Title

Learning Correlations between Internal Coordinates to Improve 3D Cartesian Coordinates for Proteins.

Journal

Journal of Chemical Theory and Computation, 19(14)

Authors

Zhang, Oufan
Lee, Seokyoung
Namini, Ashley
et al.

Publication Date

2023-07-25

DOI

10.1021/acs.jctc.2c01270

Peer reviewed

HHS Public Access

Author manuscript

J Chem Theory Comput. Author manuscript; available in PMC 2023 September 28.

Published in final edited form as:

J Chem Theory Comput. 2023 July 25; 19(14): 4689–4700. doi:10.1021/acs.jctc.2c01270.

Learning Correlations between Internal Coordinates to improve 3D Cartesian Coordinates for Proteins

Jie Li[†], **Oufan Zhang**[†], **Seokyoung Lee**[†], **Ashley Namini**[‡], **Zi Hao Liu**[‡,¶], **João M. C. Teixeira**[‡,§], **Julie D. Forman-Kay**[‡,¶], **Teresa Head-Gordon**[†,‖]

[†]Pitzer Center for Theoretical Chemistry, Department of Chemistry, University of California, Berkeley CA 94720, USA

[‡]Molecular Medicine Program, Hospital for Sick Children, Toronto, Ontario M5S 1A8, Canada

[¶]Department of Biochemistry, University of Toronto, Toronto, Ontario M5G 1X8, Canada

[§]current address: Department of Biomedical Sciences, University of Padova, Italy

[‖]Departments of Bioengineering and Chemical and Biomolecular Engineering, University of California, Berkeley, CA, USA

Abstract

We consider a generic representation problem of internal coordinates (bond lengths, valence angles, and dihedral angles) and their transformation to 3-dimensional Cartesian coordinates of a biomolecule. We show that the internal-to-Cartesian process relies on correctly predicting chemically subtle correlations among the internal coordinates themselves, and learning these correlations increases the fidelity of the Cartesian representation. We developed a machine learning algorithm, Int2Cart, to predict bond lengths and bond angles from backbone torsion angles and residue types of a protein, which allows reconstruction of protein structures better than using fixed bond lengths and bond angles, or a static library method that relies on backbone torsion angles and residue types in a local environment. The method is able to be used for structure validation, as we show that the agreement between Int2Cart-predicted bond geometries and those from an AlphaFold 2 model can be used to estimate model quality. Additionally, using Int2Cart to reconstruct an IDP ensemble is able to decrease clash rate during modelling. The Int2Cart

thg@berkeley.edu .

AUTHOR CONTRIBUTIONS

J.L. and T.H.-G. designed the project. J.L. designed and wrote the Int2Cart software. O.Z. also provided input on the neural network design and tuning of the model and valuable critiques including testing. J.L. and T.H.-G. wrote the paper and all authors provided valuable input and discussion.

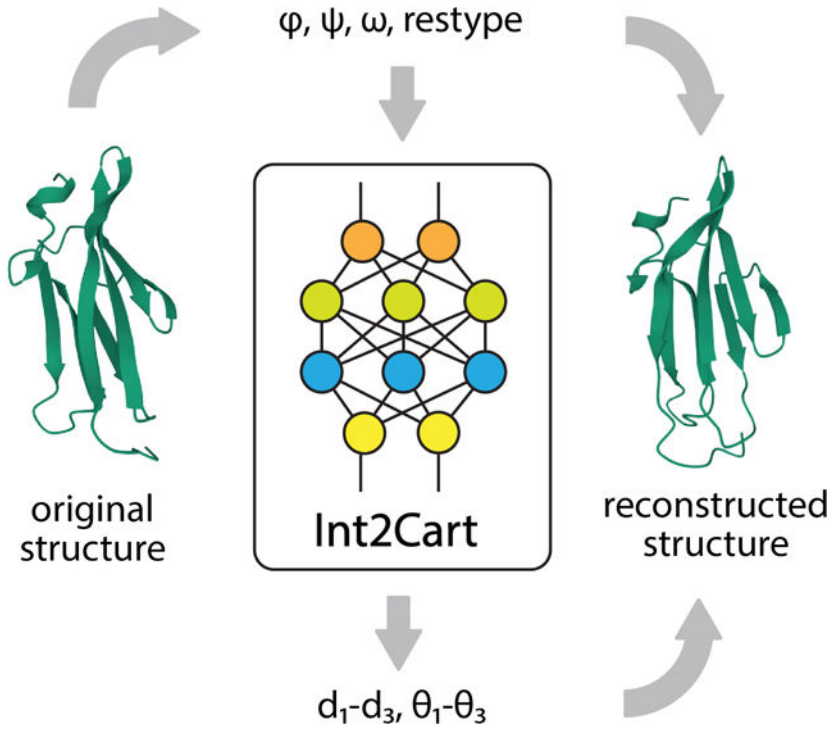
Supporting Information

- Identifiers in the AlphaFold Protein Structure Database for all AlphaFold2 proteins used in this study
- Fixed parameters for scaling predicted bond lengths and bond angles
- Int2Cart prediction accuracy on backbone bond lengths and bond angles
- Extended analysis and visualization of bond lengths and bond angles as a function of backbone torsion angles and amino acid types
- Comparison of bond angle - ω dependence among structures with different qualities
- Radius of gyration correlations between reconstructed structures using Int2Cart, or Fixed method, and the original structures
- Correlations between AlphaFold2 (AF2) structure qualities and the agreement of bond geometries between Int2Cart predictions and AF2 geometries

This information is available free of charge via the Internet at http://pubs.acs.org

algorithm has been implemented as a publicly accessible python package at https://github.com/THGLab/int2cart.

Graphical Abstract



INTRODUCTION

Biomolecular structures are described using two widely used mathematical representations: internal coordinates and Cartesian coordinates. The internal coordinate representation is defined by a set of bond lengths, bond angles, and dihedral or torsion angles, and provides a compact description in terms of the Z-matrix. In contrast, a Cartesian representation defines all of the atomic positions in Euclidean x,y,z coordinates and additionally captures the orientation of a molecule in space. Both representations are useful in certain contexts and applications. Internal coordinates can be beneficial for geometry optimizations [1] and are the preferred description for NMR structure determination and refinement as an intermediate step towards an atomistic structure. The bond lengths and bond angles are typically taken as fixed[2] in these scenarios. Cartesian coordinates are the preferred format of molecular dynamics simulations[3] and X-ray crystallography, NMR, and cryo-EM structures deposited in the Protein Data Bank (PDB) repository.[4]

Figure 1 considers the internal coordinates of a protein backbone that contains the three torsion angles $\phi(C-N-C_\alpha-C)$, $\psi(N-C_\alpha-C-N)$, and $\omega(C_\alpha-C-N-C_\alpha)$, bond lengths $N-C_\alpha(d_1)$, $C_\alpha-C(d_2)$, and $C-N(d_3)$, and bond angles $N-C_\alpha-C(\theta_1)$, $C_\alpha-C-N(\theta_2)$, and $C-N-C_\alpha(\theta_3)$; side chain information that may affect the backbone could also include $C_\alpha-C_\beta(r_1)$ and $N-C_\alpha-C_\beta(\alpha_1)$ for example. When all of these quantities are specified exactly,

the back-transformation from internal coordinates will also result in a perfect 3D Cartesian reconstruction of the protein backbone structure, using algorithms such as the natural extension reference frame (NeRF).[5] However, in certain areas of protein modelling, such as fragment-based protein folding and loop modelling,[6,7] the Cartesian reconstruction is almost universally defined by only the backbone torsions while holding the bond lengths and angles fixed at mean values to decrease the complexity of the problem. Sometimes the variations on the ω torsion angles are also ignored and taken as fixed values of 0° or 180°, due to the planar nature of the peptide bond.[8] One might assume that a protein structure can be reconstructed in Cartesian coordinates quite well utilizing fixed bond lengths and angles since they typically have quite small variations around their means. However, even small deviations from the mean of the stiff degrees of freedom can strongly influence the Cartesian reconstruction.

The origin of this error arises especially clearly from the nature of chain molecules: as the protein chain gets longer, small errors in bond lengths and bond angles can quickly accumulate and result in significant differences in the final back-transformed structure. According to a study by Holmes et al.[9] on globular proteins, the RMSD errors incurred in the internal coordinate back-transformations to C_α Cartesian positions under fixed bond lengths and angles is ~6 Å for an average 150-amino-acid protein, and can be as high as 40 Å for larger proteins.

Alternatively, one could replace the assumption of fixed bonds and bond angles with a statistical approach that uses variable bond lengths and bond angles according to sequence or structural correlations in the PDB. Given the many types of correlations that exist between the internal coordinates of globular proteins, such as the ϕ and ψ torsion angles of the Ramanchandran plot,[10] restraints on ω torsion angles as a function of ϕ and ψ,[11] and the correlation between backbone and sidechain torsion angles used in the Dunbrack rotamer library,[12] the correlations among the stiff bond lengths and angles with the flexible torsions should not be surprising. Earlier studies on the relationship between bond angles and ϕ, ψ torsion angles or amino acid types were mostly focused on the $N - C_\alpha - C$ bond angle, using both statistical methods and quantum mechanics calculation on model dipeptides.[13–16] The work by Berkholz[17] found that by using a static library for backbone bond angles dependent on backbone ϕ, ψ angles and residue types, the median RMSDs of protein reconstruction normalized to 100 amino acids is 2.85 Å. Following this pioneering study, more recent work by Roberto and co-workers have extended the correlation to all bond lengths and bond angles centered on backbone N, C_α and C atoms.[18,19] Similarly, Lundgren et al. studied the correlation between protein backbone angles, secondary structure, and sidechain orientations.,[20] and Ashraya et al. evaluated the steric-clash Ramanchandran maps conditioned on bond geometries.[21] However, none of these studies have considered the correlations in internal coordinates beyond local amino acid context and backbone geometries.

This work provides a more comprehensive machine learning approach that both quantifies and learns internal coordinate correlations within a deeper amino acid sequence context, that in turn provides a more accurate prediction of the 3D Cartesian coordinates relative to

the errors incurred under the standard assumptions of fixed bond lengths and angles. By capturing the subtle correlations observed among internal coordinates, the Int2Cart (Internal to Cartesian) algorithm reduces the reconstruction RMSD error to ~2.07 Å for test proteins normalized to 100-amino-acids, and an average RMSD of ~3.74 Å over the entire test set for globular proteins as large as 599 amino acids. While many current protein modelling algorithms have adopted pairwise distance-based constraints[22–24] or directly output 3D coordinates, thereby bypassing the internal-to-Cartesian conversions,[25,26] the applicability of the Int2Cart algorithm is multi-fold.

First, our bond geometry prediction module is capable of providing more accurate references for internal coordinates, making our method a helpful tool for structural validation and refinement. We demonstrate this Int2Cart application by showing that the agreement between bond lengths and bond angles from AlphaFold2 (AF2) predicted structures[27] and Int2cart predictions is a strong indicator of AF2 model quality. Second, torsion-angle based approaches are still widely used in loop modelling[28] and in generating conformational ensembles of intrinsically disordered proteins (IDPs).[29] We find that Int2Cart is able to reproduce a structural ensemble of the disordered Sic-1 IDP with lower RMSD error when back-calculated to experimental observables, and generates fewer undesirable steric clashes. We also envision that Int2Cart should be applicable in the development of protein force fields that could benefit from more accurate valence models of backbone bond lengths and bond angles conditioned on other geometrical or sequence features. [30]

METHODS

Dataset preparation.

We have adopted SidechainNet[31] as a preprocessed dataset that uses clustering techniques to extract protein sequences and structures with defined similarity cutoffs, to reduce bias in the original PDB structures, and to prevent information leakage from the training set to the test set relevant to assessing the machine learning generalization.[31,32] The SidechainNet dataset represents each protein by its amino acid sequence, backbone and sidechain torsion angles (ϕ, ψ, ω, χ_1, χ_2, etc), backbone bond angles $\theta_1 - \theta_3$, as well as the all-atom 3D coordinates. For this study we ignore the sidechain torsions as we are only reconstructing backbones, and supplement the protein dataset with backbone bond lengths $d_1 - d_3$ calculated from the 3D coordinates for training, validation and test sets. We also identified some θ_2 and θ_3 bond angles that were incorrect due to missing atoms in the next residue, and they were masked out along with the residue at the end of the protein chain. We used the latest available version of the SidechainNet dataset (CASP12) under 70% thinning and combined validation sets from 10% to 50% similarity cutoffs with the test set. This then defines the final test data for our algorithm while keeping track of the similarity for each individual test data point. When needed, we separated test set proteins at any broken chain positions and only retained chains longer than 50 consecutive amino acids.[32]

Our final training dataset contains 41,380 proteins with a minimum sequence length of 20 and maximum length of 4914 amino acids. Most structures in the training set have reported structural resolution < 4Å. The test set was comprised of 182 protein or protein fragments

with sequence lengths between 23 and 599 amino acids. We have additionally compiled an IDP structural ensemble comprised of 1000 conformations for the N-terminal 92 residues of the Sic1 protein[33,34] to validate the transferability of our model in a more challenging application scenario. We extracted the backbone torsions and rebuilt the Cartesian structures for each conformer under different assumptions about the bond lengths and bond angles as reported in Results. In addition, 20 randomly selected proteins from the human proteome were downloaded from the AlphaFold2 database [27] to illustrate the application of Int2Cart in validating protein structure models. The identification codes for these 20 proteins are provided in the Supplementary Information.

Neural network design.

The structure of the deep neural network, Int2Cart, is depicted in Figure 2. The recurrent neural network architecture is chosed due to its capability to capture long-range correlations in internal coordinates, such as torsion angles that are exemplified in applications including protein folding[35] and IDP modelling.[36] We utilized 3 layers of stacked bidirectional gated recurrent units (GRU) as the central component, each of which contains a hidden state h_t with its information updated by the reset and update mechanisms for each element in the input sequence through the following set of equations: [37]

$$r_t = \sigma\left(W^r x_t + U^r h_{t-1} + b^r\right) \tag{1}$$

$$z_t = \sigma\left(W^z x_t + U^z h_{t-1} + b^z\right) \tag{2}$$

$$\tilde{h}_t = \tanh\left(W^n x_t + b^{nx} + r_t \odot \left(U^n h_{t-1} + b^{nh}\right)\right) \tag{3}$$

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1} \tag{4}$$

where $[W^r, W^z, W^n, U^r, U^z, U^n, b^r, b^z, b^{nx}, b^{nh}]$ are the trainable parameters of the model, x_t is the input to the cell at the current timestep, and r_t and z_t represent the reset and update gates, which are numbers between (0, 1) that control how much information to retain in the new update vector $\tilde{h}_t$ and how the new hidden state vector h_t is composed from the update vector $\tilde{h}_t$ and the old hidden state h_{t-1}. σ denotes the sigmoid function, and $\odot$ represents element-wise multiplication. Dropout was applied to the hidden states between layers, so that $x_t^{(l)} = h_t^{(l-1)} \odot \delta_t^{(l-1)}$, where each $\delta_t^{(l-1)}$ is a Bernoulli random variable that zeros out elements in the hidden state vector with a probability defined by the dropout rate.

The inputs into the first layer of GRU cells are the ϕ, ψ and ω torsion angles and the amino acid type. Since we are using a bidirectional recurrent neural network architecture, information about previous/following residues should already be included in the hidden state at any “timestep” in an implicit way in the GRU, which is sufficient information to allow the network to make predictions accurately enough without formulating it explicitly as the input. Each torsion angle, a, was represented by a Gaussian smearing function discretized to a vector of length 180 to account for uncertainty in the data, denoted x_{ia}

$$x_{ia} = \exp\left(-\frac{\mathrm{diff}(\alpha_i, \hat{\mathrm{x}}_a)}{2\sigma^2}\right) \tag{5}$$

where α_i is the actual ϕ, ψ or ω angle and $\hat{\mathrm{x}}_a = (-180 + 2 * a)$ (both in degrees), and in this work we used $\sigma = 0.5^\circ$. The custom diff function

$$\mathrm{diff}(\alpha_i, \hat{\mathrm{x}}_a) = \min(|\alpha_i - \hat{\mathrm{x}}_a|, \min(|\alpha_i - \hat{\mathrm{x}}_a - 360|, |\alpha_i - \hat{\mathrm{x}}_a + 360|)) \tag{6}$$

ensures that the periodicity of the angles is taken into account. Each smeared torsion angle vector is further transformed through two fully-connected layers with 90 and 64 units each and Rectified Linear Units (ReLU) activation[38] to generate latent representations of the torsion angles. The residue types are encoded by a trainable embedding dictionary and formulated into latent vectors of length 64; the hidden dimension size of 64 was chosen after a careful hyperparameter search and found to be the optimal value. The torsion angle latent vectors and the embedded residue types are then concatenated and transformed together through 2 fully-connected layers with 128 and 64 units and ReLU activation and constitute the inputs into the GRU cells.

The hidden state output from the last GRU layer is connected with multiple outputs to predict the backbone bond lengths and bond angles, or optionally sidechain bond lengths and bond angles as well. Each output is a fully-connected neural network with a hidden layer of size 100 and activation ReLU, and the output has size of 1 without any activation. The raw outputs are scaled by the standard deviation and translated by the mean value of that data type in the training dataset. The means and standard deviations we used are provided in Supplementary Table 1.

Training details of the Int2Cart machine learning method.

The neural network was trained by minimizing the weighted mean square error loss function

$$\mathscr{L} = \sum_i w_i (y_i - \hat{y}_i)^2 \tag{7}$$

where w_i controls the weighting for different data types in the loss function, y_i are the predictions from the model and $\hat{y}_i$ are the actual values from the data set. In practice we used the same weighting for all the data types. Missing data targets were masked out during the training. We used the Adam optimizer[39] with an initial learning rate of 0.001 and an exponentially decayed learning rate schedule, so $\mathrm{lr}_i = \exp(-i * \alpha)$ where i is epoch number and $\alpha = 0.05$ in our case. The model was trained for a total of 100 epochs using a batch size of 128.

Building all-atom Cartesian structures from internal angle model predictions.

With the full profile of backbone torsion angles and predictions of bond lengths and bond angles from the model, the 3D Cartesian structure of the protein containing all backbone atoms is reconstructed using the SidechainNet package.[31] It utilizes the natural extension reference frame (NeRF) algorithm[5] to sequentially calculate the position of the next atom

with the positions of three previous atoms and the new bond length, bond angle and torsion angle. The all-atom backbone Cartesian structures for all the protein fragments in our test data set are built from either the Int2Cart algorithm vs. a standard baseline of using fixed bond lengths and bond angles (Fixed), or using bond lengths and bond angles from the Protein Geometry Database (PGD)[17] which uses bond geometries that depend on local torsions or amino acid type.

RESULTS

Statistical analysis of the protein training set.

Given the large collection of deposited protein structures in the PDB, we first consider a statistical analysis of protein bond lengths and bond angles when analyzed over the training set. Overall the distributions of these internal coordinate values are mostly Gaussian with relatively small standard deviations of ~ 0.01 Å for bond lengths and ~ 2.6° for bond angles. Figure 3 (a–f) depicts the deviations from the mean bond length and angle values for a given (ϕ, ψ) combination, and confirms the existence of strong correlations among the internal coordinates averaged over the training data set. Specifically, the θ_1 angle is larger for the right-hand and left-hand helix regions in the Ramanchandran plot, while the beta-sheet regions have more narrow θ_1 angles, with deviations from the mean as large as 7.5°.

The θ_2 values are strongly correlated with the ψ torsion angle, with larger angles when ψ is between −100 and 0 degrees, and smaller angles than the mean otherwise. The θ_3 values for nearly all of the (ϕ, ψ) combinations are larger than average, but have smaller angles for helix regions. Similarly, the d_1 and d_2 bond lengths show greater correlations with the ϕ torsion angle, with a preference for larger values when ϕ is between −50 and +50 degrees, in which the bond lengths change by as much as 0.02 Å. Finally the correlation for the peptide d_3 bond with the backbone torsions is weak, consistent with its partial double bond character, except for a few hot spots where it can vary up to 0.04 Å. These correlations are statistically meaningful, because the standard deviations in each bin are smaller than the mean value differences (Supplementary Figure 1), which means the statistical bias is more significant than the variance.

We have also considered the relationships between backbone ω torsion angles with bond angles (Figure 3 g–i) and bond lengths (Supplementary Figure 2), and found interesting correlations between internal coordinates and ω torsion angles. The majority of peptide bonds in proteins are in the *trans*-conformation, with ω torsion angles close to 180°. However, *cis*- peptide bonds tend to be associated with smaller θ_1 angles and larger θ_2 and θ_3 angles. This result also makes structural sense since *cis*- peptide bonds incur more steric repulsion between sidechains of two consecutive residues, and larger θ_2, θ_3 and smaller θ_1 values allow the sidechains to be more separated. On the other hand, the correlation between bond lengths and ω torsion angles are not obvious (Supplementary Figure 2). These correlations dependent on ω are also important for accurately predicting internal coordinates from backbone torsion angles as we will show later.

When we consider the observed distributions of all six internal coordinates as a function of the residue type (Supplementary Figure 3), we find that the distributions are quite similar between amino acids with only subtle differences in the shape of the peaks, with the exception of glycine, which tends to have d_1 and d_2 values that are smaller, and θ_1 angles that are larger than other residues. Proline also defines an exception, with larger d_1 values due to the formation of the five-membered ring that requires longer bond lengths. However the bond length and angle distributions as a function of backbone torsions and residue type exhibit notable variations across *all* twenty amino acids as seen in Figure 4 for the θ_1 bond angle, as well as for the other backbone bonds and angles shown in Supplementary Figures 4–8. To test whether structural resolution quality has an effect on the conclusion drawn from the statistical analysis, we further separated the training dataset by structure resolution categories of higher quality (resolution ¡= 2 Å) and lower quality (2 Å¡ resolution ¡= 4 Å) and compared the dependence of bond angles on ω torsion angles. The results are provided in Supplementary Figure 9. No significant discrepancies exist on the correlations between two groups of structures with different qualities, which supports utilizing the whole training dataset without filtering based on resolution.

Machine learning of sequence and structural correlations.

While the correlation graphs just described could serve as a source for bond lengths and angles when backbone torsion angles and residue types are provided, we are still missing the sequence-dependent correlations that are buried beneath the statistics of the single residue results. Therefore, we trained a deep neural network on the same data in order to capture the more subtle correlations among the internal coordinates conditioned on amino acid sequence. After training, the Int2Cart neural network was used to predict the test set which has low sequence and structural similarity with the training proteins. The root-mean-square error (RMSE) and Pearson correlation coefficients (R) on the test set are summarized in Supplementary Table 2. We find that the RMSE in bond length predictions are within the variance determined from the data set, while predictions on the bond angles are more successful in terms of the RMSEs that are smaller than the dataset variance.

Cartesian coordinate reconstructions.

Given the three torsion angles [ϕ, ψ, ω] for each residue over the entire protein sequence as input, we next consider how well the Cartesian coordinates are reconstructed based on whether bond and angle geometries are held fixed, using PGD, or learned from Int2Cart. Table 1 provides a general overview of the performance of the three approaches using a $C_\alpha(RMSD_{100})$ metric, which is the C_α RMSD values divided by the length of the protein and then multiplied by 100, as well as the C_α RMSD over all test set proteins regardless of length.

The reconstructed RMSDs for the Int2Cart structures are centered around lower median values of $C_\alpha(RMSD_{100})$ of 2.07 Å, and C_α RMSD of 3.74 Å over all test proteins. By contrast the Fixed model yields a median RMSD of 3.22 Å when all proteins are normalized to 100 amino acids, and the average over the entire test set is 5.39 Å. Table 1 also shows that the Int2Cart results are notably better than the PGD method which provides bond lengths

and bond angles as a function of local ϕ, ψ and amino acid type, in which the median (2.92 Å) and mean (3.32 Å) $C_\alpha(RMSD_{100})$ are much higher than that found with Int2Cart. Furthermore, to investigate the tranferability of the Int2Cart model, the test dataset was broken down into subsets that have 10%–50% sequence similarity to any protein in the training dataset, and proteins from CASP12. The results reported in Table 1 indicate that the sequence similarity to the training dataset has little effect on the reconstruction quality of the proteins. Therefore, our model is expected to be generalizable to proteins it has not seen.

To provide a more statistical view of the predictions, Figure 5a reports the distribution of RMSDs for all backbone atoms with respect to the actual PDB structure for all proteins in the test set using Int2Cart and the Fixed method, as well as the pairwise RMSDs for the test proteins (Figure 5b), and the RMSD difference between the two methods as a function of sequence length (Figure 5c). It is evident that the vast majority of the test set proteins benefit from the machine learned bond lengths and bond angles, with an average improvement of 2 – 4 Å RMSD over using Fixed bond lengths and bond angles. There is no obvious correlation between the RMSD improvements made by Int2Cart over Fixed with respect to sequence length, although the largest improvements occur in those proteins with longer amino acid sequences.

Figure 5d illustrates that proteins reconstructed by assuming fixed bond lengths and bond angles have lost significant secondary structure integrity compared to the reference structures, whereas the Int2Cart structures retain a much higher proportion of intact secondary structural elements. Beyond this anecdotal case, we performed a more extended analysis of Int2Cart and Fixed performance regarding the radius of gyration (R_g) and secondary structure recovery rate (SS-match) over the whole test dataset. Although we find that the Int2Cart Cartesian predictions have closer R_g values to the ground truth structures, the Fixed Cartesian structures still yields a comparably good result as seen in Supplementary Figure 10. Figure 5e shows that Int2Cart systematically improves upon Fixed in regards the SS-match values, defined as the proportion of helix, strand, and coil DSSP assignments[41] for each residue that matches the reference structure. It is seen that Int2Cart has a higher proportion of test set proteins that have SS-match values larger than 0.8 (Figure 5f), which translates to more than 80% of the residues having correct secondary structure assignments.

Comparison of sequence-length-dependent reconstruction quality among methods.

Due to the sequential nature of the process of modelling protein 3D structures with internal coordinates, the reconstruction error is expected to increase as the protein sequence increases in length. In Figure 6 the reconstruction error evaluated as the RMSD on the C_α atoms compared to the initial structures from the PDB are plotted as a function of sequence length, in which proteins were reconstructed using either Int2Cart-predicted bond geometries, using fixed bond lengths and bond angles, or using the local-conformation dependent Protein Geometry Database (PGD) as described in Ref.[17] Test proteins are grouped by sequence lengths with increments of 100 amino acids, and the standard deviations in each group are described by the shaded regions in Figure 6. Compared to using fixed bond lengths and bond angles, the PGD method has slight improvements in almost all sequence length ranges except around 400 amino acids. Even so, the Int2Cart has

a more significant improvement in C_α RMSD compared to PGD, suggesting its superiority is likely due to the fact that Int2Cart is able to learn deeper sequence correlations.

Ablation studies.

To understand the importance of various inputs for prediction accuracy of Int2Cart and how accuracy effects reconstructing the Cartesian structures, we performed an ablation study by training separate deep learning models using subsets of the inputs, and reconstructing structures using only predicted bond lengths, only predicted bond angles, or using both. We have also trained models with additional inputs of χ_1 torsion angles, along with r_1 and α_1 sidechain bond lengths and bond angles as additional outputs, to evaluate how including sidechain information could improve prediction and reconstruction of backbone structures. All ablation trials are reported in Table 2.

We see that the differences in predictions of the backbone bond lengths from different deep learning models are not significant, but prediction accuracy for backbone bond angles RMSE and reconstructed Cartesian structure RMSD are highly dependent on what information is available to the model. Specifically, a machine learning model that only knows about the residue types performs the worst with >5 Å in the reconstruction RMSD. Unsurprisingly based on statistical analysis of the PDB, backbone ϕ and ψ torsion angles provide more information than residue types alone, and allows the reconstruction RMSD to decrease to 4.56 Å on average. Including both ϕ, ψ and residue types further decreases the average reconstruction error to 4.29 Å. As expected from the correlation of bond lengths and bond angles with ω torsion angles as well, including exact values for ω torsion angles also significantly improves the model and allows the reconstructed structure RMSD to decrease further to 3.77 across the whole test set, and to 2.38 Å for proteins normalized to 100 amino acids. When we tested the inclusion of sidechain χ_1 torsion angles, we find that the 3D reconstruction model is even better, achieving an average reconstruction structure RMSD of 3.30 Å regardless of protein length. This is probably due to the fact that χ_1 torsion angles are indicative of avoidable steric clashes between protein backbones and side chains to create more accurate descriptions of subsequent backbone bond geometries, even though side chain atoms are not explicitly treated during structure reconstruction in this work.

To bolster these conclusions, Table 2 shows that the reconstruction quality does not depend on the direct prediction of bond lengths, as it essentially has no effect on the reconstructed structures, which may have been anticipated from the fact that bond length errors are on par with the variance. But this final ablation study provides direct evidence that accurate predictions of bond angles are of primary importance for the quality of the reconstructed Cartesian structures. In addition, using accurate ω torsion angles in the reconstruction is of great importance, since treating ω as binary greatly deteriorates the quality of reconstructed structures.

Using Int2Cart internal coordinate agreements to validate AlphaFold2 structures.

AlphaFold2 (AF2) has been a huge success in predicting atomic structures of proteins with astonishing accuracy.[26] Nevertheless its predictions have variety of quality, which is also reflected in its internal confidence estimations for each residue called the predicted

local distance difference test (pLDDT) score, with values greater than 90 indicating high confidence, and values below 50 indicating low confidence. To investigate the relationship between AF2 model quality and how much the bond lengths and bond angles in these AF2 models agree with the same Int2Cart quantities, we randomly collected 20 AF2 predicted protein structures from the human proteome, and calculated the bond lengths and bond angles using Int2Cart and the AF2 torsion angles. The results are summarized in Figure 7 and Supplementary Figures S11–S13.

On a per-residue basis, we observe strong correlation between the agreement of AF2 and Int2Cart bond geometries, and AF2 prediction confidence, as we illustrate in Figure 7(a). We see that the most confident residue predictions in AF2 models have better correlation in θ_1 values between AF2 models and Int2Cart predictions, compared to the residues with lower confidence. Figure 7(b) further discretizes the absolute differences into bins of 1° increments and shows that the residues that have a larger discrepancy between bond geometries in AF2 structures and Int2Cart predictions have on average lower quality in terms of pLDDT scores. Similar plots are generated for the θ_2, θ_3, d_1, d_2 and d_3 data where the Int2Cart and AF2 agreement is less good, but still exhibit strong correlations between geometry differences and pLDDT values (Supplementary Figures 11 and 12).

Finally, we aggregrate all three bond angle results into correlations and mean absolute differences over the entirety of all 20 AF2 protein models we have tested, and compared with their average structure confidence score. Figure 7(c–d) indicate that the agreement between Int2Cart predicted bond angle geometries and the AF2 model strongly correlates with overall model quality, thus supporting using Int2Cart for structure validations. Similar conclusions are reached for the bond lengths as given in Supplementary Figure 13.

Using Int2Cart to rebuild an IDP ensemble.

Finally we consider a test case that is quite different from the originally defined test set from SidechainNet, in which we show that our Int2Cart method can improve upon the Cartesian reconstruction of an ensemble of structures of a disordered protein compared to Fixed bond lengths and angles. Figure 8 compares the Cartesian reconstruction RMSD distributions for Int2Cart and Fixed for the Sic1 IDP ensemble, in which we find that the Int2Cart method is overall closer to the original ensemble, with a 3.1 Å average RMSD compared to the Fixed method that has a mean RMSD of 3.4 Å. We have also checked the number of steric clashes in the structures generated from these two methods. A steric clash is defined as two atoms in the structure that are closer to 0.6 times the sum of the van der Waals radii of the two atoms.[42] Out of the 1000 conformations, 73 structures generated from Int2Cart contained steric clashes, which means 92.7% of the structures are clash-free. By comparison, 102 structures generated using fixed bond lengths and bond angles contained steric clashes, which translates to 89.8% of clash-free structures. A higher proportion of clash-free structures is meaningful because typically structures containing clashes are discarded, and a method with higher proportion of clash-free structures wastes less computational resources, and supports the application of the Int2Cart algorithm to the modelling of disordered protein ensembles.

DISCUSSION AND CONCLUSION

In this work we have developed a new machine learning approach to the generic representation problem of internal coordinates (bond lengths, valence angles, and dihedral angles) and how to increase the fidelity of the back-transformation to 3D Cartesian coordinates. The Int2Cart algorithm utilizes a gated recurrent unit neural network to predict real-valued backbone bond lengths and bond angles for each residue of a complete protein sequence given its torsion angle profile. In summary, Int2Cart can reconstruct the Cartesian structure of proteins with RMSDs that are significant improvements over the fixed backbone bond lengths and bond angles that are the standard practice in a large variety of protein modelling approaches, or some recent approaches such as the Protein Geometry Database. The success of our algorithm across IDP ensembles further validates that the Int2Cart algorithm is transferable among different types of proteins, and can consistently improve the quality of Cartesian structure reconstruction. We have also exposed the potential of Int2Cart in validating structure quality by showing the agreement on bond geometries between Int2Cart predictions and values in an AlphaFold2 model has strong correlation with the AlphaFold2 pLDDT confidence metric. Possibilities in refining AF2 structures using Int2Cart will be investigated in the future.

In its current form the Int2Cart algorithm only generates backbone structures for the target proteins, although we can improve Cartesian reconstruction performance with the inclusion of the χ_1 torsion and predicting r_1 and α_1. Theoretical approaches such as the Monte Carlo Side Chain Ensemble (MC-SCE) method can utilize the backbone from Int2Cart to calculate side chain ensembles in order to complete the full structure. [42] It is also clear that there is still room for improvement in the Cartesian reconstruction of larger proteins, and the inherent scaling of error with respect to sequence length is inevitable for a deep learning model that predicts internal coordinates in a sequential manner (i.e., a GRU model). Therefore, it may be possible to improve the quality of Cartesian structure reconstruction with a distance-based neural network model, i.e., by representing the 3D coordinates of the structure directly.

Nevertheless, the model in its current form already provides a useful computational tool to greatly improve the quality of protein structures reconstructed from backbone torsion angles alone, whether globular folded proteins or disordered protein ensembles. We envision Int2Cart should see broad use in structure refinement and validation[43,44] and development of protein force fields that could benefit from more accurate valence models of backbone bond lengths and bond angles conditioned on other geometrical or sequence features. [45] Finally, the Int2Cart GRU neural network model could also be useful for other chain molecules, only requiring retraining with new data if available for systems such as nucleic acids and lipids.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

T.H.-G. and J.D.F.-K. acknowledge funding from the National Institute of Health under Grant 5R01GM127627–04. J.D.F.-K. also acknowledges support from the Natural Sciences and Engineering Research Council of Canada (2016–06718) and from the Canada Research Chairs Program.

References

(1). Baker J; Kinghorn D; Pulay P. Geometry optimization in delocalized internal coordinates: An efficient quadratically scaling algorithm for large molecules. J. Chem. Phys 1999, 110, 4986–4991.

(2). Schwieters CD; Clore G. Internal Coordinates for Molecular Dynamics and Minimization in Structure Determination and Refinement. J. Magn. Reson 2001, 152, 288–302. [PubMed: 11567582]

(3). Adcock SA; McCammon JA Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. Chem. Rev 2006, 106, 1589–1615, PMID: 16683746. [PubMed: 16683746]

(4). wwPDB consortium, Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2018, 47, D520–D528.

(5). Parsons J; Holmes JB; Rojas JM; Tsai J; Strauss CE Practical conversion from torsion space to Cartesian space for in silico protein synthesis. J. Comput. Chem 2005, 26, 1063–1068. [PubMed: 15898109]

(6). Rohl CA; Strauss CE; Misura KM; Baker D. Methods in enzymology; Elsevier, 2004; Vol. 383; pp 66–93.

(7). Das R; Baker D. Macromolecular modeling with rosetta. Annu. Rev. Biochem 2008, 77, 363–382. [PubMed: 18410248]

(8). Craveur P; Joseph AP; Poulain P; de Brevern AG; Rebehmed J. Cis–trans isomerization of omega dihedrals in proteins. Amino acids 2013, 45, 279–289. [PubMed: 23728840]

(9). Holmes JB; Tsai J. Some fundamental aspects of building protein structures from fragment libraries. Protein Sci. 2004, 13, 1636–1650. [PubMed: 15152094]

(10). Ramachandran G; Ramakrishnan C; Sasisekharan V. Stereochemistry of polypeptide chain configurations. J. Mol. Biol 1963, 7, 95–99. [PubMed: 13990617]

(11). Berkholz DS; Driggers CM; Shapovalov MV; Dunbrack RL Jr; Karplus PA Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. Proc. Natl. Acad. Sci. USA 2012, 109, 449–453. [PubMed: 22198840]

(12). Shapovalov MV; Dunbrack RL Jr A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 2011, 19, 844–858. [PubMed: 21645855]

(13). Karplus PA Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci. 1996, 5, 1406–1420. [PubMed: 8819173]

(14). Jiang X; Cao M; Teppen B; Newton SQ; Schaefer L. Predictions of protein backbone structural parameters from first principles: Systematic comparisons of calculated NC (. alpha.)-C'angles with high-resolution protein crystallographic results. J. Phys. Chem 1995, 99, 10521–10525.

(15). Schäfer L; Cao M; Meadows MJ Predictions of protein backbone bond distances and angles from first principles. Biopolymers: Original Research on Biomolecules 1995, 35, 603–606.

(16). Yu C-H; Norman MA; Schäfer, L.; Ramek, M.; Peeters, A.; Van Alsenoy, C. Ab initio conformational analysis of N-formyl L-alanine amide including electron correlation. J. Mol. Struct 2001, 567, 361–374.

(17). Berkholz DS; Shapovalov MV; Dunbrack RL Jr; Karplus PA Conformation dependence of backbone geometry in proteins. Structure 2009, 17, 1316–1325. [PubMed: 19836332]

(18). Improta R; Vitagliano L; Esposito L. The determinants of bond angle variability in protein/peptide backbones: A comprehensive statistical/quantum mechanics analysis. Proteins: Structure, Function, and Bioinformatics 2015, 83, 1973–1986.

(19). Improta R; Vitagliano L; Esposito L. Bond distances in polypeptide backbones depend on the local conformation. Acta Crystallogr. D Biol. Crystallogr 2015, 71, 1272–1283. [PubMed: 26057667]

(20). Lundgren M; Niemi AJ Correlation between protein secondary structure, backbone bond angles, and side-chain orientations. Phys. Rev. E 2012, 86, 021904.

(21). Ravikumar A; Ramakrishnan C; Srinivasan N. Stereochemical Assessment of (φ, ψ) Outliers in Protein Structures Using Bond Geometry-Specific Ramachandran Steric-Maps. Structure 2019, 27, 1875–1884. [PubMed: 31607615]

(22). Zheng W; Li Y; Zhang C; Pearce R; Mortuza S; Zhang Y. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins: Structure, Function, and Bioinformatics 2019, 87, 1149–1164.

(23). Senior AW; Evans R; Jumper J; Kirkpatrick J; Sifre L; Green T; Qin C; Žídek A; Nelson AW; Bridgland A. et al. Improved protein structure prediction using potentials from deep learning. Nature 2020, 577, 706–710. [PubMed: 31942072]

(24). Du Z; Su H; Wang W; Ye L; Wei H; Peng Z; Anishchenko I; Baker D; Yang J. The trRosetta server for fast and accurate protein structure prediction. Nat. Protoc 2021, 16, 5634–5651. [PubMed: 34759384]

(25). Baek M; DiMaio F; Anishchenko I; Dauparas J; Ovchinnikov S; Lee GR; Wang J; Cong Q; Kinch LN; Schaeffer RD et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021, 373, 871–876. [PubMed: 34282049]

(26). Jumper J; Evans R; Pritzel A; Green T; Figurnov M; Ronneberger O; Tunyasuvunakool K; Bates R; Žídek A; Potapenko A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021, 596, 583–589. [PubMed: 34265844]

(27). Varadi M; Anyango S; Deshpande M; Nair S; Natassia C; Yordanova G; Yuan D; Stroe O; Wood G; Laydon A. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022, 50, D439–D444. [PubMed: 34791371]

(28). Lee J; Lee D; Park H; Coutsias EA; Seok C. Protein loop modeling by using fragment assembly and analytical loop closure. Proteins: Structure, Function, and Bioinformatics 2010, 78, 3428–3436.

(29). Teixeira JMC; Liu ZH; Namini A; Li J; Vernon RM; Krzeminski M; Shamandy AA; Zhang O; Haghighatlari M; Yu L. et al. IDPConformerGenerator: A Flexible Software Suite for Sampling the Conformational Space of Disordered Protein States. J. Phys. Chem. A 2022, 126, 5985–6003, PMID: 36030416. [PubMed: 36030416]

(30). Touw WG; Vriend G. On the complexity of Engh and Huber refinement restraints: the angle τ as example. Acta Crystallogr. D 2010, 66, 1341–1350. [PubMed: 21123875]

(31). King JE; Koes DR SidechainNet: An all-atom protein structure dataset for machine learning. Proteins 2021, 89, 1489–1496. [PubMed: 34213059]

(32). AlQuraishi M. ProteinNet: a standardized data set for machine learning of protein structure. BMC Bioinform. 2019, 20, 311.

(33). Mittag T; Orlicky S; Choy W-Y; Tang X; Lin H; Sicheri F; Kay LE; Tyers M; Forman-Kay JD Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. Proc. Natl. Acad. Sci. USA 2008, 105, 17772–17777. [PubMed: 19008353]

(34). Gomes G-NW; Krzeminski M; Namini A; Martin EW; Mittag T; Head-Gordon T; Forman-Kay JD; Gradinaru CC Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. J. Am. Chem. Soc 2020, 142, 15697–15710, PMID: 32840111. [PubMed: 32840111]

(35). AlQuraishi M. End-to-end differentiable learning of protein structure. Cell Syst. 2019, 8, 292–301. [PubMed: 31005579]

(36). Zhang O; Haghighatlari M; Li J; Teixeira JMC; Namini A; Liu Z-H; Forman-Kay JD; Head-Gordon T. Learning to Evolve Structural Ensembles of Unfolded and Disordered Proteins Using Experimental Solution Data. arXiv preprint arXiv:2206.12667 2022,

(37). Cho K; Van Merrïenboer B; Bahdanau D; Bengio Y. On the properties of neu- ral machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 2014,

(38). Glorot X; Bordes A; Bengio Y. Deep sparse rectifier neural networks. Proceedings of the fourteenth international conference on artificial intelligence and statistics. 2011; pp 315–323.

(39). Kingma DP; Ba J. Adam: A Method for Stochastic Optimization. CoRR 2015, abs/1412.6980.

(40). Tan K; Gu M; Jedrzejczak R; Joachimiak A. The Crystal structure of the N-terminal domain of a novel cellulases from Bacteroides coprocola. PDB 2016,

(41). Kabsch W; Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers: Original Research on Biomolecules 1983, 22, 2577–2637.

(42). Bhowmick A; Head-Gordon T. A monte carlo method for generating side chain structural ensembles. Structure 2015, 23, 44–55. [PubMed: 25482539]

(43). Wilson K; Butterworth S; Dauter Z; Lamzin V; Walsh M; Wodak S; Pontius J; Richelle J; Vaguine A; Sander C. et al. Who checks the checkers? Four validation tools applied to eight atomic resolution structures. J. Mol. Biol 1998, 276, 417. [PubMed: 9512713]

(44). Kleywegt GJ On vital aid: the why, what and how of validation. Acta Crystallogr. D Biol. Crystallogr 2009, 65, 134–139. [PubMed: 19171968]

(45). Conway P; Tyka MD; DiMaio F; Konerding DE; Baker D. Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci. 2014, 23, 47–55. [PubMed: 24265211]

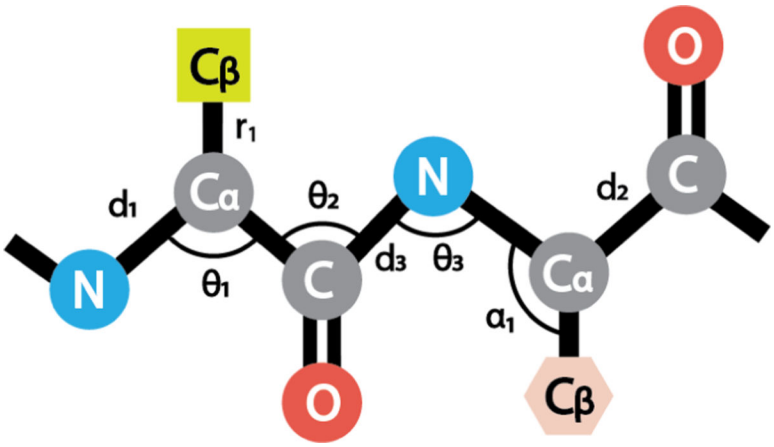


Figure 1: Schematic of the polypeptide backbone and internal degrees of freedom.
Definition of the prediction targets: backbone bond angles $\theta_1 - \theta_3$, backbone bond lengths $d_1 - d_3$, $C_\alpha - C_\beta$ sidechain bond lengths r_1 and $N - C_\alpha - C_\beta$ sidechain bond angles α_1.

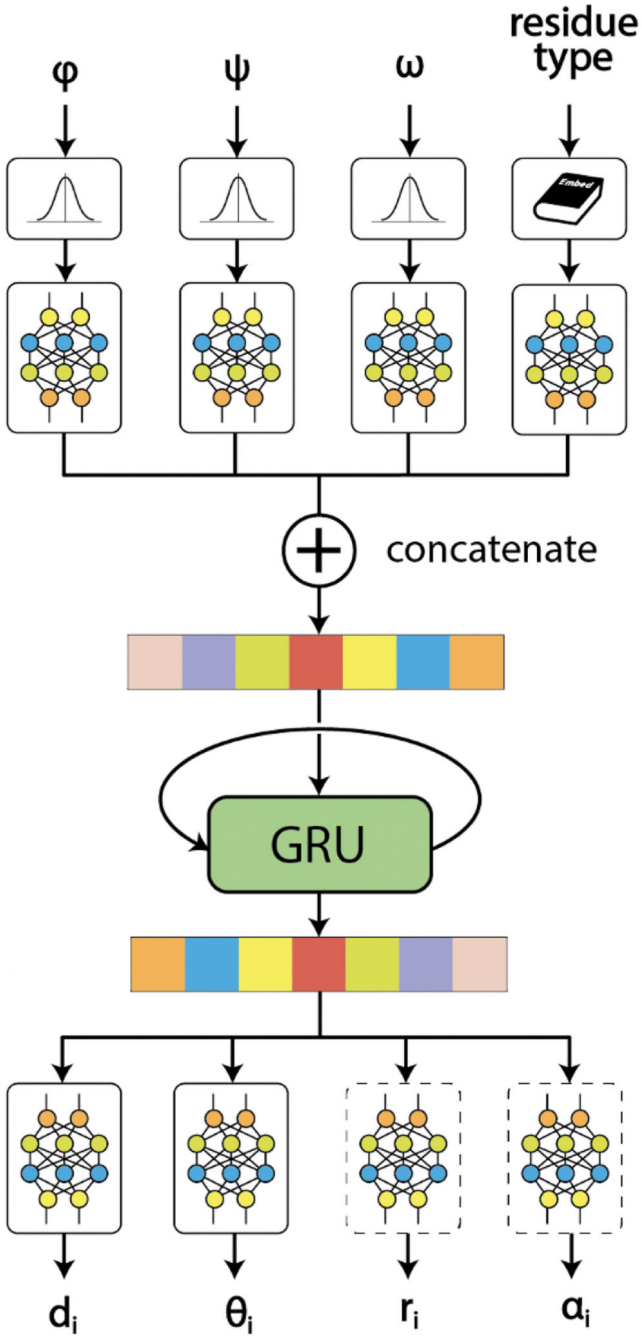


Figure 2: Schematic of the Int2Cart neural network architecture.

The neural network is a gated recurrent unit (GRU) recurrent neural network. The inputs at each timestep are the concatenated latent vectors from Gaussian-smeared ϕ, ψ and ω torsion angles and embedded residue types; variations on the Int2Cart network can include the use of χ sidechain angles as well. The latent vector output from GRU are connected with multiple output networks to predict different targets.

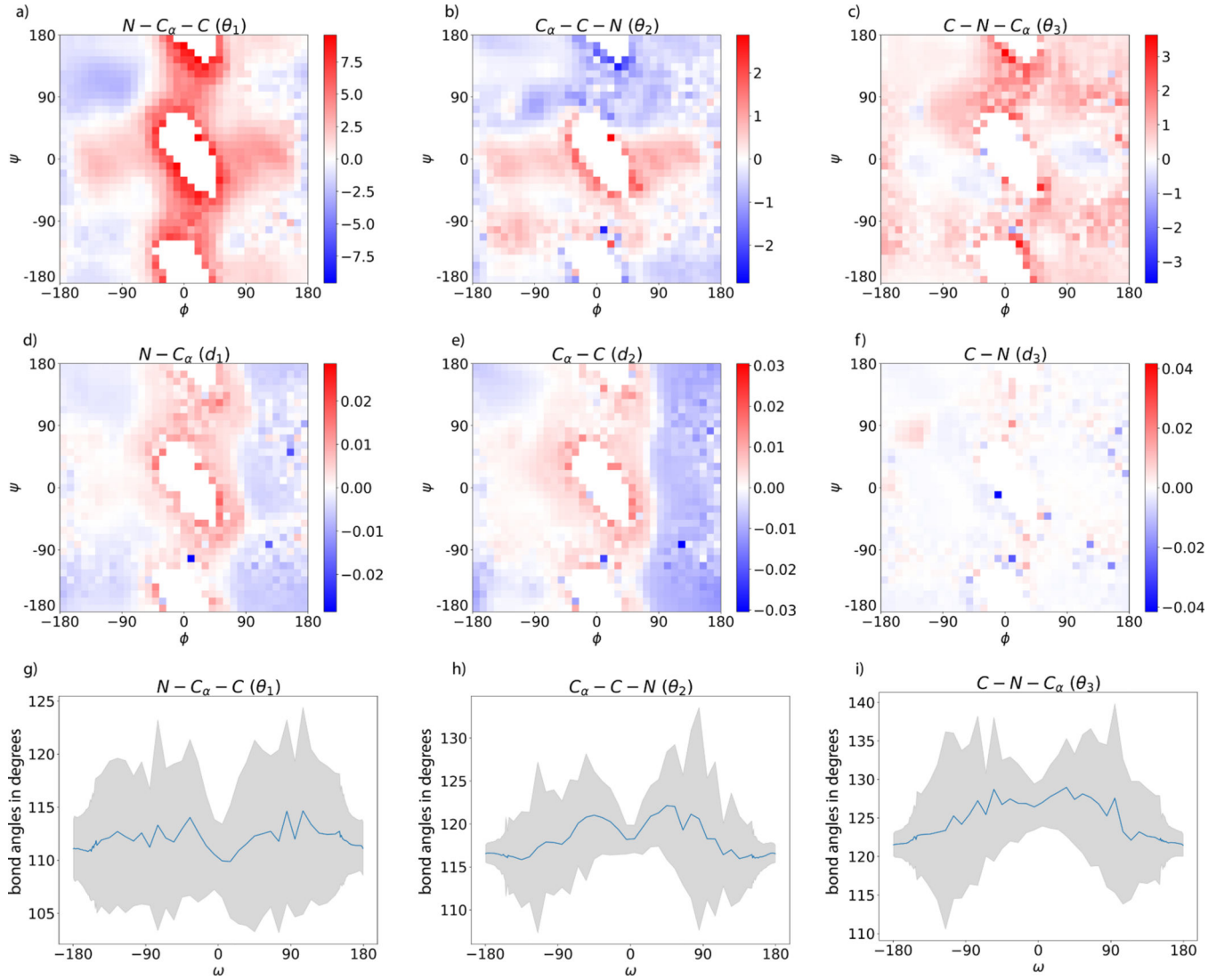


Figure 3: Variations of bond angles and bond lengths as a function of (ϕ, ψ), or ω torsion angles.

a-f) Bond angle and bond length deviations from the mean values averaged over ϕ and ψ angles of the training set. The regions of red correspond to wider angles and longer bonds while the region in blue show reduced angle and bond values relative to the mean. The bond lengths and bond angles were categorized according to ϕ and ψ angles rounded to the closest tens, and the data are aggregated by calculating the means and standard deviations in each bin. The standard deviations are provided in Figure S1. g-i) Mean values and standard deviations of bond angles as a function of ω. The blue solid line represents mean values of bond angles at specific ω torsion angles, and the gray regions correspond to one standard deviation.

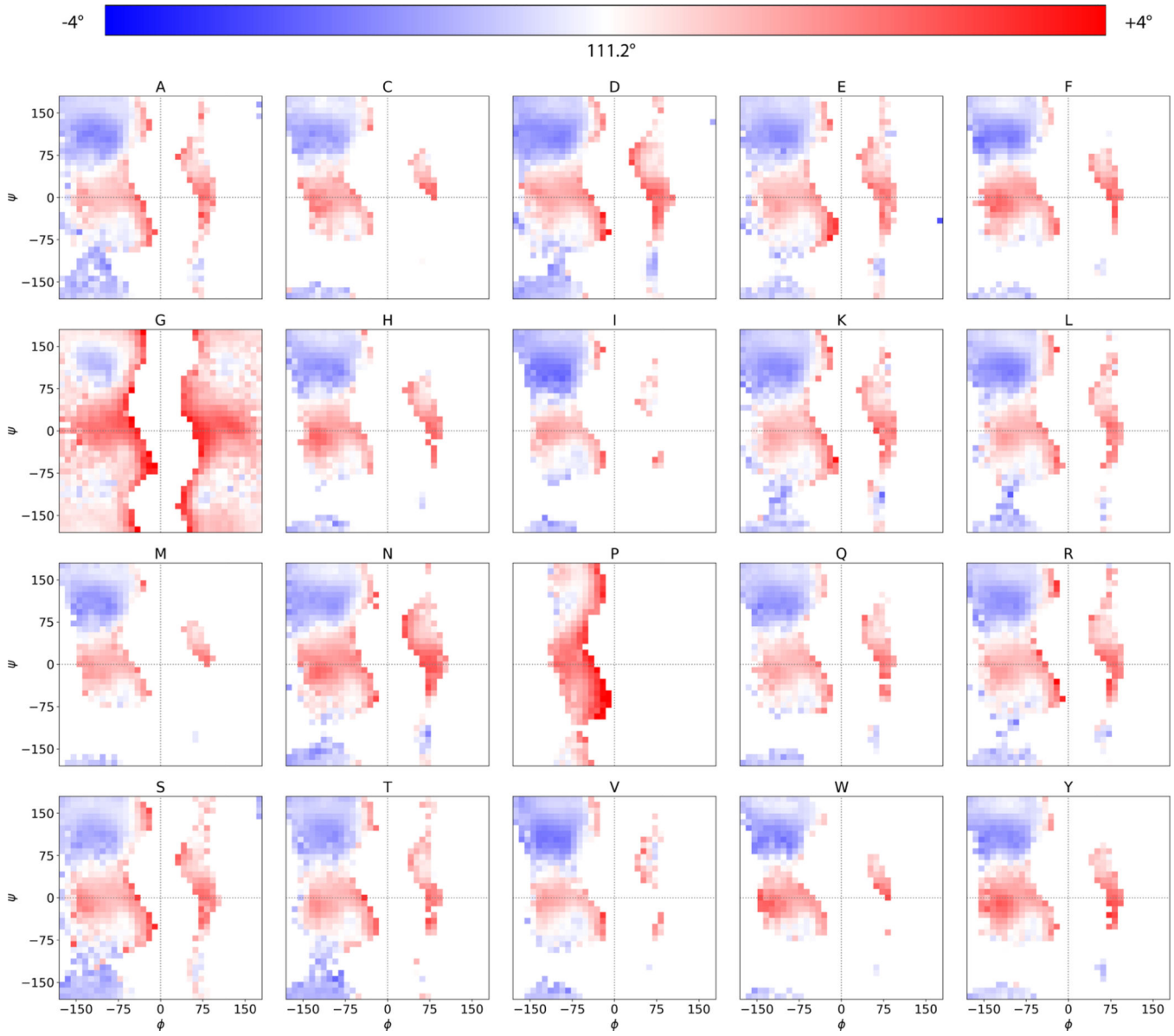


Figure 4: $N - C_\alpha - C$ bond angle deviations from the mean values averaged over ϕ and ψ angles as a function of residue type.

The regions of red correspond to longer bonds while the region in blue show reduced bond values relative to the mean. The $N - C_\alpha - C$ bond angles were categorized according to ϕ and ψ angles rounded to the closest tens.

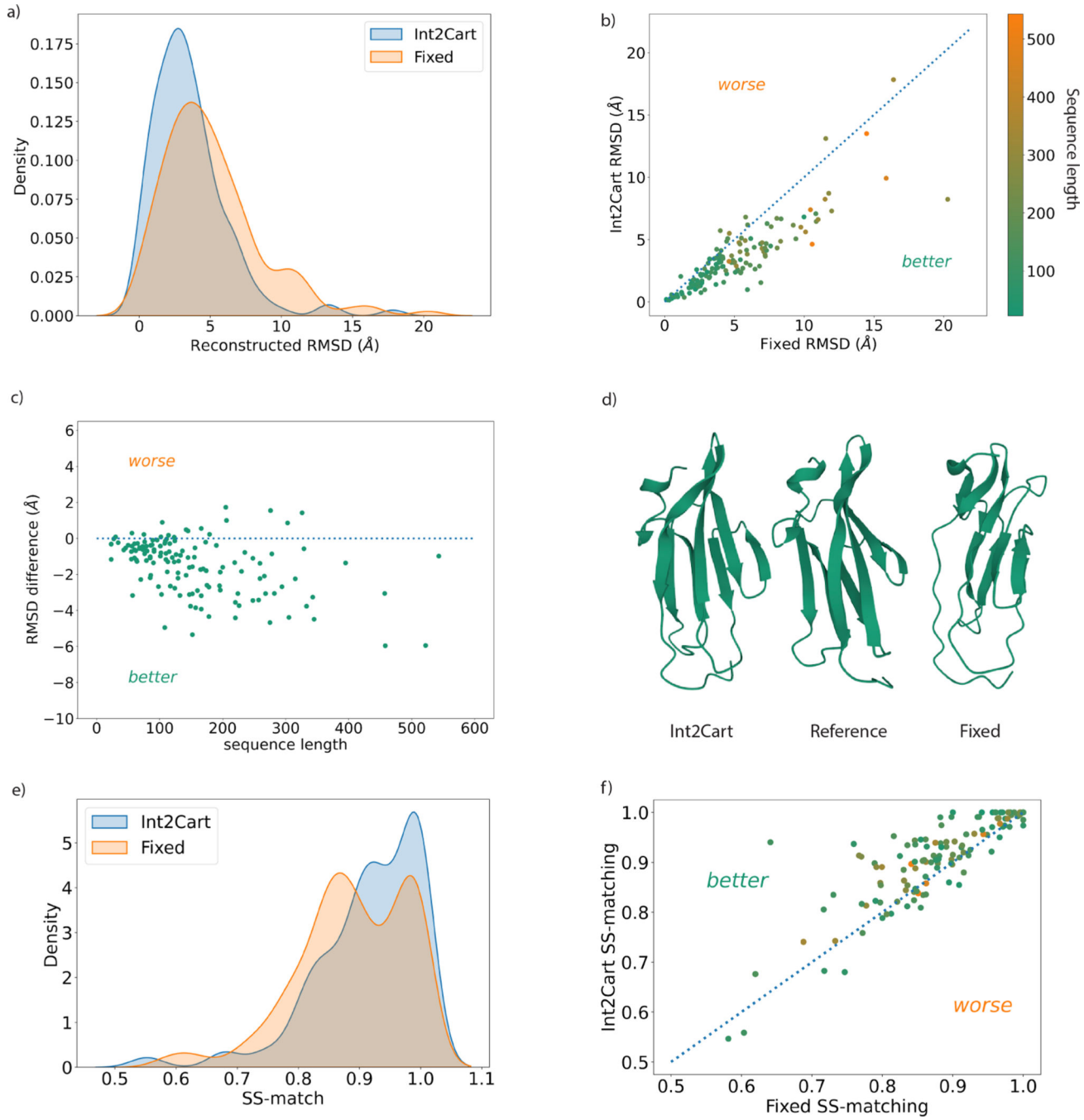


Figure 5: Comparison of 3D Cartesian reconstructions of test set proteins using Int2Cart and compared to Fixed bonds and angles.

(a) Distribution of the RMSD in reconstructed Cartesian coordinates using Int2Cart and Fixed. (b) Comparison of Cartesian reconstruction error between Int2Cart and Fixed relative to the reference structure. (c) Improvement of Int2Cart over Fixed as a function of amino acid length. (d) An example of the backbone representation using Int2Cart and Fixed for the CASP12 TBM0872 protein,[40] (e) The SS-match distribution and (f) comparison of SS-match for Int2Cart vs. Fixed across the test set.

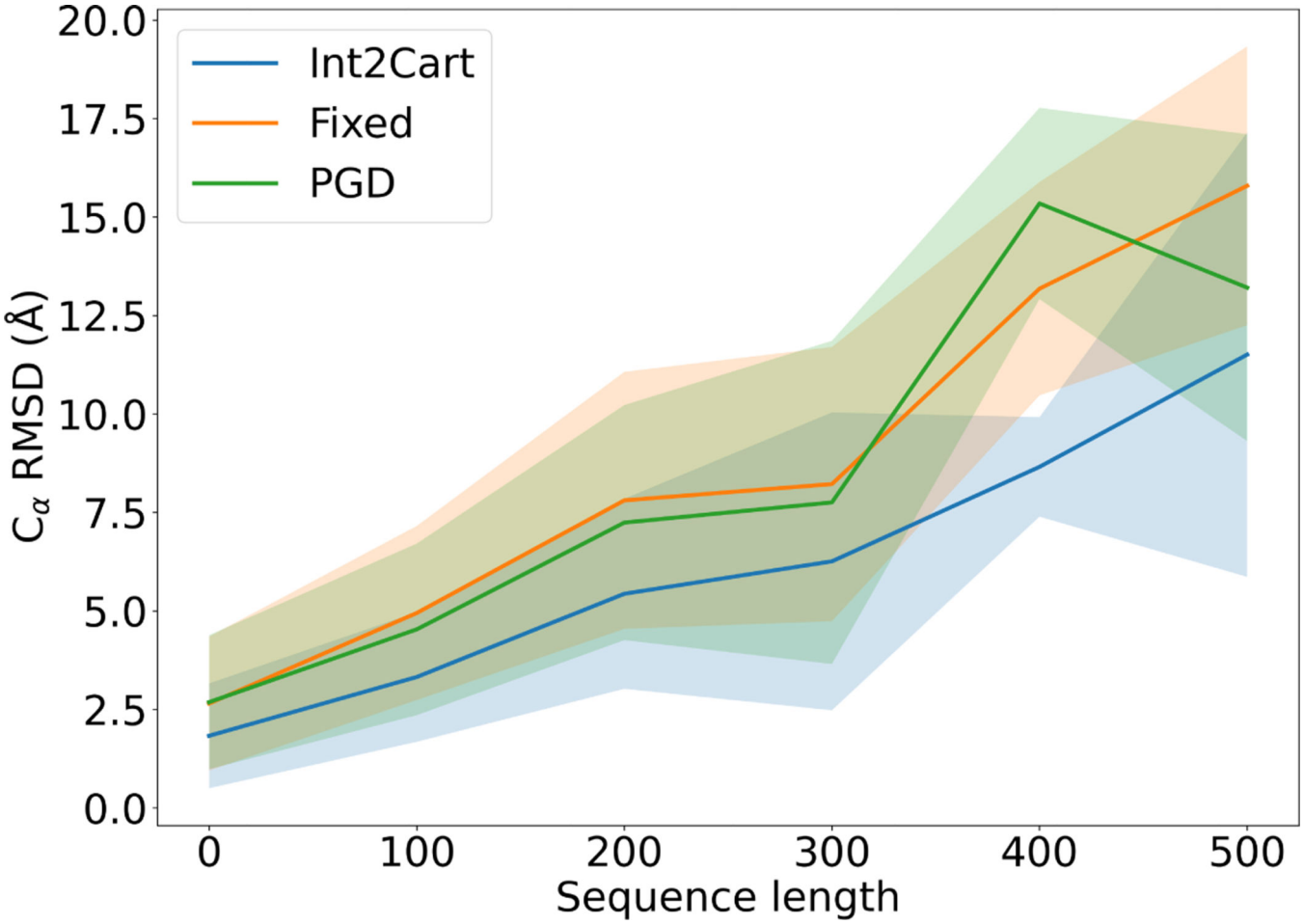


Figure 6: Comparison of reconstructed structure C_α RMSD values in the test set as a function of sequence length using different sources of bond lengths and bond angles.

The C_α RMSDs were calculated against ground truth structures after using only their torsion angles for reconstruction. Shaded regions represent 1 standard deviation. The blue line represents Int2Cart, the orange line represents fixed bond lengths and angles, and the green line is the PGD method.[17]

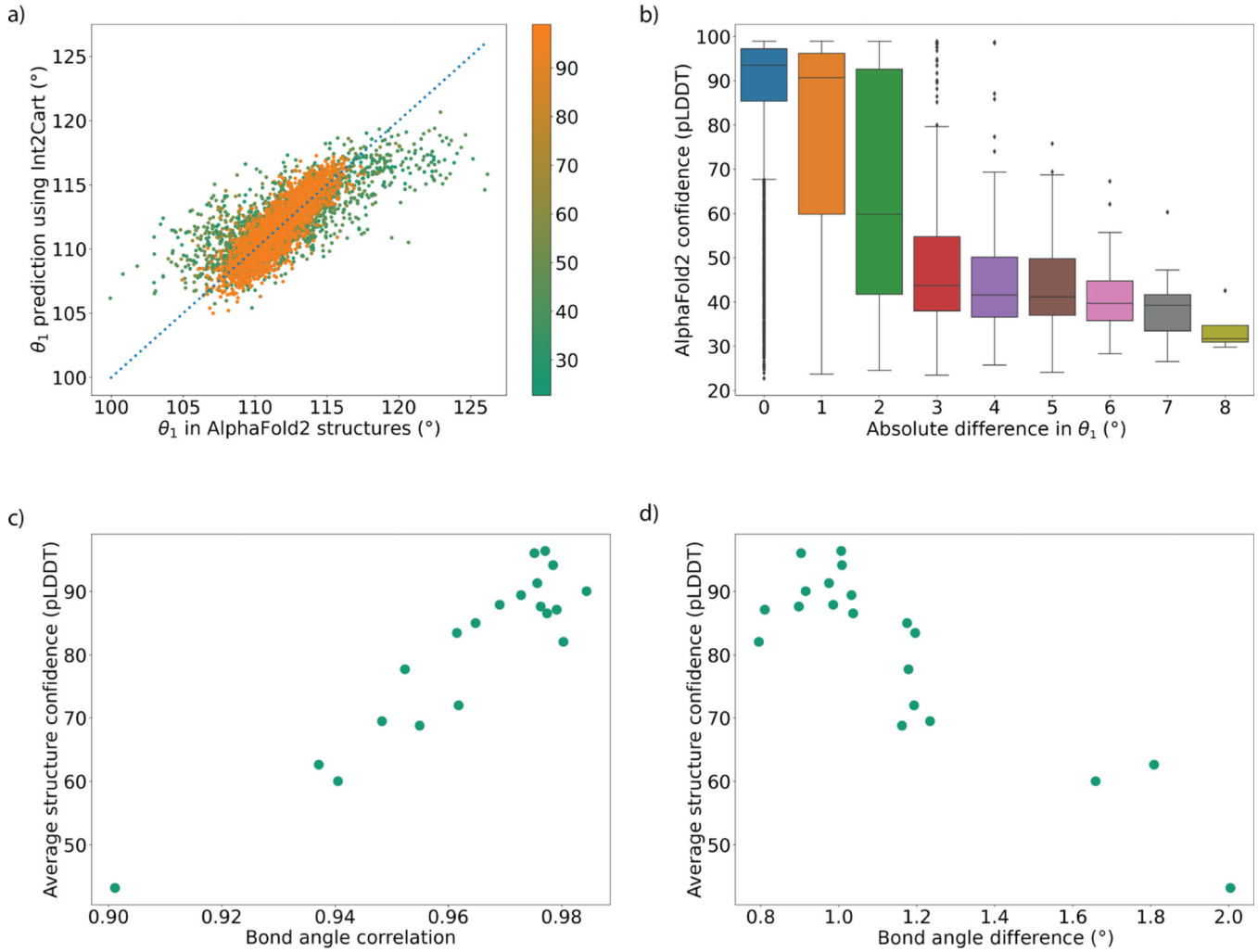


Figure 7: Correlation between AlphaFold2 (AF2) structure quality and the agreement between bond geometries from the AF2 predicted structures and Int2Cart predicted values using torsion angles from AF2 structures

(a) Correlation between θ_1s ($N - C_\alpha - C$ bond angles) from AF2 structures and Int2Cart predictions colored by AF2 pLDDT scores of the relevant residues. (b) Box plot showing distribution of AF2 pLDDT scores of individual residues based on absolute difference in θ_1 between AF2 structures and Int2Cart predictions. The boxes represent the quartiles of the distribution and the whiskers represent the rest of the distribution. Individual data points are outliers identified from the inter-quartile range. (c) Relationship between the average AF2 structure prediction confidence (pLDDT score) and all bond angle correlations between AF2 and Int2Cart in an AF2 predicted protein structure (d) Relationship between the average AF2 structure prediction confidence (pLDDT score) and all bond angle absolute difference between AF2 and Int2Cart in an AF2 predicted protein structure.

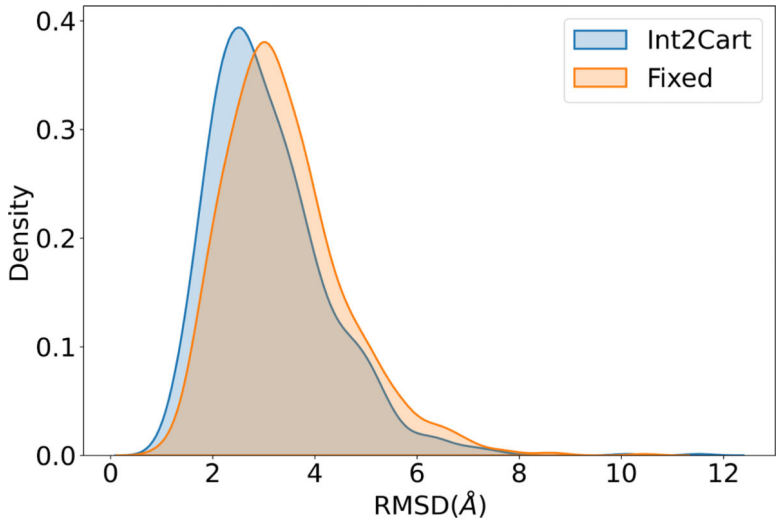


Figure 8: Comparison of distribution of reconstruction RMSD for individual conformaions in the Sic1 IDP ensemble.

Structures reconstructed with Int2Cart method on average has lower RMSD to their original structures compared with using fixed bond lengths and bond angles.

Table 1:

Quality of Cartesian reconstructed structures using Int2Cart, Fixed, and PGD methods normalized by sequence length, and Int2Cart results on different test data categories.

Accuracy is assessed in terms of the median and mean $C_\alpha(RMSD_{100})$, the root-mean-square error of the predicted C_α positions to the reference PDB structure normalized to 100 amino acids based on the test dataset. The second half of the table shows the breakdown of Int2Cart results in different similarity categories of data in the test dataset including CASP12 (which were after the time cutoff for proteins in the training dataset). All units in Å.

Method	Median	Mean±std
Fixed	3.22	3.47±1.83
PGD	2.92	3.32±1.87
Int2Cart	**2.07**	**2.38±1.36**
Test data category	**Median**	**Mean±std**
10% similarity	2.32	2.87±2.07
20% similarity	2.22	2.44±1.15
30% similarity	1.79	1.96±0.84
40% similarity	1.89	2.11±1.35
50% similarity	2.47	2.36±0.94
CASP12	2.06	2.39±1.22

Table 2:
Ablation studies of internal coordinate inputs and Cartesian coordinate reconstructions.

Upper table: predicted bond lengths and bond angles RMSEs of Int2Cart taking different internal coordinate inputs, and corresponding RMSD of the reconstructed Cartesian structure. Each ablation of the input is repeated 3 times with different initializations of the machine learning model to obtain statistically meaningful results. Standard deviations reflect fluctuations of mean values among 3 parallel experiments. Lower table: cartesian structure reconstruction RMSD using different Int2Cart predicted and Fixed combinations of bond lengths and angles and binary ω torsion angles. Standard deviations reflect range of reconstructed structure RMSDs among different proteins.

Training Model inputs	**<d> RMSE(Å)**	**< θ > RMSE (°)**	**Reconstructed RMSD (Å)**
Residue type	0.010±1E-5	1.84±0.0008	5.21± 0.04
ϕ + ψ	0.010±1E-4	1.69±0.02	4.56±0.07
ϕ + ψ + Residue type	0.010±5E-5	1.63±0.001	4.29±0.02
ϕ + ψ + ω + Residue type	0.010±1E-4	1.50±0.006	3.77±0.03
ϕ + ψ + ω + χ_1 + Residue type	0.009±1E-4	1.37±0.004	3.30±0.03

Source of bond geometries	**Reconstructed RMSD (Å)**
Predicted bond lengths and bond angles	3.74±2.94
Fixed bond lengths and predicted bond angles	3.74±2.94
Predicted bond lengths and fixed bond angles	5.38±3.70
Fixed bond lengths and angles	5.39±3.71
Fixed bond lengths, bond angles and using $0^\circ/180^\circ$ ω angles	9.52±6.49