INTRODUCTION

Many NMR, FTIR, and X-ray crystallography studies have been used to characterize the structures of crystallin protein aggregates.1–5 Cataracts is a misfolded protein disease that causes blindness and affects a large proportion of the elderly population over the age of 65 (National Eye Institute). Its treatment is invasive, expensive, and performed only when the disease has severely damaged the eye lens. Pathology of human cataractous lenses reveals that cataracts are composed of protein aggregates made up of crystallin proteins. These proteins are degenerative during adulthood and are damaged by radiation exposure and other chemical modifications that obstruct the passage of light through the lens, causing blurred vision that worsens with time.6–9 Even though the causes of cataracts are known, the aggregation mechanism and the three-dimensional structure of these aggregates remain unknown.

There are three classes of crystallins: α, β, and γ. Their solubilities and stabilities in their native states are critical for keeping the transparency of the lens for a lifetime, since crystallins are not regenerated.10–12 α-Crystallins are oligomers that exhibit chaperone activity, interact with unfolded proteins, and prevent aggregation.11–15 The β- and γ-crystallins are believed to be structural proteins involved in aggregation and share similar structures composed of antiparallel β-sheets. Their secondary structures are separated into two domains, labeled as the N-terminal domain (N-td) and the C-terminal domain (C-td), which are further clustered into four subdomains called Greek key motifs that contain antiparallel β-strands (see Figure 1A). A key biological difference between β- and γ-crystallin proteins is that β-crystallins form domain-swapped dimers while the γ-crystallins are monomeric in solution.11–3,16

Wild-type γ-crystallins are stable eye lens proteins in vitro with higher free energies of unfolding than the other crystallin proteins. The two domains of a γ-crystallin protein interact intermolecularly across a hydrophobic domain interface that helps stabilize the protein.17–18 Human γD-crystallin (HyD-Crys) is the third most abundant γ-crystallin found in young human eyes. HyD-Crys is a 173 amino acid protein exhibiting high structural homology with other γ-crystallins.10,19 Its secondary structure is composed of antiparallel β-sheets and loops that are grouped in two domains connected by a loop linker. Mutagenesis studies have revealed that single amino acid substitutions in the HyD-Crys sequence are linked to aggregation and juvenile-onset cataracts.20–26 The aggregation mechanism in the aging lens differs from that of the juvenile-onset cataracts because aging lenses present a large number of modifications such as methionine and cysteine residue oxidation, deamidation, and glycation.27–30

Fluorescence studies in triple tryptophan mutants of HyD-Crys by Kosinski-Collins et al.31 found that the C-td is more structurally stable than the N-td. Unfolding/folding experiments under physiological conditions demonstrated that the protein refolds through sequential structuring of its domains,
where the C-td folds first and the domain interface residues subsequently mediate the refolding of the N-td. Unfolding/folding kinetics studies by Flaugh et al.17,18 revealed the formation of a partially folded intermediate structure containing the C-td in its native state and the N-td in a random coil state. They further suggested that hydrophobic residues in the intercore domain are critical in the unfolding/refolding process, where the refolded C-td plays the role of a nucleating center for the N-td.

Two-dimensional (2D) ultrafast infrared spectroscopy has been used to characterize geometrically ordered structures such as hairpins, amyloid fibrils, and helical dimers for many proteins.32–38 One of the advantages of this technique is its high temporal resolution: detailed information about the kinetics of the proteins can be obtained on time scale regimes inaccessible by other techniques. It can also be applied continuously to study amyloid and protein folding kinetics on time scales of hours to months. The QM/MM computation of 2D infrared spectra has become a valuable tool in the study of the internal dynamics occurring in biological complexes such as amyloid fibrils.37,39–43 Computations are often necessary to fully interpret experimental 2D infrared (2DIR) spectra. A recent study of γ-D-crystallin using isotope labeling of domains and 2DIR spectroscopy by Moran et al.44 determined that fibril nucleation of the α-crystallin occurs exclusively in the C-td. The authors further proposed a structural model of the fibrils composed of a segment of β-sheets formed by the C-td and another segment of random structure formed by the N-td.

In this article, we present a study involving simulation of 2DIR spectra of natively folded γ-D-crystallin and comparison with the experimental results of Moran et al.44 This comparison of experiment and simulation provides new insight into differences in how the native domains are predisposed to misfolding and aggregation.

To the best of our knowledge, this is the first successful comparison of simulated and experimental 2DIR results for a protein of this size. We found excellent agreement between the two. Details of the experimental conditions may be found in ref 44.

## MATERIALS AND METHODS

### Protein Model

The H2D-Crys crystal structure at 1.25 Å resolution has been solved19 and is available in the Protein Data Bank (PDB ID 1HK0). Its sequence consists of 173 amino acids that are grouped into two domains, the N-td and C-td, each of which has two subgroups called Greek key motifs. The four subgroups are labeled as N-td1(2−40), N-td2(41−83), C-td3(88−128), and C-td4 (129−171) as shown in Figure 1A. In order to mirror the experimental studies, the S84C mutant of H2D-Cry (S84C-H2D-Crys) was created and a Met residue was added in the N-Terminus using the Modeller package.45

### Simulation Protocol

Molecular dynamics (MD) trajectories of the S84C-H2D-Crys mutant were obtained by using the NAMD package (version 2.6)46 with the CHARMM27 force field.57 Simulations were carried out in aqueous solutions (TIP3P water model)48 with cubic periodic boundary conditions. The particle-mesh Ewald protocol was used to calculate the long-range electrostatic interactions, and a cutoff distance of 12 Å was used for nonbonded interactions. The 20 ns MD simulations were performed after 2 ns of equilibration with a time step of 1 fs at T = 300 K. To compute the signals, 100 MD snapshots with a time interval of 2 ps from the last 2 ns were harvested. The simulated linear absorption (LA) and 2DIR signals were computed using the SPECTRON package.49,50

A 2DIR experiment uses three input pulses, k1, k2, k3, and k4 (where k4 satisfies k4 = ±k1 + ±k2 + ±k3), that interact with the protein sample with delays t1, t2, and t3. The signal, defined as the transmission change of the k4 beam, is recorded. The signals were computed using the fluctuating vibrational exciton (FVE) Hamiltonian,44 which is based on the Frenkel exciton model. The FVE Hamiltonian, ̂H_{FVE}, has the form

\[
\hat{H}_{FVE} = \hat{H}_{S} + \hat{H}_{E}
\]

where \(\hat{H}_{S}\) is the Hamiltonian of the system, given by

\[
\hat{H}_{S} = \sum_{m} \varepsilon_{m} \hat{\mathbf{b}}_{m}^{\dagger} \hat{\mathbf{b}}_{m} + \sum_{m \neq n} J_{mn} \hat{\mathbf{b}}_{m}^{\dagger} \hat{\mathbf{b}}_{n} + \frac{1}{2} \sum_{m} \Delta_{m} \hat{\mathbf{b}}_{m}^{\dagger} \hat{\mathbf{b}}_{m}^{\dagger} \hat{\mathbf{b}}_{m} \hat{\mathbf{b}}_{m}
\]

and \(\hat{H}_{E}\) is the interaction with the optical field, E(t), expressed as

\[
\hat{H}_{E} = -E(t) \cdot \sum_{m} \mu_{m} (\hat{\mathbf{b}}_{m}^{\dagger} + \hat{\mathbf{b}}_{m})
\]
with the fluctuating local electrostatic environment. The value $\Delta_\text{cm} = 16 \text{ cm}^{-1}$ was used for the anharmonicities. The intermode coupling $J_{\text{inter}}$ was calculated using the transition dipole coupling scheme. Once the Hamiltonian was constructed and diagonalized, the $S^{k}_{\text{xixi,xyxy}}$ signal was computed using eqs 115, 119, and 121 in ref 50:

$$
S^{k}_{\text{xixi,xyxy}}(\Omega, t_2 = 0, \Omega_t) = 2i \sum_{e_t, e_i} \langle \gamma_{x_i} \gamma_{y_i} \gamma_{x_t} \gamma_{y_t} \rangle \Omega \langle \gamma_{x_i} \gamma_{y_i} \gamma_{x_t} \gamma_{y_t} \rangle (\Omega_t)
$$

$$
\gamma_{x_i} = \gamma_{x_t} = 0, \gamma_{y_i} = \gamma_{y_t} = i(\gamma_t + \gamma')
$$

where $I(\Omega)$ and $T_{e'e'}(\omega)$ are the one- and two-exciton Green’s functions, respectively. In this study, we set the broadening factor $\gamma_t$ to $5.5 \text{ cm}^{-1}$. $\Omega_{e'e'e'(\Omega)}$ is the exciton scattering matrix with two incoming ($\epsilon_1$ and $\epsilon_2$) and two outgoing excitons ($\epsilon_1$ and $\epsilon_2$), as defined in eq 325 in ref 50.

The signals were plotted using the 2D Fourier transform of the times $t_1 \rightarrow \Omega_t$ and $t_2 \rightarrow \Omega_t$. The signals can be either nonchiral or chiral depending on the polarization configuration of the laser pulses. In the dipole approximation, there are three independent signals to leading zeroth order in $k$: $\text{xyxy}$, $\text{xyyx}$, and $\text{xxyy}$ (where $x$ and $y$ are the polarization of the laser pulses). The absorptive all-parallel signal, $\text{xxxx}$, is given by the linear combination $\text{xxxx} = \text{xyxy} + \text{xyyx} + \text{xxyy}$. Some combinations of nonchiral induced signals can be used to enhance desired features. The $\text{xyyx} - \text{xxyy}$ combination suppresses the population pathways and reveals the coherent quantum dynamics through cross-peaks. The combination $\text{xxyy} - \text{xyxy}$ highlights the population dynamics and eliminates the coherent pathways in stimulated emission. For $t_2 = 0$, we have $\text{xxxx} = \text{xyxy}$, and the combination $\text{xyyx} - \text{xxyy}$ would reveal only excited-state absorption.

We simulated the absorptive 2DIR signals $S^{k}_{\text{xixi,xyxy}}$ and $S^{k}_{\text{xixi,xyxy}}$, where $k_i = -k_1 + k_2 + k_3$ and $k_{\text{II}} = k_1 - k_2 + k_3$. We focused on the amide-I band, which corresponds to the backbone carbonyl vibrational motions. The $\text{xxxx}$ and $\text{xxyy}$ signals were compared with experiment. We further predicted the combination $\text{xyyx} - \text{xxyy}$ signal. All spectra have been plotted using a linear scale.

## RESULTS

### Linear Absorption Spectra of the yD-Crys Protein

The linear absorption (LA) signals in the IR regime for the yD-Crys protein with its respective labeled domains are displayed in Figure 1B. The labeling protocol implemented in this study is that described by Moran et al., one domain is isotopically labeled with $^{13}$C, and the other domain remains intact with $^{12}$C. In the LA signal, the unlabeled protein displays a peak centered at $1638 \text{ cm}^{-1}$. When one of the domains is isotope-labeled, the signal acquires a bimodal shape with two peaks at 1595 and $1638 \text{ cm}^{-1}$ (Figure 1B). The LA signal for the labeled C-td shows that the peak at $1638 \text{ cm}^{-1}$ moves slightly to higher frequency, making the bimodal shape more pronounced than for the labeled N-td.

### 2DIR Signals

The 2DIR signals for the S84C-HyD-Crys mutant and its respective labeled domain samples were computed at $t_2 = 0$. The experimental and simulated all-parallel $\text{xxxx}$ spectra of the mutant protein and its respective labeled domains are compared in Figure 2. For the unlabeled and labeled full protein (Figure 2, $^{12}$C Nt/Ct and $^{13}$C Nt/Ct, respectively), a pair of peaks, one negative and the other positive, are observed in each spectrum, at $\Omega_1 = 1638 \text{ cm}^{-1}$ for the unlabeled protein and $\Omega_1 = 1595 \text{ cm}^{-1}$ for the labeled protein. These frequencies correspond to a structure with antiparallel $\beta$-sheets. The negative peak results from ground-state bleach and excited-state emission pathways, while the positive peak results from excited-state absorption. The broadening of these peaks is due to contributions of other secondary structures formed in the protein. These peaks were reproduced in the simulations (Figure 2, right panels). These pairs are shifted around by $40 \text{ cm}^{-1}$, which is consistent with the effective mass replacement of the carbon. The node slopes were calculated for both of these states to compare the contributions of inhomogeneous versus homogeneous broadening (Table S1 in the Supporting Information). A node slope of 1 represents a perfectly inhomogeneously broadened peak, while a value of infinity represents a perfectly homogeneously broadened peak. The node slope of the unlabeled peak pair is in good agreement with experiment, while the node slope for the labeled full protein shows a 15% deviation from experiment. This deviation may arise because the Hamiltonian (eq 2) accounted for the isotope labeling but the MD simulations used for conformational sampling did not use a heavy carbon, which may have a slight influence on the dynamics.

![Figure 2](image-url)
For the proteins with a labeled domain, two pairs of positive and negative peaks appear in the spectra. The experimental spectra of the proteins with labeled domains (Figure 2, Nt-d and Ct-d) show two pairs of peaks near $\Omega_3 = 1640 \text{ cm}^{-1}$ and $\Omega_3 = 1605 \text{ cm}^{-1}$ in each spectrum. Differences between the experimental signals can be seen. The peak pairs in the signal for the N-td-labeled protein are less intense and more separated than those of the C-td-labeled protein (see Table S1), while the high-frequency pair in the C-td-labeled signal are elongated and broader than those in the N-td-labeled signal. For the low-frequency pair, the N-td-labeled signal shows diffused peaks, especially the negative peak, while the positive peak is well-defined and centered at $\Omega_3 = 1605 \text{ cm}^{-1}$. In the signal for the C-td-labeled protein, the low-frequency peak pair is intense and centered at $\Omega_3 = 1605 \text{ cm}^{-1}$.

Some of these features are reproduced in the simulated signals for the labeled domains (Figure 2, right panels, third and fourth rows). The two show similar profiles with two pairs of peaks separated by 40 cm$^{-1}$. In the N-td-labeled signal, the negative peaks are less broad than those in the C-td-labeled signal. The peak pairs in the labeled N-td signal are more separated than those of the labeled C-td signal, as observed in the experiment and the simulated LA spectra. Comparing the simulated and experimental node slopes, we see that the node slopes attributed to the unlabeled domains are in good agreement with experiment while those attributed to the labeled domains show a larger deviation from experiment (see Table S1). Again, we believe this larger deviation from experiment may arise because all of the MD simulations used for conformational sampling used $^{12}$C in the amide backbone.

We next turn to the $xxyy$ signals. All of the experimental spectra (Figure 3, left panels) have a similar profile as in the $xxxx$ spectra but with less intense peaks at $\Omega_3 = 1640 \text{ cm}^{-1}$ and $\Omega_3 = 1605 \text{ cm}^{-1}$. This pattern is observed in the simulated data as well. However, the peak pairs in the signal for the labeled full protein are stronger and broader than those for the unlabeled full protein, which indicates that the N-td follows different dynamics in which its eigenstates exhibit a wider frequency range despite the high homology in the backbone geometry. We note that in the simulated spectra, we are able to resolve the low intensity cross-peaks better than in the experiment because of the lack of experimental (white) noise in the simulations.

The simulated combination $xxyy = xyyx$ spectra for proteins with labeled domains (Figure 4) show the same pairs of peaks as observed in the $xxxx$ combination. However, the N-td-labeled signal shows more elongated positive peaks along the diagonal than the positive peaks in the C-td. The spectra also show cross-peaks at $\Omega_1 = 1610$ and $1660 \text{ cm}^{-1}$ that are more diffuse for the labeled C-td than the labeled N-td.

Both the experiments and simulations show clear differences in the 2DIR spectra when the N-td or C-td is labeled, despite their high structural homology. The experimental and simulated spectra show some weak cross-peaks that suggest couplings between the backbone amide states. These couplings depend strongly on the geometry of the protein. The domains are connected by a flexible linker loop, and one domain is essentially the reflection of the other domain. The secondary structure propensities for the protein (Figure 5) show reflection between the domains, especially where the propensity to form $\beta$-strands is high, except in the residues $40-41$ and $60-61$ from the N-td2 motif and their corresponding reflections in the C-td2 motif (residues $129-130$ and $147-148$). Other differences are observed in the turn propensities, where the loop between the N-td1 and N-td2 motifs is higher than that between the C-td1 and C-td2 motifs.

Figure 3. Same as in Figure 2 but for the $xxyy$ polarization configuration.

Figure 4. Simulated $xxyy = xyyx$ signals for the isotopically labeled domains.

Figure 5. Secondary structure propensities of the domains of $\gamma$D-crystallin indicating the segments of the sequence that have the probability to form $\beta$-strands or turns.
domains because the couplings are sensitive to distance and orientation. Couplings are observed from intramolecular interactions between residues that form β-sheet structures (antiparallel to the diagonal). Residues in the N-td (residue indexes 1–83) have weaker couplings than the corresponding section in the C-td (residue indexes 88–168). In the N-td, R36 is strongly coupled to T4 and K2. These surface residues are solvent-exposed, and it is known that mutant R36S is linked to congenital juvenile-onset cataracts. Interestingly, Q54 and R58 show a strong coupling with those residues in the segment where R36 is located. The image residue for Q54 in the C-td is Q143. Q54 and R58 are solvent-exposed residues that form one of the β-strands in their respective domains and belong to the set of residues in the interface core. Deamidation of Q54 and Q143 introduces a negative charge at physiological pH into the domain interface and destabilizes HγD-Crys, lowers the unfolding kinetic barrier, and leads to the formation of a partially unfolded intermediate with a structured C-td and an unstructured N-td. The R58H mutant, which results in a partially unfolded intermediate with a structured C-td and an unstructured N-td, is located at the top of the structure and exhibit some couplings. For W68, we have the pairs (W68, Y6), (W68, Y28), and (W68, Y50), and for W157 we have (W157, H88), (W157, F117), and (W157, W132). It is well-established that UV radiation may contribute to the origin of cataracts as a result of absorption by aromatic residues. These tryptophan residues W42, W68, W131, and W157 were studied using fluorescence quenching, and it was determined that the local environment around W68 and W157 originates the native-state quenching of the protein fluorescence emission. During the refolding/unfolding process, W42 and W68 unfold at lower concentrations of GdnHCl than their images residues W131 and W157, respectively. Further analysis of unfolding and refolding of the tryptophan mutants identified an intermediate state along the wild-type HγD-Crys folding pathway with a structured C-td and unstructured N-td that could be a suitable conformational model for cataract formation.

Even though the two domains are geometrically homologous, they make different contributions to the 2DIR signals. The large separation between densities of states for the pair Q54 and Q143 could be due to their exposure to the solvent and their location in the interface domain, which could change the strengths of the interactions with other residues. The tryptophan residues considered here are buried in their respective domains. Although the two domains share basically identical backbone geometries, the frequency shifts upon domain labeling suggest that the kinetics of these domains are different, where the folding in the N-td depends on the transitions occurring in the C-td, which agrees well with experimental reports.

### DISCUSSION AND CONCLUSIONS

Moran et al. used a combined approach of isotope labeling and 2DIR spectroscopy to study the γD-crystallin protein. They demonstrated that the use of this combined tool is valuable for the characterization of the kinetics and structure in proteins, helping to obtain additional information not available from standard methods. In their studies, they expressed labeled and unlabeled domains of the protein separately, and the full-length γD-crystallin protein was constituted by protein ligation using a cystine residue located in the connecting loop. They determined that C-td of the γD-crystallin protein is the leading domain to form amyloid fibrils and that the use of this covalent linkage between domains plays a role in determining the fibril structure, which is a striking result considering that each individual domain can form fibrils on its own.

We note that our simulations are limited to the computation of 2DIR spectra of the monomeric form of HγD-Crys and do not offer a direct comparison with the 2DIR kinetics experiments.

We used a computational approach to complement these results at the level of the native state of the mutant S84C-γD-crystallin monomers. The computed spectra are in excellent agreement with experiment. Although the N-td and C-td are geometrically similar with highly conserved residues in their sequences, the average coupling map analysis shows an

![Figure 6](image-url)
asymmetric contribution of some relevant residues and their corresponding image residues to the signals when one of the domains is isotopically labeled. Experiments by Mills et al.\textsuperscript{59} concluded that the isolated N-td is less stable than the full-length protein and less stable than the C-td. A computational study by Das et al.\textsuperscript{60} confirmed that the stability and the folding mechanism of the N-td are regulated by the interdomain interactions.

Our findings are consistent with the conclusions that Moran et al. stated about the role of the covalent linkage between domains in determining the fibril structure and the prevalent dynamics that the C-td has over the N-td in fibril formation. It will be interesting to use 2DIR spectroscopy to analyze the coupling between domains under different environmental conditions and with various mutations.

As stated in the Introduction, we can consider H7D-Crys in terms of two coupled domains. In a very simple approximation, it is possible to think of this as a coupled two-mode exciton system:

\[
H \approx \begin{pmatrix}
\epsilon_i & J_{ij} \\
J_{ji} & \epsilon_j
\end{pmatrix}
\]

where \(\epsilon_i\) is the mean site energy of domain \(i\) and \(J_{ij}\) is the coupling between modes \(i\) and \(j\). The eigenvalues are

\[
e_{\pm} = \frac{1}{2}(\epsilon_i + \epsilon_j \pm \sqrt{4J_{ij}^2 + (\epsilon_i - \epsilon_j)^2})
\]

When neither domain is iso-label, or both domains are labeled, \(e_x \approx e_y\) and we have (roughly) a linearly coupled homodimer. When either the N-td or the C-td is iso-label, we have a linearly coupled heterodimer.

\[
\epsilon_{\pm} = \frac{1}{2}(\epsilon_1 + \epsilon_2 \pm \sqrt{4J_{ij}^2 + (\epsilon_1 - \epsilon_2)^2})
\]

If the system is symmetric, then \((\epsilon_1 - \epsilon_2)^2 = (\epsilon_1 - \epsilon_2)^2\) and the N-td- and C-td-labeled isotopomers will produce equivalent linear absorption spectra.\textsuperscript{49} However, as seen from the calculated linear absorption spectra, this is not the case, as the peaks for the N-td-labeled isotopomer are less separated than the peaks for the C-td-labeled isotopomer. Since \(J_{ij}\) is the same for the two isotopomers, this indicates that \((\epsilon_1 - \epsilon_2)^2\) is greater for the C-td-labeled isotopomer. This is likely related to the fact that the C-td has a better-defined structure, which leads to a lower approximate site energy.

While the diagonal elements of the Hamiltonian are decreased by isotopic substitution, the off-diagonal elements remain the same because the transition dipoles of the individual amide-I modes, which determine the coupling, do not change upon isotopic substitution. This was verified experimentally by comparing the diagonal cross sections of the 2DIR spectra of the unlabeled and uniformly labeled \(^{13}\)C proteins.\textsuperscript{44}

However, as noted by Fang et al.,\textsuperscript{62} the cross-peaks are a direct measure of the off-diagonal anharmonic coupling. Since we assume zero mechanical off-diagonal anharmonicity, all of the off-diagonal anharmonicity must result from electrostatic coupling. In the weak-coupling limit, which is appropriate here, this off-diagonal anharmonicity can be approximated as

\[
\Delta_{ij} = \frac{-4\Delta_{\mu}}{\epsilon_i - \epsilon_j^2}
\]

where \(\Delta_{ij}\) is the off-diagonal anharmonicity and \(\Delta_{\mu}\) is the diagonal anharmonicity (16 cm\(^{-1}\)). Therefore, systems with a smaller \((\epsilon_i - \epsilon_j)^2\) will have stronger cross-peaks. In our case, the N-td-labeled isotopomer has the smaller value and therefore has stronger cross-peaks.

Using this very simple model, we conclude that the asymmetry in the linear absorption and 2DIR spectra is caused by the differences in the site energies in the two domains. The differences in these site energies are caused by the differing local electrostatic environments of residues within these two domains and likely is related to the secondary structure propensity in each domain.

### ASSOCIATED CONTENT

\section{Supporting Information}

Diagonal peak frequencies and FWHM node slopes for the simulated and experimental xxxx 2DIR spectra (Table S1). This material is available free of charge via the Internet at http://pubs.acs.org.

### AUTHOR INFORMATION

**Corresponding Author**

*E-mail: alamng@uci.edu.*

**Notes**

The authors declare no competing financial interest.

### ACKNOWLEDGMENTS

We gratefully acknowledge the support of the National Institutes of Health (Grants GM059230 and GM091364) and the National Science Foundation (Grant CHE-1058791). N.K.P. was supported by a National Science Graduate Research Fellowship. M.T.Z. acknowledges NIH DK79895.

### REFERENCES


