Solution Ensemble of the C-Terminal Domain from the Transcription Factor Pdx1 Resembles an Excluded Volume Polymer

Erik C. Cook,† Debashish Sahu,† Monique Bastidas,† and Scott A. Showalter*§†‡

†Department of Chemistry and ‡Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, United States

ABSTRACT: The pancreatic and duodenal homebox 1 (Pdx1) is an essential pancreatic transcription factor. The C-terminal intrinsically disordered domain of Pdx1 (Pdx1-C) has a heavily biased amino acid composition; most notably, 18 of 83 residues are proline, including a hexaproline cluster near the middle of the chain. For these reasons, Pdx1-C is an attractive target for structure characterization, given the availability of suitable methods. To determine the solution ensembles of disordered proteins, we have developed a suite of 13C direct-detect NMR experiments that provide high spectral quality, even in the presence of strong proline enrichment. Here, we have extended our suite of NMR experiments to include four new pulse programs designed to record backbone residual dipolar couplings in a 13C,15N-CON detection format. Using our NMR strategy, in combination with small-angle X-ray scattering measurements and Monte Carlo simulations, we have determined that Pdx1-C is extended in solution, with a radius of gyration and internal scaling similar to that of an excluded volume polymer, and a subtle tendency toward a collapsed structure to the N-terminal side of the hexaproline sequence. This structure leaves Pdx1-C exposed for interactions with trans-regulatory co-factors that contribute with Pdx1 to transcription control in the cell.

INTRODUCTION

Intrinsically disordered proteins (IDPs) have fascinated both the biological and physical chemistry communities because of their ability to mediate interactions that are critical for signal transduction in the cell and because of the heterogeneous conformational ensembles that ideally suit them to this function. IDPs do show some of the same structural characteristics as their folded counterparts, such as the propensity for secondary structure, globular regions, and long-range contacts; their defining feature is not necessarily the absence of such structures, but rather their transient nature in the overall solution ensemble. That said, not all IDPs display significant secondary structure elements or compaction. For example, polyelectrolytic regions and high total charge polyampholytes tend to exist as extended self-avoiding random flight chains. Substantial effort has been invested in classifying the ensemble characteristics of IDPs, with common models from the polymer physics community having emerged as especially helpful tools for this effort. For example, detailed polymer description of size and shape in the ensemble of the disordered protein NUS helped to reconcile apparent discrepancies between small-angle X-ray scattering (SAXS) and FRET data that have been described in the literature for several years.

In addition to the powerful data available from SAXS and FRET, advances in NMR spectroscopy now make it possible to collect extensive, site-resolved data that describe the solution behavior of IDPs. Having recognized this, the community has begun to call for development of advanced NMR techniques that provide constraints on statistical IDP ensembles. In our laboratory, we seek to fulfill this need through the development of 13C direct-detect NMR techniques that dramatically improve the completeness of NMR data sets for disordered systems. These efforts have allowed us to transform our early and qualitative identification of the α-helical structure in the C-terminus of the phosphatase FCP1 into a quantitative model for its solution ensemble. In addition, 13C direct-detect techniques have allowed us to quantitatively demonstrate sequence-driven enrichment in the cis-proline content of phosphorylated RNA polymerase II C-terminal domain and connect this conformational switch to readout by the downstream enzyme Ssu72.

Motivated by our work with RNA polymerase II, we have turned our attention broadly to quantitative description of proline conformational states and the impact proline enrichment has on the conformational ensembles of IDPs. Proline has long confounded NMR spectroscopists, owing to its lack of an amide hydrogen when found in polypeptide chains and

Received: October 15, 2018
Revised: December 10, 2018
Published: December 10, 2018
traditional reliance on the amide $^1$H-nucleus for NMR detection. Detection of the $^{13}$C-carbonyl nucleus removes this barrier and makes comprehensive NMR analysis of proline-rich IDPs practical. Proline-rich regions are both common in biology, where they often mediate protein–protein interactions, and of interest from a physical–chemical perspective. As proline residues lack an amide hydrogen, they possess no hydrogen bond donor with which to form stable secondary structures. Coupled to this, their reduced solvent exposure. Thus, proline paradoxically promotes "stable secondary structures. Coupled to this, their reduced glucose conditions and promotes the proteolytic degradation of Pdx1. Thus, a better understanding of the Pdx1-C structure and its relationship to function would be clinically significant.

Here, we describe the solution ensemble of Pdx1-C, which we find to be highly extended and devoid of a secondary structure outside the polyproline-II helix formed by the hexaproline cluster. Our work relies on a combination of SAXS data with measurements of backbone $^{15}$N spin relaxation, residual dipolar couplings (RDCs), and NMR paramagnetic relaxation enhancement (PRE). Significantly, we find that the extent of nonlocal contacts reported through $^{13}$C direct-detect PRE depends on whether the excitation nucleus is also $^{13}$C ("protonless" $^{13}$C, $^{15}$N-CON detection$^{15}$) or $^1$H ($^{12}$C,$^{15}$N-(HACA)-CON detection$^{16}$). The experimental PRE profiles documented here are consistent with simulated PRE profiles derived from Monte Carlo (MC) simulations using the Absinth force field,$^{17}$ as implemented in the CAMPAKI software package,$^{18}$ and indicate a high similarity to a self-avoiding polymer chain. These findings illustrate how Pdx1-C is able to maintain an accessible conformation in solution, required for interactions with SPOP and potentially other co-regulators of gene expression in the $\beta$-cell.

## MATERIALS AND METHODS

### Construct Generation.
A DNA insert coding for the C-terminal residues of Pdx1 (Pdx1-C; residues 205–283 of the human sequence) was synthesized by GeneArt (Invitrogen) and subcloned into pET-49b. The expression of recombinant Pdx1-C from the resulting plasmid yields an N-terminal glutathione-S-transferase and a 6x histidine tag, separated from Pdx1-C by a 3C Protease cleavage site. The mutant Pdx1-C C227S, S273C, required for PRE studies, was generated using site-directed mutagenesis, per manufacturer protocols for the Q5 site-directed mutagenesis kit (New England Biolabs). Plasmids encoding wild-type or mutant Pdx1-C were transformed into BL-21 (DE3) Escherichia coli for protein expression.

### Protein Expression and Purification.
To make protein with naturally abundant isotopes, as needed for NMR measurements, BL-21 (DE3) E. coli were grown in Luria broth. For all NMR experiments, the E. coli were grown in M9 minimal media supplemented with $^{15}$N-ammonium chloride and U-$^{13}$C glucose. In either case, E. coli were incubated at 37 °C under agituation until an OD$_{600}$ of 0.6–0.8 was reached. Pdx1-C expression was induced by adding IPTG to a final concentration of 0.5 mM, after which cells were incubated for 4 h at 37 °C under agituation. Cells were harvested by centrifugation and lysed by sonication. Cell debris was cleared through centrifugation and cell lysate applied to a Ni-NTA column pre-equilibrated with 50 mM Tris (pH 7.5), 500 mM sodium chloride, and 20 mM imidazole. The Ni-NTA column was then washed with 10 column volumes of 50 mM Tris (pH 7.5), 500 mM sodium chloride, 20 mM imidazole, and 0.1% Triton X-100. Proteins were eluted from the Ni-NTA column with 50 mM Tris (pH 7.5), 500 mM sodium chloride, and 200 mM imidazole. The affinity tags on Pdx1-C were cleaved with 3C-protease and buffer-exchanged into 50 mM Tris (pH 7.5) and 500 mM sodium chloride. Subsequently, Pdx1-C was separated from the cleaved affinity tags by passing the protein solution over a Ni-NTA column pre-equilibrated with the same buffer. Additional size exclusion chromatography purification was achieved through passage over a Sephacryl-200 column, which efficiently removed high molecular weight contaminants and any cleaved affinity tags that may have co-eluted from the Ni-NTA column with Pdx1-C.

### Small-Angle X-ray Scattering.
Pdx1-C samples for SAXS were buffer-exchanged into 50 mM HEPES (pH 7.0), 100 mM NaCl, and 5% glycerol (to protect against radiation damage). Samples were concentrated to $>5$ mg/mL as determined using a Direct Detect Fourier transform infrared spectrometer (EMD Millipore). SAXS data were collected at the Cornell High Energy Synchrotron Source (CHESS) on the G1 beamline. Incident radiation was estimated to be 9.9 keV, producing a flux of $8 \times 10^{13}$ photons s$^{-1}$, providing a $q$-space range of 0.01–0.7 Å$^{-1}$. Scattering from a silver behenate standard was used for $q$-axis mapping. Data collection was performed using dual Pilatus 100K-S detectors. Reduction of the 2D images to 1D scattering profiles was performed using BioXTAS RAW. For each experimental run, 50 μL of Pdx1-C solution was injected onto a Superdex 200 Increase (5 x 150 mm) column at a flow rate of 0.2 mL min$^{-1}$ for in-line SEC-SAXS data acquisition. Final reported SAXS data sets represent the average of three SEC-SAXS runs, each of which produced a total of >100 exposures stored as 20 s frames. Solvent subtraction was performed using equivalent numbers of frames of the Pdx1-C elution peak and surrounding baseline data. No signs of aggregation, interparticle effects, or radiation damage were observed. Guinier fitting was performed using the method of nonlinear least squares in BioXTAS RAW. Data fitting of the Guinier region was restricted to the data points, satisfying $qR_g < 0.8$, as has been recommended for highly flexible proteins.

### General NMR Spectroscopy.
Chemical shift assignments of Pdx1-C have been published previously and are deposited in the BMRB (access code 19596). Additional chemical shift assignments of MTS tagged Pdx1-C, used for the PRE
The sample was incubated overnight at 4 °C of MTSL, relative to the Pdx1-C molarity, was then added and buffer-exchanged into 50 mM sodium phosphate (pH 6.5), 50 mM potassium chloride, 0.01% sodium azide, and 10% (v/v) D₂O for spectroscopy. ¹³C,¹⁵N-Pdx1-C and ¹³C,¹⁵N-(HACA)CON spectra were acquired on this paramagnetic sample as described below. Following data acquisition, the MTSL radical was quenched using a 10× molar excess of sodium ascorbate. Identical NMR spectra were acquired under diamagnetic conditions. All spectra for the PRE measurements were collected with 32-scans of signal averaging and a 1024 × 512 data matrix. The sweep width for the direct ¹³C-dimension was set to 18 ppm, centered at 173 ppm, while the indirect ¹⁵N-dimension had a 40 ppm sweep width centered at 124 ppm.

NMR Spin Relaxation. ¹⁵N R₂ and R₁ spin relaxation parameters were determined using [¹H/¹H]-start CON experiments previously developed in our laboratory. The R₂ and R₁ spin relaxation experiments were collected as pseudo-3D spectra with a data matrix of 1024 × 512 points for each of the eight interleaved acquisitions. Spectra were collected with 64-scans of signal averaging per FID. The R₂ data were collected as a randomized series of ¹³C,¹⁵N-CONs with relaxation delays set to 50, 150, 250, 350, 450, 550, 650, and 750 ms. For the R₁ data set, the relaxation delays were 16, 48, 80, 112, 144, 176, 208, 240, 272, and 304 ms. The sweep widths for the relaxation experiments were set to 19 ppm for the direct ¹³C-dimension, centered on 173, and 28 ppm for the indirect ¹⁵N-dimension, centered on 120 ppm.

RDC Measurements. NMR samples were prepared by dissolving Pdx1-C at a concentration of 1.1 mM in dilute buffer [50 mM sodium cacodylate, pH 6.5, 50 mM potassium chloride, 5 mM TCEP, 0.01% sodium azide, and 10% (v/v) D₂O] prior to transfer into a Shigemi tube. The aligned sample was generated by swelling a neutral 7% polyacrylamide gel in D₂O prior to transfer into a Shigemi tube. The aligned sample was then collected with 32-scans of signal averaging per FID.

The Journal of Physical Chemistry B

DOI: 10.1021/acs.jpca.8b10051
Figure 1. SAXS characterization of Pdx1-C reveals an extended and disordered structure in solution. (A) Raw scattering data for Pdx1-C at concentrations ranging from 3.8 to 9.8 mg/mL. (B) Raw data in the Guinier region (red circles) and Guinier fits (black lines) for Pdx1-C. Data were fit to the Guinier approximation in Python using the method of nonlinear least squares. (C) Kratky plots generated from the data in panel (A) show a linear increase as a function of $q$ in the high-$q$ region, which is suggestive of a disordered structure in solution. (D) Pairwise distance distribution plot, which demonstrates few features and suggests an extended, disordered structure for Pdx1-C in solution.

where $R_{ij}^p$ is the PRE due to the spin probe in the paramagnetic form, $R_0$ is the intrinsic transverse relaxation rate of the protein in the absence of the spin probe, and $\tau$ is the total INEPT transfer time in the preparation element of the pulse sequence. To estimate $R_0$, we used the average full-width at half-height for all analyzed peaks in the spectrum. For each site, the distance separating the excitation $^1H$ or $^{13}C'$ and the unpaired electron is calculated as

$$ r = \left[ \frac{K}{R_0^p} \left( 4\zeta^2 + \frac{3\zeta^2}{1 + \omega_0^2 \text{Ho}^2 / \gamma^2} \right) \right]^{1/6} $$

where $\omega_0$ is the Larmor frequency of the excitation nuclear spin and $\tau_c$ is the rotational correlation time, estimated from the molecular weight of Pdx1-C using previously described methods that account for the internal flexibility of IDPs.$^{32}$ The parameter $K$ is composed of fundamental constants

$$ K = \frac{1}{15} S(S+1) \gamma^2 g^4 \beta^2 $$

where $\gamma$ is the gyromagnetic ratio of the nuclear spin, $g$ is the electronic $g$ factor, and $\beta$ is the Bohr magneton. The parameters and constants used in our calculation were set to $R_{2H} = 13$ s$^{-1}$, $R_{2C} = 8.1$ s$^{-1}$, $\tau = 9.0 \times 10^{-3}$ s, $\tau_c = 1.2 \times 10^{-9}$ s, $\omega_0^H = 2\pi \times 500 \times 10^6$ s$^{-1}$, $\omega_0^H = 2\pi \times 126 \times 10^6$ s$^{-1}$, $K^H = 1.23 \times 10^{-34}$ cm$^6$ s$^{-2}$, and $K^C = 7.778 \times 10^{-34}$ cm$^6$ s$^{-2}$.

The sulfur atom in the side chain of C227 or C273 was used as an estimate for the location of the unpaired electron in MTS. Distances between the sulfur atom and each C' or H$_2$ were used to estimate the reduction in the peak intensity using previously described methods.$^{3,31}$ The scaling profiles were calculated as the average distance between two residues in 3-dimensional space over all simulation frames, $\langle R_{ij} \rangle$, as a function of the sequence separation in the primary structure, $j - i$, using

$$ \langle R_{ij} \rangle = \left( \frac{1}{Z_{ij}} \sum_{m \in I} \sum_{n \in J} |r_m - r_n| \right) $$

where $i$ and $j$ are residue indexes; $m$ and $n$ are atoms in residues $i$ and $j$, respectively; $Z_{ij}$ is equal to the number of unique pairwise distances; and $r$ indicates the position vector for an atom in a particular residue.

## RESULTS AND DISCUSSION

**SAXS of Pdx1-C.** Intrinsically disordered regions by definition adopt multiple conformations in space and/or time, necessitating characterization of both their average ensemble characteristics and the details of individual high-probability structural features. To begin this study of Pdx1-C, we performed SAXS measurements with the aim of determining what features are present in its solution ensemble at the nm length scale. In order to control for aggregation or other interparticle interactions in the beamline, we made measurements at three starting concentrations (3.8, 7.6, and 9.8 mg/mL) prior to loading on a Superdex 200 Increase (5 × 150 mm) column from which the eluate solution flowed directly into the X-ray beamline for SAXS analysis. Data were averaged over the set of frames spanning the SEC peak determined to have approximately the same radius of gyration ($R_g$). For each run, buffer subtraction was achieved by utilizing a matched number of frames from the SEC chromatogram baseline. The raw scattering data are consistent with preservation of the same assembly state in solution across the full concentration range (Figure 1A), and the low-$q$ region in particular suggests that Pdx1-C is not aggregating under our conditions. Analysis of the data in the Guinier region of the scattering data (Figure 1B) is consistent with a solution $R_g$ of $27.4 \pm 0.8$ Å for the 3.8 mg/mL data set, which will be the source of all discussed statistics and fitting parameters going forward (see Table S1 for a full summary). The fitted $R_g$ value is not consistent with the expectation for a cooperatively folded
83-residue monomeric protein, but is reasonable for a monomeric yet disordered chain (and with the molecular weight determined by in-line MALS following exit from the SEC column; data not shown). Similarly, Kratky analysis suggests that Pdx1-C lacks a closed volume in solution, as evidenced by the divergence of the dimensionless Kratky data at high-\(q\) (Figure 1C). In summary, the SAXS profile of Pdx1-C is consistent with the presence of an ensemble of unfolded, but monomeric, conformers in solution.

Fourier transformation of SAXS data reveal additional details of the conformational ensemble adopted by proteins in solution. Pairwise distance distribution analysis of Pdx1-C reveals a smooth distribution, best described by \(R_g = 26.2\) Å (in excellent agreement with the Guinier analysis) and a maximum intramolecular distance of 95.7 Å. The pairwise distance distribution of Pdx1-C is otherwise surprisingly smooth and devoid of identifiable features (Figure 1D). For comparison, the experimental \(R_g\) compares well to the prediction for an excluded volume (EV) random chain (\(R_g^\text{EV} = 27\) Å) and is only slightly extended compared with a re-calibration of the Flory scaling law intended to model IDP behavior (\(R_g^\text{IDP} = 24.6\) Å). These findings are unsurprising, given the high proline enrichment of Pdx1-C and, nearly complete lack of aromatic or large-chain aliphatic residues, modest fraction of charged residues (FCR = 0.20), and very low net charge per-residue (NCPR = −0.012); all of which are expected to favor a disordered conformation in solution.

**Absence of a Secondary Structure Indicated by Chemical Shift and RDC Measurements.** The absence of a stable tertiary structure and the adoption of a relatively extended average conformation are not necessarily predictive of the secondary structure present at individual motifs within a polypeptide chain. Our previous work has produced complete Pdx1-C chemical shift assignments that are consistent with the absence of significant \(\alpha\)-helix or \(\beta\)-strand conformation at any position in Pdx1-C (reproduced in Figure 2A). While our chemical shift data suggest the absence of a secondary structure, they do not rule out the possibility of other persisting and/or highly populated local structural features at variance with respect to the random-coil structure, which might otherwise be well-described as enriched tertiary contacts.

As an orthogonal measure of local order, we collected RDC data for Pdx1-C in the presence of compressed polyacrylamide gels. RDCs are frequently relied on in contemporary investigations of IDP structure because they allow comparison of the relative orientations sampled by individual bond vectors in the polypeptide chain. An additional strength of RDC measurements in the context of protein disorder is their insensitivity to transient long-range interactions, such as those observed in the solution ensemble of \(\alpha\)-synuclein. Consistent with our general strategy of reliance on \(^{13}\)C direct-detect spectroscopy, we have taken inspiration from the prior literature on scalar coupling measurement and developed variants of the \(^{13}\)C\(^{15}\)N-CON for this study. In this way, we have generated pulse programs that are fully described in the Supporting Information, which enabled measurement of \(^{1}D_{\text{NH}}, ^{1}D_{\text{C,H}}, ^{1}J_{\text{NC}}, \) and \(^{1}J_{\text{C,C}}\) (see Figures S6–S9 for representative spectra). For a random flight chain, the \(^{1}D_{\text{NH}}\) RDC profile takes on a bell-shape, with the depth in the center determined by the overall length of the polypeptide. The \(^{1}D_{\text{NH}}\) RDCs for Pdx1-C shown in Figure 2B are consistent with an inverted bell-shaped profile.

Similarly, the \(^{1}D_{\text{C,H}}, ^{1}D_{\text{NC}}, \) and \(^{1}J_{\text{C,C}}\) feature a bell-like character that is also consistent with a coil conformation for Pdx1-C in solution. Taken together, our SAXS, chemical shift, and RDC data suggest that Pdx1-C adopts an extended and highly random ensemble in solution, with little to no evidence for globularity.

**Backbone NMR Spin Relaxation of Pdx1-C.** Conformational rearrangements of proteins often occur on the nanoseconds–microseconds timescales; unsurprisingly, the backbone dynamics of IDPs on these timescales are particularly extensive and have been heavily studied by NMR. Therefore, in order to probe the backbone dynamics of Pdx1-C, we utilized our previously published \(^{13}\)C direct-detector \(^{15}\)N spin relaxation experiments. In the presence of uniform \(^{13}\)C-enrichment, passive scalar couplings can introduce sine-wave distortions onto the exponential decays used to measure \(^{15}\)N spin-relaxation. As previously described, our pulse sequences were designed to minimize the influence of passive coupling; the overall quality of the recorded decays is evident in representative decay curves that span the observed range of systematic distortions (Figure S10). \(T_1\) and \(T_2\) relaxation times are sensitive to both global reorientation and local fluctuations of the chain, as has been extensively modeled in studies of cooperatively folded proteins. For IDPs, where separation of global and local structure can lose its meaning, a variety of analysis techniques have been adopted. In the case of Pdx1-C, the \(T_1\) and \(T_2\) profiles are largely featureless (Figure 3), yet the depressed transverse relaxation times, compared to a folded protein with an equivalent number of amino acids, are again consistent with overall disorder and a lack of collapsed structure. Interestingly, it does appear that the
dynamics in the N-terminal half of Pdx1-C are nonuniform. Residues 210–225 appear to be more flexible than their neighbors, consistent with its inclusion of seven glycine residues. Similar variance in relaxation times in correlation with a biased amino acid representation has been observed in previous studies of IDP dynamics. In contrast, a reduction of $T_1$ and $T_2$ relative to baseline can suggest high contact density or the presence of a secondary structure. While there is a slight depression in the $T_1$ profile near residues 205–210, there is little other evidence for such contacts throughout the chain. In summary, the spin relaxation data suggest a local structure near residues 205–225 that is distinct from the remainder of the Pdx1-C chain, but overall, these data are consistent with the picture of an extended and highly dynamic backbone in solution.

**PRE Data for Pdx1-C.** PRE is a powerful measurement for constraint of IDP ensembles because of its ability to detect minor populations that feature long-range contacts or locally collapsed states. For example, sensitivity to minor states enabled detection of pathological but low-population conformations in the solution ensemble of $\alpha$-synuclein. An additional feature that enhances the value of PRE as a structure constraint for expanded systems is that the high gyromagnetic ratio of the electron, compared with those of nuclear spins, leads PRE to produce distance constraints for contacts out to a range of 20 Å. Here, PRE was measured through attachment of the MTSL spin-label to the native cysteine residue C227, or to an engineered C227S/S273C double mutant.

Consistent with our preference for $^{13}$C direct-detection displayed throughout this study, PRE measurements were recorded in the $^{13}$C,$^{15}$N-CON format. We note that CON-based detection allows measurement of a PRE intensity ratio for most proline residues, but because of spectral overlap, the hexaproline sequence at residues 239–244 was excluded from analysis. Unique to this phase of the study, spectra were acquired in both the proton-start $^{13}$C,$^{15}$N-(HACA)-CON, whereas panels (B,D) report data collected using the carbon start $^{13}$C,$^{15}$N-CON. All data bars represent the average intensity ratio over duplicate measurements. The dashed line indicates the predicted profile for a self-avoiding random flight chain as described in the Materials and Methods. The solid line represents the back-calculated PRE profile from CAMPARI simulation trajectories. (E) Net charge per residue (NCPR) is plotted to reveal local clusters of positive (blue) and negative (red) charge along the primary structure of Pdx1-C. NCPR is calculated as the average charge over a 5-residue window and assigned to the central residue of the window.

Figure 4. PRE data for Pdx1-C demonstrate nonuniform local chain collapse and the absence of long-range contacts. (A,B) PRE profiles for MTSL conjugation to C227. (C,D) PRE profiles for MTSL conjugation to C273, in the C227S/S273C double-mutant. The profiles in panels (A,C) report data collected using the proton-start $^{13}$C,$^{15}$N-(HACA)-CON, whereas panels (B,D) report data collected using the carbon start $^{13}$C,$^{15}$N-CON. All data bars represent the average intensity ratio over duplicate measurements. The dashed line represents the back-calculated PRE profile from CAMPARI simulation trajectories. (E) Net charge per residue (NCPR) is plotted to reveal local clusters of positive (blue) and negative (red) charge along the primary structure of Pdx1-C. NCPR is calculated as the average charge over a 5-residue window and assigned to the central residue of the window.
As stated above, Pdx1-C possesses a native C227 residue that was exploited for our initial PRE measurements. The data for 1H-start (Figure 4A) and 13C-start (Figure 4B) PRE from C227-MTSL reveal two intriguing features. Recall that perturbations in the spin-relaxation profile near C227 suggested the possibility of globular contacts in the N-terminal half of Pdx1-C. Consistent with this observation, C227-MTSL showed short- and medium-range contacts consistent with a globule or compaction bias in the ensemble of Pdx1-C (Figure S2). Moreover, PRE studies showed short- and medium-range contacts consistent with a coil-like ensemble in solution. The radius of gyration was determined from the ensemble of CAMPARI-generated molecular simulations that will be discussed below.

Strikingly, there were no observable long-range contacts between the N-terminal and C-terminal half of Pdx1-C for either construct; all structural perturbations or deviations from random coil behavior are limited to relatively local contacts. This observation was surprising, given that the NCPR indicates a tendency to negative charge near C227 and positive charge near the C-terminus of the protein that provided our initial motivation for selecting C227S/S273C as a second target for PRE studies. On a more local scale, although the N-terminal half of Pdx1-C displays a higher local NCPR, it is noteworthy that this arises mostly from reduced charge mixing compared to the C-terminal half. The degree of segregation in charge has been quantified using the Pappu laboratory in the form of a parameter (κ) that represents charge asymmetry over a sliding window. The N-terminal and C-terminal halves of Pdx1-C have a κ of 0.331 and 0.132. Thus, one possible explanation for the more collapsed structure of the N-terminal region in Pdx1-C is that it is driven by electrostatic interactions between the strongly negative patch near C227 and the positively charged patch consisting of residues 207–213.

CAMPARI Simulations of Pdx1-C. The analysis of SAXS and NMR data presented above left us with several molecular hypotheses that can best be tested through computational modeling of the Pdx1-C solution ensemble. Recently, MC simulations performed in the CAMPARI software package, utilizing the ABSINTH implicit solvation model, have been used to characterize the effects of phosphorylation on the ensemble conformations of Ash1 and the physical properties that contribute to liquid–liquid phase separation of NCID.51,52 Similarly, we have recently succeeded in modeling the C-terminal region of the phosphatase FCP1 using this methodology, yielding results that were consistent with both SAXS and NMR measurements. Therefore, we chose to conduct MC simulations in ABSINTH to generate an ensemble of Pdx1-C for analysis in the context of the current data, with particular emphasis on reproduction of the PRE data sets described above.

Both our SAXS and PRE data indicated that Pdx1-C adopts a coil-like ensemble in solution. The radius of gyration was determined from the ensemble of CAMPARI-generated structures and was determined to be 25.0 ± 3.4 Å, while slightly smaller than expected, this average is indistinguishable within experimental uncertainty from the SAXS-derived Rg of 27.4 ± 0.77 Å. The simulated PRE data, derived from our MC calculation, is plotted as a solid black line in Figure 4A–D. The MC data show extensive short-range contacts between residue 227 and the proximal N-terminal basic patch, but no long-range interaction with residues C-terminal to the central polyproline stretch (Figure 4A,B). Interestingly, PRE experiments performed in either a carbon-start or proton-start format were in excellent agreement with their respective simulated PRE data. This suggests that adjusting the gyromagnetic ratio term in the Battiste–Wagner equation and solving for the peak intensity change was enough to account for the differences in the width of the relaxation enhancement zone surrounding the site of MTSL attachment in carbon-start and proton-start experiments. Similarly, simulated PRE data of Pdx1-C C227S, S273C were in agreement with experimental PRE data (Figure 4C,D). Taken together, these data preclude any persistent long-range intramolecular interaction between the N- and C-terminal halves of Pdx1-C.

The coil-like character of our ensemble is intriguing, given that Pdx1-C features a low fraction of charged residues and net charge per residue, both of which are sequence features that often correlate with collapse to a disordered globule state.5 We note that this general tendency was established for sequences depleted in proline residues, an abundance of which will tend to disfavor chain compaction in the disordered state. Given Pdx1-C’s high enrichment in proline, we next turned our attention to the potential impact of these residues on the chain’s dimensions. Proline is the only residue which appreciably populates the cis-peptide isomerization state in solvated polypeptides (∼5% prolines are cis),53 because Pdx1-C is 22% proline, there is a high probability of a nontrivial amount of cis-proline in any single conformer that contributes to the overall ensemble. Improved MC methods described by Radhakrishnan et al., which were utilized here, can reproduce experimental proline isomer distributions.54 Our MC data showed significant cis-proline in Pdx1-C, with the highest probabilities localized in the center of the protein (Figure 5A). To study the effects of cis-proline on the conformational ensemble of Pdx1–C, we computed two additional MC ensembles that describe the (experimentally hypothetical) states in which all prolines are either completely in the trans or completely in the cis configuration (though small fluctuations in the ω-torsion angles were allowed; see Materials and Methods). Theses all-trans or all-cis MC ensembles were used to compare with the unbiased ensemble, where the ω-bond was allowed to flip between cis and trans isomers for all proline residues. Asphericity versus Rg was determined for each model of the unbiased, all-cis, and all-trans Pdx1-C ensembles, as shown in Figure 5B–D (left panel). In the unbiased trajectory, the mean Rg of the distribution was 25.0 ± 3.4 Å, as described above, and the mean asphericity A = 0.378 ± 0.165 (Figure 5B). Locking all the prolines into a cis-peptide configuration does not change the mean Rg or asphericity, which were 24.8 ± 3.5 Å and 0.370 ± 0.164, respectively. The all-cis proline Pdx1-C trajectory yields a slight decrease in the Rg to 22.5 ± 3.1 Å, but this decrease is not significantly different within error from the mean in the unbiased trajectory; the same was true for the asphericity, which produced a mean value of 0.36 ± 0.16. Thus, even in an all cis-Pro context, the global chain dimensions are only marginally different from those in the state with native isomer distribution. Fitzkee and Rose have elegantly demonstrated that, consistent with expectations from polymer theory, a local nonrandom structure generally has no impact on global coil-like chain dimensions.55 Consistent with this polymer view, our data demonstrate that proline isomerization cannot significantly alter the global dimensions of the Pdx1-C chain.
While proline trans-to-cis isomerization will not collapse Pdx1-C on average, it remains likely that the persistence length of a polypeptide enriched in proline will be greater than that expected for a random polypeptide sequence. The scaling of internal distances for each of the three trajectories was calculated and is shown in Figure 5B. Internal distances are calculated as the average distance between all atoms in a pair of residues versus the sequence separation of those two particular residues. Scaling profiles for EV chains and a Flory random coil (FRC) are shown as references. The internal scaling profile of the unbiased and all-trans proline trajectories was not observably different, but the all-cis trajectory showed a significant overall reduction in the internal scaling parameter.

The analysis above demonstrates that on the global scale, Pdx1-C adopts an open and extended ensemble, inconsistent with globular collapse. On the other hand, our MC ensemble provides us with the opportunity to investigate whether Pdx1-C features segments of its primary structure that collapse into more globular states. Indeed, the PRE data recorded with MTSL attached to C227 suggest the existence of just such a cluster of local interactions. Figure 6 shows a contact map of the 310 K replica from the trajectory of Pdx1-C with unbiased proline isomerization. To construct this map, distances were measured for each pairwise $C_a$ to $C_a$ in a model and averaged over all the models in the trajectory. Small distances are depicted in red, medium distances in yellow, and large distances in blue (quantitative scale bar shown in Figure 6). The most surprising elements of the Pdx1-C primary structure is a 6-residue polyproline segment encompassing residues 239–244. This polyproline region appears to separate Pdx1-C into two structurally isolated segments; few contacts are indicated between residues bridging this segment.

### CONCLUSIONS

NMR is well-suited to characterize the conformational ensembles adopted by IDPs in solution, even in the absence of a stable secondary structure. Here, we have presented extensive NMR characterization of the highly proline-enriched IDP Pdx1-C, relying exclusively on the $^{13}$C direct-detect methodology our laboratory has developed. Toward this broader objective, we have introduced four new RDC experiments, which are available for download from our website. Further, PRE measurements are reported in both the $^{15}$N/$^{13}$C-CON and $^{15}$N/$^{13}$C-(HACA)CON detection format, yielding additional structure constraints. We note that it appears the $^{13}$C,${}^{15}$N-(HACA)CON is more sensitive to long-range contacts, which will merit further investigation with an IDP that features persistent long-range contacts in its solution ensemble. Here, our PRE data show that Pdx1-C is highly extended overall, with a slight tendency to a globular structure at the N-terminus.
We set out to describe the solution ensemble of Pdx1-C, with the purpose of building necessary information regarding the unbound structure of this critical regulatory region of the full transcription factor. Pdx1-C is known to be the target of the E3-ubiquitin ligase substrate adaptor protein SPOP, which promotes Pdx1-C degradation under low blood glucose conditions.\(^\text{18}\) Often in the IDP community, there is a tendency to seek confirmation of a secondary structure, such as partially stable α-helix motifs, in the unbound ensemble of IDPs, as these are presumed to be the site of interactions with biological partners. In the case of SPOP, available crystal structures of the protein’s MATH recruitment domain bound to client protein-derived peptides demonstrate that those peptides bind in an extended format;\(^\text{56}\) there is no secondary structure for one to seek in the unbound ensemble. However, the SPOP binding motif within Pdx1-C is embedded within the region we seek in the unbound ensemble. The SPOP binding strategy reported here should be of general benefit to future efforts toward determination of IDP ensemble properties.

**REFERENCES**


**ACKNOWLEDGMENTS**

This work was supported by a US National Science Foundation CAREER grant MCB-0953918 and grant MCB-1515974 to S.A.S., and NIH NIRS predoctoral fellowship F31GM101936 to M.B. We would like to thank the Penn State ScholarSphere team (http://scholarsphere.psu.edu) for their help establishing permanent and stable web-hosting of our pulse programs. This work is based upon research conducted at the Cornell High Energy Synchrotron Source (CHESS), which is supported by the National Science Foundation and the National Institutes of Health/National Institute of General Medical Sciences under NSF awards DMR-1332208 and DMR-0936384, using the Macromolecular Diffraction at CHESS (MacCHESS) facility, which is supported by award GM-103485 from the National Institute of General Medical Sciences, National Institutes of Health.

**ASSOCIATED CONTENT**

[1] Supporting Information The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.8b10051.

Table of SAXS acquisition and analysis parameters, representative Pdx1-C NMR spectra, timing diagrams for all RDC pulse programs, representative spectra from all pulse programs, and representative spin relaxation decays (PDF)

**AUTHOR INFORMATION**

**Corresponding Author**
*E-mail: sas76@psu.edu. Phone: 1-814-865-2318.*

**ORCID**
Scott A. Showalter: 0000-0001-5179-032X

**Notes**
The authors declare no competing financial interest.


(35) Novák, J.; Haba, N. Y.; Chill, J. H.; Židek, L.; Sklenář, V. 4D Non-Uniformly Sampled HCBCACON and 1J(NC)-Selective HCBCANCO Experiments for the Sequential Assignment and

