE (left) and of RNRE (right) generated SCN0s, averaged over 100 runs per 52 two-state proteins, arranged by fold type, at five levels of *Q* (SCN0 completion). **Bottom:** C_SCN0 of DNPE (left) and of RNRE (right) generated SCN0s, averaged over 100 runs per 52 two-state proteins, arranged by fold type, at five *Q* levels.

Table E1 Attributes of phi-restrained generated transition-state (PTS) structures and their SCN0s.

PDB id	*n* PTS	*n* SCN0	% SC0	NS SCN0_RCO	PTS SCN0_RCO	NS C_SCN0	PTS C_SCN0	NS SCLE	PTS SCLE
1lmb4 α	127	127	48.4	0.054	0.034	0.480	0.174	6	0.354
1bf4 αβ	74	74	53.1	0.106	0.105	0.326	0.095	22	12.230
2ptl αβ	126	126	35.2	0.115	0.093	0.345	0.109	28	8.405
1imq α	16	16	36.5	0.139	0.076	0.424	0.100	26	3.250
1ten β	90	90	37.8	0.249	0.170	0.183	0.055	105	34.133
1shf β	33	33	42.2	0.252	0.157	0.191	0.074	47	15.909
1bk2 β	31	31	35.6	0.253	0.171	0.248	0.070	49	10.645
1aps αβ	29	29	40.1	0.267	0.188	0.254	0.073	79	26.034
1fmk β	147	147	35.2	0.274	0.152	0.179	0.039	48	10.401
2ci2 αβ	184	184	27.6	0.278	0.095	0.266	0.023	46	4.804
Correlation with folding rate				-0.716	-0.802	0.679	0.576	-0.878	-0.805
p-value				0.020	0.005	0.031	0.082	0.001	0.005
Correlation with native state				1.0	0.803	1.0	0.831	1.0	0.912
p-value					0.005		0.003		0

n PTS: number of structures in the phi-generated transition-state ensemble (TSE) from [Paci05].
n SCN0: number of unique native shortcut networks.
% SC0 = percentage of native shortcuts found in a PTS structure averaged over a TSE.
NS = Native-state
PTS = average for TSE structures
RCO = Relative Contact Order
SCN0_RCO = RCO computed on a SCN0
C_SCN0 = network clustering coefficient computed on a SCN0
SCLE = number of long-range (sequence separation > 10) native shortcuts

Table E2 Results of subset testing on folding rate correlations with SCN0_RCO and with C_SCN0. The 52 data points supporting these two correlations reported in section 4.1 are sampled uniformly at random without replacement 50 times to create datasets of size 10. The following tables record the number of sample datasets having correlations with p-value < 0.05 (top), and with p-value < 0.01 (bottom).

Fold. rate corr.	ND	Q=0.1	Q=0.2	Q=0.3	Q=0.4	Q=0.5	Q=0.6	Q=0.7	Q=0.8	Q=0.9	Q=1.0
SCN0_RCO	DNPE	0	1	2	2	5	21	25	31	33	28
	RNPE	2	2	2	3	7	15	21	29	34	30
	RNRE	28	30	29	29	29	29	29	29	29	29
C_SCN0	DNPE	12	11	12	11	13	12	12	10	9	5
	RNPE	12	9	9	9	10	10	8	6	4	2
	RNRE	11	4	5	6	8	8	8	8	8	7

Fold. rate corr.	ND	Q=0.1	Q=0.2	Q=0.3	Q=0.4	Q=0.5	Q=0.6	Q=0.7	Q=0.8	Q=0.9	Q=1.0
SCN0_RCO	DNPE	0	0	0	0	0	5	8	12	13	12
	RNPE	0	0	0	1	1	4	5	10	12	9
	RNRE	8	10	10	9	10	9	11	12	12	12
C_SCN0	DNPE	6	4	4	2	3	4	4	3	1	1
	RNPE	3	3	3	2	3	3	2	1	1	1
	RNRE	4	2	1	0	1	1	0	0	0	0

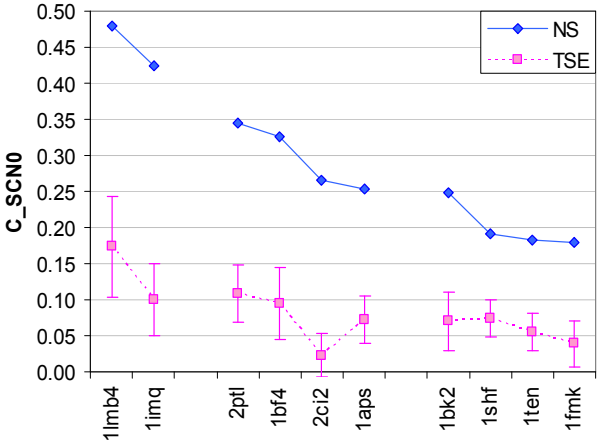


Fig. E1 SCN0 clustering coefficient (C_SCN0) of native (NS) and PTS (TSE) structures. Error bars denote one standard deviation about the mean.

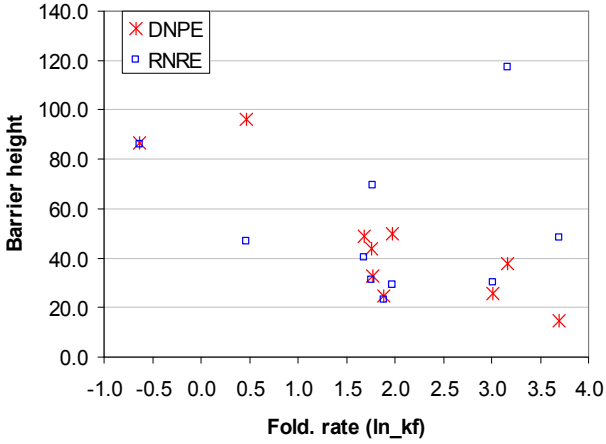
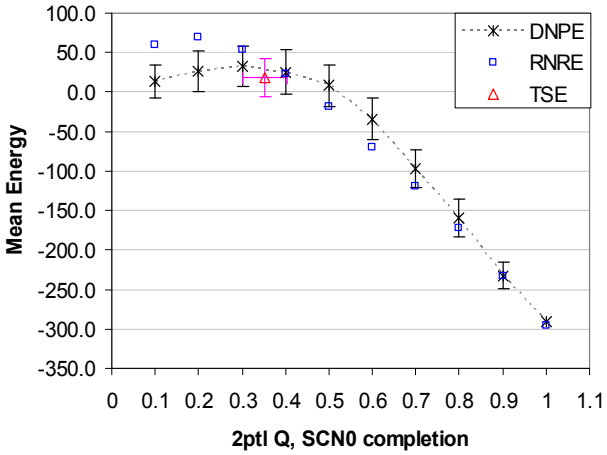
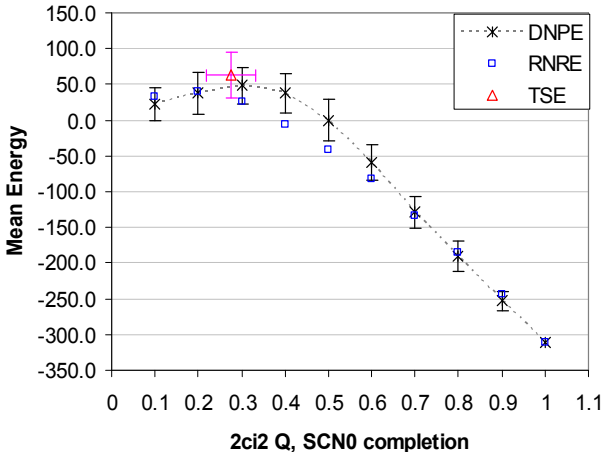
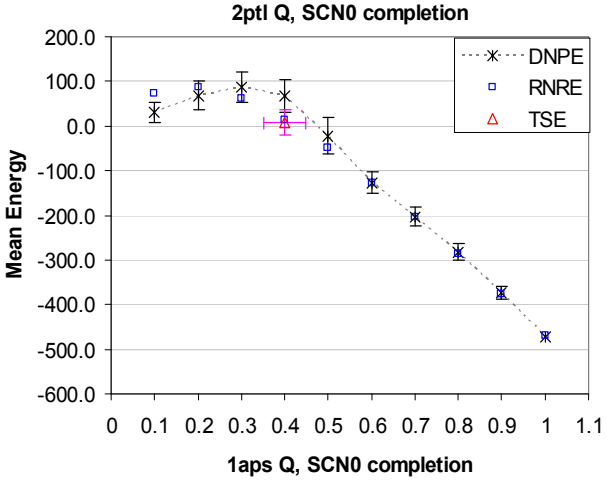
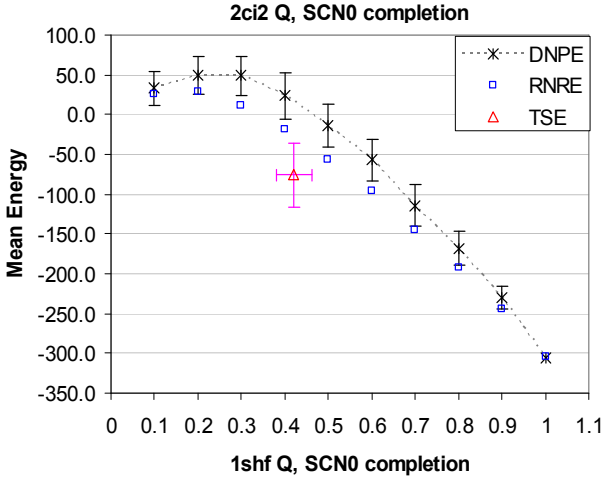


Fig. E2 Dispersion of DNPE and RNRE barrier heights (peaks in Fig. E3) against folding rate for the 10 PTS (TSE) proteins. Pearson correlation between DNPE barrier height ($Q = 0.260 \pm 0.052$) and fold. rate is -0.865 (p-value = 0.001). There is no significant correlation (-0.559, p-value 0.117) between RNRE barrier height and fold. rate even after excluding the one outlier at the top-right of the plot.









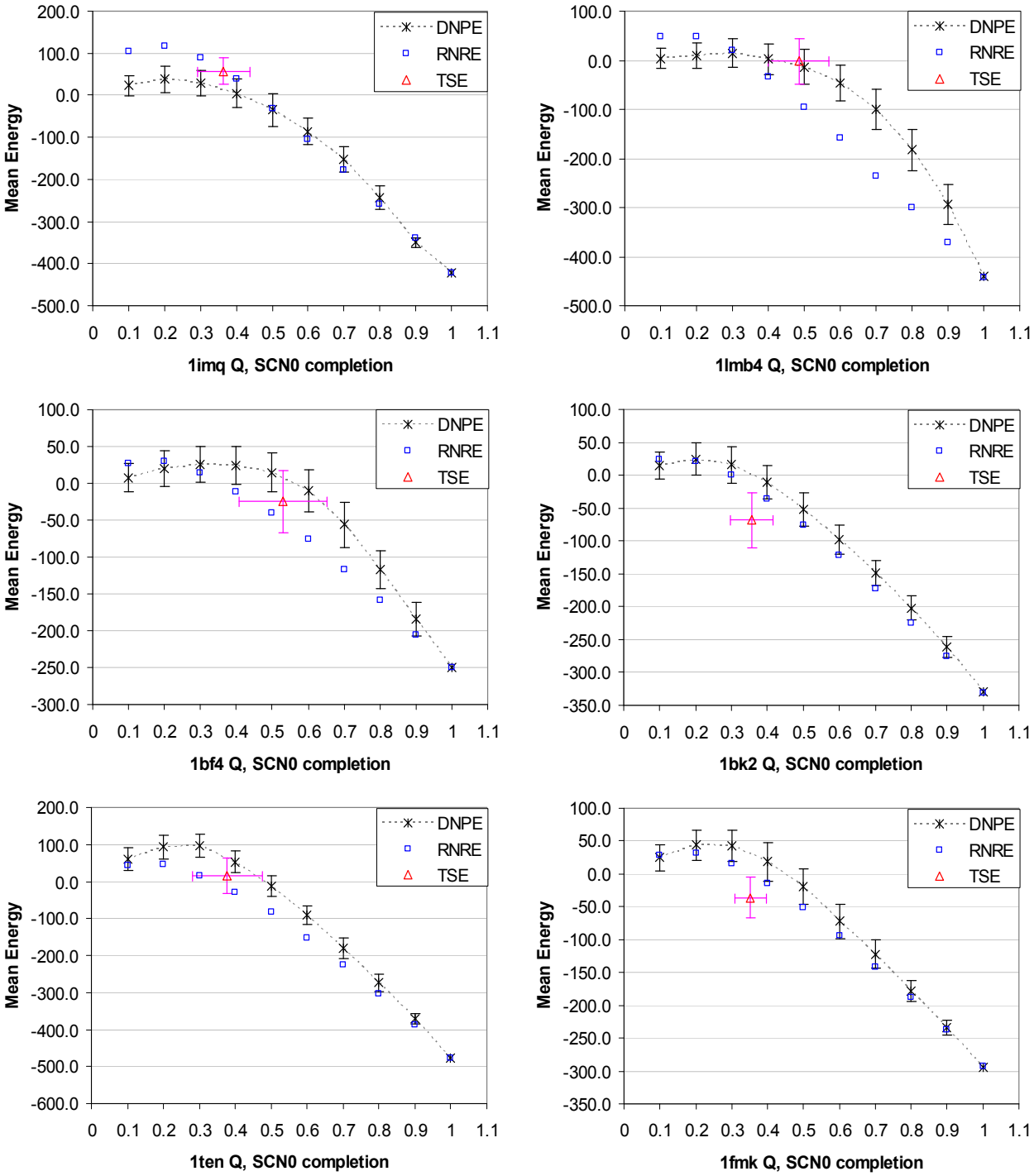
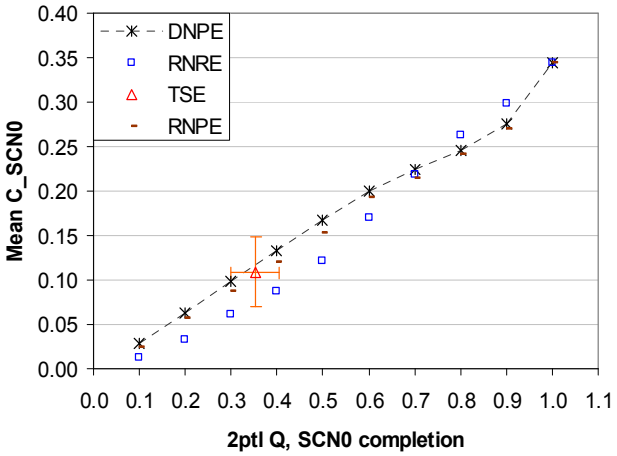
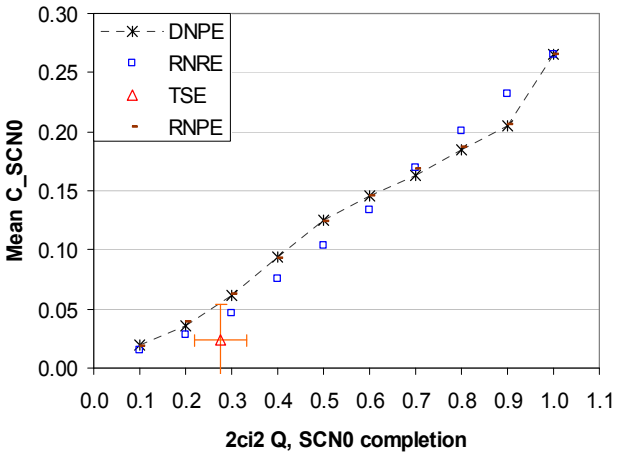
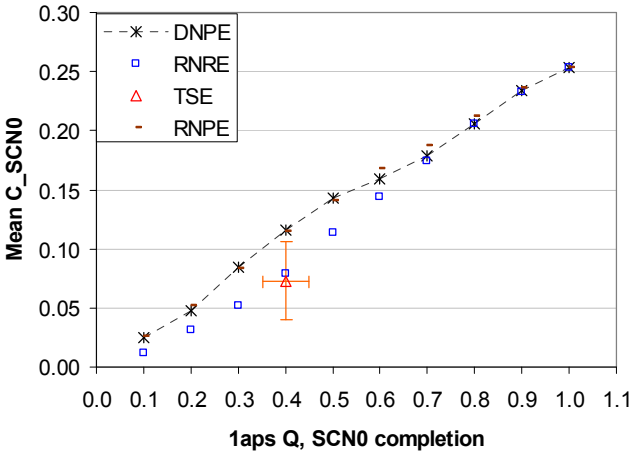
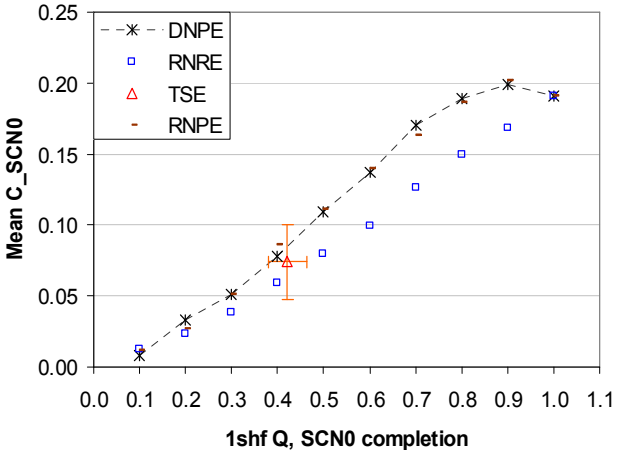


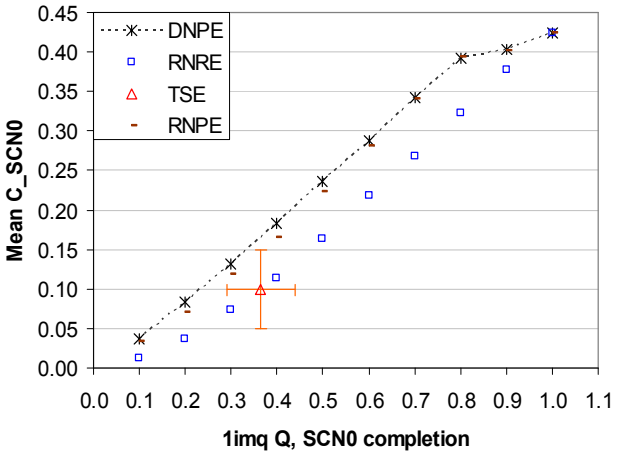
Fig. E3 ND energy profiles for the 10 PTS(TSE) proteins. The TSE point in each plot marks the average energy of PTS structures, at PTS *Q* (Table 3). Error bars denote one standard deviation about the mean.

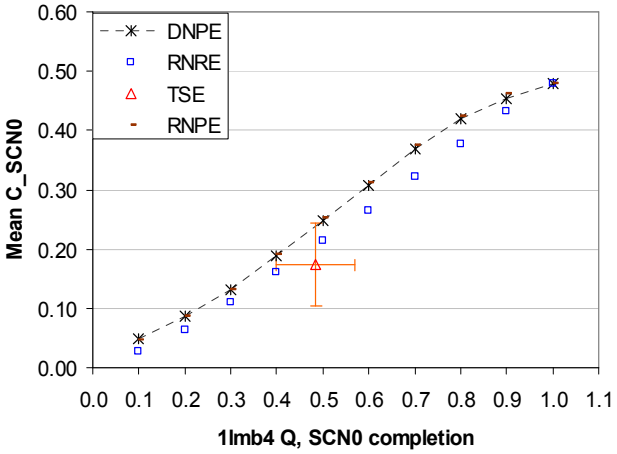

DNPE
RNRE
TSE
RNPE
Mean C_SCN0
2ptl Q, SCN0 completion


DNPE
RNRE
TSE
RNPE
Mean C_SCN0
2ci2 Q, SCN0 completion


DNPE
RNRE
TSE
RNPE
Mean C_SCN0
1aps Q, SCN0 completion


DNPE
RNRE
TSE
RNPE
Mean C_SCN0
1shf Q, SCN0 completion


DNPE
RNRE
TSE
RNPE
Mean C_SCN0
1imq Q, SCN0 completion


DNPE
RNRE
TSE
RNPE
Mean C_SCN0
1lmb4 Q, SCN0 completion

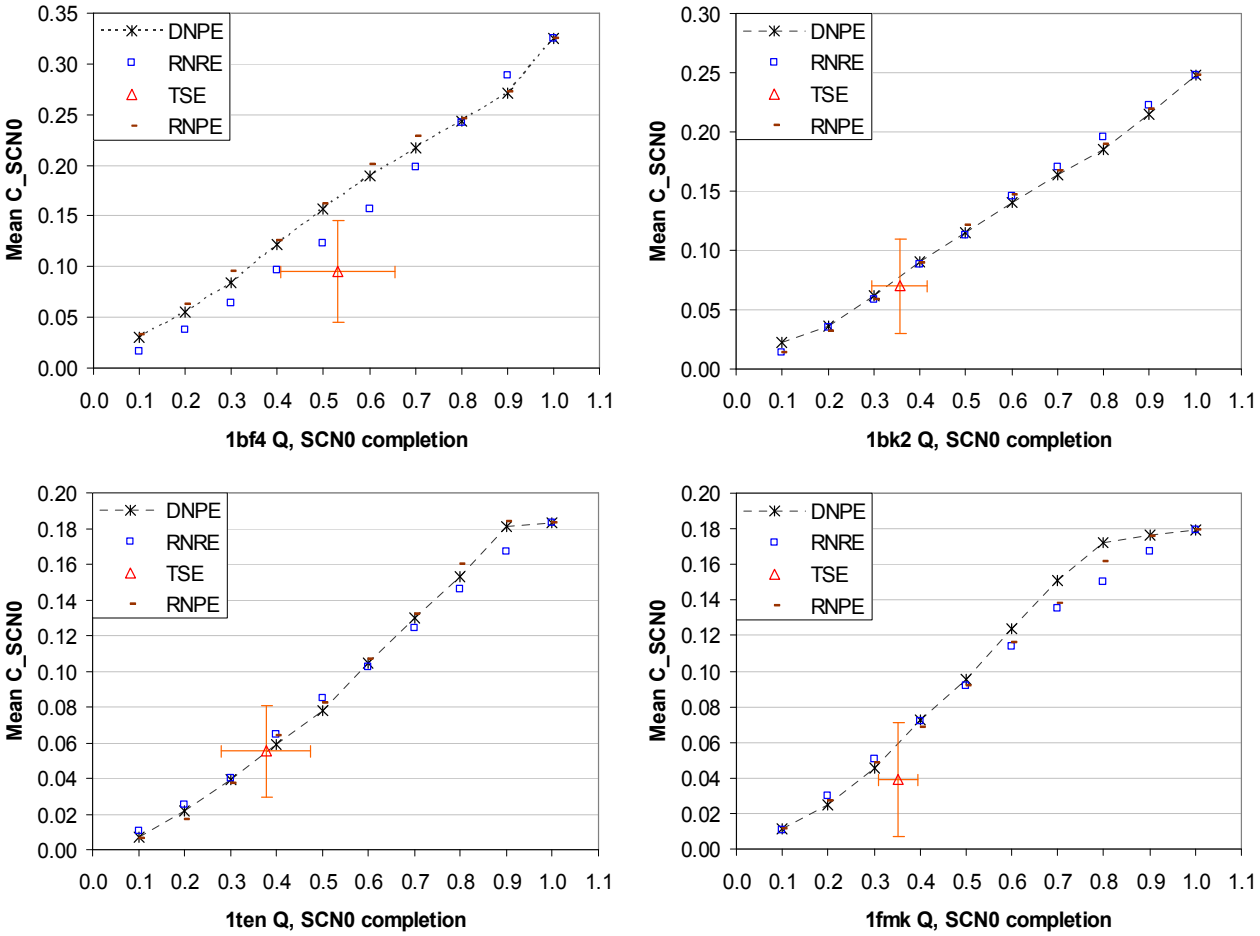
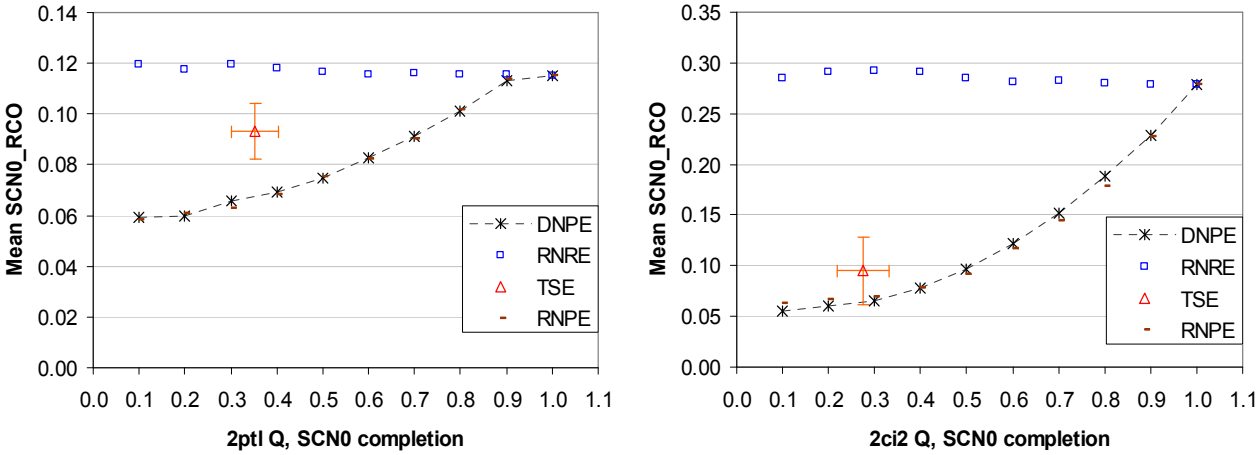
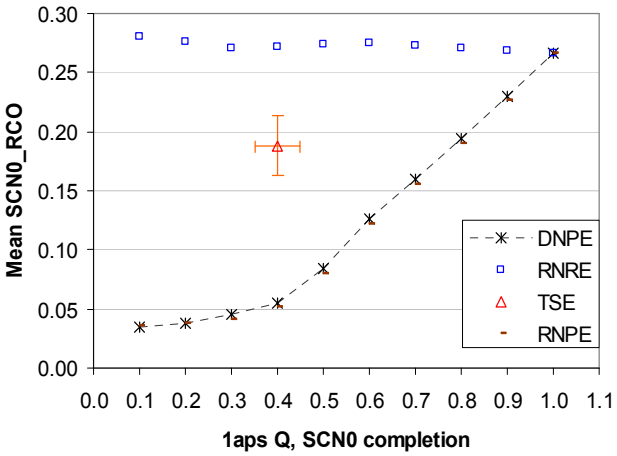
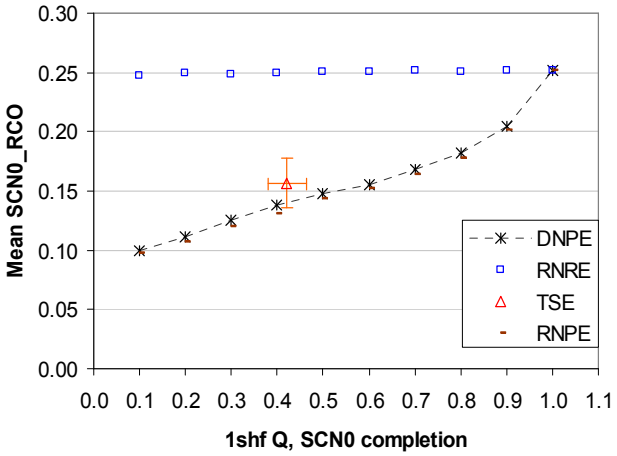
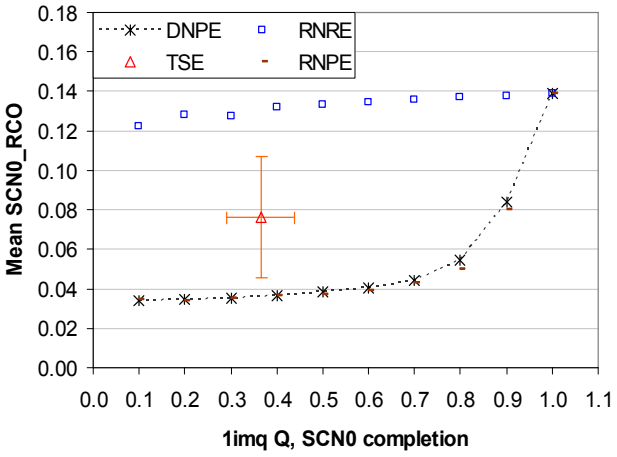


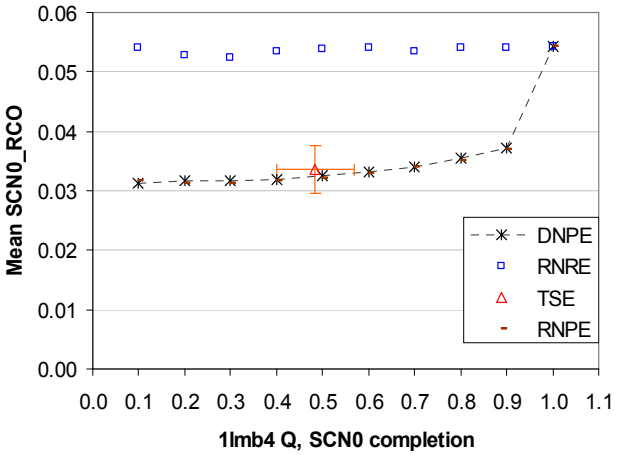
Fig. E4 *C* of ND generated SCN0s (averaged over 100 independent runs) for the 10 PTS (TSE) proteins. The TSE point in each plot marks the average *C* of PTS SCN0s at PTS *Q* (Table 3). Error bars denote one standard deviation about the mean.

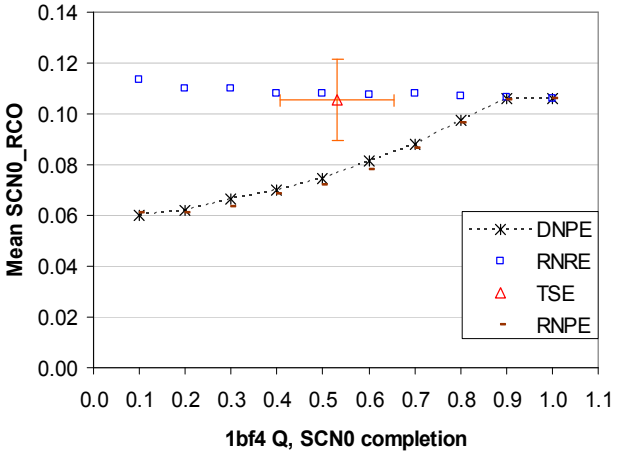


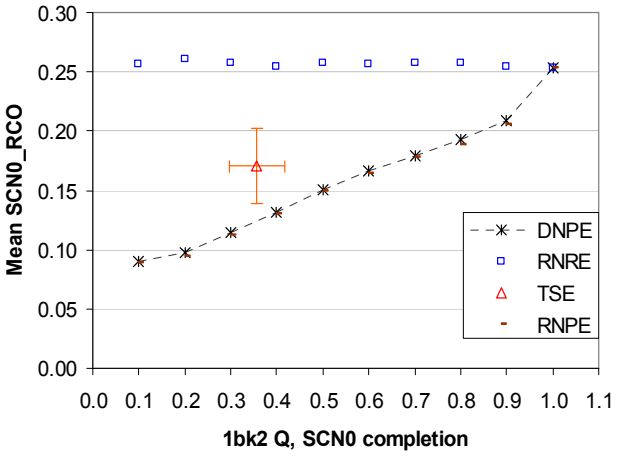

Mean SCN0_RCO
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1
DNPE
RNRE
TSE
RNPE
1aps Q, SCN0 completion


Mean SCN0_RCO
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1
DNPE
RNRE
TSE
RNPE
1shf Q, SCN0 completion


Mean SCN0_RCO
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1
DNPE
RNRE
TSE
RNPE
1imq Q, SCN0 completion


Mean SCN0_RCO
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1
DNPE
RNRE
TSE
RNPE
1lmb4 Q, SCN0 completion


Mean SCN0_RCO
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1
DNPE
RNRE
TSE
RNPE
1bf4 Q, SCN0 completion


Mean SCN0_RCO
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1
DNPE
RNRE
TSE
RNPE
1bk2 Q, SCN0 completion

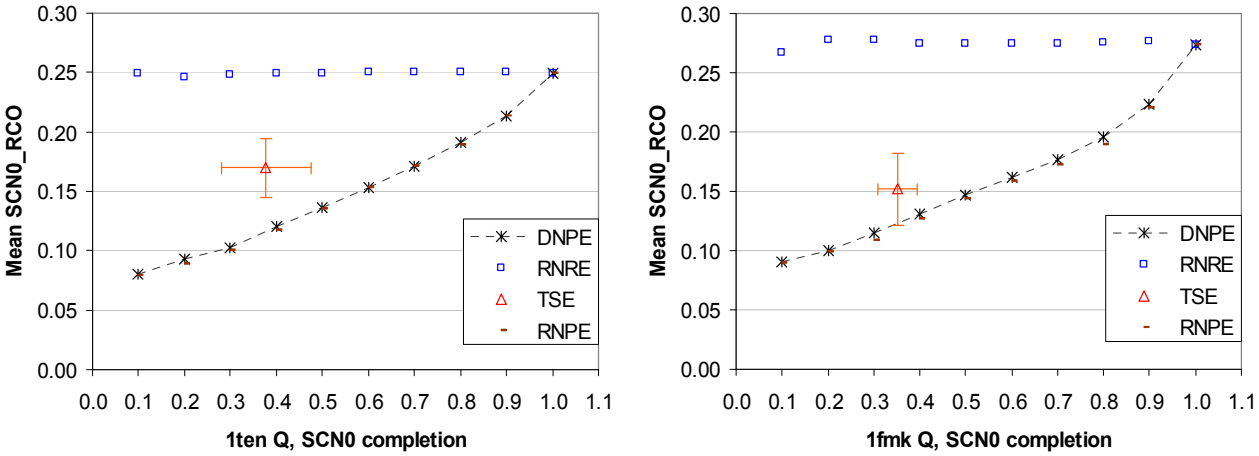


Fig. E5 RCO of ND generated SCN0s (averaged over 100 independent runs) for the 10 PTS (TSE) proteins. The TSE point in each plot marks the average RCO of PTS SCN0s at PTS *Q* (Table 3). Error bars denote one standard deviation about the mean.

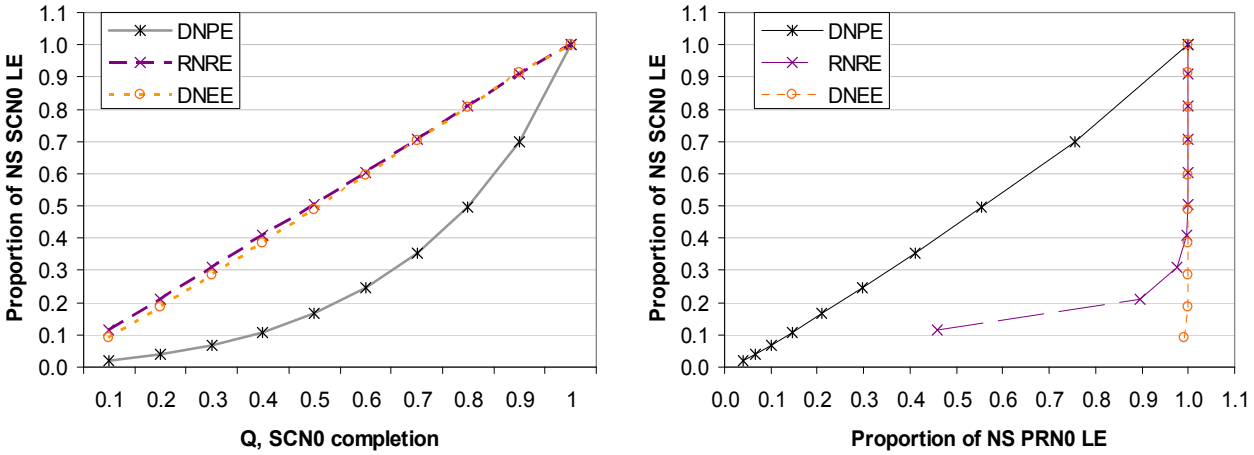


Fig. F1 Left: Proportion of long-range (sequence distance > 10) native-state shortcuts (NS SCN0 LE) in ND generated SCN0s. **Right:** Proportion of NS SCN0 LE as a factor of the proportion of NS PRN0 LE generated by ND. The proportions are the average over 100 independent ND runs per 52 two-state proteins.

Table F1 Comparison of long-range native shortcuts present in ND-TS SCN0s with those in PTS.

PDB id	(1) DNPE *Q* bh	(2) DNPE SCLE	(3) Match	(4) RNRE SCLE	(5) Match
1lmb4 α	0.3	1	1	6	0.33
1bf4 αβ	0.3	13	1	22	0.82
2ptl αβ	0.3	12	0.5	28	0.43
1imq α	0.2	0	0	25	0.40
1ten β	0.3	64	0.75	105	0.70
1shf β	0.2	21	0.76	46	0.54
1bk2 β	0.2	12	0.67	48	0.58
1aps αβ	0.3	11	1	78	0.68
1fmk β	0.2	16	0.56	46	0.43
2ci2 αβ	0.3	12	0.67	46	0.24

(1) *Q* where DNPE barrier height (bh) is measured, also locates the TS for a protein.
SCLE = long-range native shortcuts
(2) Number of SCLE present in at least 10% of DNPE generated SCN0s at DNPE *Q* bh.
(3) Proportion of SCLE in (2) which is also SCLE in at least 10% of phi-generated TS structures (PTS).
(4) Number of SCLE present in at least 10% of RNRE generated SCN0s at DNPE *Q* bh.
(5) Proportion of SCLE in (4) which is also SCLE in at least 10% of phi-generated TS structures (PTS).

Table F2 Secondary structure element (SSE) pair analysis of the DNPE long-range native shortcuts (SCLE) from Table F1 for three model proteins. SCLE are identified by node id (nid), which starts at 0 and ends at *N*-1.

2PTL	SCLE	SCN0	PTS
SSE pair	nid-nid	/100	/126
1S-3S	5-20	11	126
1S-3S	5-22	15	121
1S-3S	6-19	18	126
1S-3S	6-12	13	126
1S-3S	7-18	15	86
1S-3S	8-19	23	81
6T-9S	45-61	13	2
7S-9S	46-59	19	3
7S-9S	46-60	12	2
7S-9S	46-61	16	1
7S-9S	47-59	13	9
7S-9S	48-60	18	3

1S-3S is the N-terminal β-hairpin
7S-9S is the C-terminal β-hairpin

1APS	SCLE	SCN0	PTS
SSE pair	nid-nid	/100	/29
1S-7S	6-51	10	17
1S-7S	12-47	10	25
2T-6T	15-44	11	15
4T-7S	33-52	10	6
4T-7S	34-51	23	7
4T-8T	34-53	13	6
5S-7S	35-52	19	8
5S-7S	36-49	25	26
5S-7S	36-51	23	23
5S-7S	37-48	28	27
5S-7S	38-49	23	27

5S-7S is the two central β strands; it corresponds to β2-β3 in the naming scheme by [Vend01]

2CI2	SCLE	SCN0	PTS
SSE pair	nid-nid	/100	/184
4T-7S	27-46	11	2
5S-7S	28-47	11	11
5S-7S	29-48	10	26
5S-7S	30-47	12	27
5S-7S	30-49	14	100
5S-7S	31-48	10	104
5S-7S	31-50	11	114
5S-7S	32-49	15	102
5S-7S	32-51	10	42
5S-7S	33-50	10	88
7S-9S	47-62	19	4
7S-9S	47-64	16	3

5S-7S is the parallel β-strand pair.