January 2015

COMPUTATIONAL MODELLING OF PROTEIN FIBRILLATION WITH APPLICATION TO GLUCAGON

Hamed Tabatabaei Ghomi
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Hamed Tabatabei Ghomi

Entitled

COMPUTATIONAL MODELLING OF PROTEIN FIBRILLATION WITH APPLICATION TO GLUCAGON

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Dr. Markus A. Lill
Chair

Dr. Chiwook Park

Dr. Carol B. Post

Dr. Elizabeth M. Topp

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Dr. Markus A. Lill

Approved by: Dr. Val Watts 12/3/15

Head of the Departmental Graduate Program Date

COMPUTATIONAL MODELLING OF PROTEIN FIBRILLATION WITH APPLICATION TO GLUCAGON

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Hamed Tabatabaei Ghomi

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2015

Purdue University

West Lafayette, Indiana

ای الهه ای ترانه ام

آشیانه ام

ای برای شاد زیستن بهانه ام

قهرمان قصه های عاشقانه ام

با نسیم نام تو

بهار می شود

بذر شعرهای تازه در میان خاک

بی قرار می شود

ACKNOWLEDGEMENTS

I am indebted to my advisors, Dr. Lill and Dr. Topp who taught me many things about science, and about life. Also, I thank my advisory committee, Dr. Park, Dr. Post and Dr. Kihara for all their support.

I am grateful to my lovely wife, Elaheh. I owe all my success to her support and sacrifice. I am also grateful to my parent for their support and encouragement.

And finally, I owe my deepest and greatest gratitude to God, the Most Beneficent, the Most Merciful.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

ABSTRACT

Tabatabaei Ghomi, Hamed. Ph.D., Purdue University, December 2015. Computational Modelling of Protein Fibrillation with Application to Glucagon. Major Professor: Markus A. Lill.

A computational method to model the steric zipper of amyloid fibrils (FibPreditor) is developed. The method generates an ensemble of structures for the steric zipper by a number of geometric operations and presents the most energetically favorable candidates as models of steric zipper. The method is shown to successfully reproduce a number of experimentally determined fibril structures.

FibPredictor is then applied to model the steric zipper of glucagon fibrils. Phosphate ester derivatives of glucagon are designed based on these models as soluble and stable prodrugs or active alternatives for glucagon.

A number of penta-peptide chaperones are also designed as excipients to delay glucagon fibrillation. Although penta-peptides can delay glucagon fibrillation, they are less effective compared to phosphorylation of glucagon.

CHAPTER 1. INTRODUCTION

1.1 Computational Modelling of Amyloid Fibrils

Amyloid fibrils have been associated with many important pathological conditions such as Alzheimer's disease and type II diabetes. Amyloid fibrils also pose an important challenge in peptide and protein drug delivery as a major degradation pathway and have gained importance as bio-nanotubular scaffolds and triggerable drug delivery platforms [1–6]. The rational design of drugs that inhibit fibrillation, the design of stable formulations of peptide and protein drugs and the development of bio-nanotechnological fibril devices all depend on understanding the structure of amyloid fibrils[1]. However, experimental amyloid fibril structure determination is difficult [7]. Computational method thus are specifically useful to predict the structure of amyloid fibrils and study their dynamics and energetics[8].

There have been a few successful attempts to generate de novo computational models for some specific amyloid fibrils [9–11], many computational studies on the mechanisms of fibril formation[8], and many methods to predict aggregation-prone regions and amyloid forming sequences[8,12–14]. Nonetheless, a method for modelling any class of amyloid fibrils starting from its sequence has been lacking until now. In this dissertation, a computationally fast and general computational procedure, FibPredictor, is proposed to generate structural models for any amyloid fibril, starting from its sequence.

1.2 Glucagon Fibrillation

Glucagon is a 29-residue peptide hormone secreted by pancreatic α-cells which plays an important role in glucose metabolism. Currently, it is used for the emergency treatment of hypoglycemia and as a muscle relaxant for endoscopy procedures [15]. Due to poor water solubility of this peptide in neutral pH it has to be solubilized in acidic pH. However, it is not stable even in acidic solution and comes out of solution forming irreversible, insoluble amyloid fibrils. Amyloid fibrils are highly stable protein constructs formed by long β-sheets known as β-spines which interact side-by-side by entanglement of their side chains forming a "steric zipper" [16,17].

Glucagon amyloid fibrils formation compromises the potency of drug, generates toxic effects and increases solution viscosity which causes difficulty in delivering the formulation using an infusion pump or injection pen [15]. Because of these solubility issues, glucagon is currently formulated as a lyophilized powder that is reconstituted just prior to administration, and any leftover solution is discarded immediately [18]. The inconvenience and the risk of needle exposure and dosing error associated with the current formulation has led to underutilization of glucagon despite its safety and efficacy for treatment of insulin-induced hypoglycemia [18]. Moreover, glucagon solubility issues has hindered development of closed loop artificial pancreas device. An artificial closed loop pancreas device can administer insulin and glucagon automatically in response to fluctuations in blood glucose and can significantly improve quality of life for insulin dependent diabetic patients [15]. It is impractical to use the lyophilized formulation for an artificial pancreas, which requires that an adjustable amount of glucagon solution be administered instantaneously in response to fluctuations in blood glucose. Therefore, formulating

glucagon as stable solution not only promotes its utilization for the current uses but also is a major step for expanding glucagon's therapeutic benefits. Nonetheless, in spite of many attempts to solubilize glucagon and inhibit glucagon fibrillation such as modifying glucagon's chemical structure [19,20], controlling solution conditions (e.g., pH, ionic strength) [21–24] and using stabilizing additives (e.g., cyclodextrins) [25], to date a stable solution formulation of glucagon is not yet available in clinic.

Stable phosphorylated glucagon derivatives are introduced in this dissertation as pro-drugs or active alternatives to glucagon which are soluble in neutral pH and do not show any fibrillation for at least for one month. Penta-peptide chaperones are also tested as an alternative method to delay glucagon fibrillation.

1.3 Outline

Chapter 2 presents penta-peptide chaperones to delay glucagon fibrillation. Although penta-peptides delay glucagon fibrillation, their effects are limited compared to alternative approaches, such as the one presented in chapter 5. Chapter 3 presents a number of statistical potentials for protein structure prediction. One of these statistical potentials is then used in the software presented in chapter 4. Chapter 4, presents a computational method for modelling the steric zipper of amyloid fibrils. This computational method is applied in chapter 5 to design phosphate ester derivatives of glucagon as stable and soluble pro-drugs or active alternatives to glucagon.

I have performed the computational studies in Dr. Markus A. Lill's lab, and the experimental part in Dr. Elizabeth M. Topp's lab.

CHAPTER 2. PENTAPEPTIDE CHAPERONES TO INHIBIT GLUCAGON FIBRILLATION

2.1 Introduction

This chapter presents small peptide chaperones, to inhibit glucagon fibrillation. Small peptides have previously used in other cases of problematic amyloid β-fibrils and a number of natural and non-natural peptides have been shown to successfully inhibit fibrillation [26–29]. We go in the same direction and design small peptide chaperones to inhibit glucagon fibrillation. Due to the particular restrictions in case of glucagon such as high hydrophobicity and unavailability of the atomic structure for the fibril, we use a design approach which differs from that of our predecessors. The peptides introduced in this paper, successfully delayed glucagon fibrillation in spite of their simple structure and small size. These peptides provide a starting point for further investigation of small peptide chaperones for inhibiting β-fibril formation.

2.2 Materials and methods

2.2.1 Peptide design

A few natural and non-natural small peptide chaperones have shown to inhibit β-fibril formation in Alzheimer amyloidosis. The natural peptides were designed using the hydrophobic fragment of the target fibrillating protein as template. Proline residues were then incorporated into the template sequence for their known β-breaker properties due to

their special geometric and hydrogen bonding characteristics [26–28]. Although this design approach is shown to be successful in other cases, it cannot be used to design peptides to inhibit fibrillation of our target, glucagon. Glucagon hydrophobic region is too hydrophobic and not soluble at all. Incorporation of proline residues would only aggravate this water insolubility resulting in insoluble peptide chaperones. The rational structure-based design approach applied in case of previous non-natural peptides that inhibit fibrillation is also not possible in the case of glucagon. Those non-natural peptides were designed specifically to interact with the steric zipper region of their target fibrils and prohibition of zipper formation inhibited fibrillation in those cases [29]. This design approach depends on availability of atomistic details of the zipper structure which is not at hand for glucagon. Due to the challenges of β-fibril structure determination, a three-dimensional atomistic structure of glucagon fibril is not yet available [7,30].

Since the template-based and structure-based design approaches were not possible in case of glucagon, we aimed at global screening of peptides for their ability to interfere with glucagon fibrillation. In order to limit the screening set, we focused on penta-peptides, the shortest natural peptides with known fibrillation-inhibition properties [26]. However, even for penta-peptides, there are 20^5 = 3,200,000 candidates, and a comprehensive screening was impractical. Fractional factorial design was used to design a small set of peptides covering the whole penta-peptide space. Fractional factorial design approach has previously used for designing small but information-rich sets of peptides [31,32] and theoretically, a set designed in this way provides a fast and cheap way to screen the whole peptide space for hits.

Two numerical descriptors for amino acids (tciz1 and tciz2) introduced by Muthas et al [32] were used. The two descriptors are the first two principle components of a number of different descriptors for amino acids. These two descriptors have been shown to capture most of the variance in peptide sets and are calculated for many natural and unnatural amino acids [32]. There are two sets of these variables, one calculated based on only amino acids, the other calculated based on a larger set of natural, unnatural and derivatized amino acids [32]. The latter set was used due to its larger scope and extendibility to unnatural amino acids in later studies.

Describing each amino acid with two descriptors, each penta-peptide was described with ten residue-position-specific variables (table 1). A 2^{10-6} fractional factorial design table was used (obtained from [33]) (table 1). This fractional factorial design table assumes two levels for each descriptor, and hence the values for the descriptors should be discretized in two levels. As the current study was focused on natural peptides, the positivity or negativity of the variable calculated based on both natural and unnatural amino acids could not be used for discretizing the values into two levels. Therefore, the average value of the tciz1 and tciz2 for natural amino acids was set as the zero point and all the descriptors were transformed accordingly. The negativity and positivity of the transformed values was the criteria to discretize the variables in two levels: positive or negative.

Having two descriptors each with two discretized levels, amino acids were categorized into four classes: positive-positive, negative-positive, positive-negative and negative-positive (table 2). Representative amino acids were chosen to represent each category (table 2). T (negative-negative), F (positive-negative) and Q (negative-positive) were

shown to be important in glucagon fibrillation in our MD simulation studies [34] and thus, were chosen to represent their corresponding categories. H was chosen for the (positive-positive) class as it can participate in various types of interactions and its interaction versatility may facilitate the peptide-glucagon interaction.

Substituting the representative amino acids in table 1, we obtained our set of peptides (Table 3). From the sixteen peptides of this set, H6 and H11 were excluded due to their very high insolubility which interferes with the experiments.

2.2.2 Sample Preparation

Glucagon at 1.6 mg/mL in 3.2 mM HCl, 0.9% NaCl (w/v) (pH 2.5) was centrifuged at 14,000 rpm for 5 min and filtered through 0.1 μm filters to eliminate any insoluble particles. For water-soluble peptides (h1, h3, h4, h8, h9, h10, h12, h13, h14, h15 and h16), 100 μL of the filtered glucagon sample was quickly transferred to a 96-well black flat bottom microtiter plate in duplicate or triplicate depending on peptide availability and incubated with 40 μL of 10 mg/ml solution of peptide in buffer and 50 μM ThT final concentration. For peptides with less water-solubility (h2, h5, h7), 10 μL of 40 mg/ml solution of peptide in DMSO was used. The final volume was adjusted to 200 μL using the buffer as mentioned above. Two control triplicates of glucagon and ThT without peptide, one without and the other with 10 μL DMSO were also prepared as standards. Samples of peptides with ThT but without glucagon were also prepared as negative controls as described above. Buffer was used to adjust the final volume of the control samples to 200 μL. The plate was sealed with a clear sealing tape. Fluorescence

measurements were carried out in a BioTek Synergy 4 Multi-Detection microplate reader (BioTek Instruments, Winooski, VT) as described below.

2.2.3 ThT Assay

The fluorescence intensity of ThT was measured over 24 hours every 15 minutes at 23°C with 5 s automixing before each reading with the excitation and emission wavelengths set to 440 nm and 482 nm. Fluorescence signals exceeding 100,000 (overflow) were re-set to 100,000 for graphing purposes.

2.2.4 Intrinsic Fluorescence Assay

The excitation and emission wavelengths were set to 295 nm and 355 nm, respectively, to look at the fluorescence of Trp25. Peptides do not have tryptophan in their sequence and therefore do not interfere with glucagon signal. Measurement was carried out for 24 h at 23°C at 15-min intervals preceded by 5 s automixing before each reading. Very high fluorescence signals exceeding 100,000 (overflow) were re-set to 100,000 for graphing purposes.

2.2.5 Partial Least Square Regression

Partial least square regression (PLSR) is a common linear modelling technique for QSAR modelling and is superior to multiple linear regression (MLR) due to its ability to build

reliable models with numerous collinear and noisy variables. In this method, latent independent and dependent factors are constructed aiming at maximizing correlation between the variations of the independent and dependent ones [35].

2.2.6 MD Simulations

Initial structures of the N-terminal (residues 1–8) and C-terminal (residues 22–29) fragments were generated from reported NMR structures of glucagon (Protein Data Bank (PDB) ID: 1KX6)[36]. Three different models were selected as starting configurations for MD simulations, and are referred to as models 1, 5, and 10 in keeping with the numbering in the ensemble of NMR models in the original PDB file. In simulations of the interactions of two molecules of either the 1–8 fragment or the 22–29 fragment, the molecules were initially placed close to one another with arbitrary relative initial orientation, maintaining at least a 4 Å distance between any two atoms in the two fragments. Combining the conformations of the three NMR models for each fragment, three starting configurations were generated for each of the N-terminal and C-terminal fragment simulations. Specifically, starting configurations for both the N-terminal fragment (1, 2, 3, 4, 5, 6, 7 and 8) and the C-terminal fragment (22, 23, 24, 25, 26, 27, 28 and 29) simulations were: model 1 with model 5, model 1 with model 10, and model 5 with model 10. All simulations were performed on capped peptides (i.e., N-terminus acetylated and C-terminus amidated) and the side chains of His residues in the N-terminal fragments were doubly protonated to represent the most likely state in solution at pH 2.5. To simulate the interactions of peptide fragments, the molecules were solvated in a preequilibrated octahedron of TIP3P water molecules, with a minimum distance of 10 Å between the octahedron boundary and solute atoms [37]. Production simulations were

performed in an NPT ensemble using the AMBER-99SB force field with periodic boundary conditions and an integration time step of 2 fs, applying the particle mesh Ewald method to treat electrostatic interactions[38]. All bonds involving hydrogen atoms were constrained using the SHAKE algorithm [39] and van der Waals interactions were truncated at a distance of 10 Å. A Langevin thermostat [40] with collision frequency of 1 ps^{-1} was used to maintain the temperature at 298 K, and pressure was maintained at 1 atm using isotropic position scaling with a pressure relaxation time of 2 ps. The N-terminal 1–8 fragment simulation was neutralized by the addition of one Cl^- ion per fragment. In a simulation, the water molecules with constrained peptide(s) first were energy minimized. The system was then gradually heated from 0 K to 298 K over a 20 ps MD simulation period. The system was then equilibrated at constant temperature and pressure for 200 ps and final production runs performed for 100 ns. Snapshots were saved every 0.05 ns, resulting in 2000 snapshots for each production simulation.

A contact between residues from two molecules was identified if a distance <5 Å was observed between any pair of atoms. Only contacts formed between two different peptide molecules were analyzed, and not those within a single strand. All MD snapshots of the simulations were considered for contact analysis. The frequencies of observing contacts were first analyzed for the three separate simulations of two molecules, and then averaged over all three simulations to obtain a single mean contact frequency. The α-helix content of each snapshot was analyzed using the DSSP software [41,42]. In simulations of two interacting peptides, snapshots were analyzed separately for each peptide. For each amino acid, the percentage of snapshots in which it was part of an α-helix substructure was computed for all simulations, and the mean structural content of the

various single and two peptide molecule(s) simulations was computed. Molecules were visualized in PyMOL [43–45] and the graphs were generated using Python and matplotlib [46].

2.3 Results

2.3.1 Glucagon interactions by MD simulation

MD simulations were performed to provide insight into structural changes and early interactions involved in glucagon fibrillation. The α-helix content of fragment 1–8 was negligible in simulations of either one or two molecules, the latter allowing for effects of interaction on secondary structure (Figure 1, *A* and *B*). In contrast, the C-terminal fragment 22–29 formed α-helices in both one- and two-molecule simulations (Figure 2, *C* and *D*), with greater α-helix content in simulations of two molecules. To mimic the experimental conditions, MD simulations were repeated in the presence of 0.9% NaCl for a system containing two N-terminal fragments (model 1 with model 10) and two C-terminal fragments (model 1 with model 10). The simulations were performed for 15 ns and compared to the first 15 ns of the salt-free simulations. Though a slight increase in secondary structure was observed in the presence of salt, the difference in the α-helix content was minimal (data not shown).

In light of the experimental evidence that C-terminal interactions are involved in the early stages of fibrillation[34], we aimed to identify the critical contacts for the C-C-terminal interactions. When analyzing the contacts between amino acids, the C-terminal fragment 22–29 showed at least one contact in >94% of snapshots for all models tested. To highlight the preferred side-chain interactions, the 10 most frequently observed contacts averaged from three independent simulations of two molecules of the C-terminal fragment 22–29 were identified (Figure 2). Hydrophobic interactions between amino

acids are most frequently observed and account for eight of the 10 most frequent interactions. In particular, Trp-25 participates in four of the 10 most frequent interactions, i.e., with Phe-22, Val-23, Leu-26, and Met-27. Amino acids adjacent to Trp-25 also participated in hydrophobic contacts. Phe-22, for example, is engaged in five out of the top 10 most frequent interactions, four with hydrophobic or aromatic residues. An aromatic T-shaped interaction between Phe-22 and Trp-25 is also among the most frequent contacts[34].

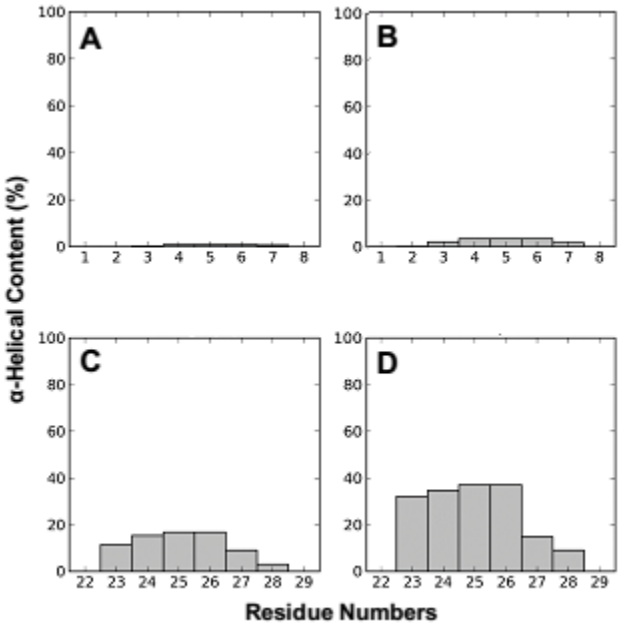

Figure 1 Simulations of α-helix content of glucagon-derived peptides: (A) a single molecule of fragment 1-8, (B) two molecules of fragment 1-8, (C) a single molecule of fragment 22-29 and (D) two molecules of fragment 22-29. The α-helix content for each amino acid in the fragment is shown.

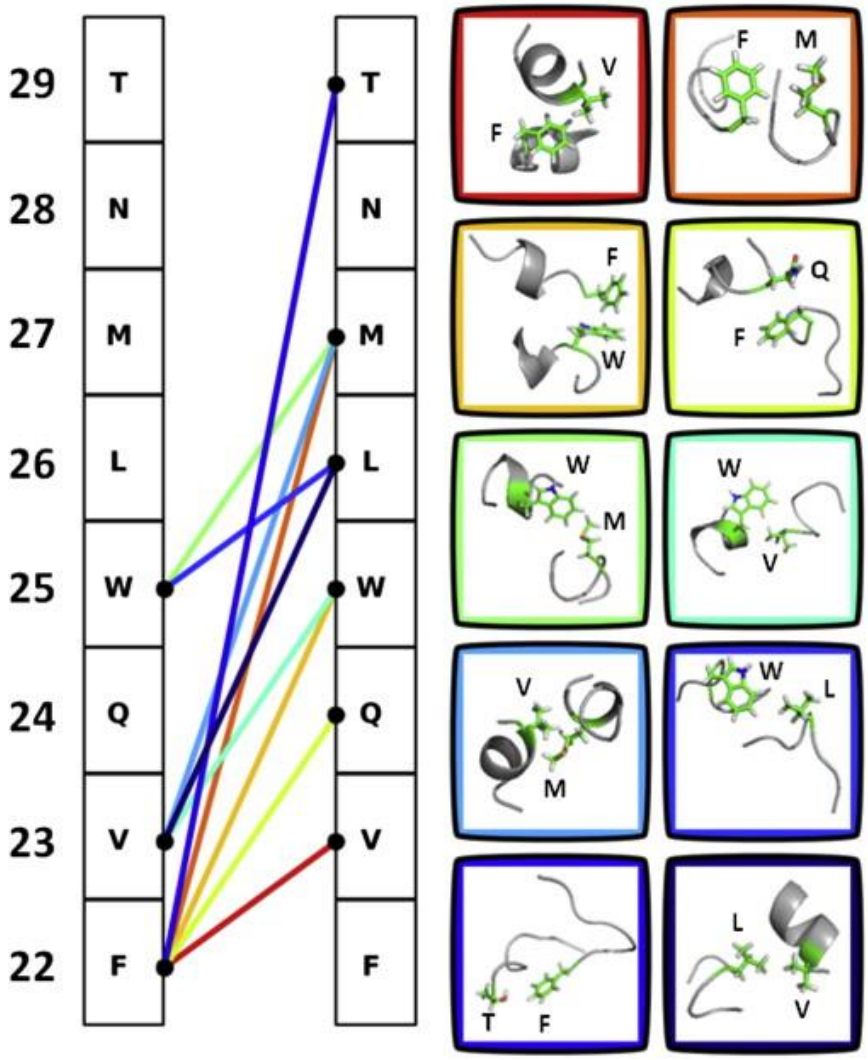

Figure 2 Identification of critical contacts for the C-terminal interactions in glucagon fibrillation under acidic conditions. The 10 most frequent contacts observed in simulations of two molecules of glucagon fragment 22–29 are shown. Each line represents one of the 10 interactions, which are ordered from red to blue based on frequency. The amino acid residues are indicated by their single letter code with residue numbers on the left.

2.3.2 Glucagon Fibrillation

The ThT fluorescence signal increases upon interaction of ThT molecules with an amyloid β-fibril and allows following fibrillation. Tryptophan fluorescence signal drops as tryptophan residues get buried upon peptide aggregation and thus, provides a second complementary method to conform ThT results [34]. ThT and tryptophan intrinsic fluorescence of glucagon were followed over 24 hours in presence and absence of each peptide, in order to investigate the fibrillation inhibitory effects of the peptide set.

ThT fluorescence graphs (Figure 3) show that glucagon fibrillation starts with a lag time followed by a sudden log phase and ends reaching a plateau. This is a known pattern and is previously reported and explained by us and others [34,47]. This pattern shows that once the fibrillation passes the lag time, it fast goes to completion. Any effort to stop or reverse the fibrillation is better to be focused on elongation of the lag time. Tryptophan fluorescence graphs (Figure 1), although less clearly, show a general pattern similar to ThT: a lag time and a sudden drop indicating a fast aggregation. However, the tryptophan fluorescence patterns are less clear and definitive compared to ThT fluorescence graphs. Nonetheless, the sudden drop in the tryptophan fluorescence, if identifiable, happens usually close to the time that ThT fluorescence surges and verifies the lag time identified by ThT fluorescence.

Although the lag time-log phase-plateau pattern is generally preserved across DMSO- and water-soluble peptides, the shape of the graphs are slightly different between these two groups. The difference is most salient in the standard glucagon samples with no peptide, where the DMSO containing standard show a less definitive plateau compared to the other standard not containing DMSO. Also the lag time for the glucagon standard is

shorter in presence of DMSO compared to the no DMSO standard sample. This indicates that DMSO interferes with the fibrillation process and make it slightly faster. None of the peptides show fibrillation of their own (graphs not shown) and the glucagon containing samples were the only ones which showed fibrillation.

Although the general patterns of ThT and tryptophan fluorescence graphs were preserved in presence or absence of peptides, the lag time varied significantly in presence of peptides as discussed in detail in the next section.

2.3.3 Fibrillation lag time extension

Figure 4 shows the lag time difference between the standard glucagon without any peptide and glucagon in presence of each of the peptides. All of the peptides studied affected the fibrillation lag time and except H7, all of them elongated the lag time. The most effective peptide H8 (QFFTQ) elongated the fibrillation lag time for more than 700 minutes resulting in a total lag time of nearly 1000 minutes. H8 is water soluble and its solubility in addition to its effectiveness make it a promising hit for glucagon fibrillation inhibition.

There is considerable variation between the effects of different peptides on the fibrillation lag time. This variation shows the lag time elongations are not due to general presence of any peptide, but are in fact the sequence-specific. Note that the sequence of these peptides are composed of only four different residues and the variation between the sequences is very limited. Nonetheless, even this limited sequence variety results in considerable divergence in fibrillation inhibitory effects. Variation in inhibitory effects of the DMSO-soluble peptides shows the effect is not reducible to simple physical

properties such as hydrophobicity and supports sequence-specificity of the inhibitory effects.

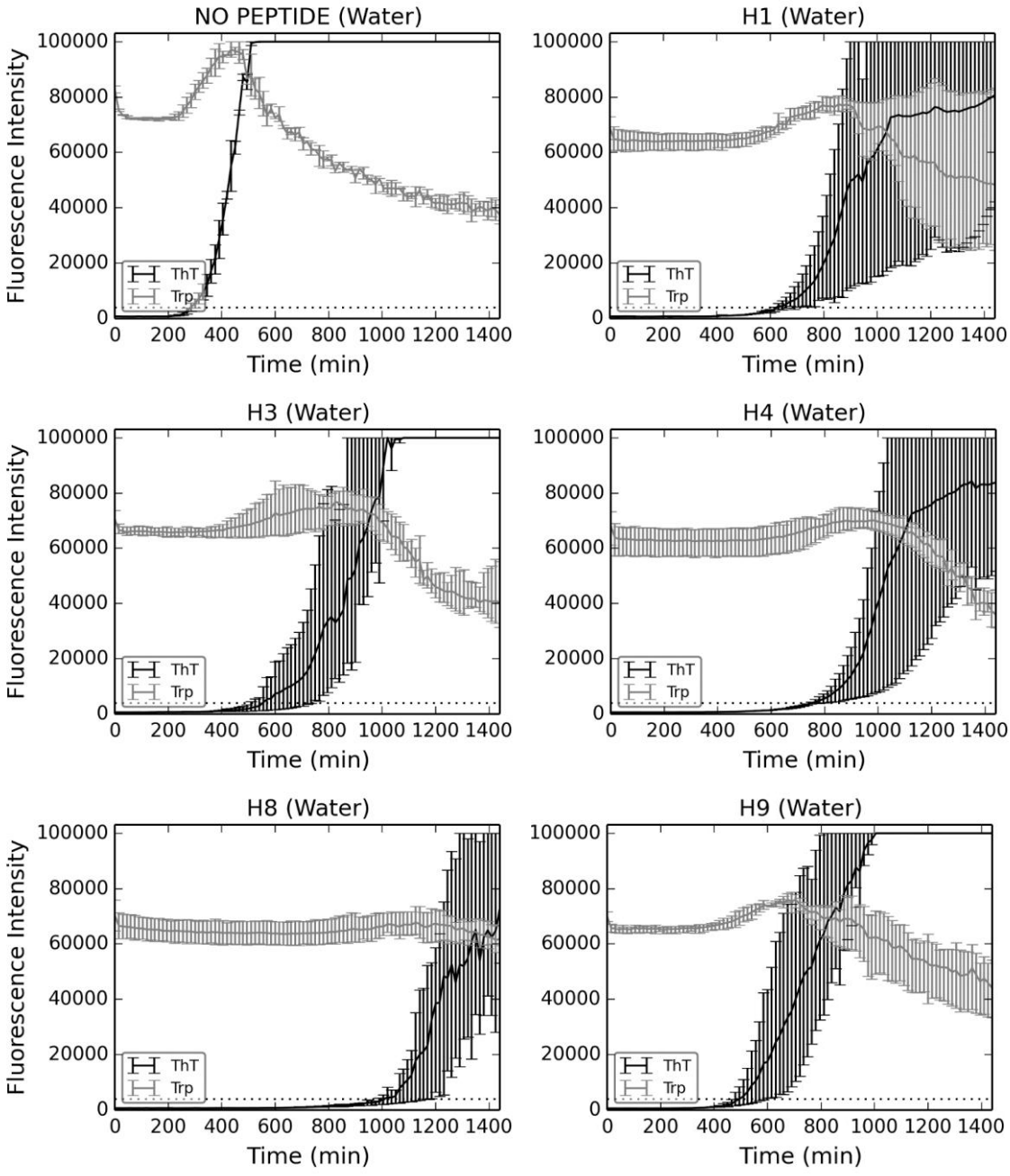

Figure 3 ThT and Tryptophan fluorescence over time. The dotted y=4000 line indicates the cutoff for end-of-lag-time identification.

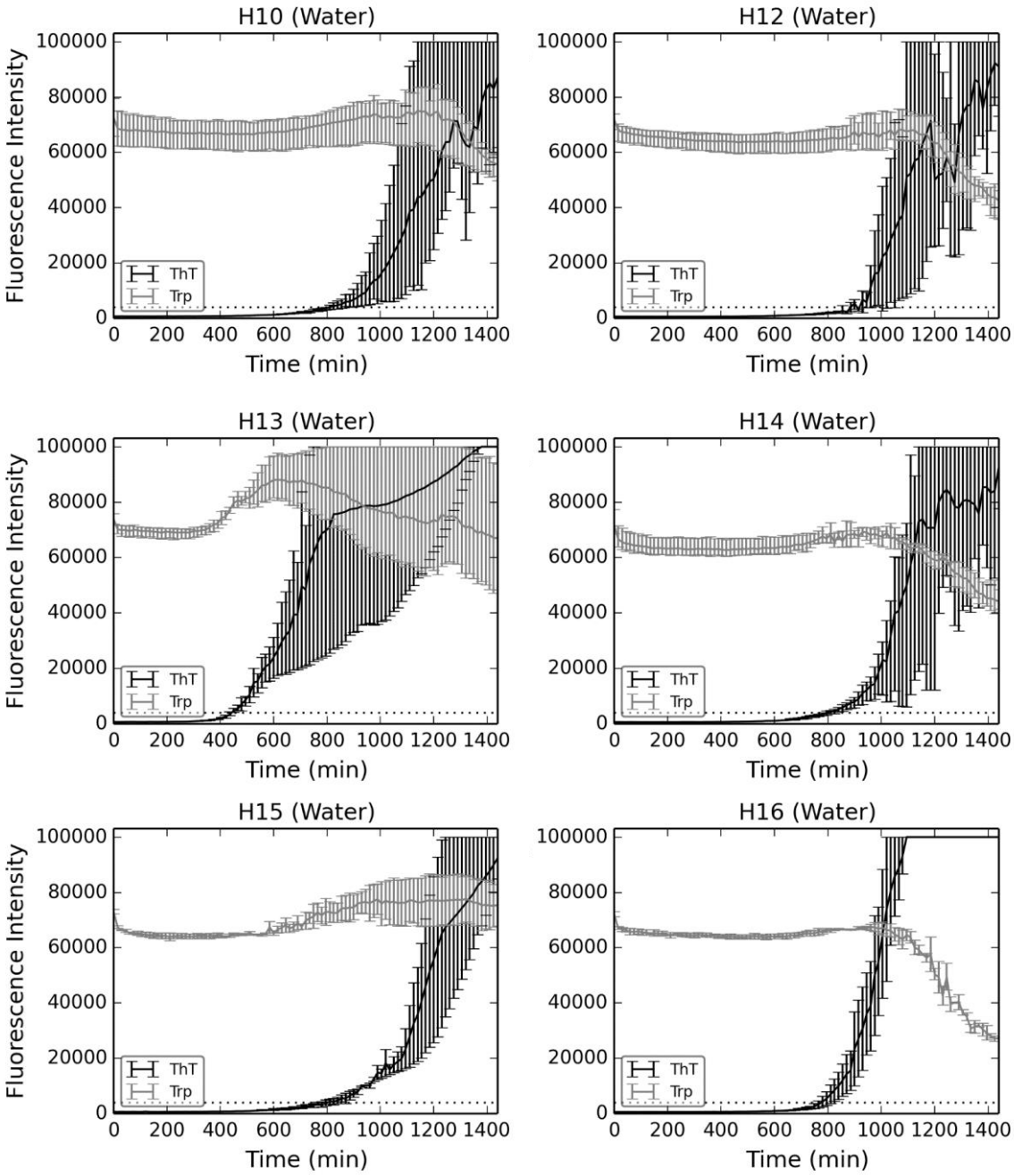

Figure 3 continued

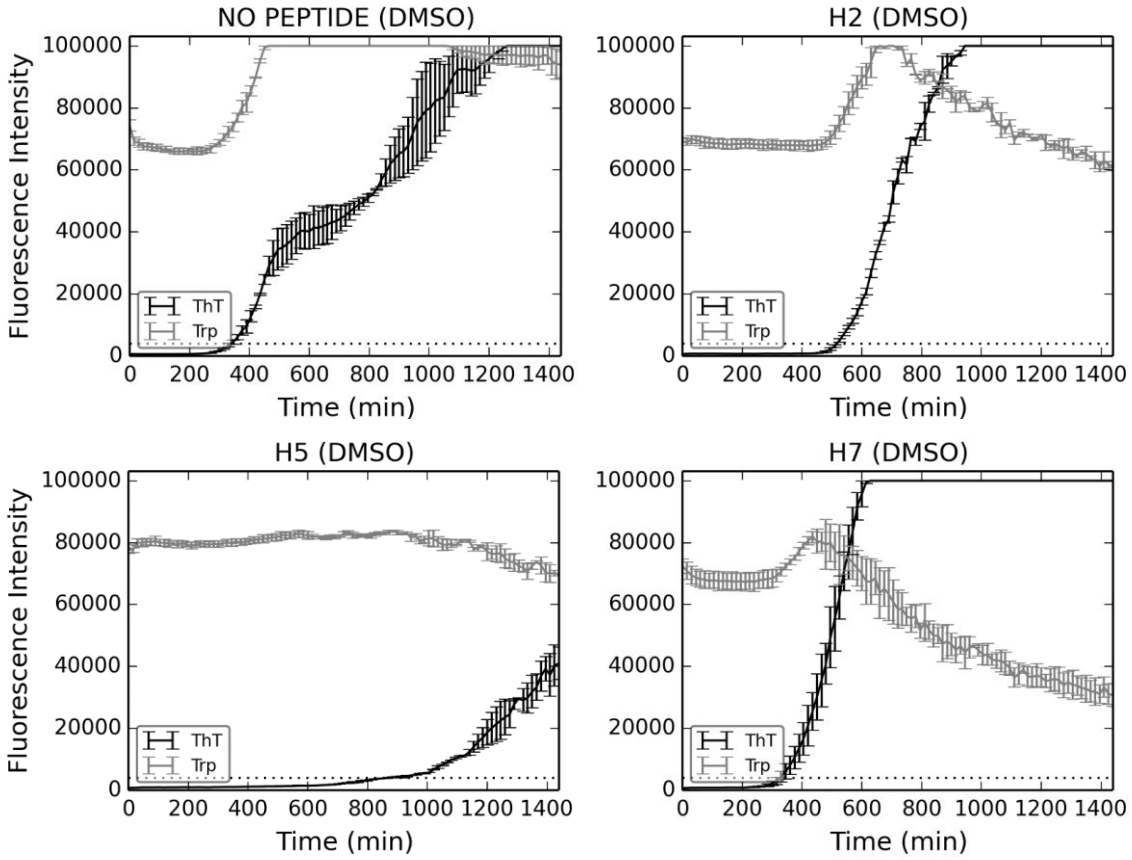

Figure 3 continued

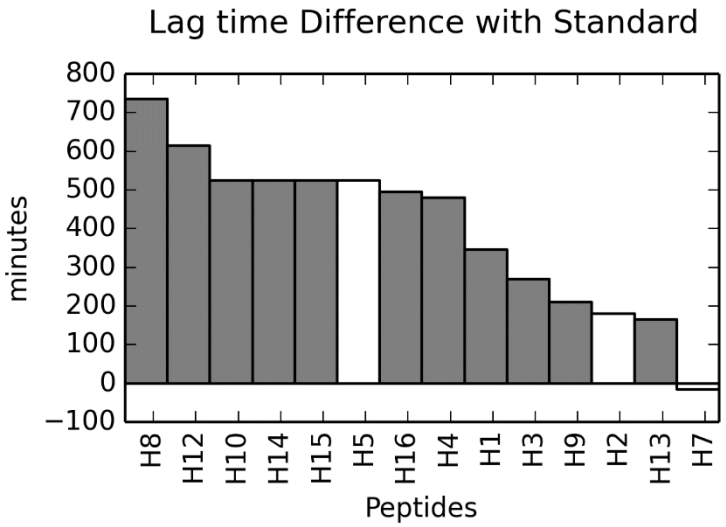

Figure 4 The glucagon fibrillation lag time difference between the samples containing various peptides and the standard no-peptide samples. Bars related to water soluble and DMSO soluble peptides are colored in grey and white respectively. Water soluble and DMSO soluble peptides were compared with their corresponding standard samples.

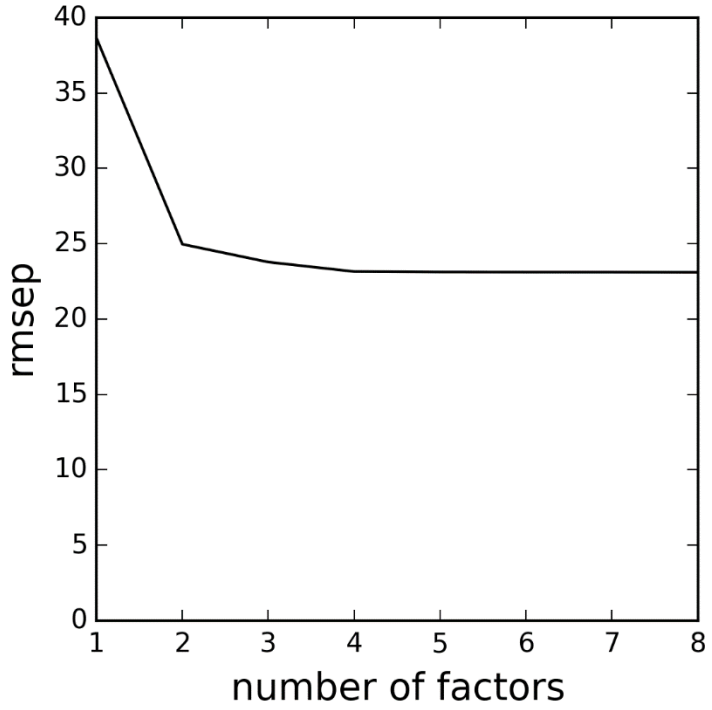

Figure 5 Root mean squared error of the predicted lag time differences with experimental values vs number of factors in the PLS model

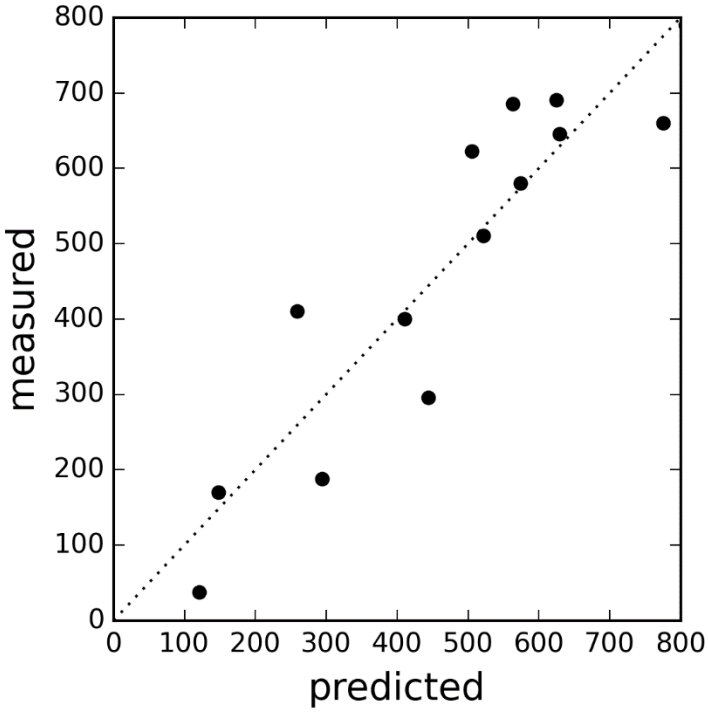

Figure 6 Measured lag time difference vs predicted lag time difference

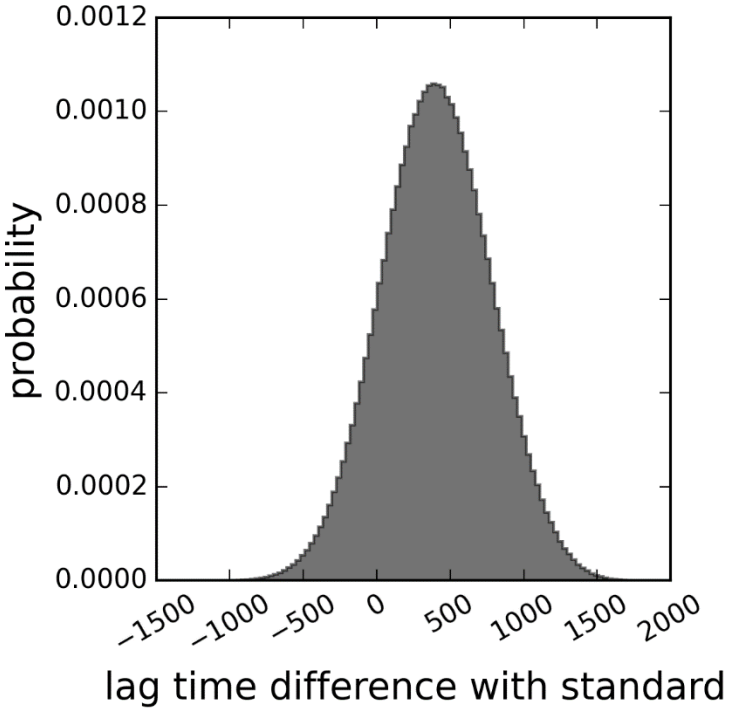

Figure 7 Distribution of calculated lag time differences with standard for all possible penta-peptides

2.3.4 PLS model

Figure 5 shows the root mean squared error between the predicted lag time differences with standard and the experimental measurement (rmsep) vs the number of latent independent factors included in the PLS model. The figure shows improvement in rmsep up to three factors. Figure 6 shows the predicted lag time difference with standard vs the measured values for the PLS model with three latent factors. Based on these results the PLS model with three latent factors was used to predict the lag time difference with standard for all possible penta-peptides. The results are shows in Figure 7. According to these predictions, there are many candidate penta-peptides that can delay fibrillation, and the study suggests hit penta-peptides, such as QFFTQ. However, the maximum delay will be limited to around two thousand minutes (≈33 hours) which is not enough for practical uses of glucagon.

2.4 Conclusions

Protein aggregation poses an important challenge for therapeutic formulation of proteins [48]. Glucagon fibrillation is an example where the therapeutic benefits of a peptide drug is significantly limited by low stability of its formulation. Many attempts to stabilize glucagon have not yet resulted in its soluble formulation in clinic. This studies the potential of small peptide chaperones, more specifically penta-peptides, to inhibit glucagon fibrillation. We also suggested a hit penta-peptide sequence: QFFTQ. The fibrillation inhibitory effects of these peptides is sequence-specific. This opens a path towards developing more effective and more potent glucagon fibrillation peptide inhibitors. However, the limited delay predicated for these penta-peptides is not enough

to solve glucagon formulation problem. Other paths (such as the chemical modification of glucagon presented in the next chapters) should be pursued.

The methods and the results presented in this paper have implications even beyond glucagon fibrillation. Amyloid β-fibrils are involved in many serious pathological conditions and are important drug targets [49]. The present work and its predecessors show that small peptide chaperones have the potential to successfully inhibit amyloid β-fibrillation and underscore the importance of small peptide chaperones for drug development. However, peptide drug design is challenging due to the combinatorially large number of candidate sequences. The peptide design approach used in this study can provide a general guideline for the initial steps of designing small peptide libraries.

Table 1 2^{10-6} fractional factorial design table; each row corresponds to one peptide

Amino acid 1		Amino acid 2		Amino acid 3		Amino acid 4		Amino acid 5	
tciz1	tciz2	tciz1	tciz2	tciz1	tciz2	tciz1	tciz2	tciz1	tciz2
-1	-1	-1	-1	-1	-1	-1	-1	+1	+1
+1	-1	-1	-1	+1	-1	+1	+1	-1	-1
-1	+1	-1	-1	+1	+1	-1	+1	-1	-1
+1	+1	-1	-1	-1	+1	+1	-1	+1	+1
-1	-1	+1	-1	+1	+1	+1	-1	-1	+1
+1	-1	+1	-1	-1	+1	-1	+1	+1	-1
-1	+1	+1	-1	-1	-1	+1	+1	+1	-1
+1	+1	+1	-1	+1	-1	-1	-1	-1	+1
-1	-1	-1	+1	-1	+1	+1	+1	-1	+1
+1	-1	-1	+1	+1	+1	-1	-1	+1	-1
-1	+1	-1	+1	+1	-1	+1	-1	+1	-1
+1	+1	-1	+1	-1	-1	-1	+1	-1	+1
-1	-1	+1	+1	+1	-1	-1	+1	+1	+1
+1	-1	+1	+1	-1	-1	+1	-1	-1	-1
-1	+1	+1	+1	-1	+1	-1	-1	-1	-1
+1	+1	+1	+1	+1	+1	+1	+1	+1	+1

Table 2 Four classes of amino acids based on the two-level discretization of tciz1 and tciz2 variables.

tciz1	tciz2	Amino acids
-1	-1	A, T, S, C
+1	-1	V, L, I, M, F, W
-1	+1	N, D, Q, E
+1	+1	K, H, R, T

Table 3 training set; each row corresponds to one peptide

Peptide code	aa 1	aa2	aa3	aa4	aa5
H1	T	T	T	T	H
H2	F	T	F	H	T
H3	Q	T	H	Q	T
H4	H	T	Q	F	H
H5	T	F	H	F	Q
H6	F	F	Q	Q	F
H7	Q	F	T	H	F
H8	H	F	F	T	Q
H9	T	Q	Q	H	Q
H10	F	Q	H	T	F
H11	Q	Q	F	F	F
H12	H	Q	T	Q	Q
H13	T	H	F	Q	H
H14	F	H	T	F	T
H15	Q	H	Q	T	T
H16	H	H	H	H	H

CHAPTER 3. ARE DISTANCE-DEPENDENT STATISTICAL POTENTIALS CONSIDERING THREE INTERACTING BODIES SUPERIOR TO TWO-BODY STATISTICAL POTENTIALS FOR PROTEIN STRUCTURE PREDICTION?

3.1 Introduction

Protein structure prediction still represents a significant challenge to computational biophysics. Recently developed statistical scoring functions have proven to be a valuable tool for identification of the native structure among a typically large set of candidate structures [50,51]. These potentials are typically based on the assumption that the total free energy of a protein structure can be computed by the sum of all pairwise free energies ($\Delta G(r_{ij})$)

$$\Delta G_{tot} = \sum_{i<j} \Delta G(r_{ij}) \qquad \text{Eq. 1}$$

where *i* and *j* are either interacting bodies e.g. individual atoms of the protein or representative points for each amino acid, e.g. the Cα atom etc. The pairwise free energies are often calculated based on the pairwise distribution function ($P(r_{ij})$) between a specific pair of atom types or amino acids, *i* and *j*

$$P(r_{ij}) = \frac{1}{Z}\exp\left(\frac{-\Delta G(r_{ij})}{RT}\right) \qquad \text{Eq. 2}$$

where Z is the partition function, R is the gas constant and T is the temperature. Therefore, the inverse Boltzmann equation used to calculate $\Delta G(r_{ij})$ would be:

$$\Delta G(r_{ij}) = -RT\ ln\ P(r_{ij}) - RT\ ln\ Z \qquad \text{Eq. 3}$$

Typically $\Delta G(\mathrm{r_{ij}})$ (the potential of mean force (PMF)) is computed with respect to a reference state R representing a hypothetical system with uniform and unbiased interactions between the different atom types or amino acids. The relative free energy between a pair of atoms or residues *i* and *j* with respect to this reference state is then computed by

$$\Delta G(r_{ij}) = -RT \ln \frac{P(r_{ij})}{P_R(r_{ij})} - RT \ln \frac{Z}{Z_R} \qquad \text{Eq. 4}$$

The pairwise distribution function can be computed by measuring the frequency of pairs of atom types at a given distance using databases of experimentally solved protein structures [52]. An early example of such potential functions is developed by Samudrala and Moult. Their function models potentials of atomistic interactions based on the pairwise distance between two interacting bodies [51].

The underlying assumption of Eq. 1, that the total free energy of a protein structure can be computed by the sum of all pairwise free energies, however, is not physically justified. More precisely, the exact free energy of a system is determined by the statistical mechanical relationship between the *N*-body distribution function and the free energy[51].

$$\Delta G(r_1, \ldots, r_N) = -k_B T \ln P(r_1, \ldots, r_N) - k_B T \ln Z \qquad \text{Eq. 5}$$

Thus Eq. 1 neglects correlation effects between multiple atoms or amino acids in a protein. In order to model these higher order interactions a number of multi-body contact based statistical potentials have been developed. Most of these statistical potentials are based on Delauney tessellation - a geometric technique to identify the neighboring bodies [53–57], although some other geometric approaches have also been investigated [58]. To the best of our knowledge none of these multi-body potentials look into the details of distance between interacting residues. Also they usually use very coarse-grained representations of interactions e.g. interaction between residues and do not model interactions between various atom types. Based on this discussion, we asked the question if we can model details of three-body interactions using distance dependencies between pairs of pairwise interactions between atomistic interacting bodies. More precisely we hypothesized that considering the presence of a third body adds valuable information to statistical potentials based on interaction pairs. This additional information of multi-body interactions may improve the scoring process and consequently the identification of native protein structures.

The importance of three-body terms in determining the stability of globular forms of polymers has been established long ago, and by analogy their inclusion in statistical potentials for protein native structure detection has been conjectured [59]. The importance of multi-body interactions in protein folding has been shown independently using other computational methods [57,60], which makes the idea of building a multi-body distance-based statistical potential for protein structure determination seem even more promising.

In this study, we generated a distance-based quasi-three-body statistical potential for atom-based interacting bodies and analyzed if we can identify dependence between multiple pair-wise interactions. We investigated the effect of the distance from a third body on the pairwise distance of two interacting partners.

We developed statistical potentials describing the simultaneous interaction of three bodies that represent important physical elements of the protein and used it to differentiate native protein structures from decoys. Those elements characterize either physicochemical properties of the protein, which we call the physicochemical elements throughout the paper (hydrogen-bond acceptors and donors, negatively and positively charged, hydrophobic, and aromatic groups), Amber atom types, or amino acid Cα atoms. We assumed that the presence of the third interacting element affects the pairwise distribution function of the other two interacting elements by altering the energetically optimal distance between the two interacting bodies. We also used three simple counting scoring functions (counting hydrophobic centers or Cα's within a certain distance from each other and counting the number of hydrogen bonds) in order to investigate if using more sophisticated and computationally costly methods perform better compared to very simplistic approaches.

In order to assess the performance of different scoring functions, we tested the functions' ability to separate decoys from native protein structures. Three different decoy sets were utilized to evaluate the performance of the scoring functions for protein structure prediction [61]. The performances of our quasi-three-body scoring functions were compared to existing method including FoldX, DFIRE2, dDFIRE, GOAP, Rosetta, and simple counting methods.

3.2 Materials and Methods

3.2.1 Assigning the properties to proteins

Statistical potentials were derived between different elements characterizing the physicochemical and structural properties of a protein structure. Physicochemical properties of a protein were defined as hydrophobic (H), hydrogen-bond donor (D), acceptor (A), and aromatic (R) properties and formally charged functional groups (P for positively charged, N for negatively charged,). For a given protein, the physicochemical elements were assigned as follows: Hydrophobic elements were assigned to carbon and sulfur atoms that are not bonded to an oxygen or nitrogen atom. For assigning hydrogen bond donors and acceptors, hydrogen atoms were added to the protein structure using Open Babel 2.3.1. Hydrogen bond acceptor and donor physicochemical elements were included in the generation of statistical potentials only if they form intra-protein hydrogen bonds, discarding unpaired hydrogen bonds. The following criteria were used to define hydrogen bonds: The distance between a donor group and the acceptor atom must be closer than 4.6 Å, the angle between donor heavy atom, donor hydrogen and the acceptor heavy atom needs to be in the range 120-180°, and the angle between acceptor lone-pair, acceptor heavy atom and the donor hydrogen must be smaller than 45°. Acceptor and donor elements were then assigned to the acceptor and donor heavy atoms. It should be noted that although only acceptor or donor groups that are engaged in hydrogen bonds are considered in the analysis, triplets can freely contain one partner independent of the other as well as both partners. An aromatic physicochemical element was assigned to the center of each aromatic ring, i.e. to the side chains of Phe, Tyr, His and Trp. Negatively and positively charged physicochemical elements were assigned to Glu, Asp, Arg and

Lys specific side chain atoms. Scoring functions constructed from physicochemical elements are denoted by "Phys_" in their names. We also generated statistical scoring functions based on analyzing quasi-three-body and two-body interactions using all heavy atoms classified by Amber99 atom types. These scoring functions are denoted by "Amb_". Two additional scoring functions are based on quasi-three-body and two-body interactions among the Cα atoms of all residues; no classification with respect to amino acid attributes was used. These scoring functions are denoted by "Ca_". Throughout this paper we call the scoring functions resulting from the quasi-three-body approach as *quasi-three-body* scoring functions (denoted by the suffix "_3b_score") to differentiate them from the *two-body* scoring functions (denoted by suffix "_2b_score") resulting from pair-wise distance distributions. Physicochemical elements, Amber atom types and amino acid Cα atoms were assigned using in-house software.

3.2.2 Protein database for generation of statistical potential

To generate the statistical potentials, 1000 non-redundant protein structures were chosen from the PDB databank by clustering proteins into groups based on their pairwise sequence similarity and picking a representative from each group using the online tool VAST.

3.2.3 Interacting Pairs and Triplets

For each set of properties (physicochemical elements, Amber atom types and Cα atoms) pair-wise and quasi-three-body statistical potentials were derived for all possible combinations of properties. For pair-wise potentials the frequency of each pair of properties *A* and *B* as a function of distance is stored in histograms (F_i^{AB}) where i represents one of 32 distance bins with a bin width of 0.25 Å. Distances between 2 to 10

Å are considered in our analysis. Throughout the paper, parentheses around vectors and matrices like those around (F_i^{AB}) refers to the vector or matrix as a whole, and lack of these parentheses denotes an element in that vector or matrix

For quasi-three-body interactions, we extend pair-wise statistical potentials to triplets of interacting properties using a novel geometric approach. The three distances AB, AC, and BC unambiguously describe the relationship of the triplet of interacting elements A, B, and C (Fig. 8-B). The corresponding histogram would require for each triplet of properties data sampling for 32^3=32,768 bins (32 bins per distance). Obtaining sufficient experimental data for such a large number of bins is impractical. To address this sampling problem, we reduced the dimensionality of the triplet by reducing the description of triplet interactions to two distances spawning from a center point (Fig. 1-A). As a consequence, three different pairs of distances (Fig.8-A) with different center element can be formed which constitute different statistical potentials, i.e. (AB, BC) with center B, (BA, AC) with center A, (AC, CB) with center C. Therefore each triplet is defined by its center and the two other elements (Fig. 8-A). Consequently, the full three-body statistical potential is reduced to two pairs of conditional pair-wise interactions, which we named quasi-three-body potentials throughout this study.

Each of the properties is used as the center of the triplet, and all of the combinations of other properties that form triplets with this center are computed. For example, for six different physicochemical elements 126 triplets were formed. A two-dimensional distance matrix (F_{ij}^{ABC}) for each triplet ABC (center: B) is computed with a distance range from 2.0 to 10 Å, and a bin size of 0.25 Å. The bin number for distance *AB* and *BC* are i and j. ($F_{R,ij}^{ABC}$) is the same matrix populated based on the reference state described

below. A vector (F_i^{AB}) is used to store distance data in a similar way for two-body interactions. ($F_{R,i}^{AB}$) stores the distance data for pair-wise interactions in the reference state. (F_{ij}^{ABC}) , ($F_{R,ij}^{ABC}$) , (F_i^{AB}) , ($F_{R,i}^{AB}$) vectors and matrices of each triplet or pair are normalized to one to give the probabilities (P_{ij}^{ABC}) , $\left(P_{R,ij}^{ABC}\right)$, ($P_i^{AB}$) , ($P_{R,i}^{AB}$).

3.2.4 Statistical potential and definition of reference state

The quasi-three-body and two-body statistical potentials are derived from the elements of the distance matrix P_{ij}^{ABC} and vector P_i^{AB} using Boltzmann inversion:

$$B_{ij}^{ABC} = -RT \ln \frac{P_{ij}^{ABC}}{P_{R,ij}^{ABC}} \qquad \text{Eq. 6}$$

and

$$B_i^{AB} = -RT \ln \frac{P_i^{AB}}{P_{R,i}^{AB}} \qquad \text{Eq. 7}$$

for interacting triplets and interacting pairs respectively. *R* is the gas constant, *T* is the temperature, and (B_{ij}^{ABC}) and (B_i^{AB}) are matrices of the individual quasi-three- and two-body interaction terms of a statistical potential. If $P_{R,ij}^{ABC}$ is equal to or less than $4 \cdot 10^{-6}$ or if $P_{R,i}^{AB}$ is equal to or less than $2 \cdot 10^{-4}$, B_{ij}^{ABC} or B_i^{AB} are set to zero respectively in order to avoid artificially high values due to division by a value close to zero.

A randomized state with no specific interactions between the protein-describing elements is generated to serve as reference state. In generating the random state we adopted a shuffling approach [62]: randomized state matrices ($P_{R,ij}^{A'B'C'}$) are generated by assigning each triplet ABC from a protein structure with given distance bins i and j to the same distance bins but randomized properties A', B', and C'. For example, a donor-acceptor-

aromatic triplet with distance bins i=6, j=15 will be assigned to the same distance bins i=6, j=15 in the random $P_{R,ij}^{A'B'C'}$ matrix where A'B'C' might be any random triplet of properties such as acceptor, positively charged, hydrophobic, etc.. In this way, the random state matrices for different triplets preserve the shape and associated interaction distances of the proteins used in the analysis. As the reference state has a protein-like shape, the resulting scoring function will not be biased towards decoys solely by having a protein-like shape. Using an ideal gas to generate the random matrix would not remove the inherent shape and density dependency of the statistical potential from the protein shape. In other words, the ideal gas reference state produces a random spherical distribution of properties, and all protein structures, native and decoy, would already vary significantly from this reference state due to having a protein-like shape.

For the Ca_score scoring function, there is only one type of triplet or doublet which makes the use of the randomization method described above infeasible. In this scoring measure, for each protein, a 1 Å grid is overlaid onto the protein structure. The protein's shape is reproduced by those grid points whose x, y and z coordinates of a grid point fall between the x, y and z coordinates of any two Cα's of the protein respectively. Then the same number of Cα atoms of a protein are randomly distributed onto those grid points that cover the shape of the protein with a minimum distance of 1.5 Å between any two Cα atoms. This distribution generates a pseudo protein corresponding to each protein structure and is used as the reference state for that protein.

3.2.5 Smoothed Potential

smthd_Phys_2b and _3b potentials were generated by smoothing Phys_2b and _3b using a cubic spline. Every other bin was been used as a knot and the fitted cubic spline was then used to calculate the values for the other bins.

3.2.6 Scoring

The total scores result from the summation of sub-scores corresponding to all individual pairs or triplets in a protein: Matrices (C_{ij}^{ABC}) or vectors (C_i^{AB}) are constructed for each protein by counting the number of observations for each triplet ABC or pair AB in the distance interval corresponding to bin *ij* or bin *i*, respectively. The sub-score for each triplet ABC (S_{ABC}) or pair (S_{AB}) is then calculated using the following formula:

$$S_{ABC} = \sum_{i,j} \left(\frac{C_{ij}^{ABC} \cdot B_{ij}^{ABC}}{d_i^2 \cdot d_j^2}\right) \qquad \text{Eq. 8}$$

and

$$S_{AB} = \sum_{i} \left(\frac{C_i^{AB} \cdot B_i^{AB}}{d_i^2}\right) \qquad \text{Eq. 9}$$

d_i is the AB distance and d_j is the BC distance in angstroms. Division by $d_i^2 \cdot d_j^2$ and d_i^2 normalizes the frequency of observing interacting bodies with respect to their distance from the central body of the triplet. The total quasi-three-body and two-body scores are then calculated by summing over all quasi-three-body or two-body sub-scores.

3.2.7 Other scoring functions used for comparison

To evaluate the performance of our statistical scoring functions for identifying the native protein structure, the following existing scoring functions and some simplistic counting methods were used for comparison:

3.2.7.1 Simple Counting Methods

Two dominant interaction types are often considered to be main forces for the stability of proteins: the hydrophobic effect and hydrogen bonding [63]. For comparison with our statistical scoring function, these two underlying forces are represented in two very simplistic counting methods to differentiate native structures from decoys. The number of hydrophobic atoms within 5 Å distance of each other (count_Phob_score), and the number of hydrogen bonds formed (count_H_score), were considered. The final simple counting scoring function measures the compactness of the protein by counting the number of Cα's within 5 Å distance of each other Cα atom (count_Ca_score).

3.2.7.2 Conventional Scoring Functions

Four widely used scoring functions, DFIRE2, dDFIRE, GOAP, FoldX and Rosetta (called conventional scoring functions in this paper) are tested for comparison. Details of these scoring functions is as follows and more can be found in the cited references:

FoldX: FoldX uses an empirical scoring function that calculates the free energy by linear combination of several empirical terms describing various energetic contributions to the stability of protein structures (e.g. van der Waals energy, hydrogen bond energy etc.):

$$\Delta G = a.\Delta G_{vdw} + b.\Delta G_{solvH} + c.\Delta G_{solvP} + d.\Delta G_{wb} + e.\Delta G_{Hbond} + f.\Delta G_{el} + g.\Delta Gk_{on} + h.T\Delta S_{mc} + k.T\Delta S_{sc} + l.\Delta G_{clash} \quad \text{Eq. 10}$$

in which a, b,…,l are relative weights of different energies and T is temperature. ΔG_{vdw} represents van der Waals interactions and is calculated based on experimental data of vaporizing amino acids from water. ΔG_{solvH} and ΔG_{solvP} represent desolvation energies of hydrophobic and polar groups respectively and are calculated based on experimental data

on transferring amino acids from aqueous to organic solvents. ΔG_{wb} represents the energy of water molecules forming more than two hydrogen bonds with the protein. ΔG_{Hbond} represents hydrogen bonding energies and is computed based on data resulted from engineered double mutant cycles. ΔG_{el} is the electrostatic interaction energy and is computed using Coulomb's law. ΔGk_{on} is an additional electrostatic component between atoms of different polypeptide chains. ΔS_{mc} and ΔS_{sc} are entropic penalties for restraining the backbone and side chains in a certain conformation and is calculated based on results of statistical analyses on protein structures. ΔG_{clash} is a measure of the energy penalty associated with steric clashes between different atoms.

FoldX can be used to investigate the destabilizing/stabilizing effects of point mutations on protein structure. The executable of FoldX (version 6.0) was downloaded from foldx.crg.es.

Rosetta: Rosetta scoring function includes a combination of statistical and physical scoring terms. The terms of the scoring function include residue solvation, residue pair interactions, strand-pairing, arrangement of strands into sheets, helix packing, radius of gyration, C_β density which is related to solvation, steric repulsion, preferred torsions in the Ramachandran map, Lennard-Jones interactions, hydrogen-bonding, solvation, electrostatic and disulfide interactions of various residues, energies of different rotamer states, and unfolded state reference energy. Details on these terms in Rosetta can be found in the cited references. We used Mini-Rosetta 3.3 downloaded from rosettacommons.org

DFIRE and dDFIRE: DFIRE potential stands for Distance-scaled Finite-Ideal gas Reference potential and is a statistical energy function based on distances observed

between pairs of atom types in known protein structures. The atom types are residue specific which resulted in a total of 167 atom types. The pair energy is calculated using the following equation:

$$\bar{u}^{DFIRE}(rij) = \begin{cases} -RT \ln \dfrac{N_{obs}(i,j,r)}{(\frac{r}{r_{cut}})^{\alpha}(\frac{\Delta r}{\Delta r_{cut}})N_{obs}(i,j,r_{cut})}, & r < r_{cut} \\ 0, & r \geq r_{cut} \end{cases} \quad \text{Eq. 11}$$

in which R is the gas constant, T is temperature (300 K), α equals 1.61, N_{obs}(i,j,r) is the number of (i,j) pairs within the sphere with radius r observed in the structure database, r_{cut} is 14.5 Å, and Δr (Δr_{cut}) is the bin width at r (r_{cut}).

dDFIRE potential stands for dipolar DFIRE. The difference between DFIRE and dDFIRE is that the latter takes the angles between interacting dipoles into consideration thus accounting for dipole-dipole interactions.

The executables were downloaded from sparks.informatics.iupui.edu/yueyang/download/index.php?Download=dDFIRE1.1-bin.tbz and sparks.informatics.iupui.edu/yueyang/download/index.php?Download=DFIRE2.1-bin.tbz.

GOAP: A plane is associated with each heavy atom defined by the heavy atom and its two neighbor bonded heavy atom. A local coordinate system ($\vartheta_x, \vartheta_y, \vartheta_z$) is defined based on this plane. Two polar angles ψ and θ and a torsional angle χ are defined based on this

coordinate system (for details look at the cited reference).The GOAP potential, then, is defined as follows:

$$E(r_{ab},\theta_a,\psi_a,\theta_b,\psi_b,\chi) = -RT\,\frac{P^{obs}(r_{ab},\theta_a,\psi_a,\theta_b,\psi_b,\chi)}{P^{exp}(r_{ab},\theta_a,\psi_a,\theta_b,\psi_b,\chi)} \qquad \text{Eq. 12}$$

where a and b represent atom types of the two interacting partners, r_{ab} is the distance, $P^{exp}(r_{ab},\theta_a,\psi_a,\theta_b,\psi_b,\chi)$ is the probability observed in the reference state and $P^{obs}(r_{ab},\theta_a,\psi_a,\theta_b,\psi_b,\chi)$ is the probability observed in known protein structures. It should be noted that GOAP benefits from the DFIRE reference state and uses different equations for indifferent cut-offs. For details please refer to the cited reference.

3.2.8 Decoy Sets

Three different decoy sets from Decoys 'R' Us version 1.3 (dd.compbio.washington.edu) [61] were used to test the performance of the various scoring functions for differentiating native protein structures from decoys. These decoy sets differ by the type of proteins and the method employed to generate the decoys. The details of these decoy sets are as follows.

hg_structural: This set contains decoys for 29 globin proteins. For each protein, comparative modeling with all other globins in the set was performed to generate decoys for each of the proteins; hence each globin set contains 28 decoy structures in addition to the native structure. All structures were energy minimized using ENCAD 22.

vhp_mcmd: This set focuses on the thermostable domain of villin (1vii). 6255 structures were selected from snapshots of five 100 ns MD simulations, four of them producing

decoy sets and one of them which is based on the X-ray structure generating native-like structures. The decoy trajectories were generated starting from conformations obtained from a coarse-grained MC simulation. All the structures were energy minimized with MM/GBSA using CHARMM. The set contains 1251 native and 5004 decoy structures.

fisa: This set is generated from four small alpha-helical proteins (1fc2, 1hdd-C, 2cro, and 4icb). The main chains for the decoys were modeled by fragment-insertion simulated annealing and Bayesian scoring functions based on fragments from proteins with similar local sequences. Then the SCWRL software package was used to model the side chains. All the structures were energy minimized using CHARMM22b. 500 decoys for each of these four proteins were generated (a single file ackcalb11-min.pdb related to 4icb was missing so one of the sets has 499 decoys).

Considering the differences between these decoy sets, different strategies were used to calculate the area under the curve (AUC) in their corresponding scoring experiments (see section 3.2). Unlike vhp_mcmd which has only one sub-set (1vii), hg_structural and fisa have 29 and 4 sub-sets respectively with only one native structure in each sub-set. Hence, while the number of native structures found was used for calculating the AUC for vhp_mcmd, we used the number of sub-sets with identified native structure to calculate the AUCs for fisa and hg_structural. For hg_structural, identification of a structure with RMSD less than 2 Å with the native structure was considered equivalent to identification of the native structure.

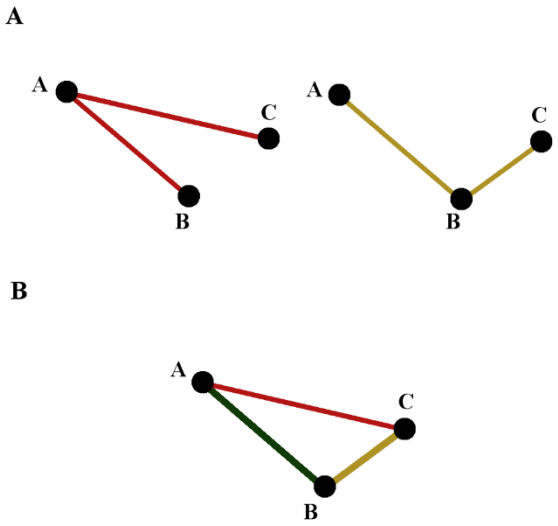

Figure 8 A) ABC triplet (AB,BC) with center B (yellow) and BAC triplet (BA, AC) with center A(red). Each triplet is defined with a center and two other points, hence it comprises two distances. Changing the point considered as the center will lead to a a different triplet as the new triplet would have one common and one different distance compared to the previous triplet. B): Three distances observed in a triplet: AB (green), BC (yellow), and AC (red). By choosing a center, there will be only two distances in each triplet (Fig.1-A).

3.3 Results and Discussions

Scoring functions are named based on prefixes and suffixes introduced in the Materials and Methods section. For a brief description of the scoring functions please refer to Table 4.

3.3.1 Quasi-three-body pseudo-potentials

We first wanted to investigate if the presence of a third interaction site or body C does have any effects on the pairwise interaction of two other bodies A and B. If such an influence is not present, then AB_i (AB pair having distance corresponding to bin i) and BC_j (BC pair having distance corresponding to bin j) would be independent variables for all bins i, j. In such a case a cut of (P_{ij}^{ABC}) along a specific j, would generate a contour that reproduces the pattern of probability density (P_i^{AB}) multiplied (or scaled) by the value of P_j^{BC} for that specific bin j, i.e. $P_{ij}^{ABC} = P_j^{BC} \cdot P_i^{AB}$. If C has no influence on the interaction profile of AB, this similarity in contour should be observed for any i, j. Therefore multiple contours of (P_{ij}^{ABC}) for different j should have the same pattern with different scaling factors. This pattern should also match that of the corresponding pairwise interactions. However, observing different patterns in the contour maps and those also which differ from the corresponding pairwise pattern, would mean that the (P_i^{AB}) distribution is influenced by j (BC distance) which implies a statistical dependency of AB_i and BC_j. Dependency between AB_i and BC_j means that there is higher order information in (P_{ij}^{ABC}) not implied in either (P_i^{AB}) or (P_j^{BC}) which could be used in differentiating native from decoy structures.

Using the abovementioned strategy, we can visualize the existence of any dependency between AB_i and BC_j in (B_{ij}^{ABC}). (B_{ij}^{ABC}) and (B_i^{AB}) are calculated based on (P_{ij}^{ABC}) and (P_i^{AB}) using Eq. 6 and 7 respectively as described in detail in Materials and Methods and are the statistical potentials used in our scoring functions (see Eq. 8 and 9). Figures 9 and 10 are graphs of (B_{ij}^{ABC}) and (B_i^{AB}) for a number of representative triplets ((B_{ij}^{ABC}) graphs for all of the triplets can be found in figure 12). The examples of (B_{ij}^{ABC}) shown in figure 9 represent the effects of the presence of a third body on the potential of interaction between hydrogen-bond donors (D) and acceptors (A) and the (B_{ij}^{ABC}) shown in figure 10 represent such effects on interactions between two hydrophobic (H) elements. The contours of each three-dimensional plot are also shown on each side of the graph. The (B_i^{AB}) corresponding to pairwise interactions is shown with the red line overlaid on the contours on each side of the graph. For HHX potentials (X referring to the third physicochemical elements) we see very high positive peaks at distances less than 3.5 Å (see figure 12) which can be attributed to van der Waals clashes. These peaks overwhelm the scaling of the rest of the graph which makes observation of discernible patterns difficult. In order to examine the pattern of contours in HHX potentials, the first 8 bins were ignored. The trimmed potentials were then re-plotted (figure 10). It is noteworthy that the potential from the beginning bins (~2.5-3.5 Å) dominates the whole potential for nearly all of the triplets.

Despite small fluctuations the (B_{ij}^{ABC}) graphs in figures 9 and 10 have contours that generally follow the same pattern. Also the pattern of contours is the same as the pairwise interaction potentials. This demonstrates lack or weakness of higher order information in

(B_{ij}^{ABC}) and shows the potentials have not been highly influenced by introducing the third interacting body. Lack of higher order information observed in the ADX triplets could be related to the fact that the AD interactions are dominated by backbone-backbone interactions leading to the formation of secondary structure elements of the protein, thus they are less susceptible to the presence of a third interacting body.

In addition to visual comparison of patterns in the quasi-three body and two-body distance-dependent statistical potentials, we aimed to quantify the lack of difference between those patterns. Using the underlying quasi-three body and two-body probability distribution functions, we performed Kolmogorov-Smirnov tests (K-S tests). K-S test compares a test sample with a reference sample and identifies if they originate from the same probability distribution. In our study, the null hypothesis to be tested states that the two samples, i.e. the pair-wise distribution functions and the corresponding slices of the quasi-three body function, originate from the same probability distribution. The null hypothesis is tested against a certain significance level where a typical value of 0.05 is used in this study [64,65]

In detail, all two-body contours ($\hat{P}_i^{AB}$) and ($\hat{P}_j^{AB}$) were obtained from bins 4 to 32 for each three-body probability distribution P_{ij}^{ABC}. Each of the contours was normalized. The contours were tested against their corresponding two-body distributions (P_i^{AB}) and (P_j^{BC}) obtained from an analysis of the same protein database. The first three bins were excluded from the analysis since they cover distances between 2.0 and 2.75 Å for which typically only few if any observations were made in the database. For each triplet, there are a total of 58 pattern comparisons, 29 for each of the two pairwise interactions

embedded within a triplet. Out of 126 triplets, only 9 triplets violate the null hypothesis in more than 10% of the comparisons and only two triplets in more than 20% of the comparisons (Table 5). 77 triplets do not violate the null hypothesis at all, i.e. the probability distributions of quasi-three body interactions are identical to the corresponding pair-wise interactions for all slices with a significance level of 0.05. The results show that higher order information is not established for almost 93% of all triplets. All of the triplets that violate the null hypothesis in more than 10% of all comparisons (Table 5) contain positive-negative (PN) or negative-only interactions (NNN). Whereas these results may be interpreted as engagement of charged atoms in higher order interactions, it should be noted that interaction triplets containing two charged atoms are relatively rare compared to all other triplets studied, and that the small sample size of those triplets might at least contribute to the relatively frequent violation of the null hypothesis.

Figure 11 shows the statistical potential for triplet APN that displays the most significant higher order interactions based on the KS-test. The shallow maximum in the region AP=3.0-5.0 and PN=5.0-8.0 might represent an instance of higher order interactions in this potential map.

3.3.2 Quasi-three-body scoring functions

KS-test analysis demonstrated that there are only a few three-body potentials with significant higher order interactions. Consequently, a significant improvement in scoring performance is not expected between quasi- three-body distance-dependent potentials and their two body counterparts. The following study was designed to support this argument in a practical application setting. We constructed statistical potentials to test whether or

not the influence of a third body on the interaction profile between two particles would improve the performance of the potential in its ability to discriminate native-like structures from decoy structures. The graphs resulted from using various scoring functions tested for vhp_mcmd, hg_structural and fisa decoy sets can be found in Supplementary Material (figures 13, 14 and 15). The area under the curve (AUC) (ranging 0 to 1) of these graphs are plotted in figure 16 and can be used for comparison between different scoring functions.

In general the scoring functions developed in this study are very successful in identifying native structures from decoys. Our scoring functions perform perfectly on the fisa decoy set displaying highest AUC (equal to the ideal scoring function). Also these scoring functions have very good performances which are comparable to or better than the conventional scoring functions for vhp_mcmd and hg_structural decoys sets. This observation is important as it supports our idea of using protein structure prediction as a practical test case for comparing quasi-three body and pairwise atomistic statistical potentials. Although Ca_2b_score and Ca_3b_scores are not as successful as the rest of our pairwise and three body scoring function for two of the decoy sets, it is hard to identify one representation of the interacting bodies which always leads to superior performance. Also there is not a significant difference between scoring performance of smthd_Phys_score and Phys_score and they almost overlap. Simple counting methods have good performances in fisa and show better or comparable results compared to conventional scoring functions.

The general linear correlation in figure 16-A implies similar scoring performances of three-body and pairwise functions. In fact except one case (Ca_3b_score for fisa), we do

not observe a significant improvement or deterioration in scoring performance by switching between pairwise and three-body functions.

Examples of approximate time needed for calculations of two main steps for pairwise and quasi-three body scores are shown in table 6. The first step is populating (C_{ij}^{ABC}) or (C_{ij}^{AB}) and normalizing them with distance squared (d^2). The second step is multiplication of distance-normalized (C_{ij}^{ABC}) by (B_{ij}^{ABC}) to calculate S_{ABC} and S_{AB} (equations 8 and 9). Although pairwise scoring takes less time, quasi-three-body scoring is still extremely fast. For instance, the total time needed for qusi-three-body scoring of the largest protein tested (153 amino acids) is less than 4 sec.

3.3.3 Correlations between different scoring functions

In order to investigate potential correlations between the various scoring functions tested in this study, we calculated the Pearson correlation coefficient between scores obtained from the various scoring functions for each subset in all decoy sets. Specifically, a high correlation between two-body and quasi-three-body scores can be additional evidence for the lack of higher order information in quasi-three-body potentials. Figure 17 graphically shows correlations for selected subsets of various decoy sets. The correlation heat maps for all subsets of all decoy sets can be found in the Supplementary Material (Figure 18). Figure 17 shows correlation between the scoring functions for the vhp_mcmd decoy set (represented by 1vii). In general, all of the corresponding two-body and quasi-three-body scoring functions show very high correlations (> 0.8) with each other. Excluding the Ca_3b_score and Ca_2b_score scoring functions, the remainder of the two-body and quasi-three-body statistical scoring functions and the four studied conventional scores

(dDFIRE, DFIRE2, GOAP, Rosetta, FoldX) are highly correlated (>0.7 with DFIRE2, dDFIRE, FoldX, and >0.6 with Rosetta, GOAP). Correlation coefficients between various scoring functions for the hg_structural decoy set (as represented by 2pgh-A subset in figure 17) follow the same general pattern as the vhp_mcmd. We again see high correlations (>0.8) between two-body and quasi-three-body scores in this decoy set. There is much less correlation among the physicochemical element-based statistical scoring functions, dDFIRE, DFIRE2, GOAP, FoldX and Rosetta for the fisa decoy set. Similar to vhp_mcmd and hg_structural decoy sets scores based on Cα's are weakly correlated with our other scores (figure 17) although they are highly correlated with each other (>0.7). The high correlation (>0.6) between two-body and quasi-three-body scores is repeated in this decoy set.

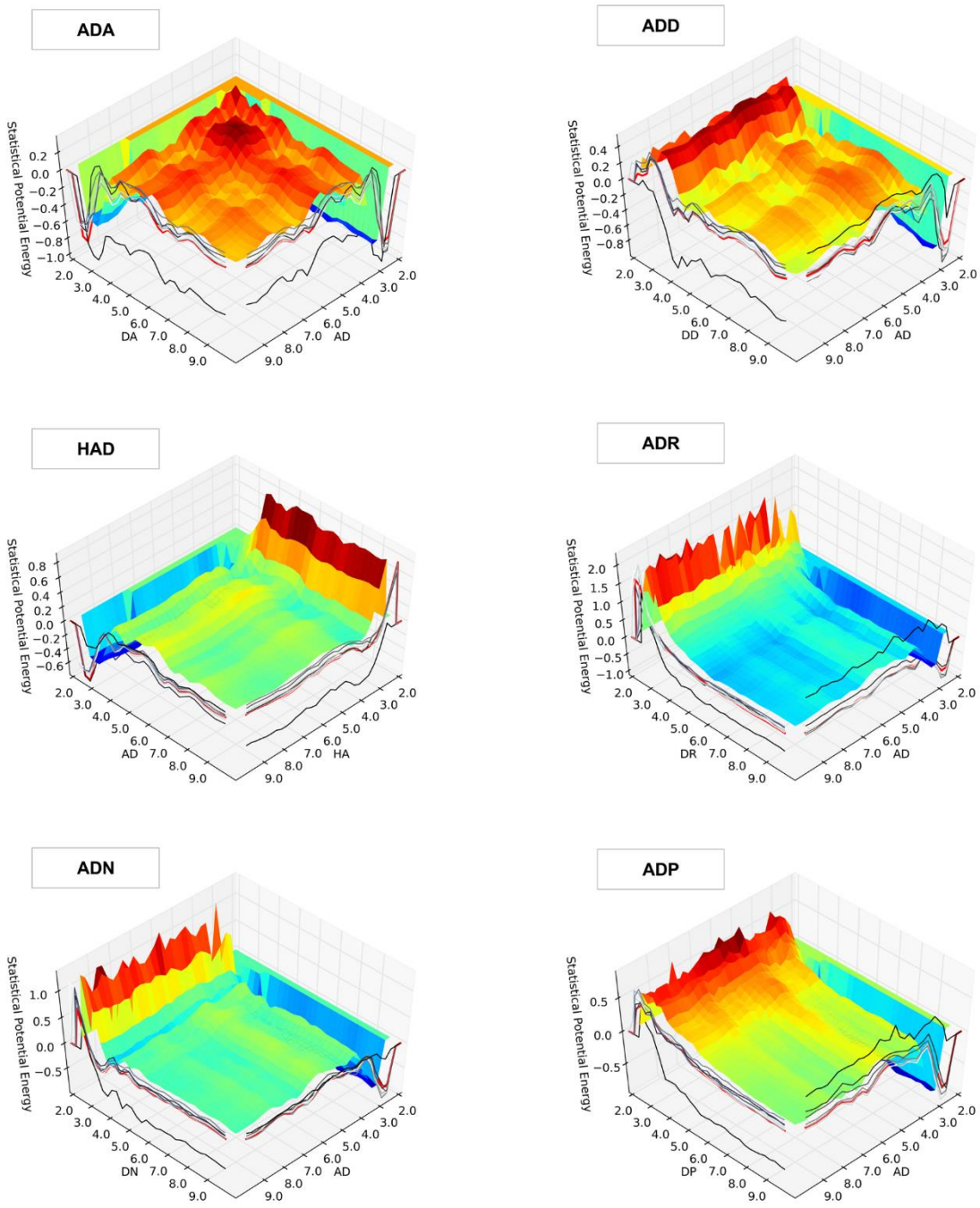

Figure 9 Graphs of (B_{ij}^{ABC}) (quasi-three-body statistical potential for interacting triplet ABC with distance bins i and j) of a number of representative triplets (indicated by the three letter code on top of each graph). Only interactions with pairwise distances between 2 to 10 Å are considered. The contours of each plot (darker colors for bins with larger distances) are shown on each side of the graph. The corresponding (B_i^{AB}) (two-body pseudo-statistical potential for interacting pair AB with distance bins i) is shown by a red line overlaid onto the contours. These quasi-three-body pseudo-potentials show the effects of the presence of a third body on the potential of interaction between hydrogen bond donor (D) and acceptor (A) elements

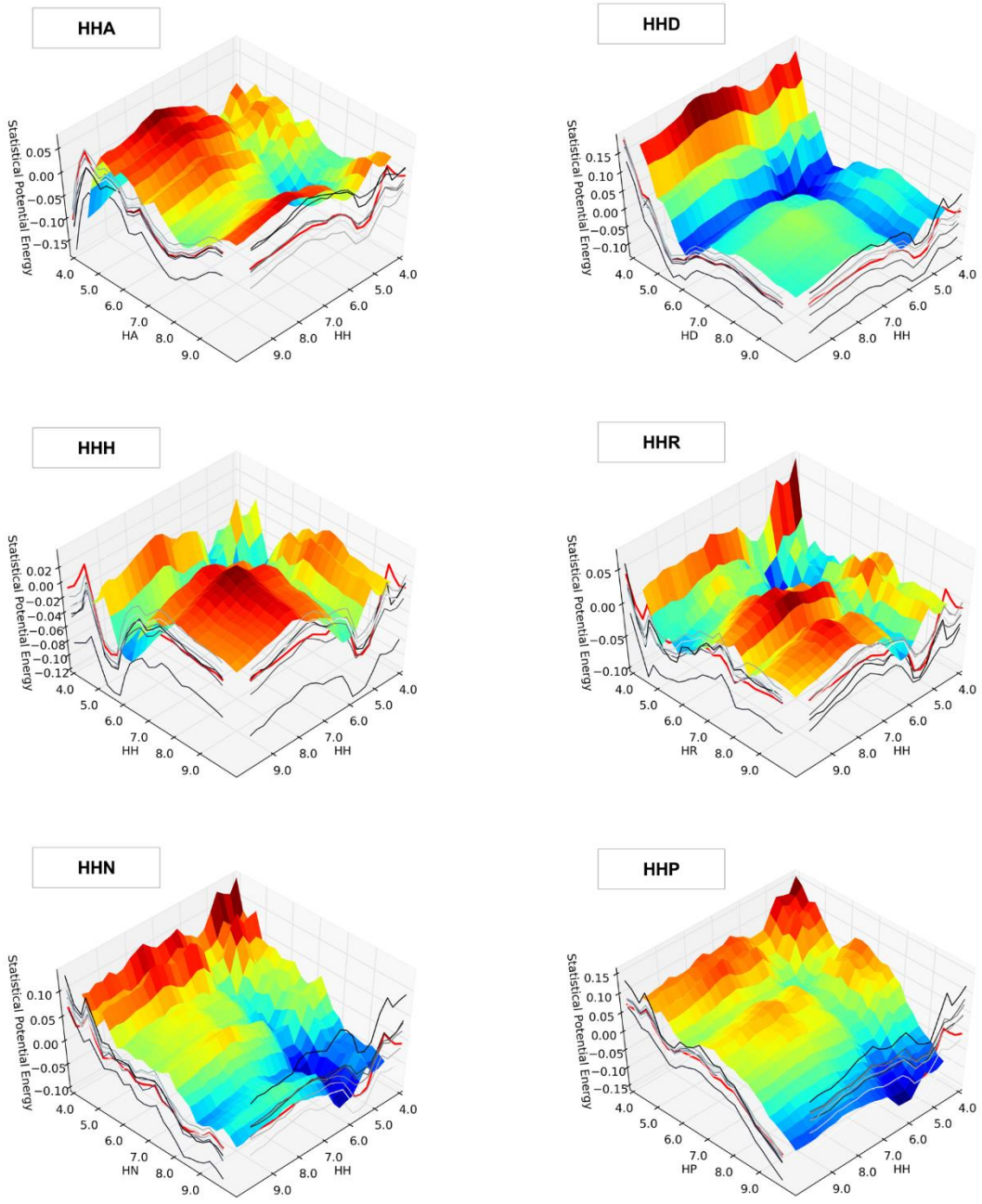

Figure 10 Graphs of (B_{ij}^{ABC}) (quasi-three-body statistical potential for interacting triplet ABC with distance bins i and j) of a number of representative triplets (indicated by the three letter code on top of each graph). Only interactions with pairwise distances between 2 to 10 Å are considered. The contours of each plot (darker colors for bins with larger distances) are shown on each side of the graph. The corresponding (B_i^{AB}) (two-body pseudo-statistical potential for interacting pair AB with distance bins i) is shown by a red line overlaid onto the contours. These quasi-three-body pseudo-potentials show the effects of the presence of a third body on the potential of interaction between two hydrophobic (H) elements.

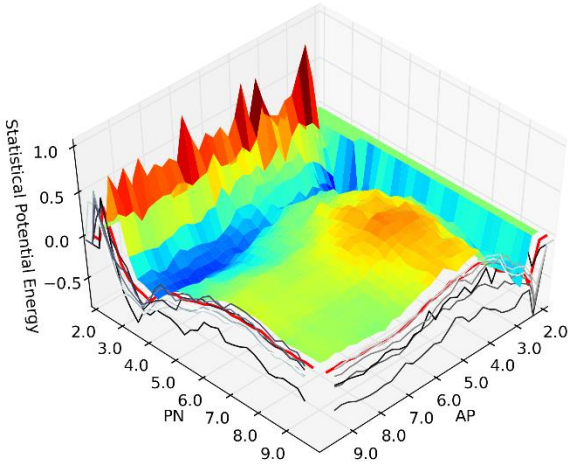

Figure 11 Graph of (B_{ij}^{APN}) (quasi-three-body statistical potential for interacting triplet APN with distance bins i and j). Only interactions with pairwise distances between 2 to 10 Å are considered. The contours of the plot (darker colors for bins with larger distances) are shown on each side of the graph. The corresponding (B_i^{AP}) and (B_i^{PN}) (two-body pseudo-statistical potential for interacting pair AP and PN respectively with distance bins i) is shown by a red line overlaid onto the contours. APN shows the most significant higher order interactions compared to pairwise interactions.

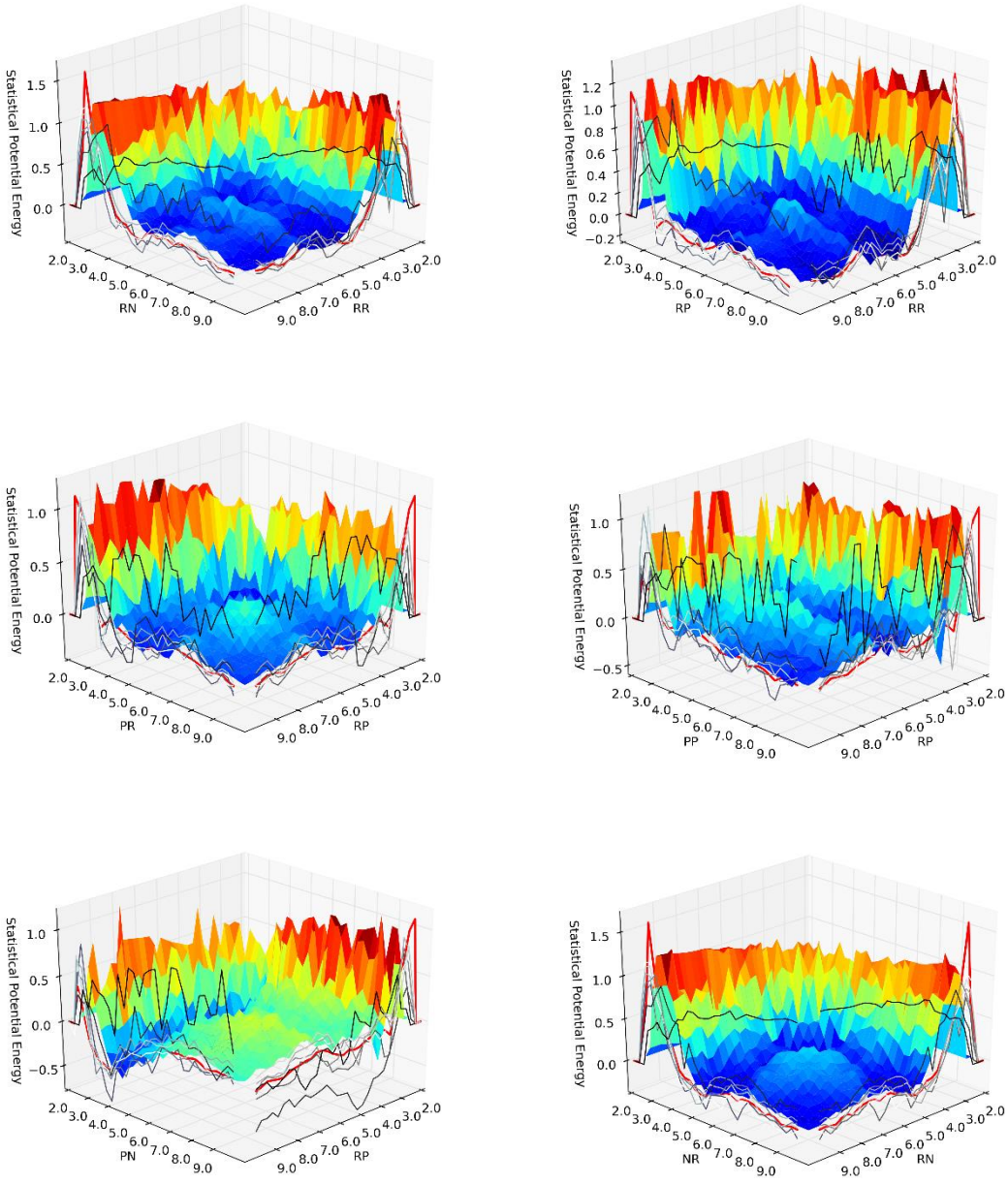

Figure 12 Graphs of (B_{ij}^{ABC}) of a number of representative triplets (indicated by the three letter code on top of each graph). Only interactions with pairwise distances between 2 to 10 A are considered. The contours of each plot (darker colors for bins with larger distances) are shown on each side of the graph. The corresponding (B_i^{AB}) (two-body pseudo-statistical potential for interacting pair AB with distance bins i) is shown by red line overlaid onto the contours.

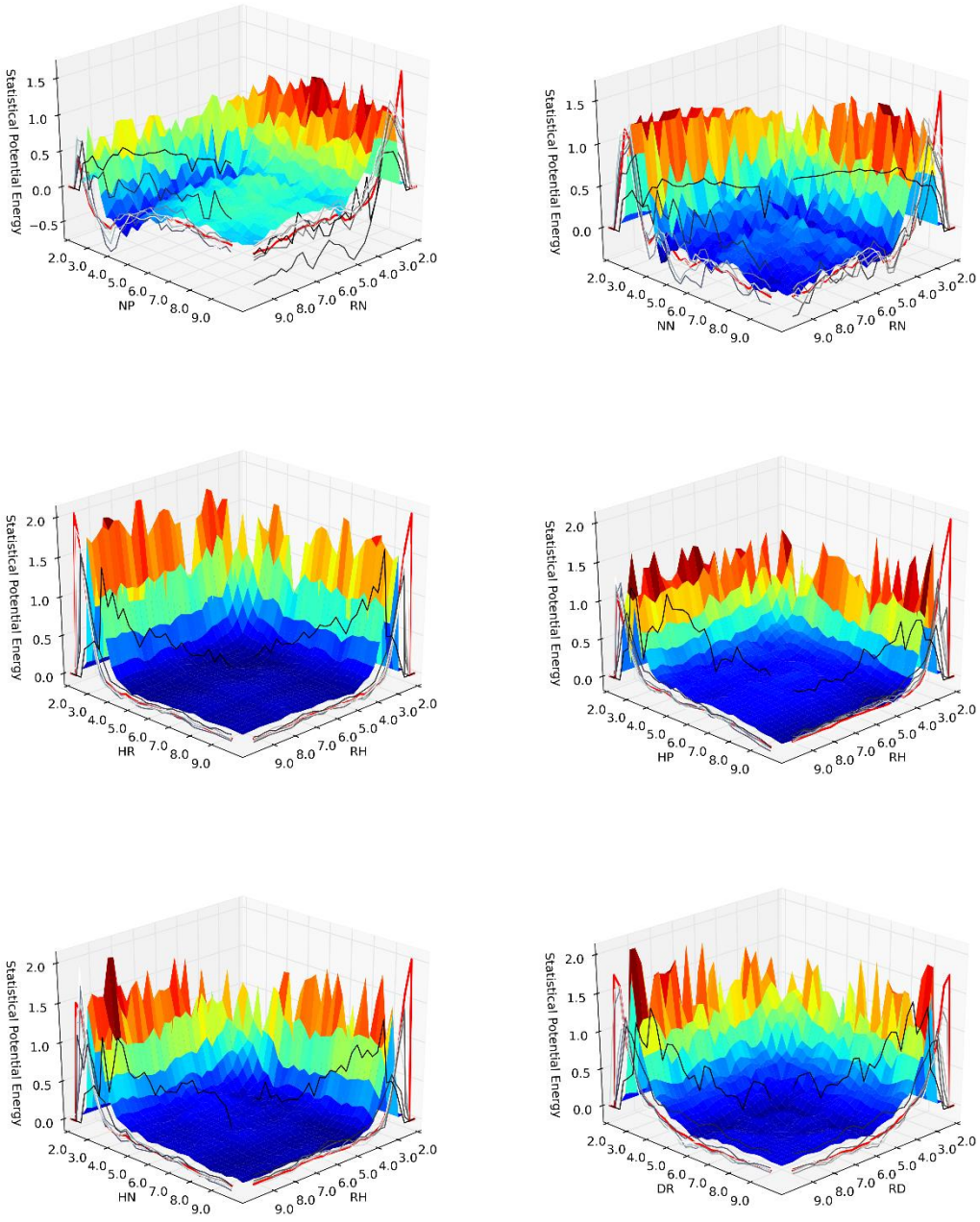

Figure 12 continued

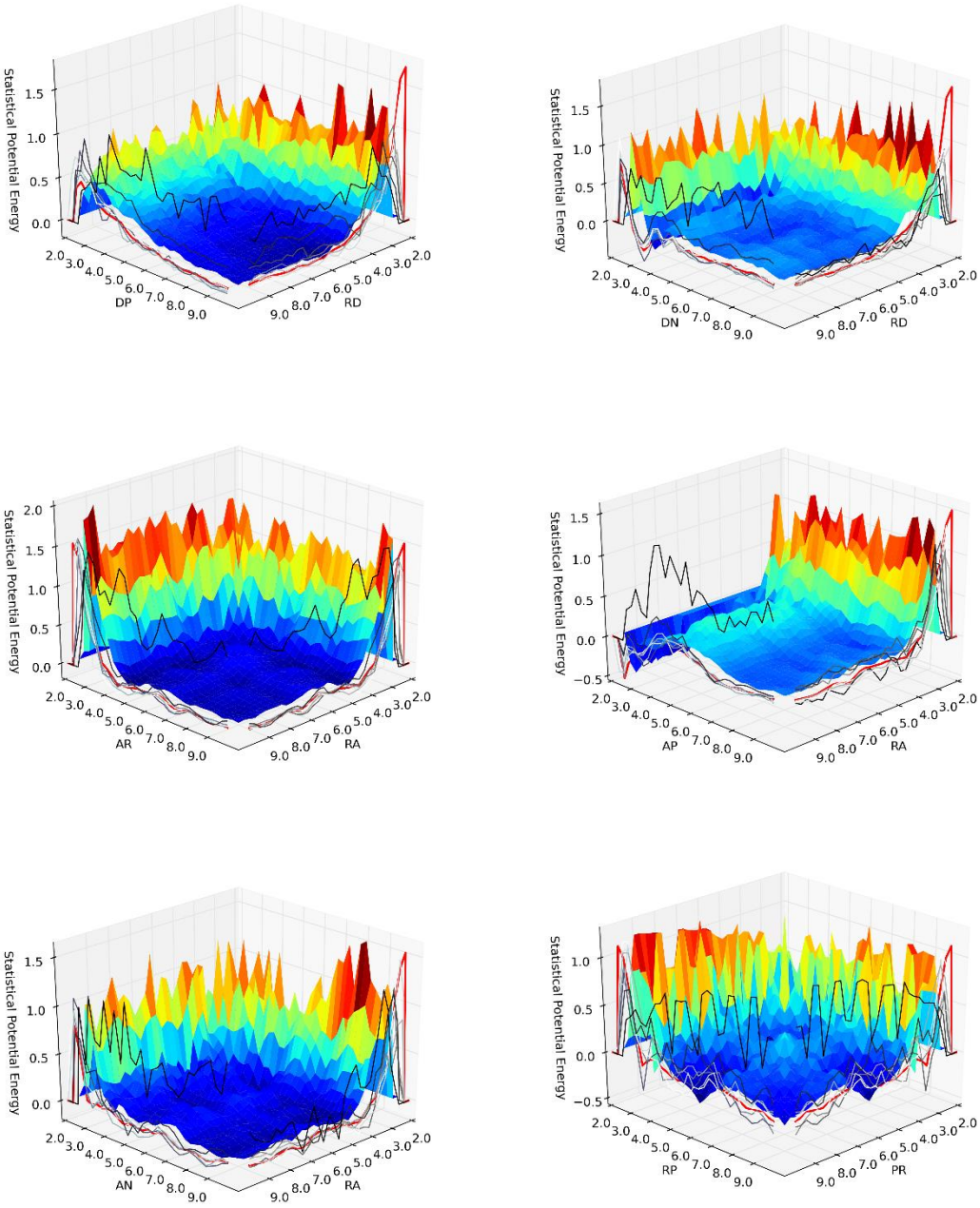

Figure 12 continued

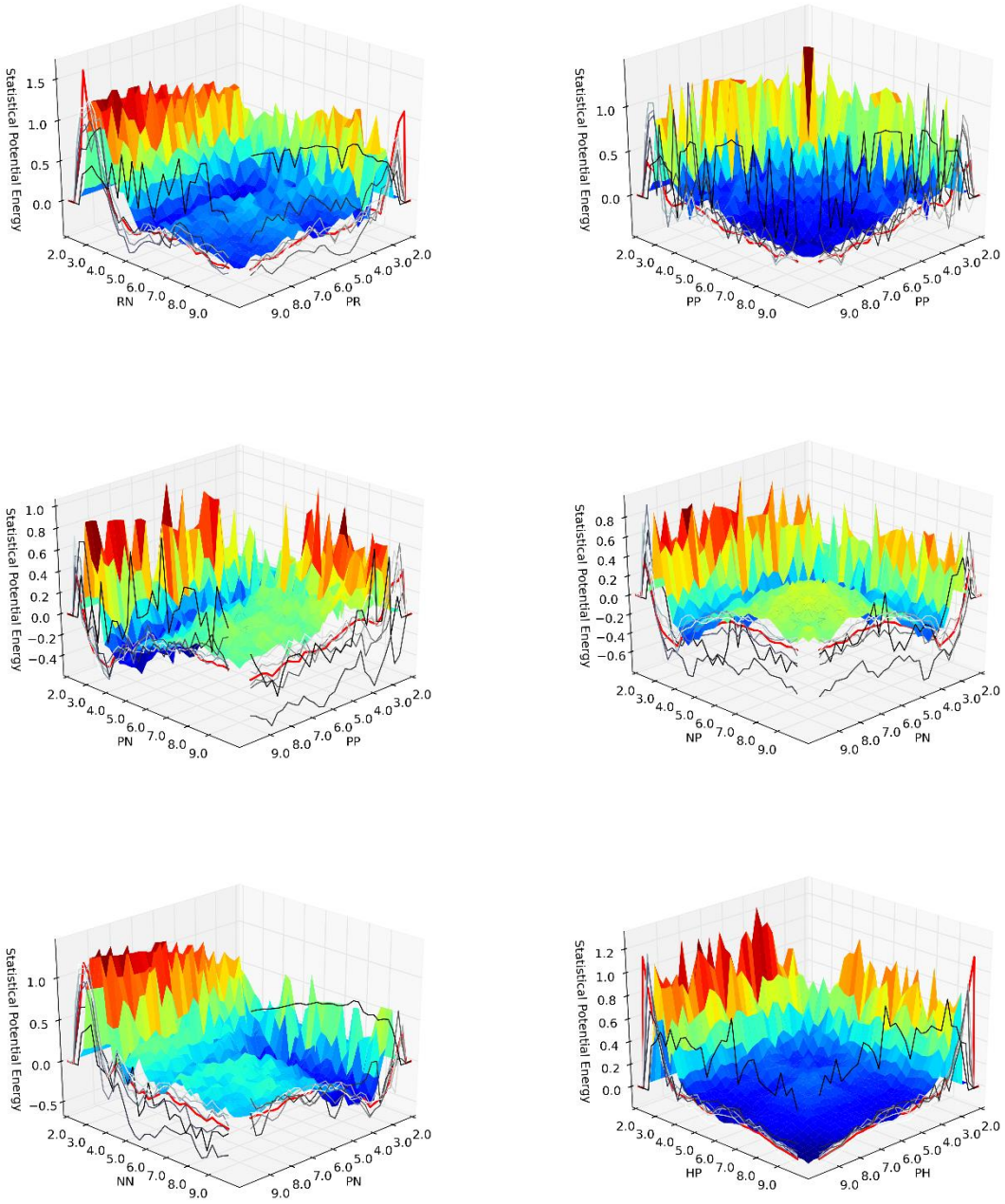

Figure 12 continued

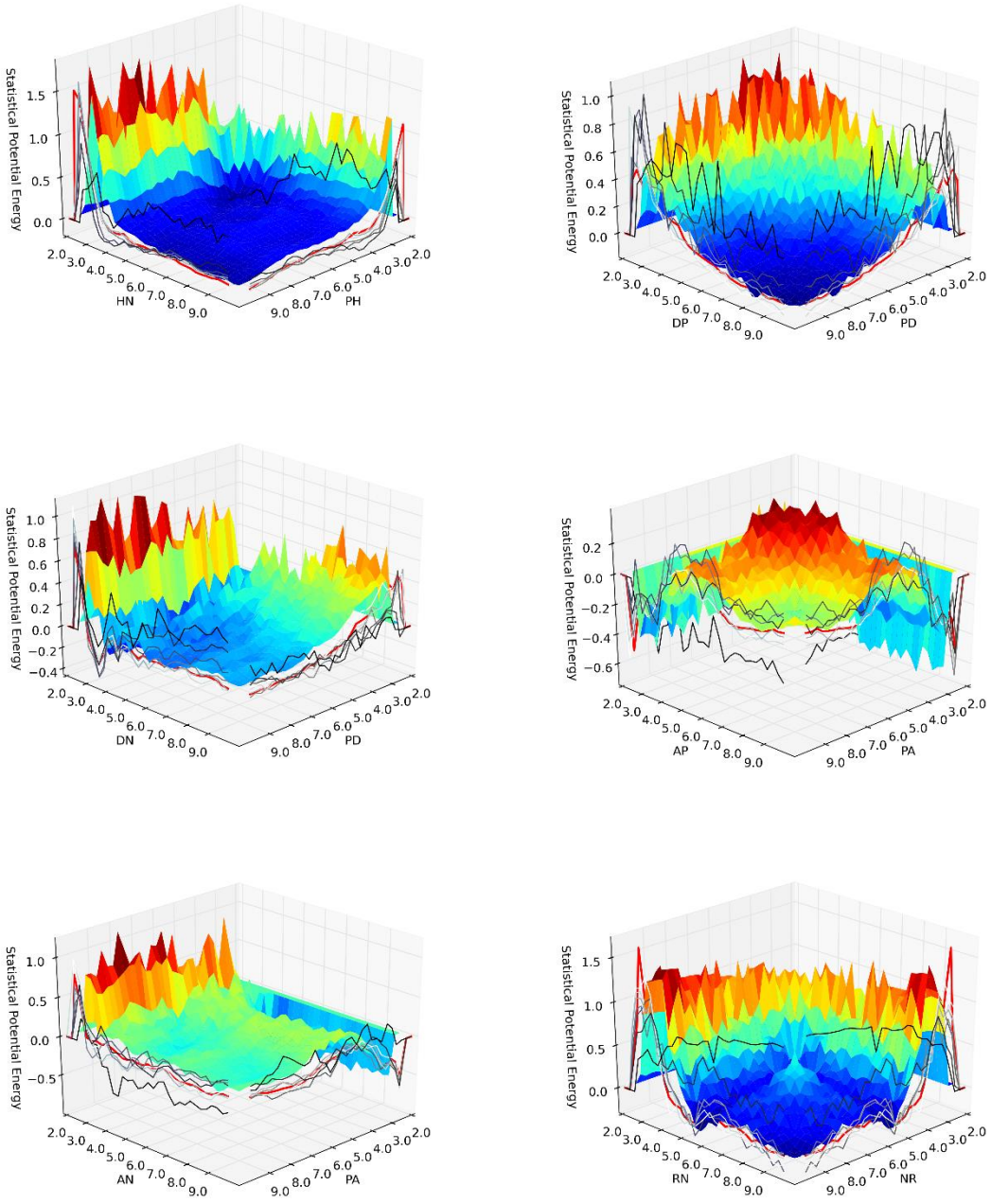

Figure 12 continued

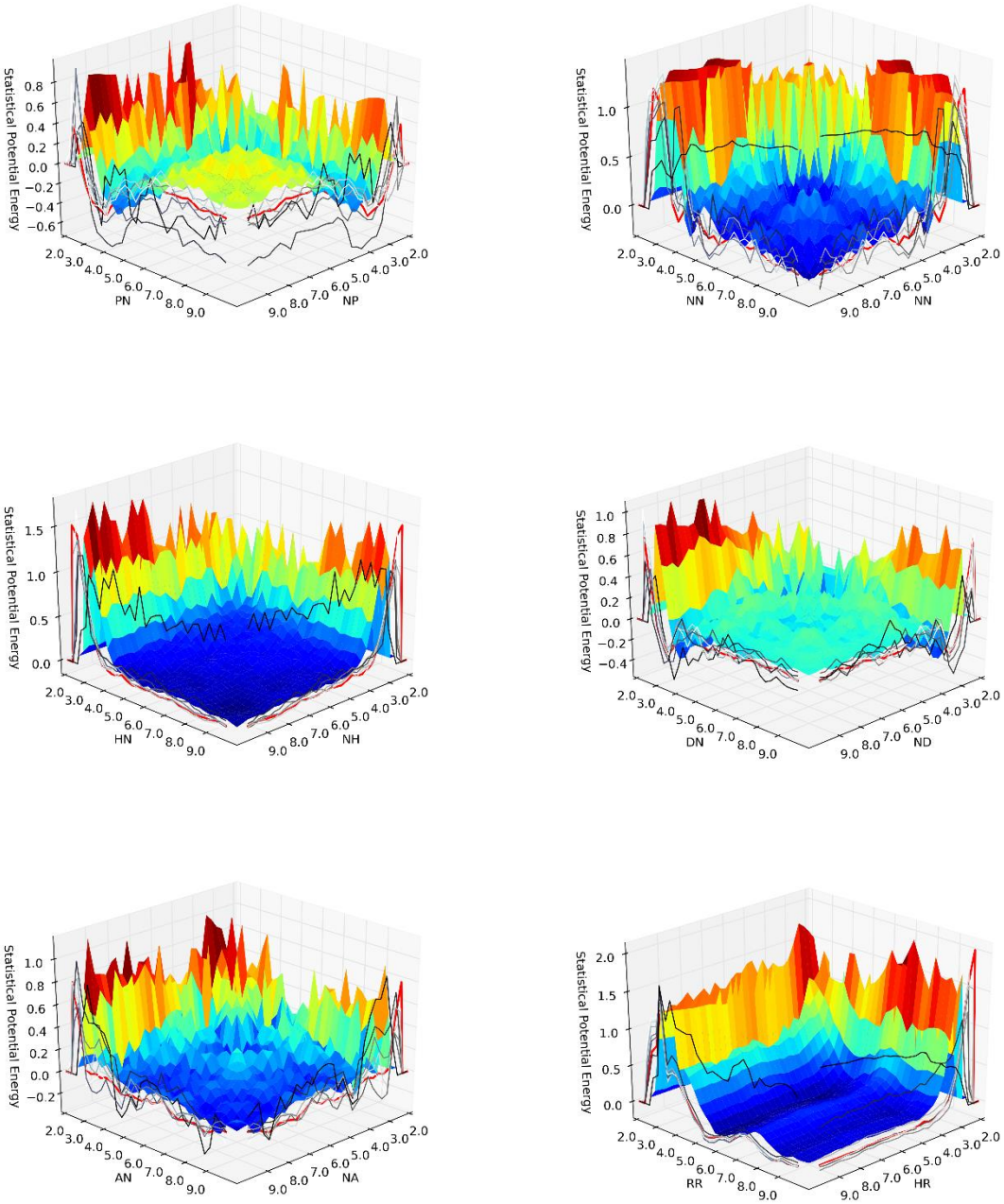

Figure 12 continued

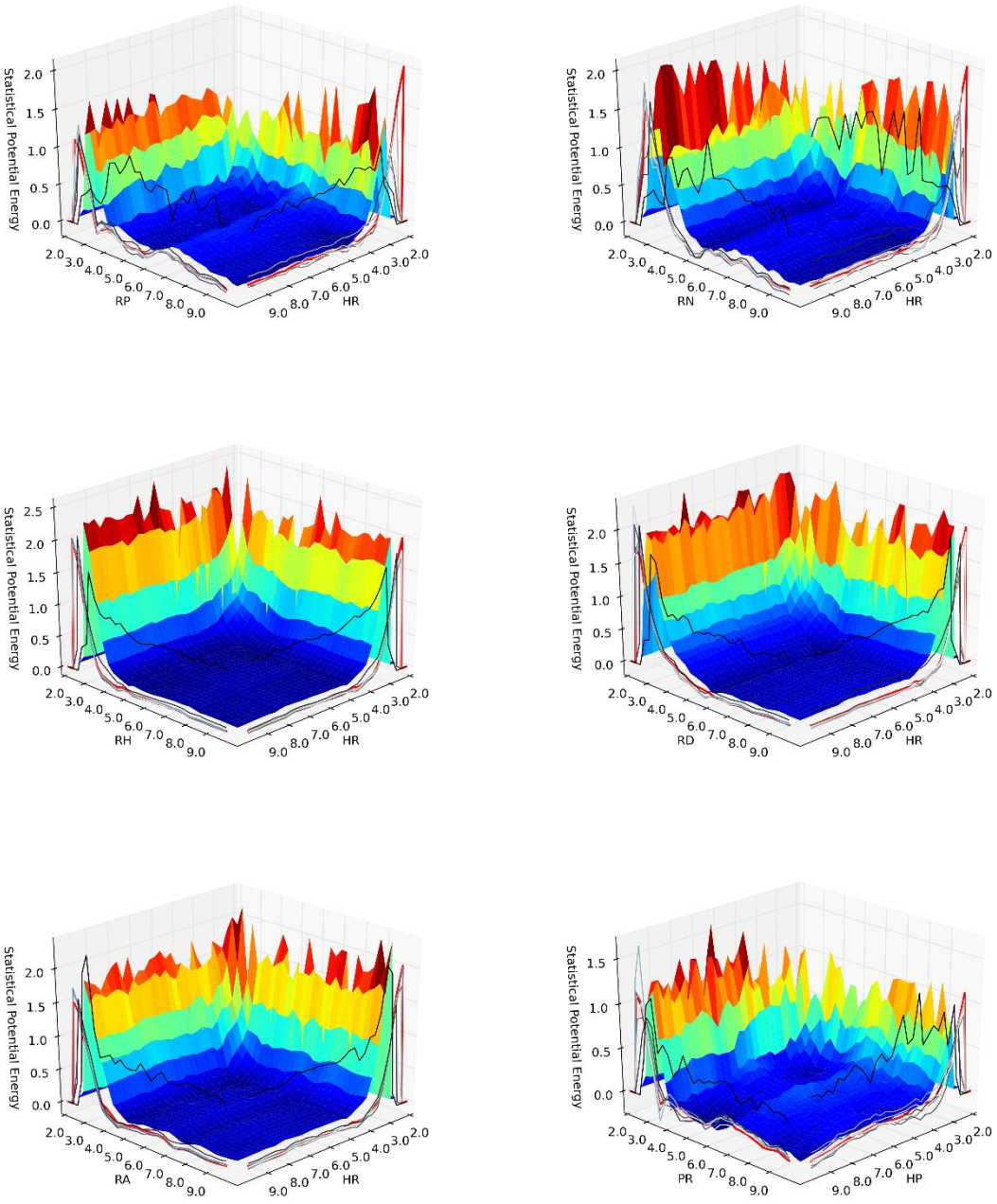

Figure 12 continued

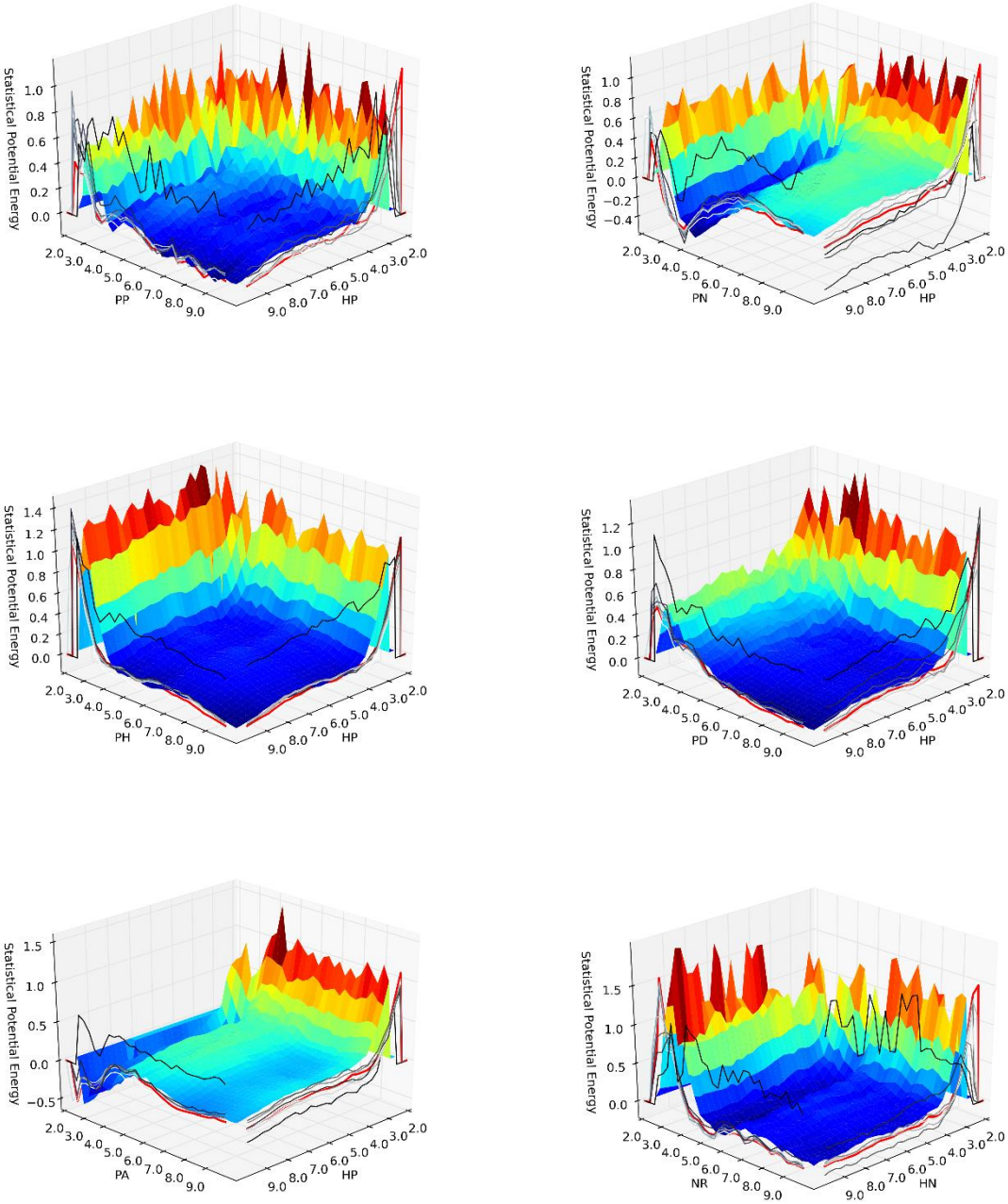

Figure 12 continued

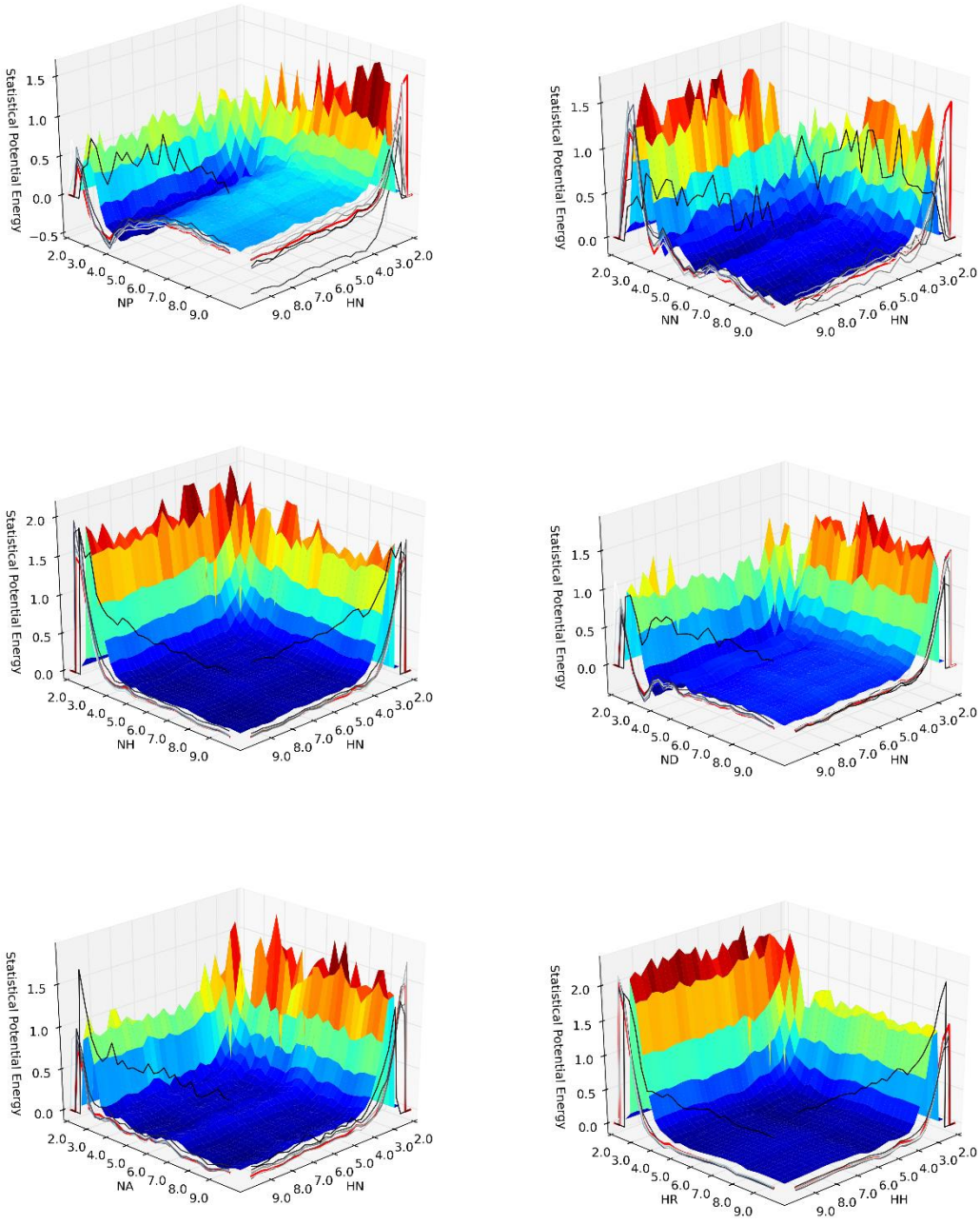

Figure 12 continued

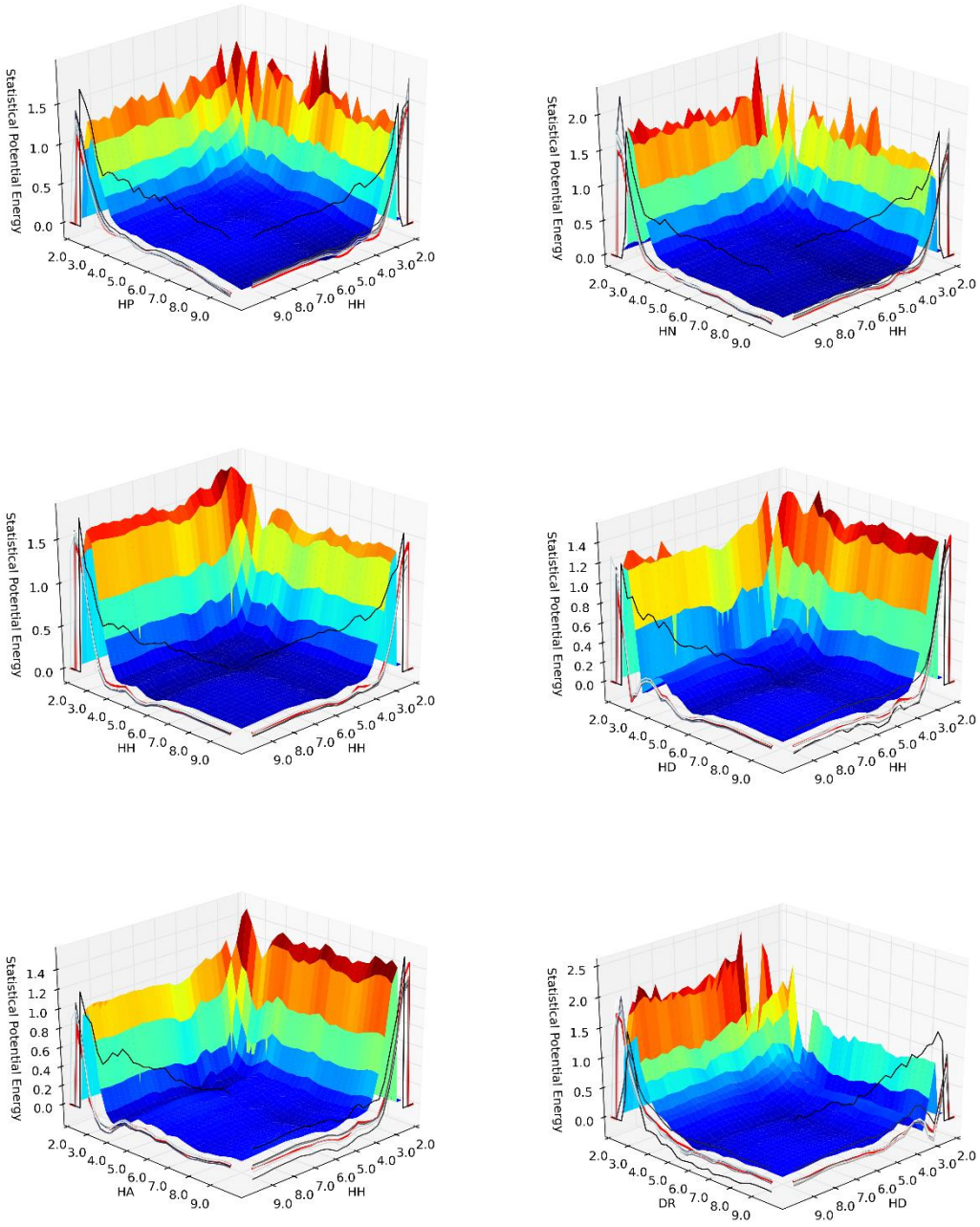

Figure 12 continued

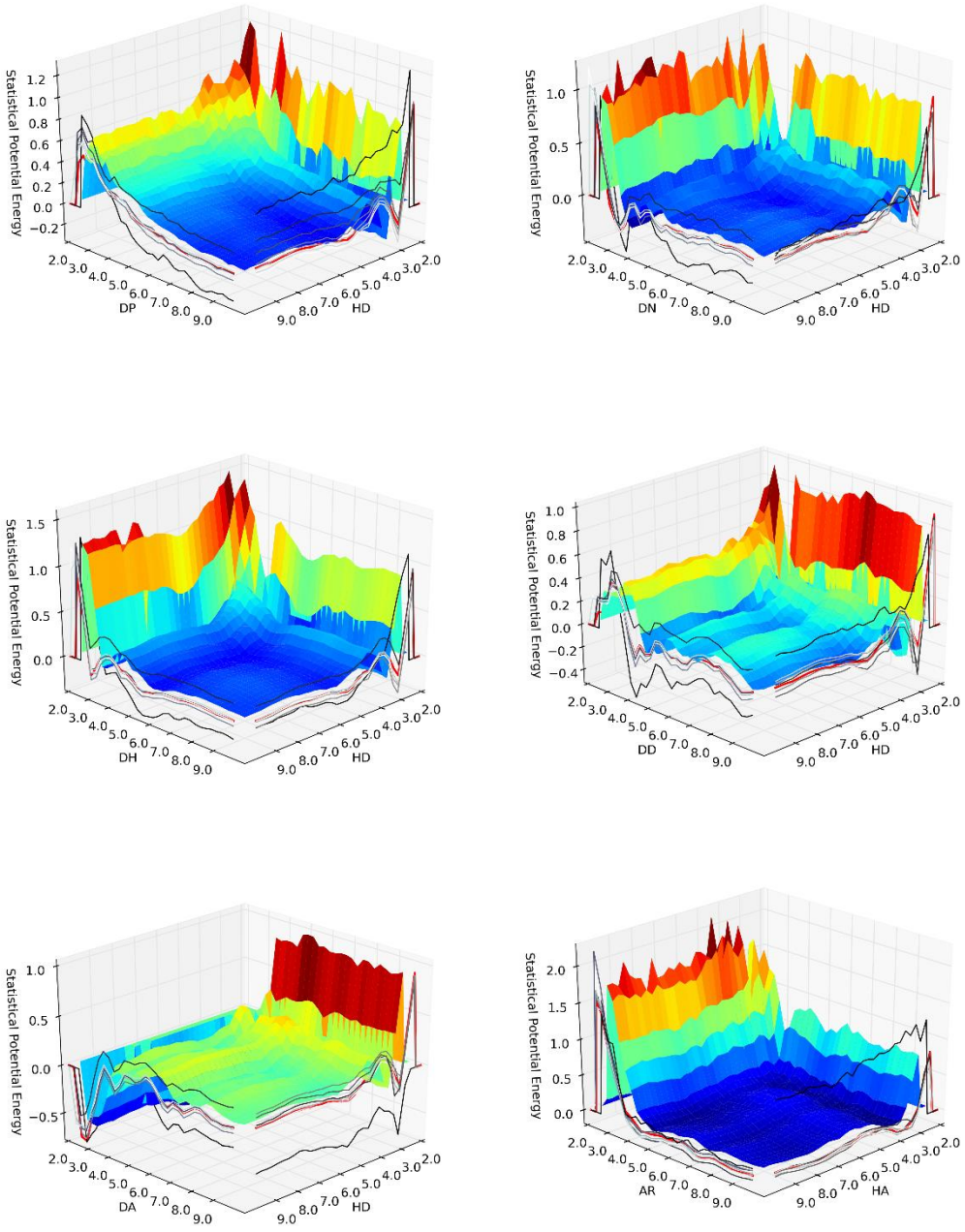

Figure 12 continued

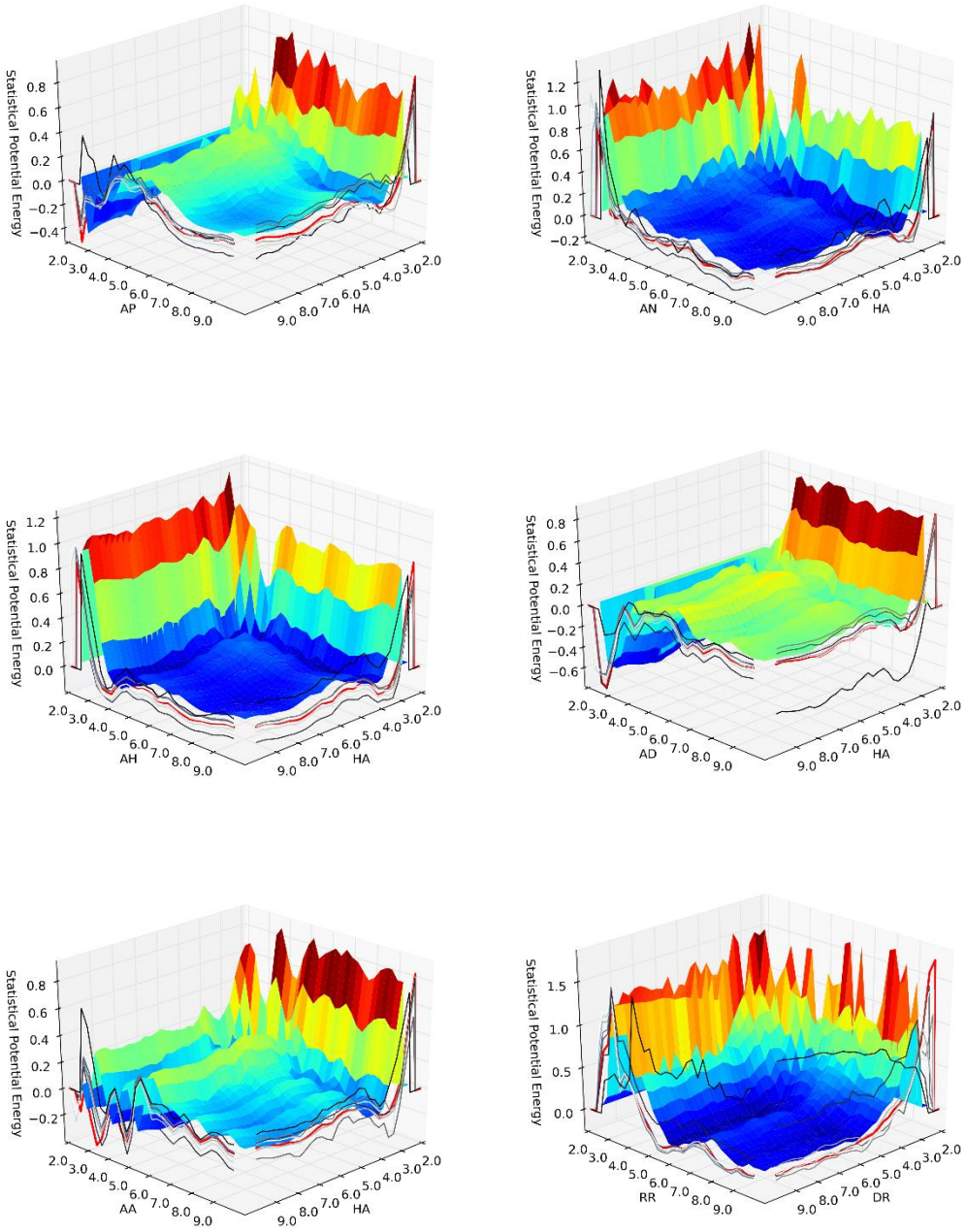

Figure 12 continued

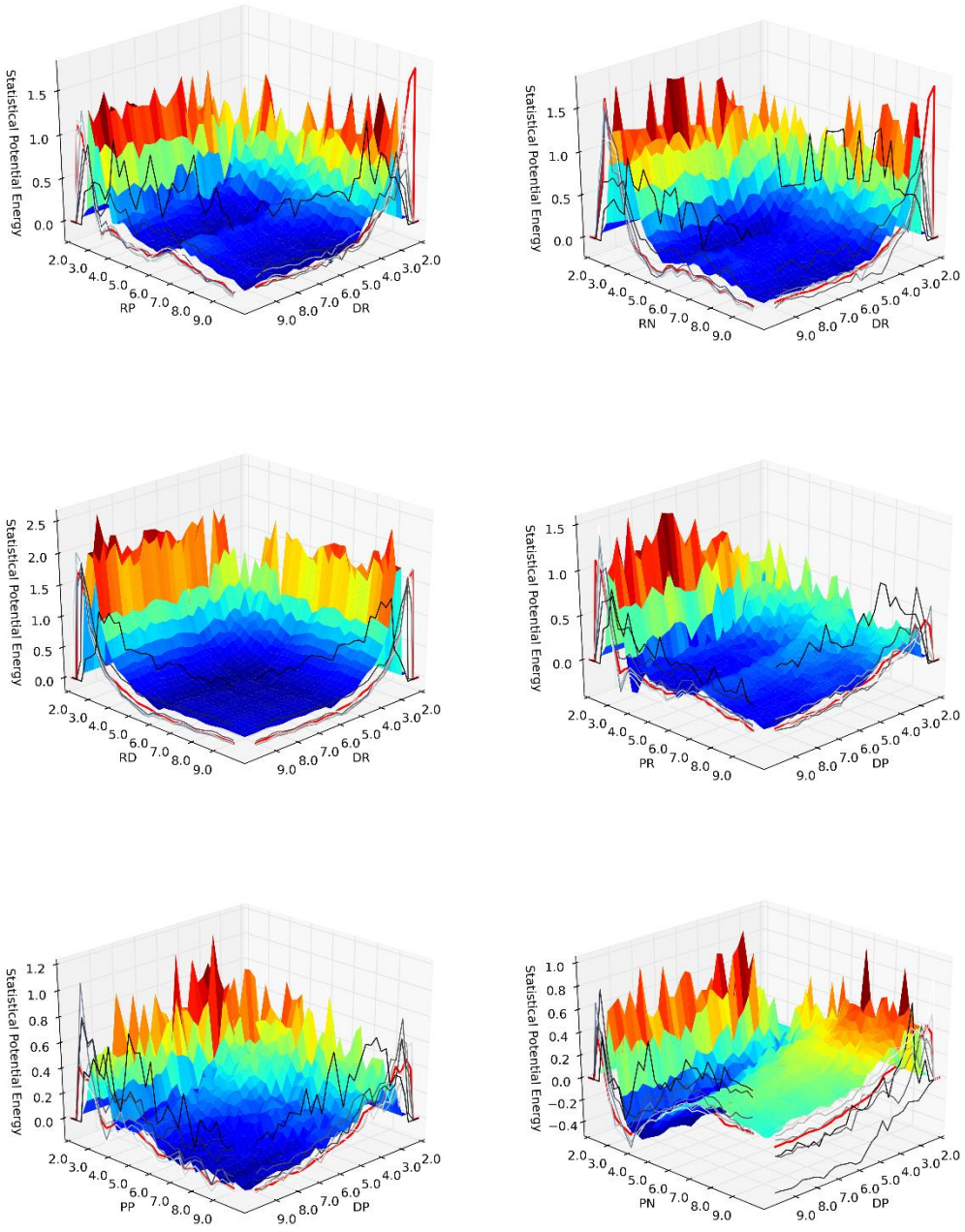

Figure 12 continued

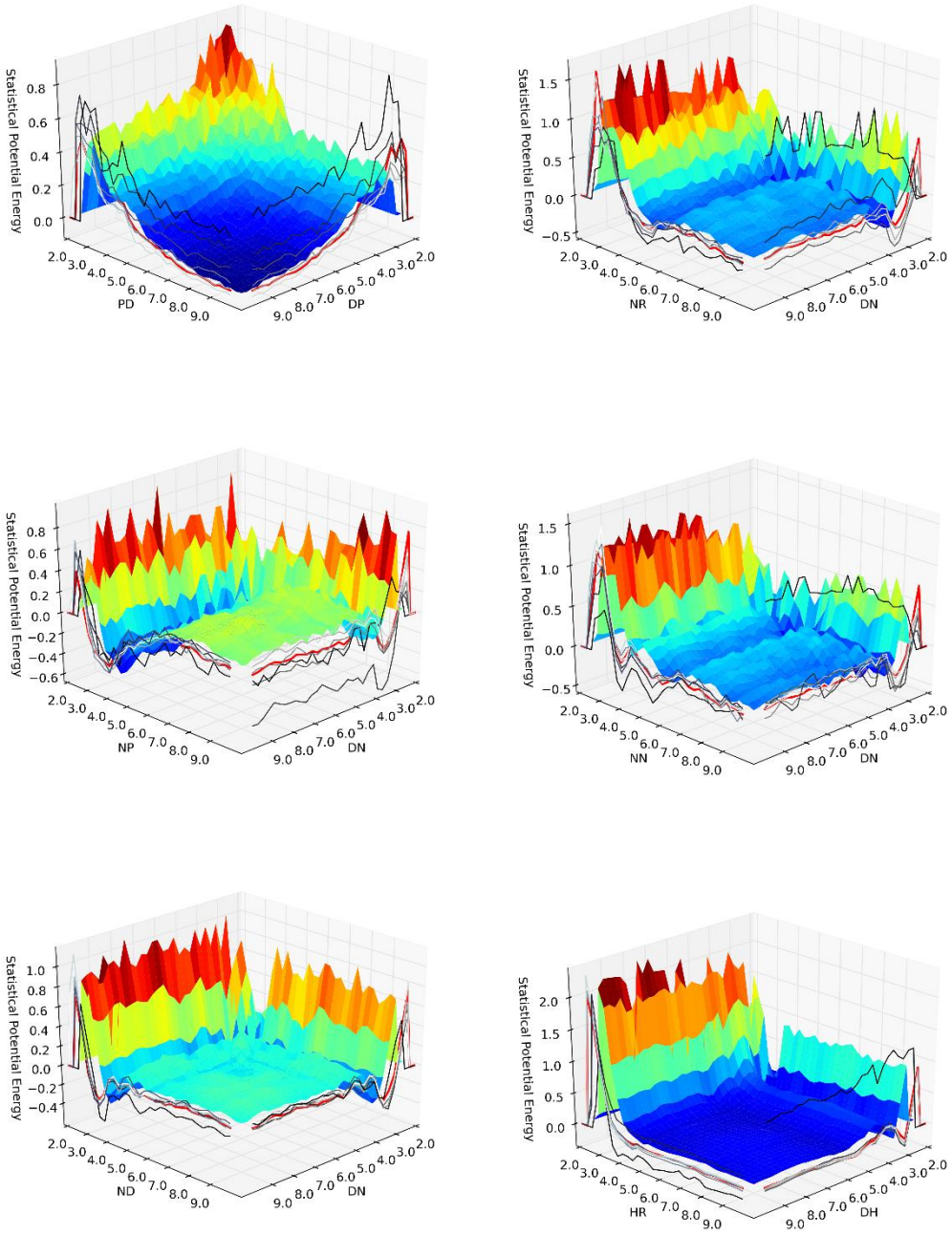

Figure 12 continued

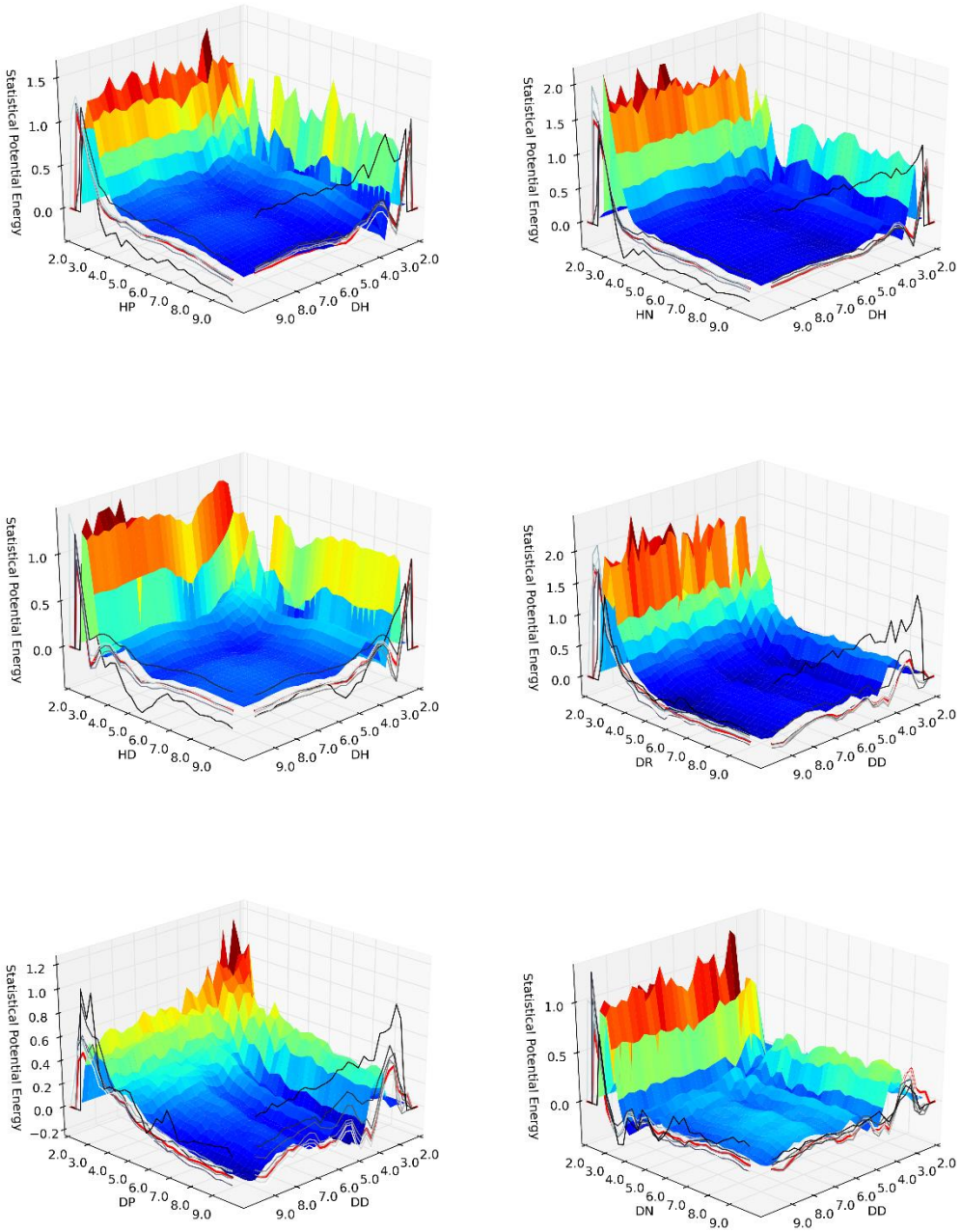

Figure 12 continued

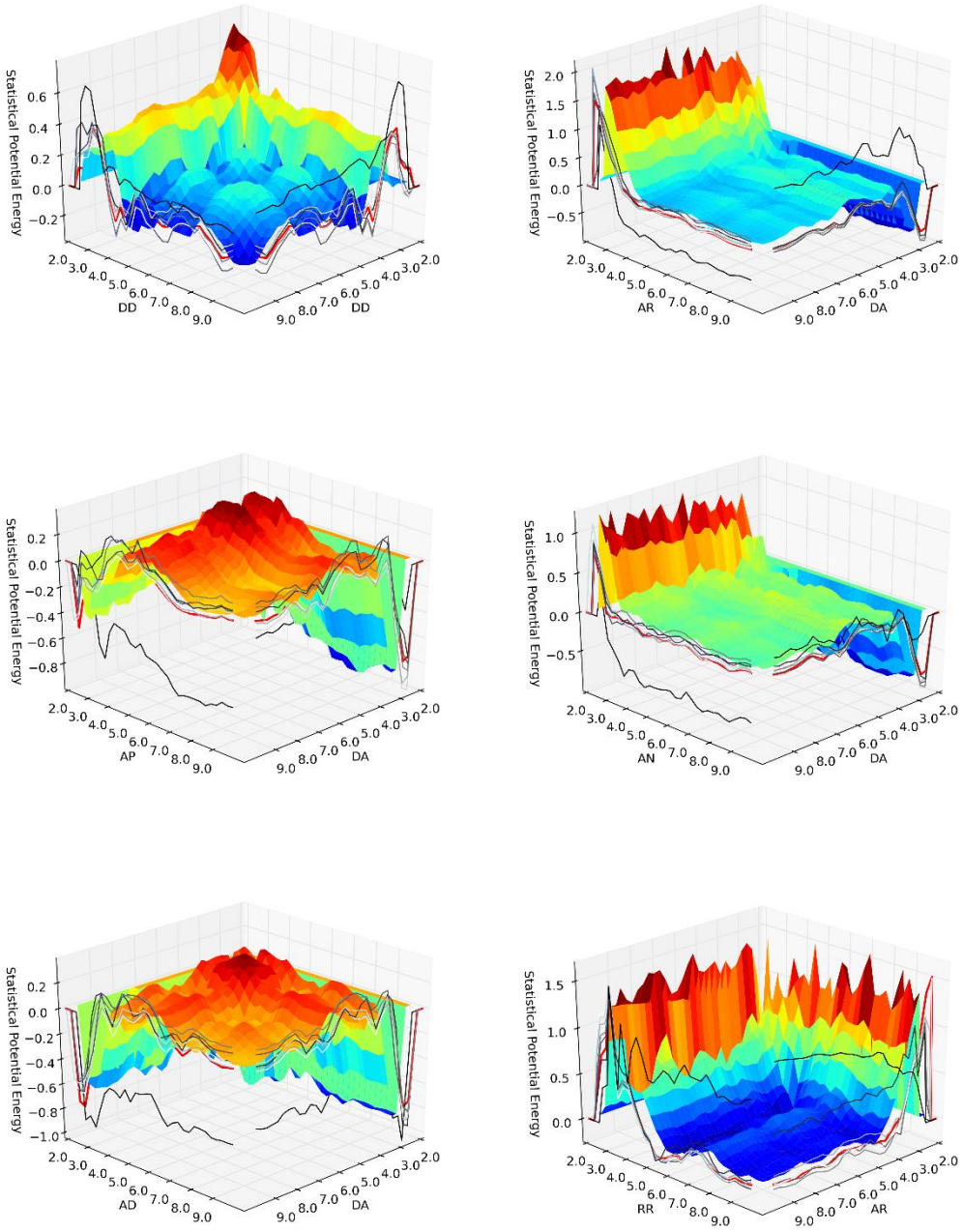

Figure 12 continued

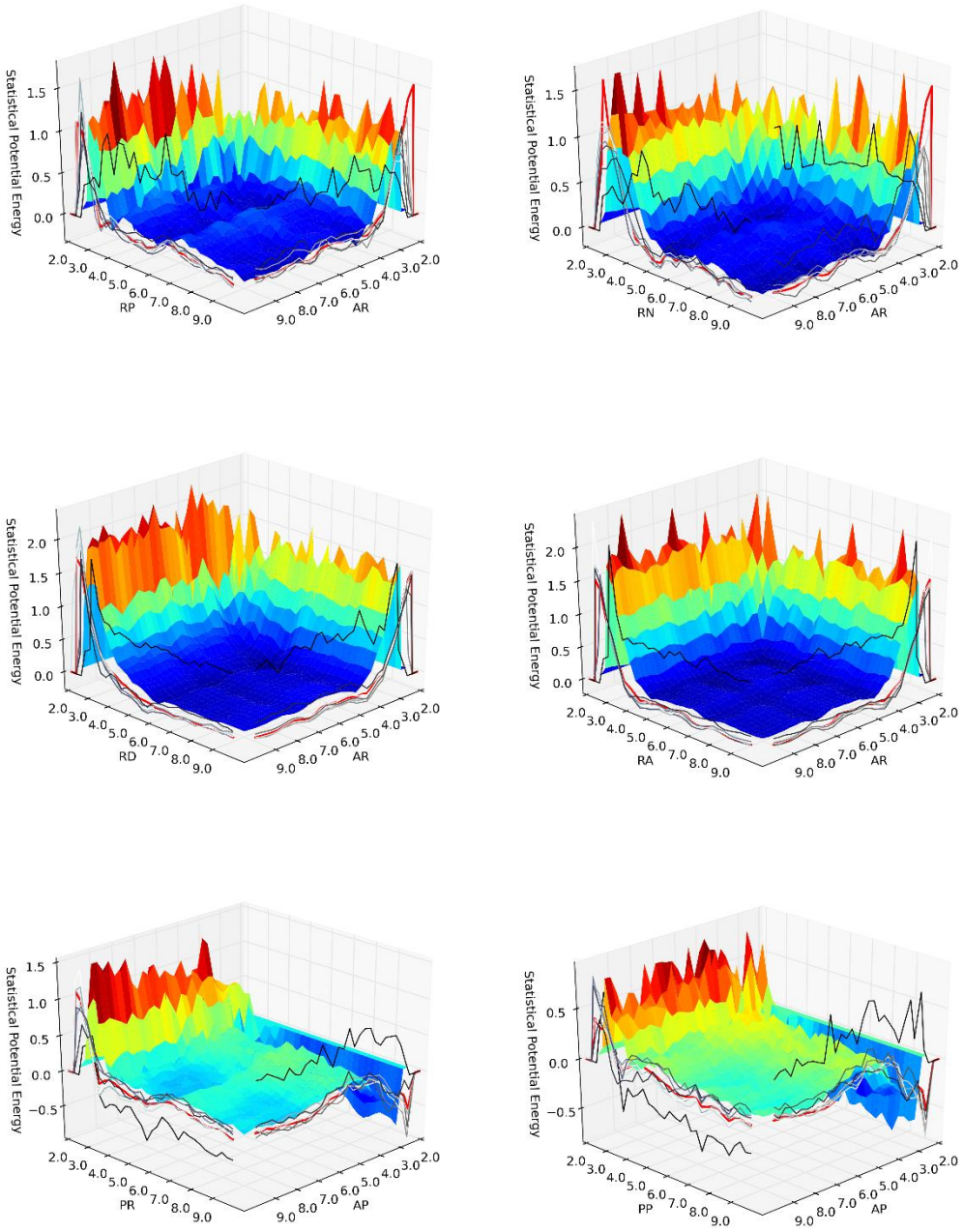

Figure 12 continued

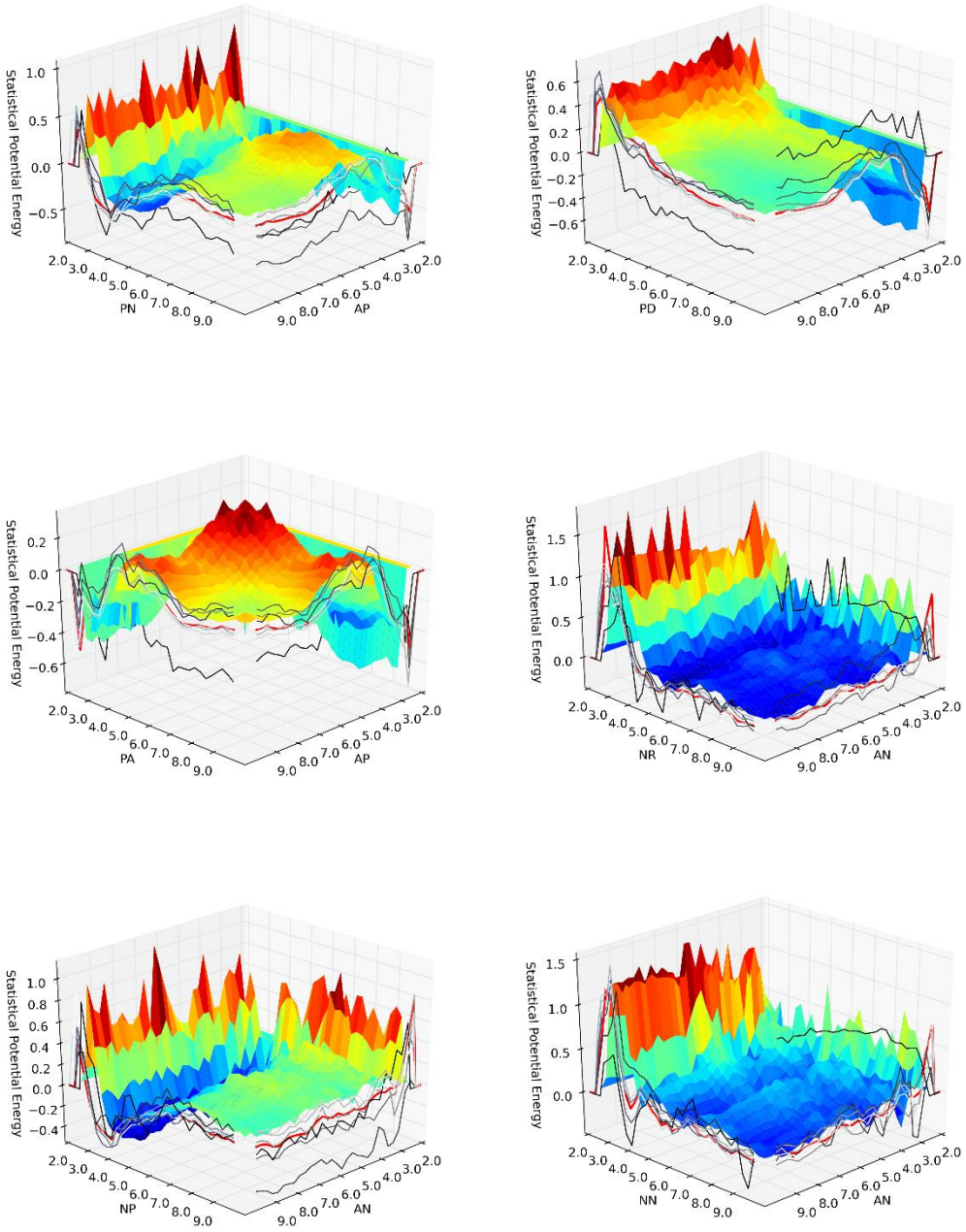

Figure 12 continued

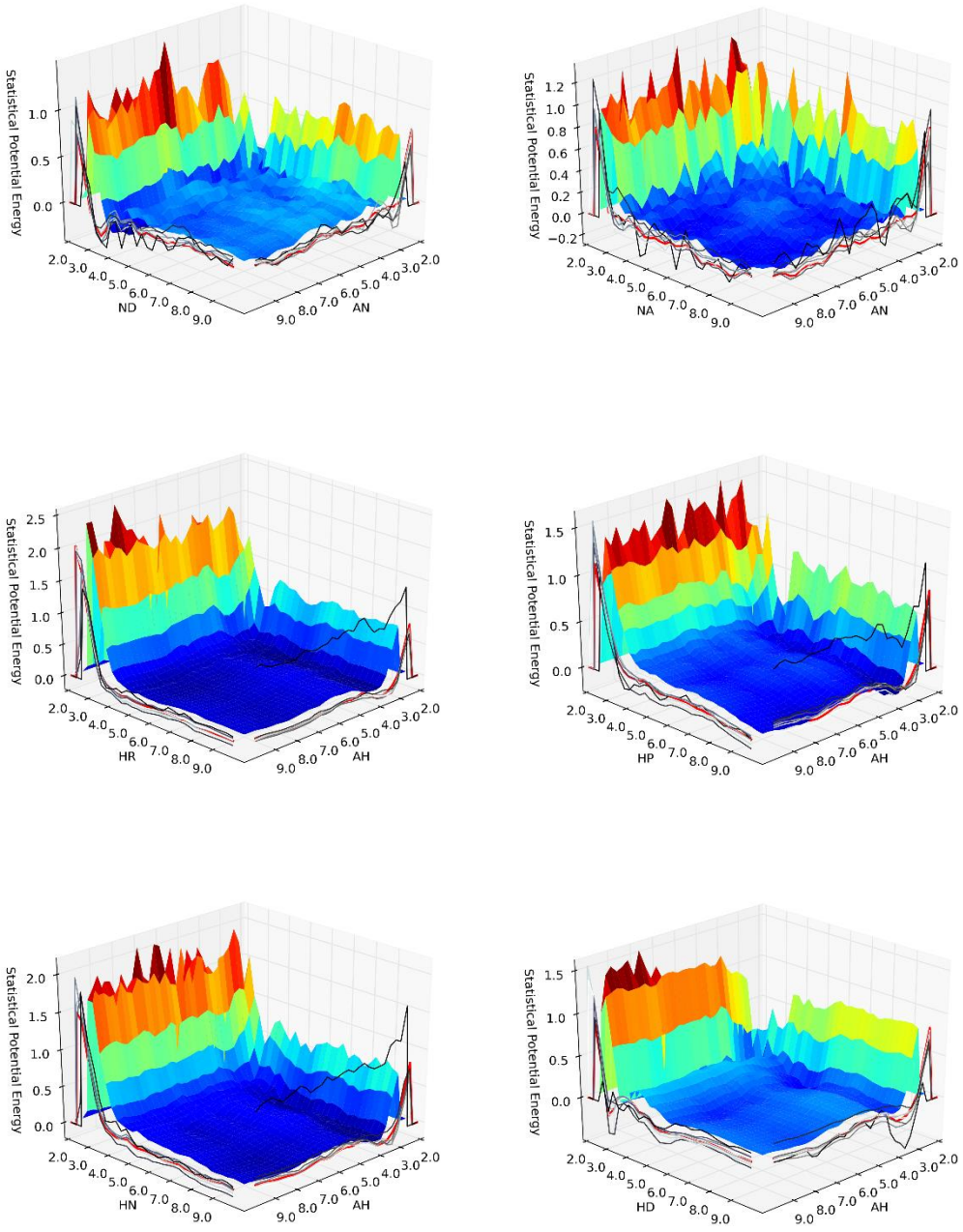

Figure 12 continued

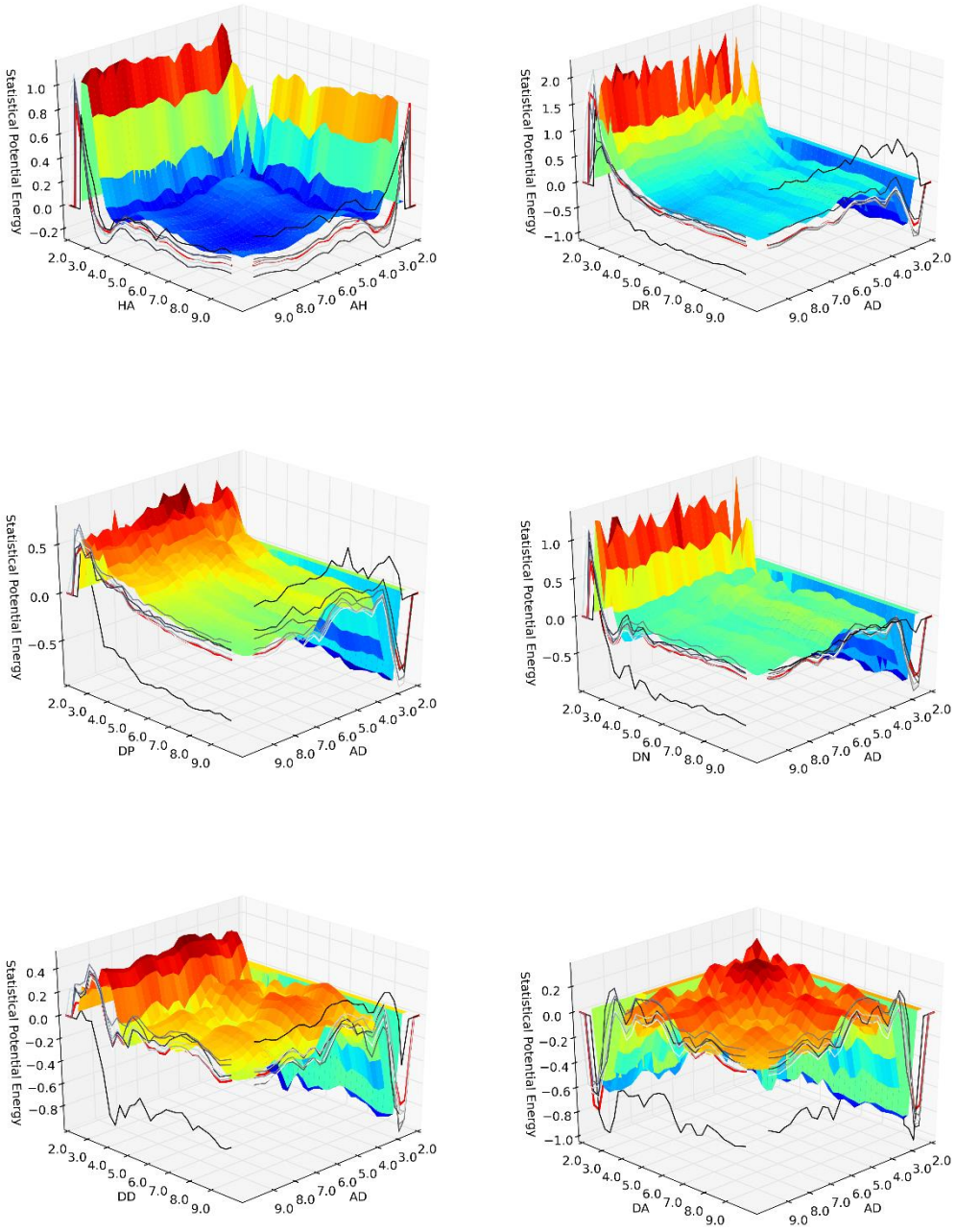

Figure 12 continued

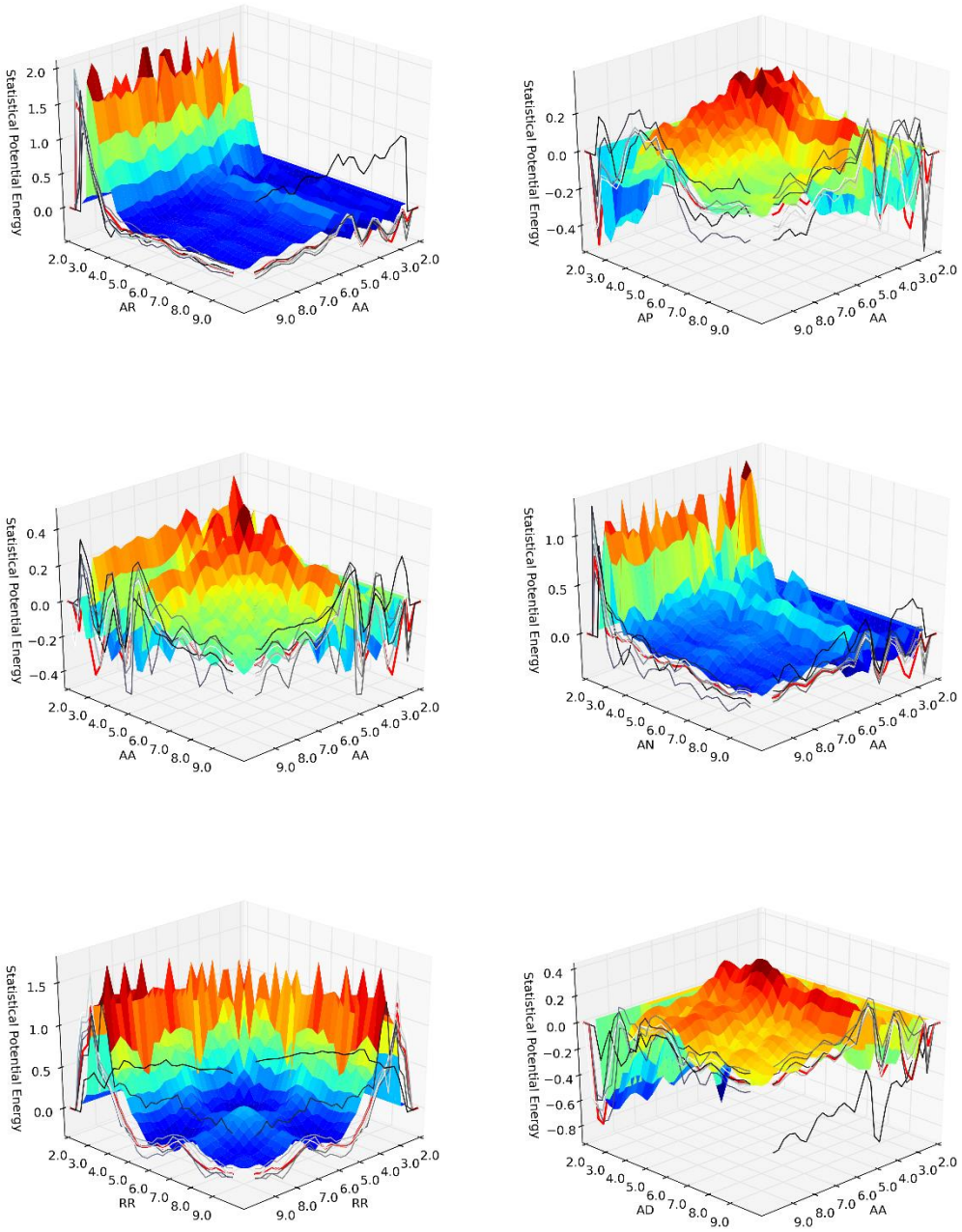

Figure 12 continued

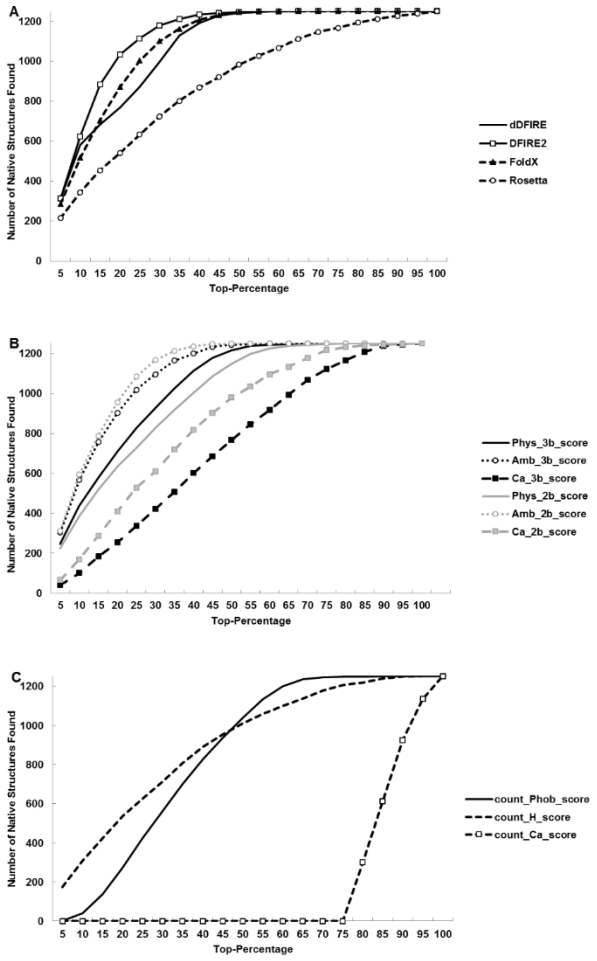

Figure 13 Number of sub-sets in vhp_mcmd decoy set that their native structure is ranked among various top percentages of structures, by A) dDFIRE, DFIRE2, FoldX, Rosetta, B) two-body and quasi-three-body scoring functions, and C) by simple counting methods

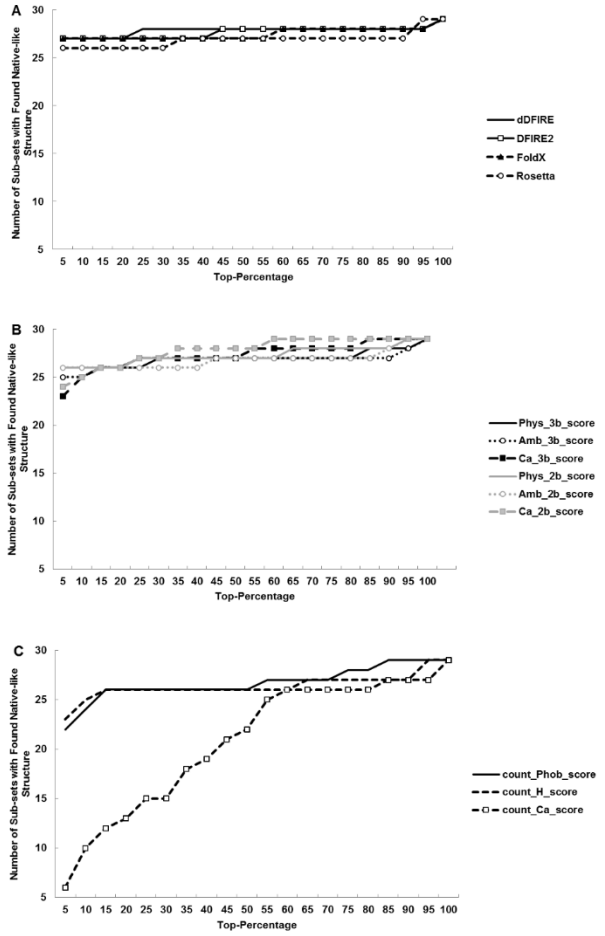

Figure 14 Number of native structures in hg_structural decoy set ranked among various top percentages of structures, by A) dDFIRE, DFIRE2, FoldX, Rosetta, B) two-body and quasi-three-body scoring functions, and C) simple counting methods.

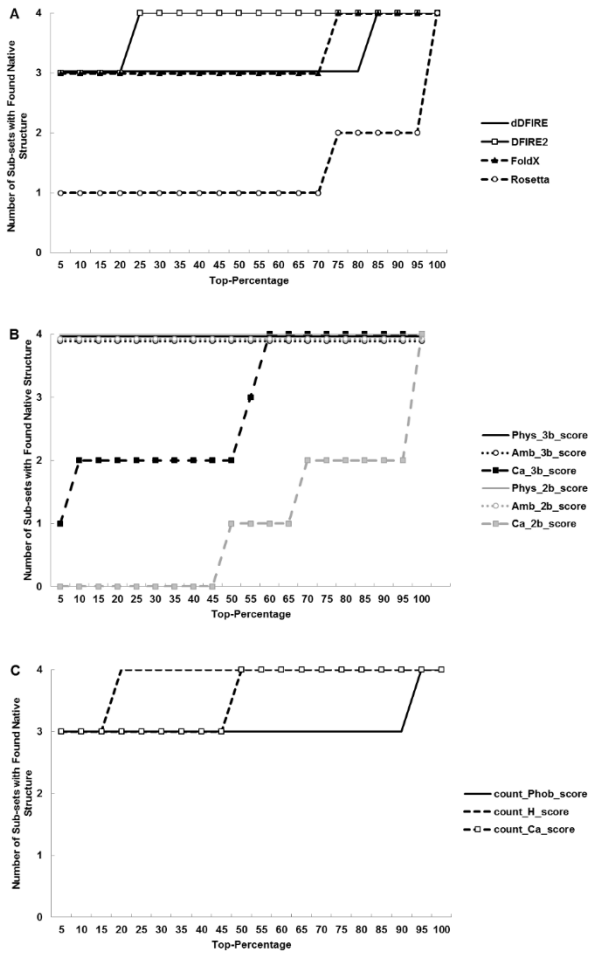

Figure 15 Number of sub-sets in fisa decoy set that their native structure is ranked in among various top percentages of structures, by A) dDFIRE, DFIRE2, FoldX, Rosetta, B) two-body and quasi-three-body scoring functions, and C) simple counting methods.

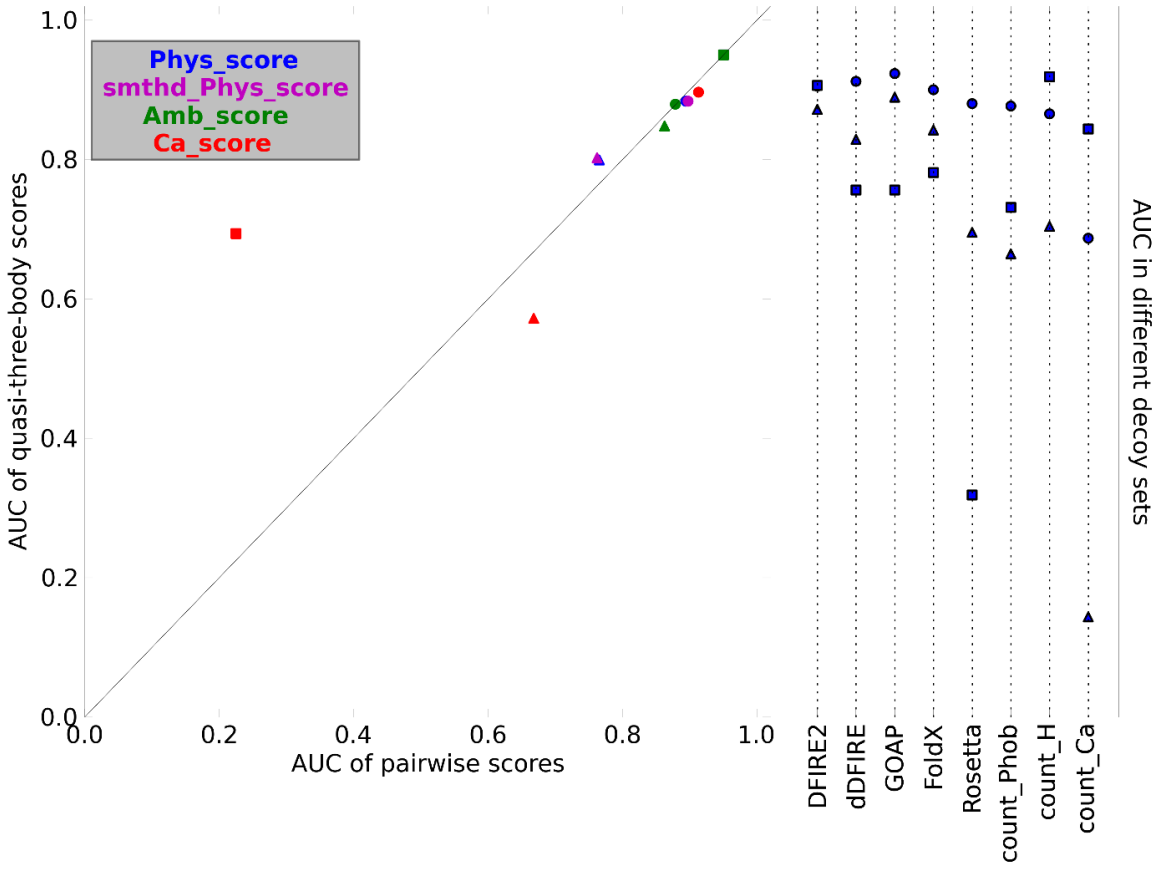

Figure 16 A) AUCs resulted from quasi-three-body scores vs AUCs of their two-body scores in different decoy sets tested. Comparison of pairwise to quasi-three body scoring functions shows little differences in structure-prediction quality. B) AUCs resulted from conventional scoring functions and simple counting methods. Result for vhp_mcmd, fisa, and hg_structural are represented by ▲, ■ and ●.

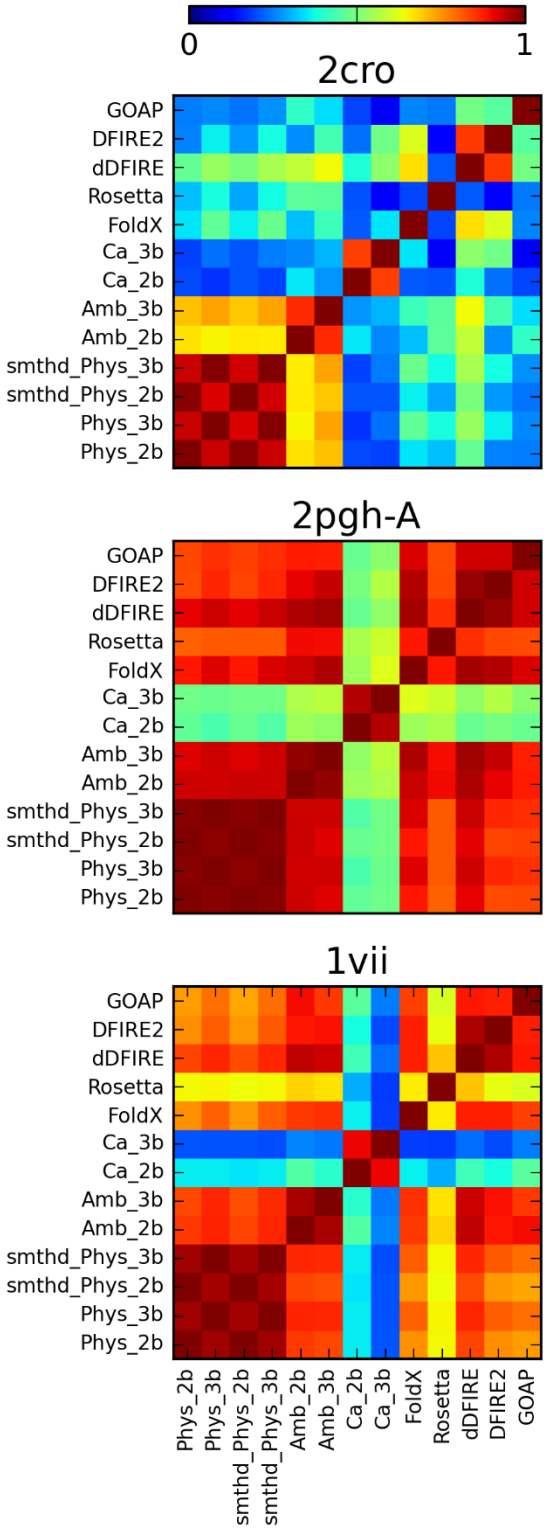

Figure 17 Pearson correlation coefficient among various scoring functions in 2cro from fisa, 2pgh-A from hg_structural and 1vii representing vhp_mcmd.

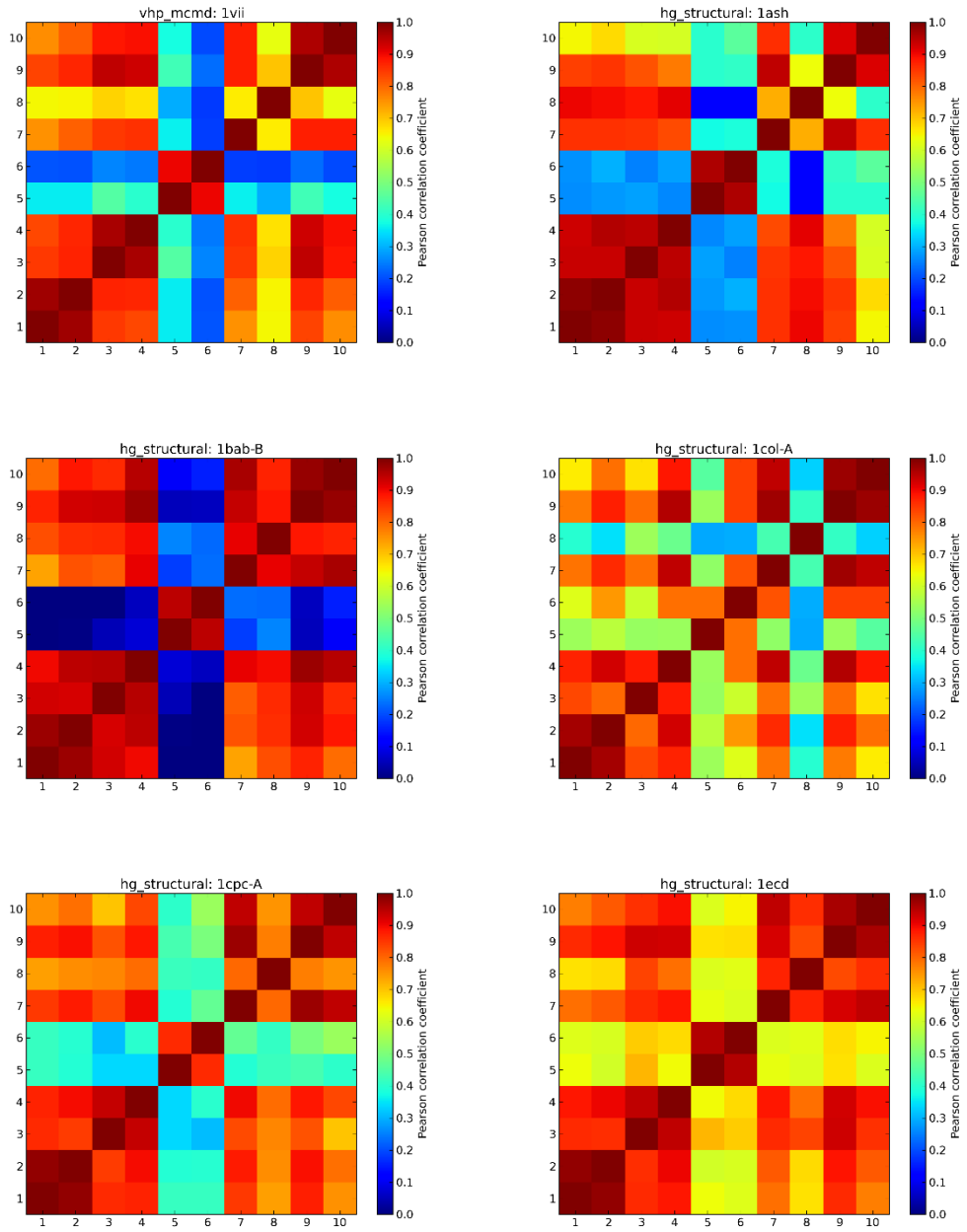

Figure 18 **:** Pearson correlation coefficient among 1- Phys_2b_score 2- Phys_3b_score 3- Amb_2b_score 4- Amb_3b_score 5- CALPHA_2b_score 6- CALPHA_3b_score 7- FoldX 8- Rosetta 9- dDFIRE 10- DFIRE2. The title of each graph shows ‘decoy set : subset of the decoy set’

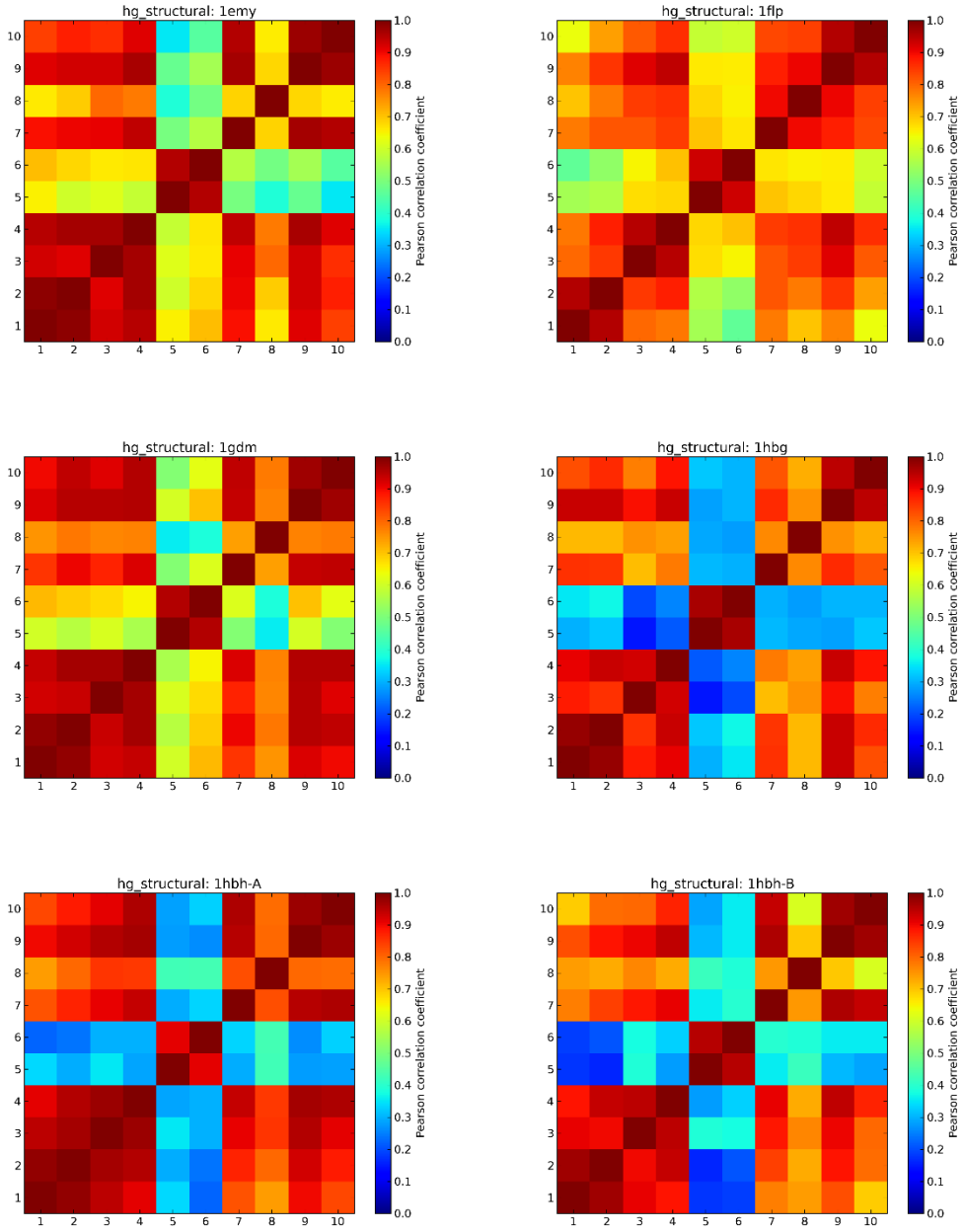

Figure 18 continued

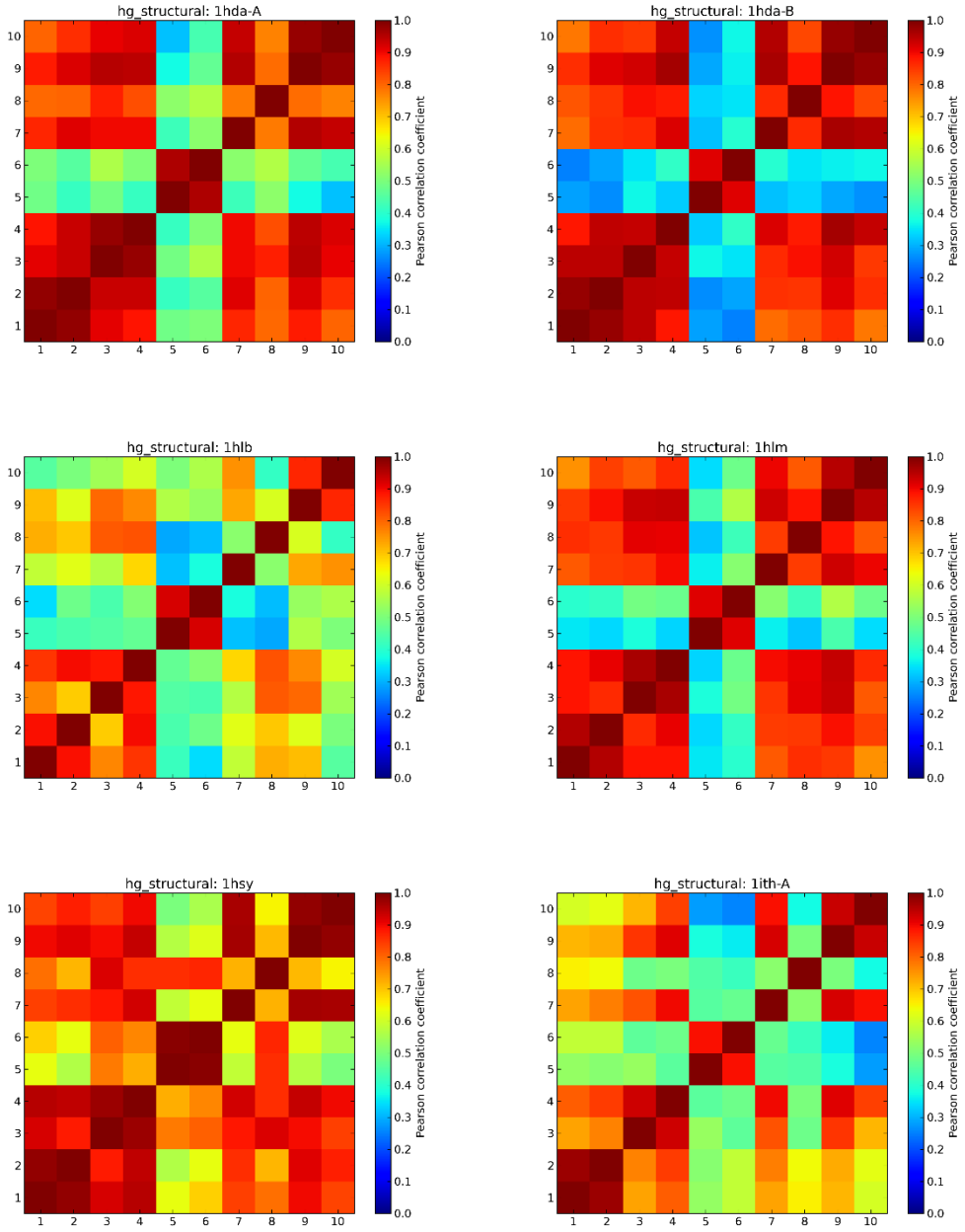

Figure 18 continued

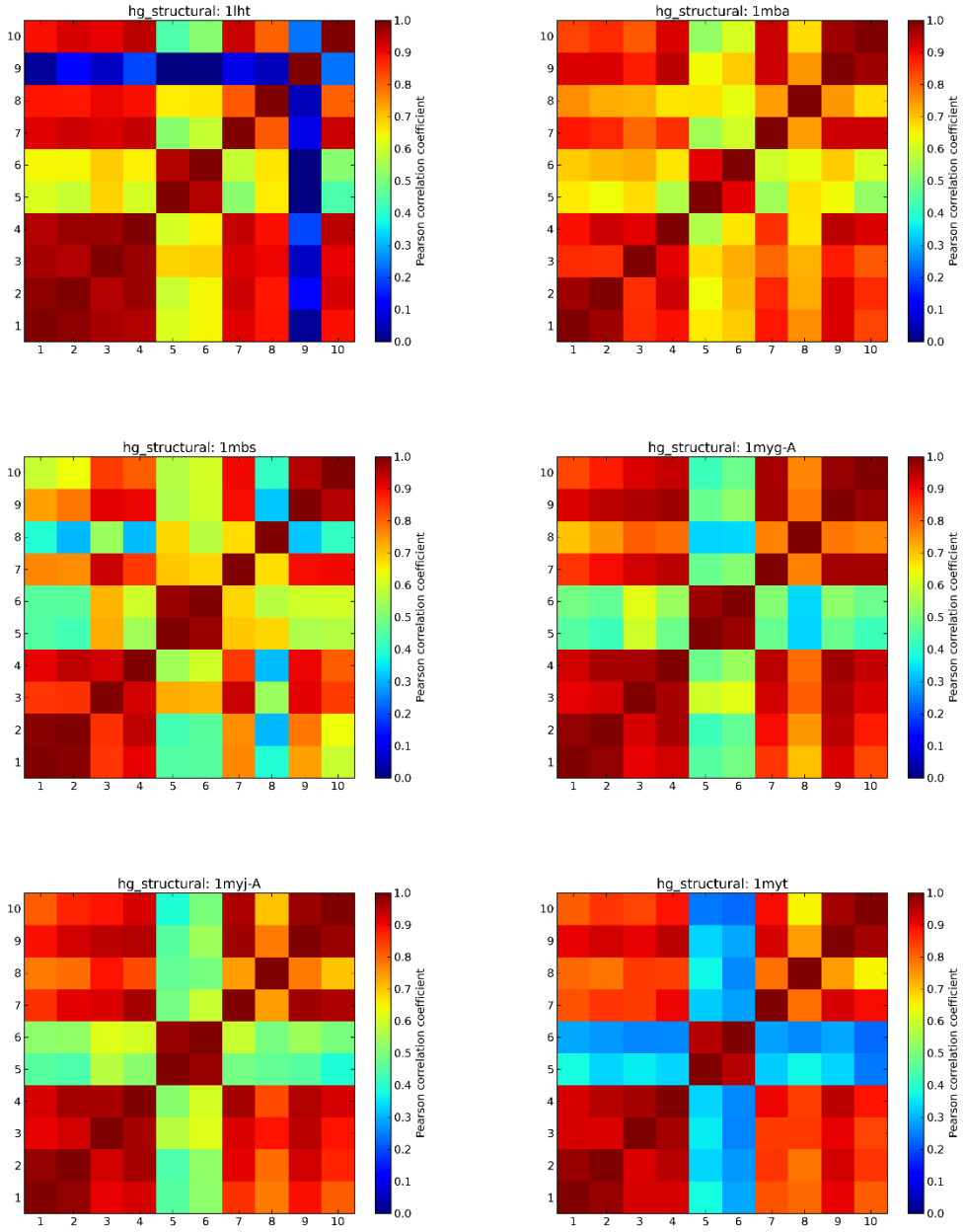

Figure 18 continued

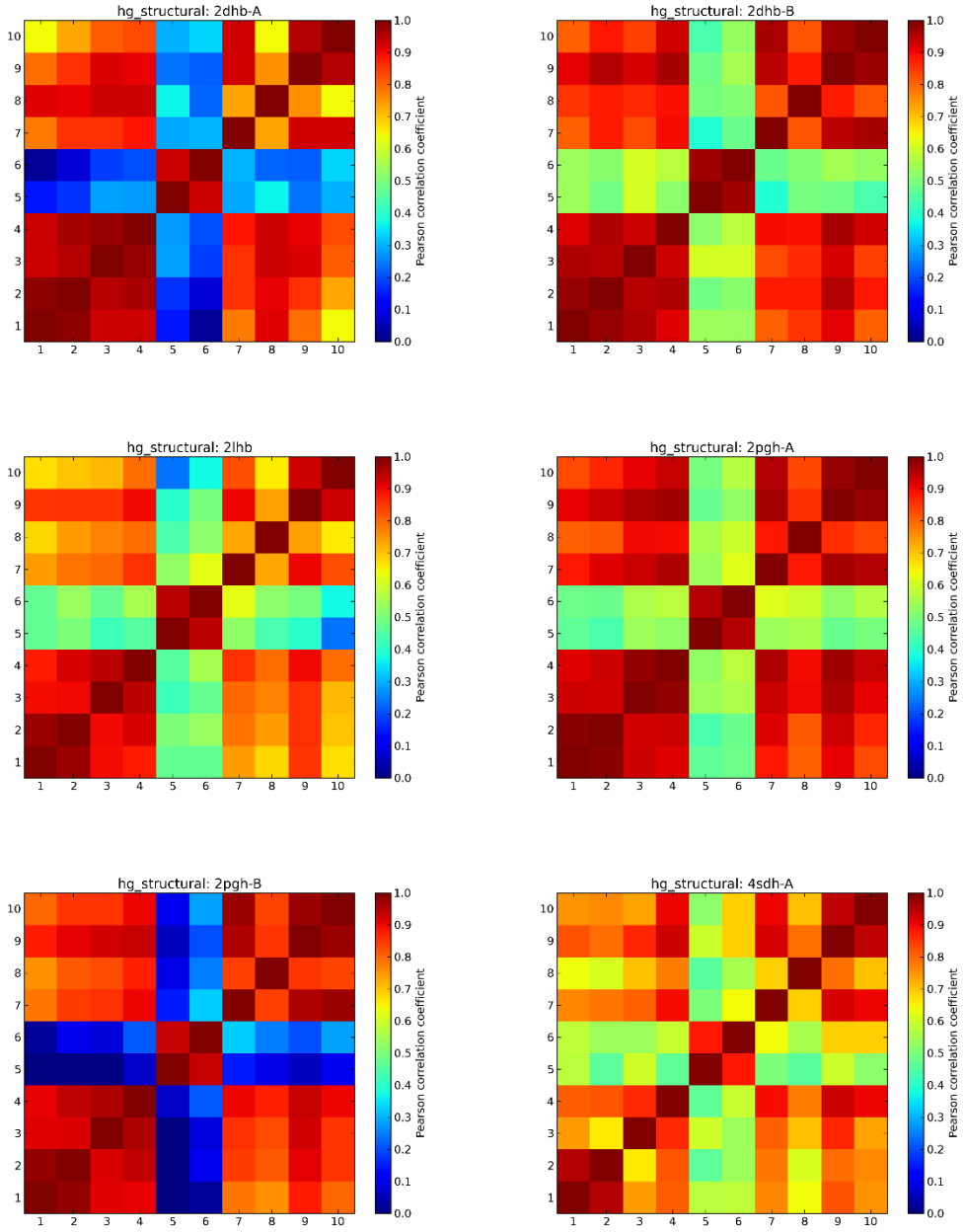

Figure 18 continued

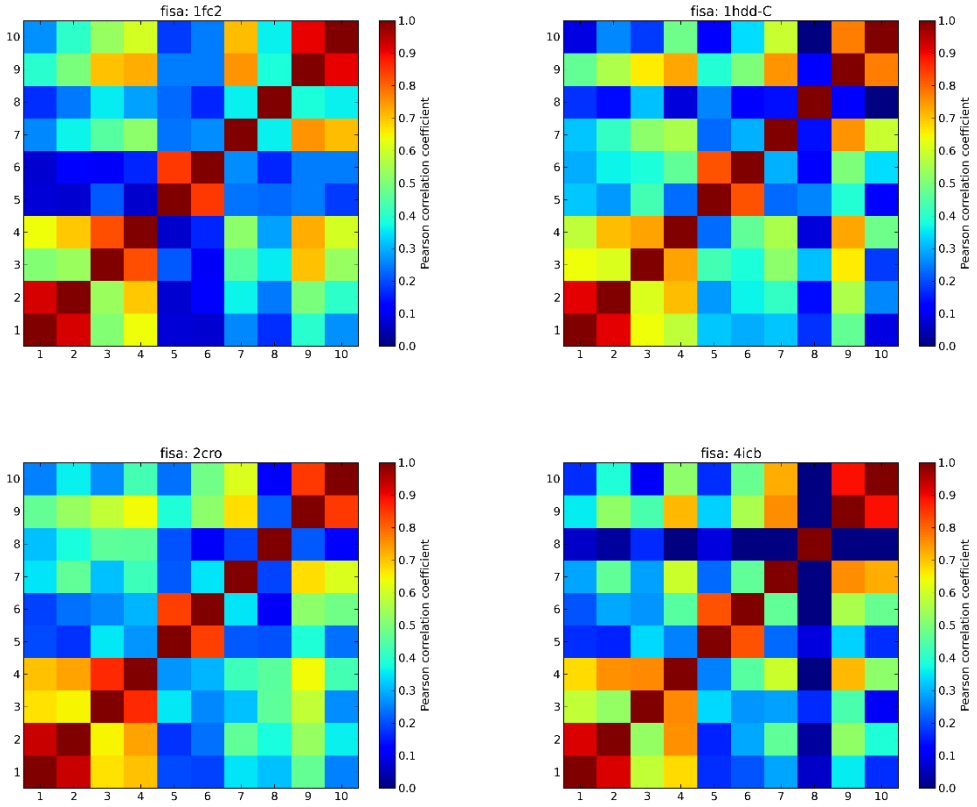

Figure 18 continued

3.4 Conclusion

Calculating the free energy of a protein system using statistical potential derived from pair-distribution functions is not physically justified. Theoretically, for using statistical potentials to compute the total free energy of a protein system, the energies should be expressed in terms of multi-body interaction terms. Various multi-body potentials have been developed based on this theoretical argument. Higher order information and better scoring performances are reported for those multi-body potentials[57,66–68]. To the best of our knowledge, however, all of those multi-body potentials are based on a coarse-grained representation of interacting bodies and are contact-based. In this study, we asked the question if higher order information is also important for distance-dependent statistical potentials that are based on an atomistic representation of the interacting bodies.

Our results indicate that the multi-body interaction energies are dominated by pairwise interactions, with small contributions from higher order interactions, resulting in the lack of significant difference between pairwise and quasi-three-body potentials. In contrast to our initial hypothesis, we have seen in the majority of cases a lack of distance dependency between two pairs of interacting bodies constituting quasi-three-body statistical potentials. Higher order interactions can only be established in few triplets modeling interactions between charged atoms. In other words, besides charge interactions, considering the effect of the distance of a third interacting body on the pair distribution function of two other interacting bodies utilizing the methods we presented in this study adds negligible additional information to the statistical potential. Considering scarcity of charged bodies in protein structures compared to other types of interacting bodies, similar performance of two-body and quasi-three-body scores is not surprising.

This similar performance can be attributed to the lack or weakness of such higher order information in quasi-three-body potentials. We see a very high correlation between corresponding quasi-three-body and two-body scores which is in line with similarities in the patterns of contour maps observed in quasi-three-body and two-body potentials.

The scoring functions developed in this study show higher or comparable performances with the four conventional scoring functions tested. We also obtain good results for many systems from simple counting scoring functions designed to model hydrophobic or hydrogen bond interactions. High performance of these simple counting approaches can be attributed to the decoy sets not being sufficiently challenging. Also it can imply that hydrogen bonding and hydrophobic effect can adequately be used to differentiate native structures from decoys in many protein systems. It is not surprising considering the importance of these interactions in the protein folding process[69].

Table 4 Brief description of the scoring functions generated throughout the study.

Score Title	Description
count_Ca_score	Number of Cα's within 5 Å distance of each other
count_Phob_score	Number of hydrophobic atoms within 5 Å distance of each other
count_H_score	Number of hydrogen bonds formed
Phys_2b_score	Two-body score based on physicochemical elements
Phys_3b_score	Quasi-three-body score based on physicochemical elements
smthd_Phys_2b_score	Phys_2b smoothed by a cubic spline
smthd_Phys_3b_score	Phys_3b smoothed by a cubic spline
Amb_2b_score	Two-body score based on AMBER atom types
Amb_3b_score	Quasi-three-body score based on AMBER atom types
Ca_2b_score	Two-body score based on Cα atoms
Ca_3b_score	Quasi-three-body score based on Cα atoms

Table 5 Triplets with more than 10% violations of null hypothesis in KS-test: Normalized contours of quasi-three-body joint probability distributions are compared with the corresponding two-body probability distribution using KS-test. Out of 126 triplets, nine triplets violate null hypothesis that distributions are the same for more than 10% of all distance slices.

Triplet	Number of violations of null hypothesis with significance level of 0.05
APN	43
DPN	12
HPN	11
PPN	7
RPN	7
NNN	6
RPN	6
NPN	6
HNP	6

Table 6 Examples of time needed for calculations of main steps of quasi-three-body and pairwise scoring (precision of 1 msec)

protein	decoy set	#amino acids	Populating (C_{ij}^{ABC}) and normalizing with d^2	Populating (C_{ij}^{AB}) and normalizing with d^2	Calculating S_{ABC}	Calculating S_{AB}
NATIVE_132	vhp_mcmd	36	0.124	0.002	0.001	0
2cro	fisa	65	0.399	0.005	0.001	0
1emy	hg_structural	153	2.916	0.033	0.001	0

CHAPTER 4. FIBPREDICTOR: A COMPUTATIONAL METHOD FOR RAPID PREDICTION OF AMYLOID β-FIBRIL STRUCTURES

4.1 Introduction

In this chapter, a computationally fast and general computational procedure, FibPredictor, is proposed to generate structural models for any amyloid fibril, starting from its sequence. Despite the efficiency of the algorithm, the generated models are accurate in generating experimental structures among the top-5 ranked models, for providing a description of the structural landscape available to an amyloid fibril forming sequence and can be used as initial structures for more sophisticated computational studies. FibPredictor is available at http://nanohub.org/resources/fibpredictor.

The following two-step procedure of Fibpredictor was developed to generate amyloid fibril structures: For a given protein sequence, an ensemble of candidate amyloid fibril structures is generated comprehensively representing the amyloid fibril conformational space accessible to that specific sequence. This ensemble contains representative structures from all eight classes of amyloid fibrils. These eight classes are described in table 7 and figure 19. Further details can be sought at the cited references [16,30]. Using a scoring function developed for protein-structure prediction, the most energetically favorable candidate structures are then identified comprising the suggested computational models of amyloid fibrils.

None of the individual steps of this computational procedure includes time consuming and computationally expensive methods, such as molecular dynamics simulations, so the procedure is computationally efficient and easy to implement. Validity of the approach is demonstrated by reproducing the experimentally determined structures of six amyloid fibrils.

4.2 Materials and Methods

4.2.1 Input for Fibpredictor

The minimum necessary input for the program Fibpredictor are the sequences of all strands within each of the interacting β-spines. β-spines are β-sheets which interact with each other side-by-side to form the full amyloid fibril. The sequences of each individual strand within each β-spine can be identical or different, covering various cases of amyloid fibrillation. The number of β-strands and their length should be the same for both β-spines.

4.2.2 Generating the structural ensemble

Figure 20 summarizes the procedure for generating the structural ensemble of amyloid fibrils. First, coordinates of the backbone atoms (Cα, C, O and N) are generated for one of the β-spines as a regular β-sheet. These strands should have the same number of residues. Two separate sets of coordinates are generated, one as a parallel and the other as an anti-parallel β-sheet. For each set of coordinates, the normal vector of the approximate sheet formed by all Cα's is determined. This normal vector is calculated by averaging over normal vectors of all planes formed by any three Cα's. This vector is then tilted and elongated randomly within a user-defined range of values for tilt angles and elongation length. This process defines a translation vector, which will be used to place the

backbone of the second β-spine. The ranges for tilt angles and elongation lengths in the current study were set to 45 degrees and 3.5 - 14 Å, respectively, but can be adjusted by the user to the target amyloid fibril specifications. Note that in order to adjust these specifications, no knowledge about the details of the structure of the target amyloid fibril is necessary. Instead, the length of the side chains in their fully extended conformations can be used to set the maximum distance between the sheets. The length of the fully retracted conformations of the side chains, on the other hand, can be used to set the minimum distance between the sheets. Initial hypotheses on the probable types of interactions between certain amino acids on the first and the second sheet can be used to limit the range of the tilting angles.

Multiple translation vectors are randomly generated to create structural options for the second β-spine relative to the first β-spine. It is necessary to sample relative positions for the two β-spines which can lead to proper entanglement of the side-chains, creating a strongly interacting steric zipper. Fifty translation vectors were used for this study. For each of these translation vectors, the backbone atom coordinates of the first β-spine are copied along the translation vector to generate the backbone coordinates of the second β-spine. In addition to simple copying of the coordinates, rotation operations are performed on the second β-spine to generate other members of the eight potential classes of amyloid fibrils. Rotation around the z-axis generates similar or different directionalities of the β-spine (classes two, three, six and eight) and rotation around the x-axis generates face-to-face or face-to-back steric zippers (classes one, three, five and eight). In summary, the different copies of the second β-spine are generated by simple copying or by additional x-rotation, z-rotation or zx-rotations. These rotations result in four different amyloid fibril

classes for each of the initial parallel and anti-parallel backbone coordinates comprising all eight classes of amyloid fibrils. Table 7 shows the initial β-sheet conformations and rotations used to generate each of the eight classes.

Each of these initial backbone structures is then passed to the side-chain prediction program SCRWRL4 which adds all sidechains to the backbone using a rotamer library aiming to minimize the SCWRL4 scoring function [70].

4.2.3 Scoring the ensemble structures

We tested three different scoring functions to identify the most energetically favorable candidate structures in the ensemble: GOAP, Amb_3b, and the SCWRL4 internal scoring function. The SCWRL4 internal scoring function is used by SCWRL4 to predict the energetically lowest side chain orientations [70]. SCRWL4 uses a rotamer library and calculates the self-free energy and the pair-wise free energy of the different rotamers using a scoring function including terms describing intra- and intermolecular interactions such as hydrogen bonding and van der Waals interactions. For more details the reader is referred to [70].

GOAP[71], is a statistical scoring function widely used in homology modelling, especially as part of the homology-modelling software MODELLER[72]. GOAP defines a plane with each heavy atom and two other neighboring bonded heavy atom and associates a local coordinate system ($\vartheta_x, \vartheta_y, \vartheta_z$) with this plane. Two polar angles ψ and θ and a torsional angle χ are then defined using this coordinate system. The GOAP potential, then, is calculated as shown in Eq. 12.

Amb_3b is a statistical scoring function developed in our lab, which has shown better performance than a number of conventional scoring functions including Rosetta and

FoldX in differentiating native from decoy protein structures in three different protein structure ensembles [73]. Interacting partners are represented as AMBER atom types. The total energy of the protein structure is then determined using a pre-calculated quasi-three body statistical potential as shown below:

$$Amb\_3b = \sum_{ABC} \sum_{i,j} \left(\frac{C_{ij}^{ABC}.\ B_{ij}^{ABC}}{d_i^2.\ d_j^2}\right) \qquad \text{Eq. 13}$$

Where A, B and C represent interacting partners, ABC refers to any possible quasi-three body interaction, i and j refer to the discretized distance between the first and second, and second and third interacting partners, d_i and d_j represent the interaction distances in angstroms, C_{ij}^{ABC} is the frequency of each triplet ABC in the distance interval corresponding to i and j and B_{ij}^{ABC} is the pre-calculated quasi-three body potential for ABC interaction in distance i and j.

4.2.4 FibPredictor usage and GUI

A graphical user interface (GUI) was developed for FibPredictor (Figure 21) allowing the user to specify the options of the software and export the results. For complete details on usage the reader is referred to user's manual available on https://nanohub.org/resources/fibpredictor/supportingdocs. The most important options are described in more detail in the following.

4.2.4.1 Sequences of the first and the second sheets:

FibPredictor models amyloid fibrils as two β-sheets parallel or antiparallel to each other. Each β-sheet consists of two or more β-strands. The user can enter the sequences of the

strands of the first (top box) and the second (bottom box) β-sheet in one-letter amino-acid code.

4.2.4.2 Sense of the β-sheets:

Amyloid fibrils can be formed by parallel or anti parallel β-sheets. FibPredictor can generate both types of backbone structure for amyloid fibrils, but the user can limit the modelling to only one types if experimental data on the sense of the target amyloid fibril does exist.

4.2.4.3 Scoring function:

Either the Amb_3b or GOAP scoring function, or both, can be chosen for ranking the ensemble of generated amyloid fibril models. The SCWRL4 internal scoring function is always used internally in FibPredictor as part of the side-chain optimization using SCWRL. Amb_3b is computationally more efficient than GOAP and can be used for initial modeling studies. A consensus scoring scheme using all of the three available scoring functions may allow for the most robust ranking of the structure models.

4.2.4.4 Rotations:

Rotations of one β-sheet with respect to the other are used to generate the various classes of amyloid fibrils. All types of rotations should be chosen unless experimental data allow some amyloid classes to be eliminated.

4.2.4.5 Number of randomly generated models (Rand. models):

This variable specifies the number of translation vectors generated to place the second sheet relative to the first sheet, randomly somewhere in the chopped cone (Figure 20).

With increasing number of random vectors, the chance of obtaining good models increases at the cost of reduced computational efficiency.

4.2.4.6 Top models:

This variable determines the number of top-ranked structures provided as output to the user, based on the selected scoring function. This output allows the user to perform a more focused analysis of the predicted amyloid fibril models.

4.2.4.7 Minimum distance between the sheets:

This variable specifies the minimum distance between the sheets and should provide enough space between the sheets to accommodate side chains in the steric zipper conformation in a fully entangled conformation (Figure 20 and Figure 22-A).

4.2.4.8 Distance variation between the sheets:

This variable specifies the variation between minimum and maximum distance of the two sheets when generating the amyloid fibril structure models. The distance separation for each model will be a random number within this range (Figure 20 and Figure 22-A).

4.2.4.9 Angle variation between the sheets:

This parameter specifies the maximum horizontal translation of the second sheet with respect to the first sheet for investigating different entanglements of sidechains between the sheets (Figure 20 and Figure 22-B).

4.2.5 Validation

In order to demonstrate the ability of Fibpredictor to correctly model amyloid β-fibrils, we aimed to reproduce the experimentally determined structure of six β-fibrils. The

corresponding PDB-IDs of the six structures are 3OVL (class1)[74], 3HYD (class 1), 2ONV (class 4), 3OW9 (5), 2OMQ (class 7) and 2ONA (class 8) [16]. Despite the increasing number of amyloid fibril structures deposited in the protein data bank, only a small fraction are suitable for validating our structure prediction method because many of the deposited structure lack either the β-sheet or the steric zipper portion of amyloid fibrils.

Using an in-house program based on BioPython, all computational models generated for each of these six systems were superimposed on their corresponding reference PDB structure and their root mean square deviation (RMSD) from the experimental structure was calculated for all heavy atoms.

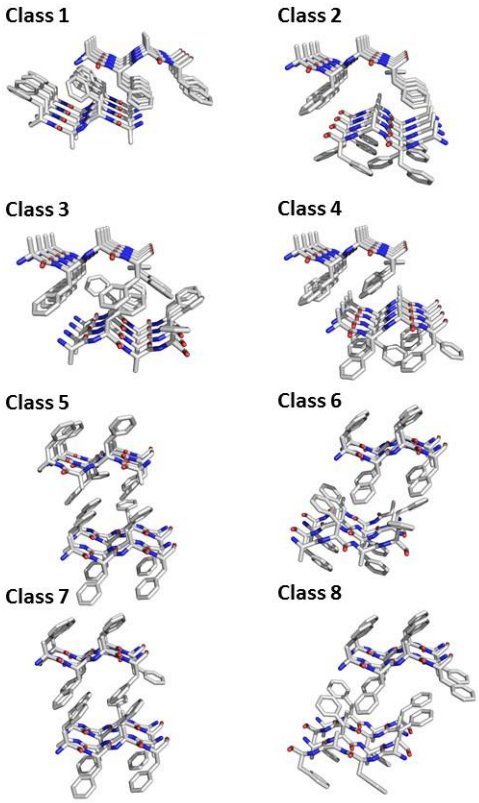

Figure 19 Eight classes of amyloid fibrils. Molecular models do not represent any natural fibril and are only presented to highlight the different classes. For more details refer to Table 1 and reference [16].

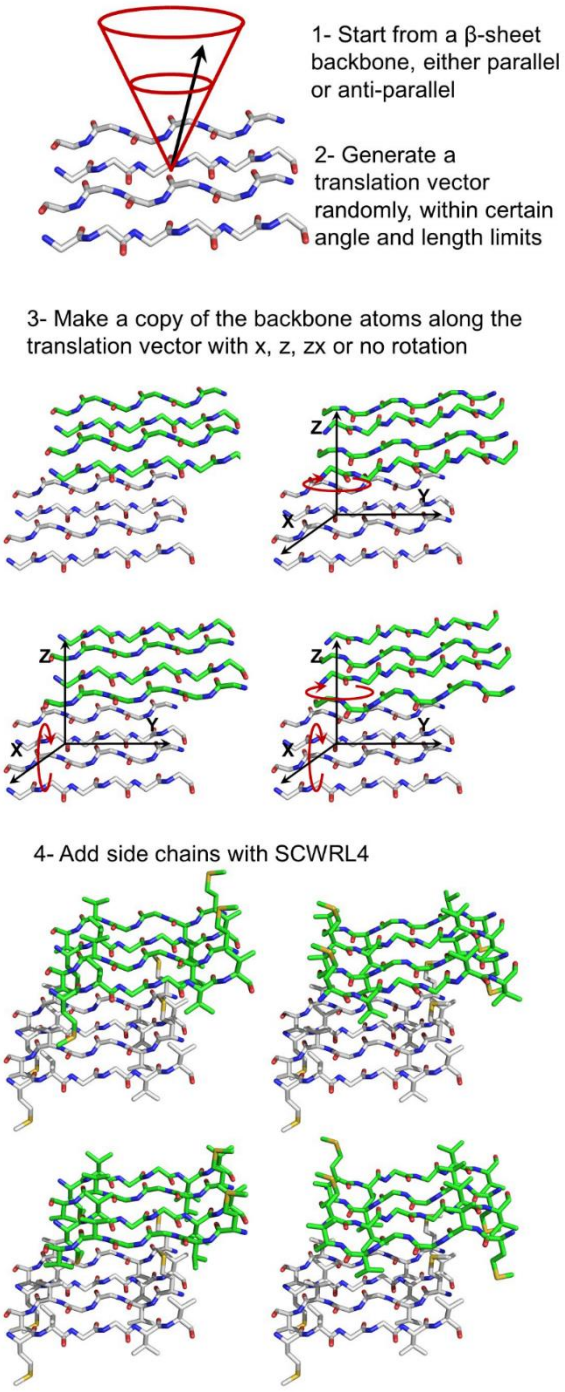

Figure 20 Procedure for generating computational candidate models for amyloid structures (example PDB ID: 2ONA). Multiple translation vectors are generated randomly and for each candidate structure four separate structures are generated using rotation operations.

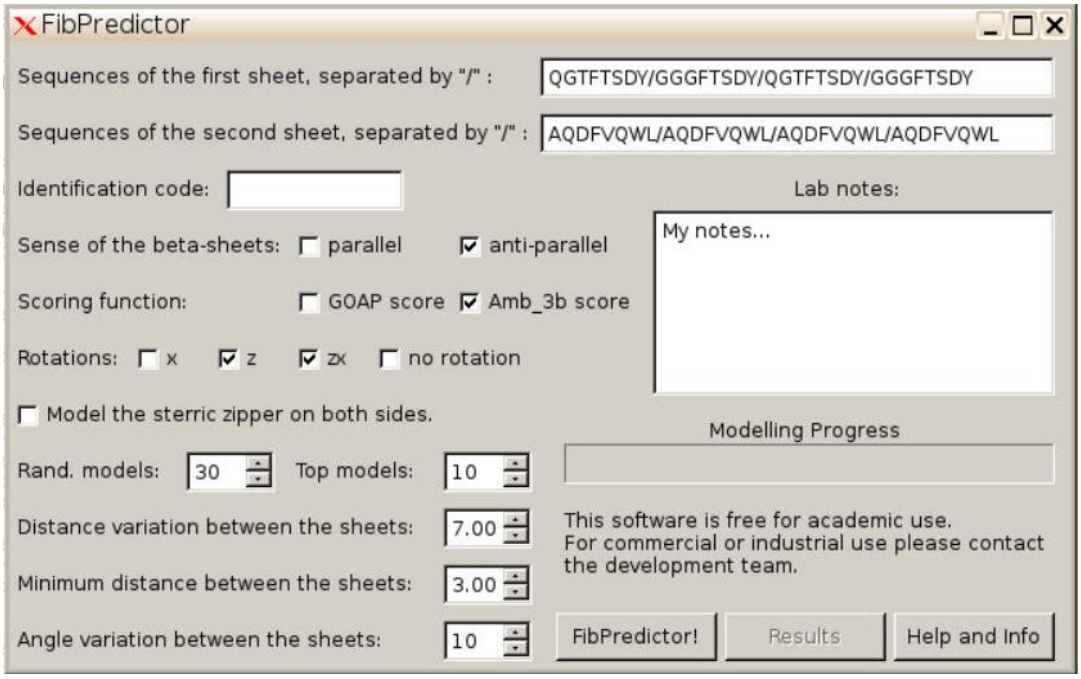

Figure 21 Fibpredictor GUI

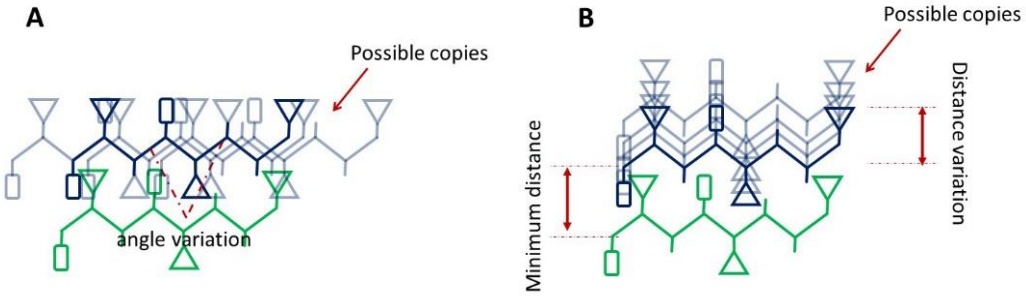

Figure 22 Minimum distance, distance variation (A) and angle variation (B) parameters in FibPredictor. The green schematic represent the initial β-sheet. The blue schematics represent the copied β-sheets.

4.3 Results and discussion

Figure 23 shows the modelled structures with the lowest RMSD, superimposed on their reference PDB X-ray crystal structures. For all of the six amyloid fibril test systems, FibPredictor generates structures with an RMSD less than 2.5 Å from the reference PDB structure. This demonstrates the feasibility of the computational sampling procedure to generate ensembles which contain fibril structures very close to the experimentally observed structure.

In order to investigate the accuracy of the three tested scoring functions for identifying the native fibril structures in the generated ensembles, an enrichment analysis was performed. For each protein system, the ensemble structures belonging to the class of the amyloid reference structure were ranked according to the three different scoring functions. The percentage of near-experimental structures (RMSD <3 Å) identified as a function of scoring rank of all predicted structures was plotted in the enrichment graphs shown in Figure 24; the underlying scatter plots are shown in Figure 25. The performance of an ideal scoring function and that generated by a random ranking of the structures are also shown for comparison. We observe that the results of the scoring functions usually are significantly better than random selection and sometimes even approach the ideal enrichment. This means that the scoring functions are generally successful in identifying the correct fibril structures. Overall, GOAP is the most successful scoring function in four of the test systems and Amb_3b is the best scoring function for the other two fibril systems. Within the correct class of amyloid fibril, the first native-like model is identified among the top 5 ranked structures with both GOAP (3OVL, 3HYD, 2ONV, 3OW9, 2OMQ) and Amb_3b (2ONA, 3HYD, 2ONV, 3OW9, 2OMQ). For the remaining tested

amyloid fibrils, the first native-like model appears among the top 10 structures for both GOAP and Amb_3b.

Although the scoring functions were successful in enriching native-like structures among the top-ranked structures within one class, they failed to differentiate between classes. Figure 26, for example, displays GOAP scores as a function of RMSD for the 2OMQ system. The graphs of the other scores and amyloid fibril systems follow the same general pattern (Figure 25). Although there are small differences between various classes, there are always predicted structures with favorable scores which belong to classes other than that of the reference structure and thus have high RMSD. This, however, does not necessarily mean that the scoring functions failed in identifying favorable structures, as structural polymorphism is widely observed in amyloid fibrils [75–77]. Hence, it is likely that the structure represented by the reference PDB is only one of several amyloid fibril structures energetically accessible to the peptide sequence. Structures with favorable scores but high RMSD may represent other polymorphs of the β-fibril as they display steric zipper interactions of potentially similar strength as the crystallized form of the fibril.

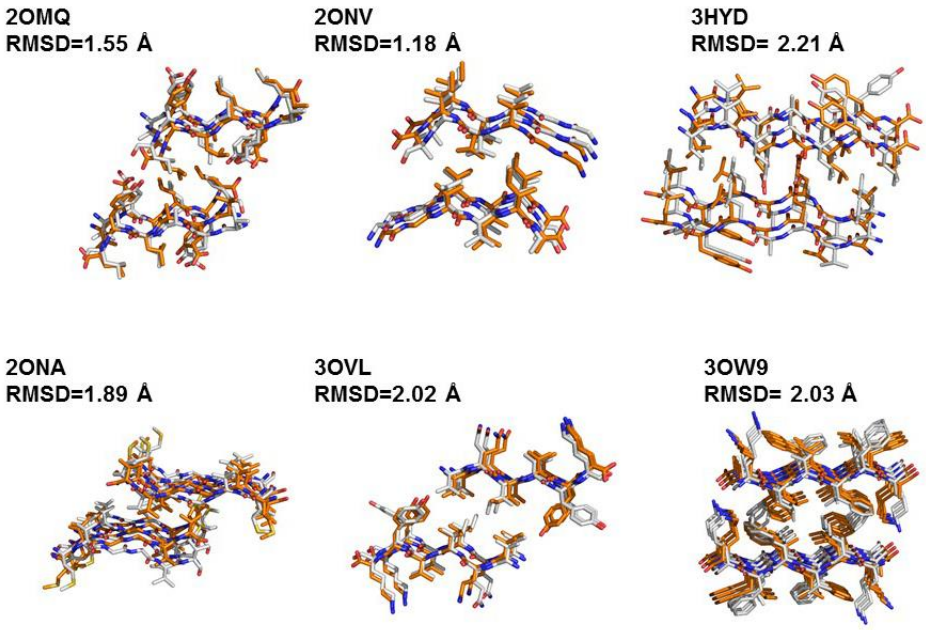

Figure 23 Predicted structure (carbon atoms in white) with the lowest RMSD value superimposed to their experimental reference PDB structure (orange) for the six fibril structures investigated in this study.

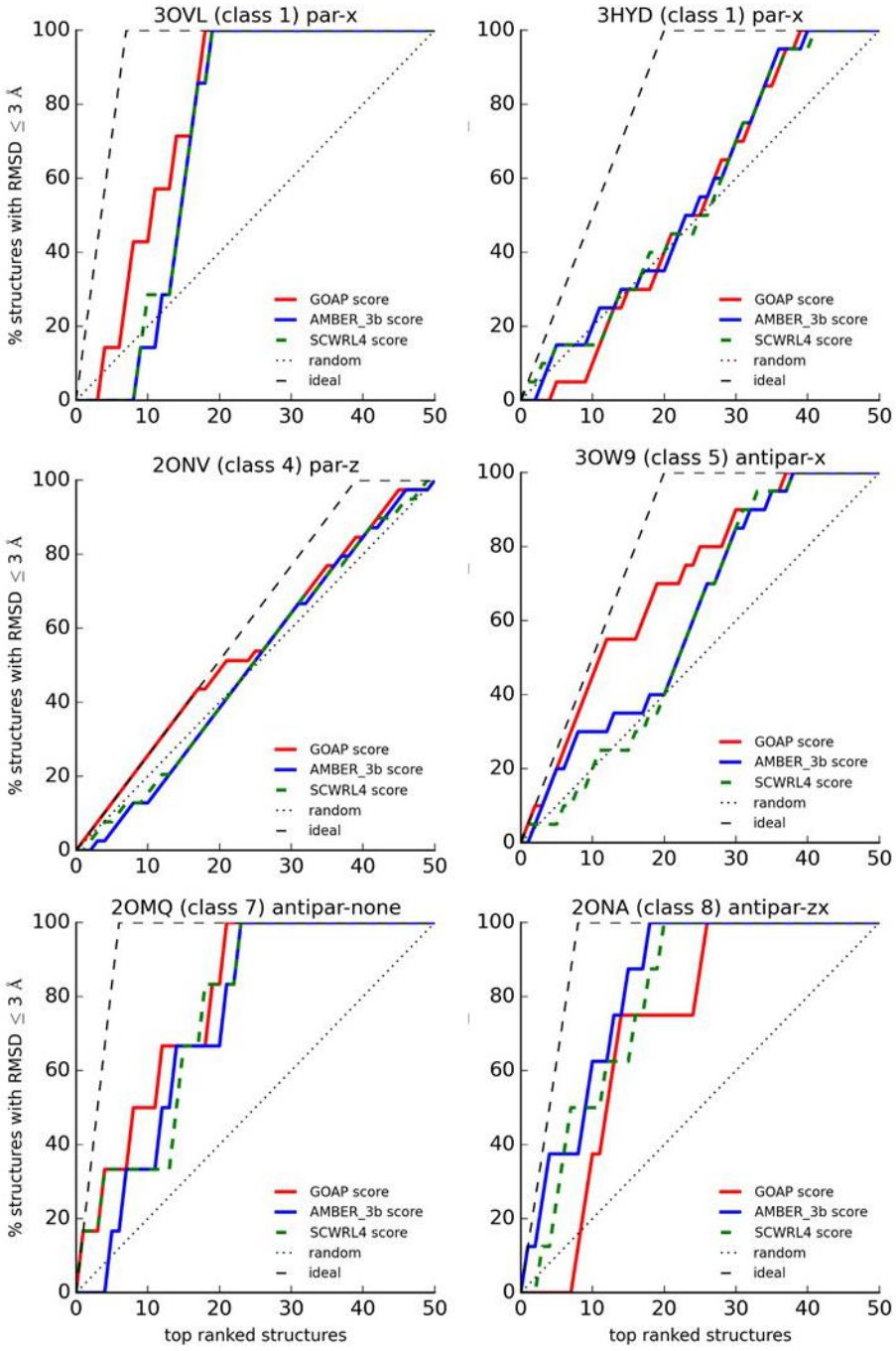

Figure 24 Enrichment plots for the four amyloid classes using three different scoring functions, showing the percentage of identified near-experimental fibril structures as a function of ranked ensemble structures. The reference PBD ID, its amyloid class, sense of the initial sheet (parallel (par) or anti-parallel (antipar)) and applied rotation operations (none, z, x or zx) are included in the title of each graph.

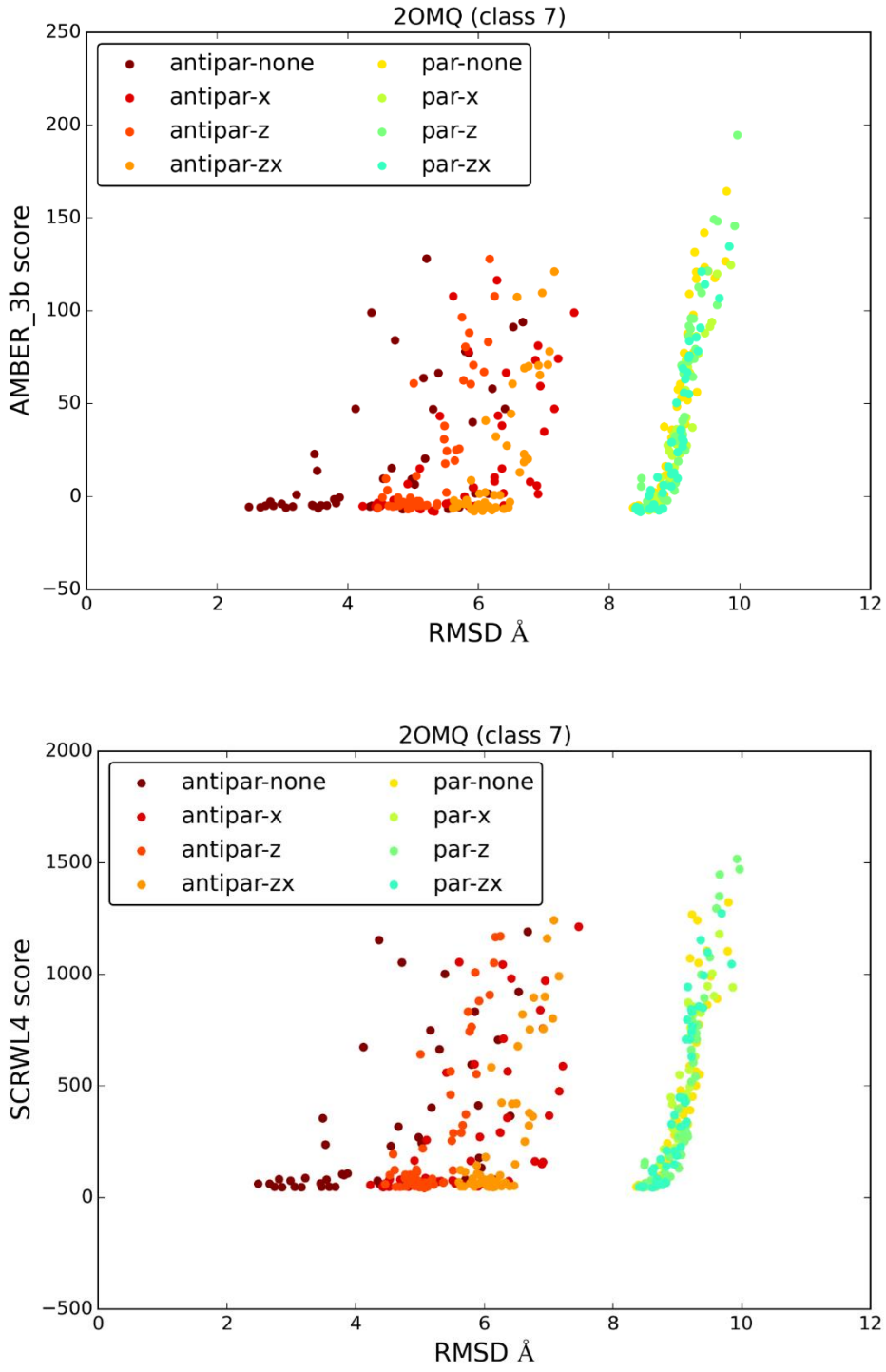

Figure 25 Various scores vs. RMSD for all of the six amyloid systems tested. The triangle shows the score of the reference PDB structure.

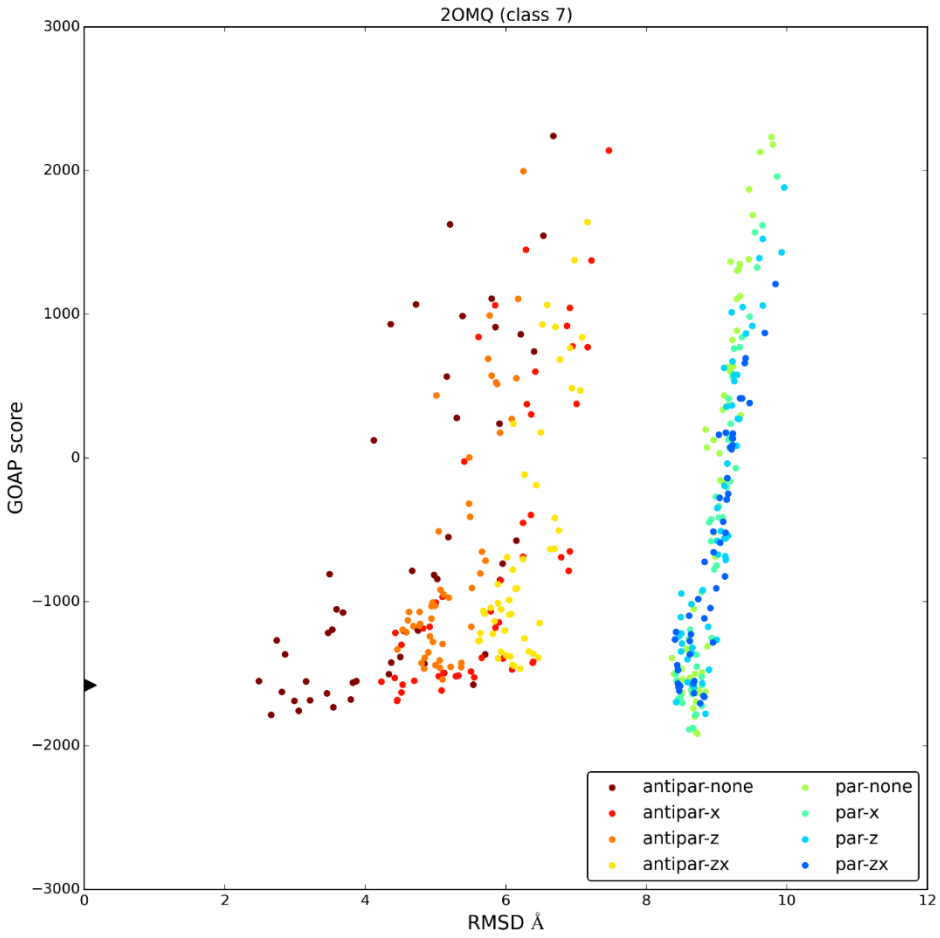

Figure 25 continued

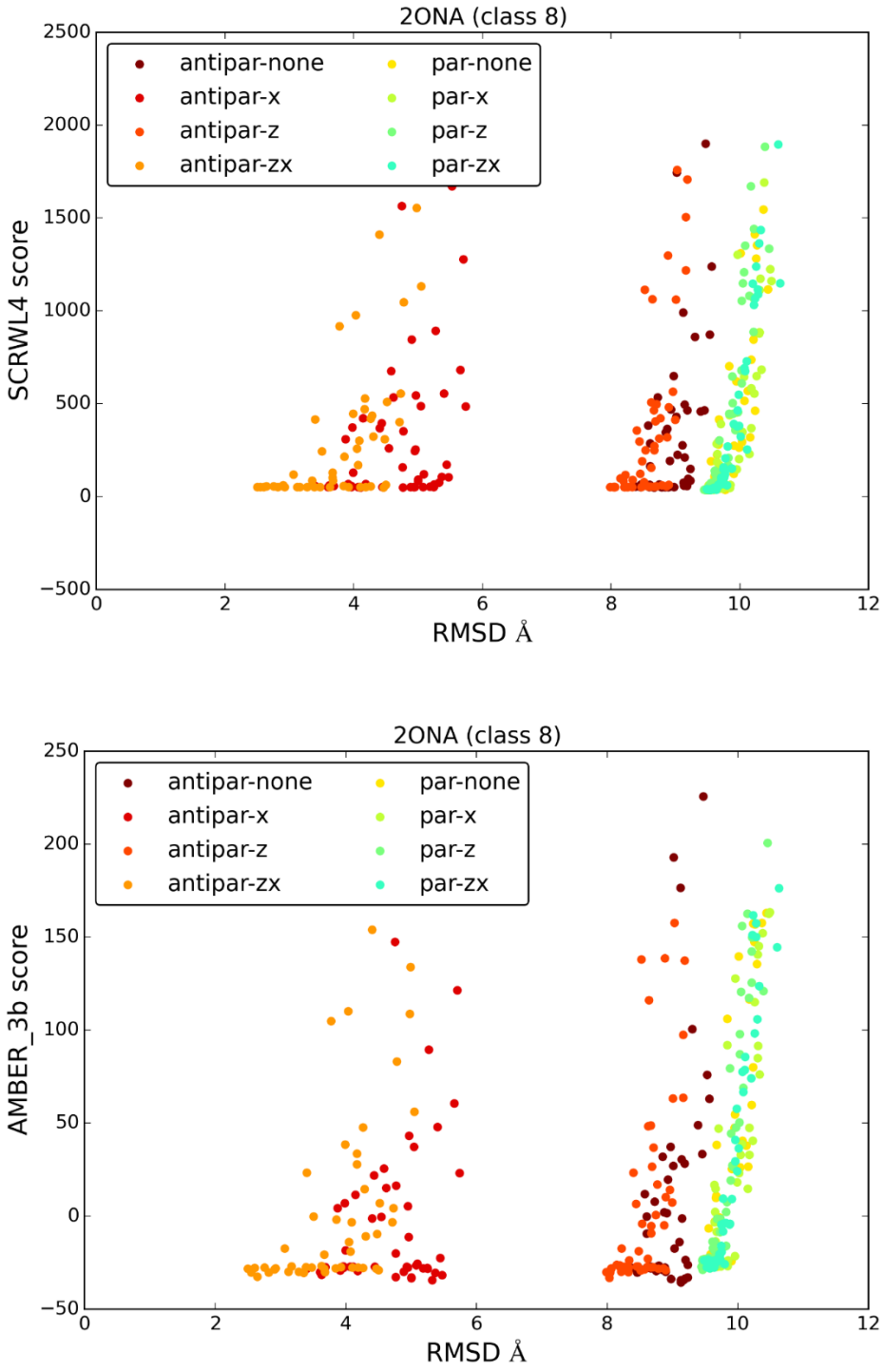

Figure 25 continued

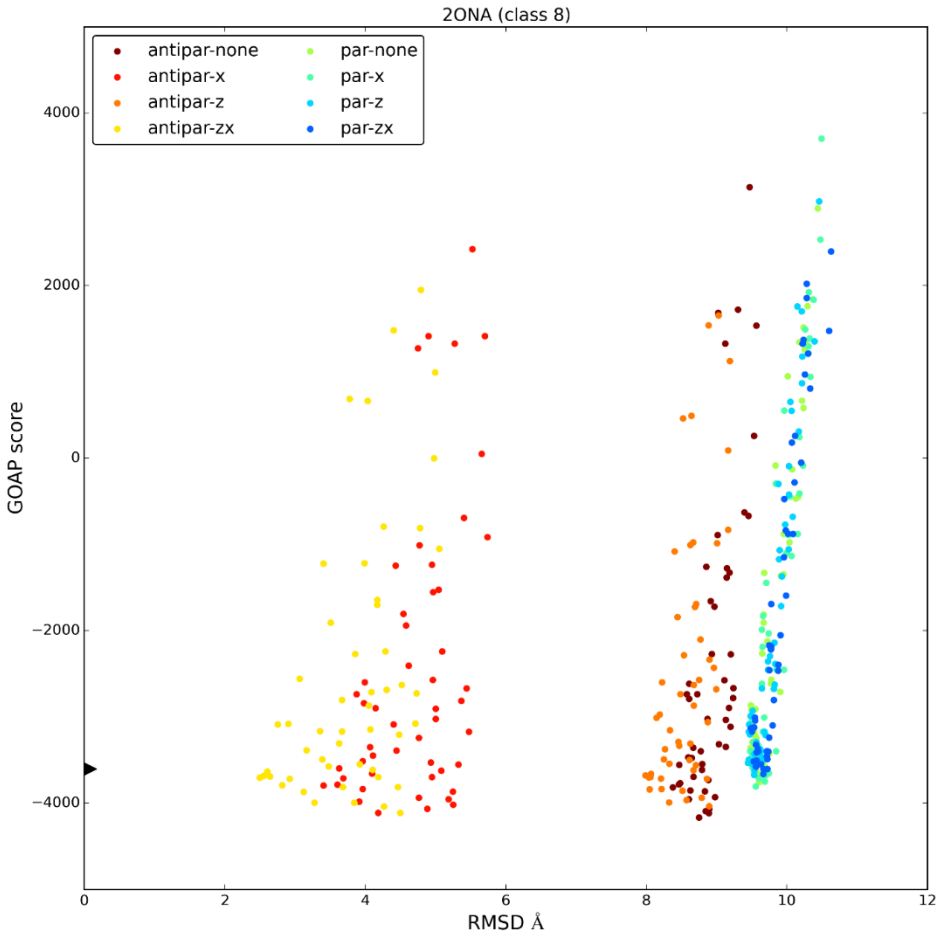

Figure 25 continued

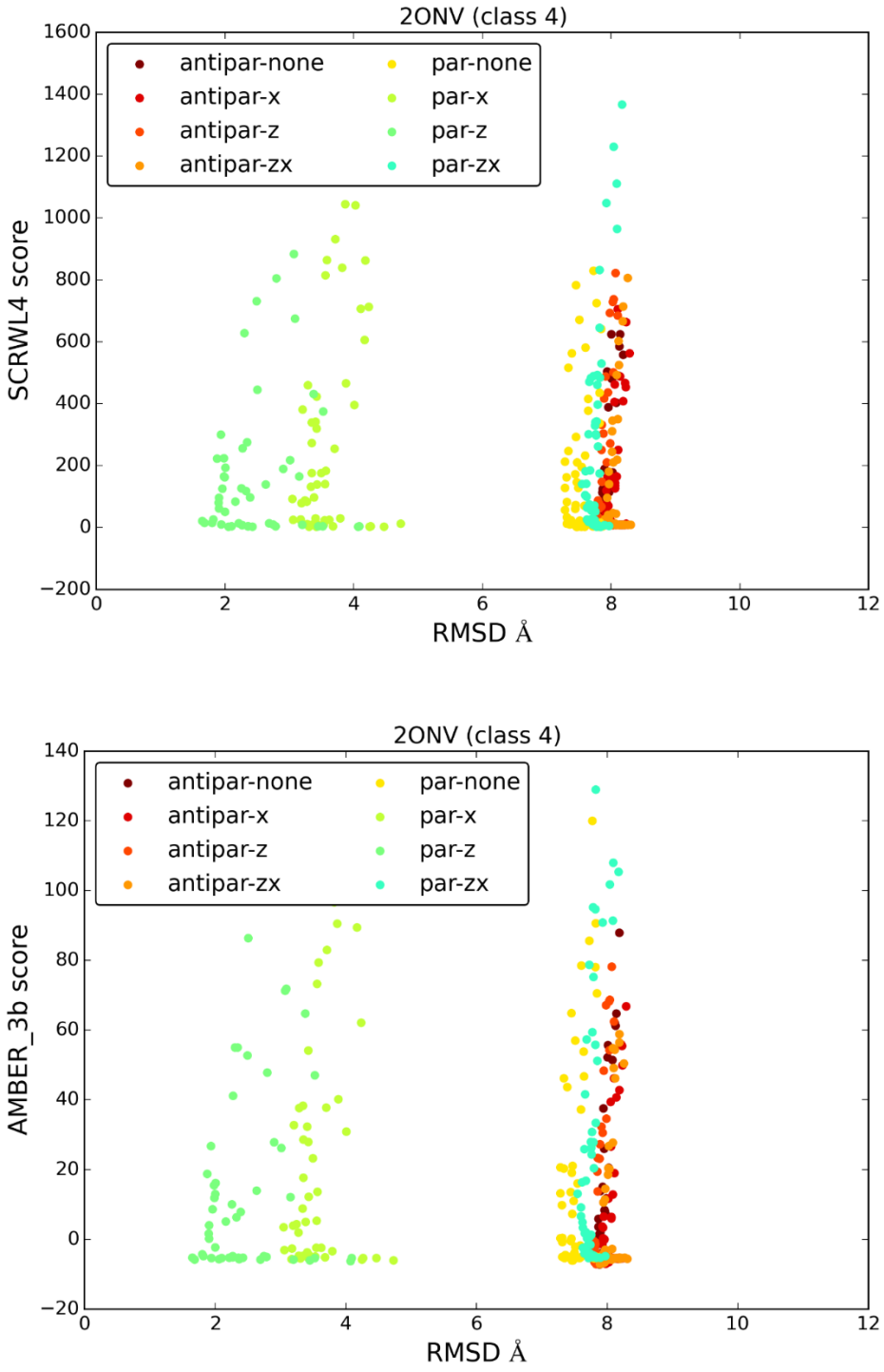

Figure 25 continued

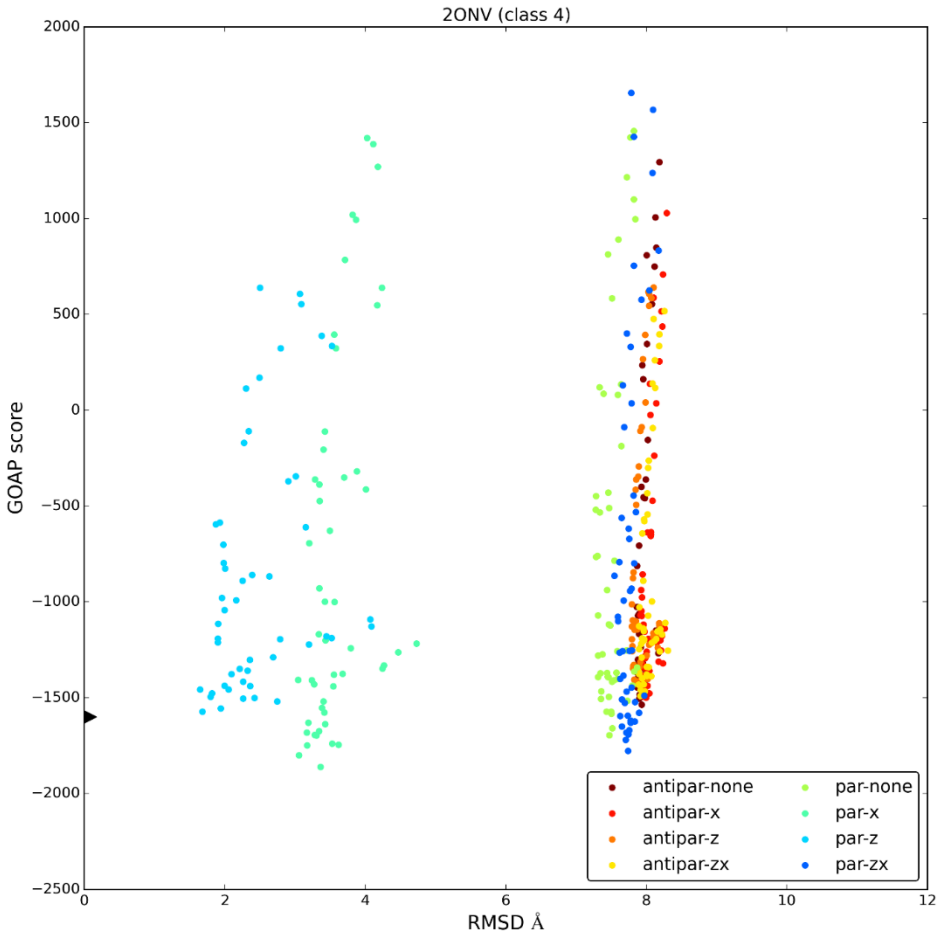

Figure 25 continued

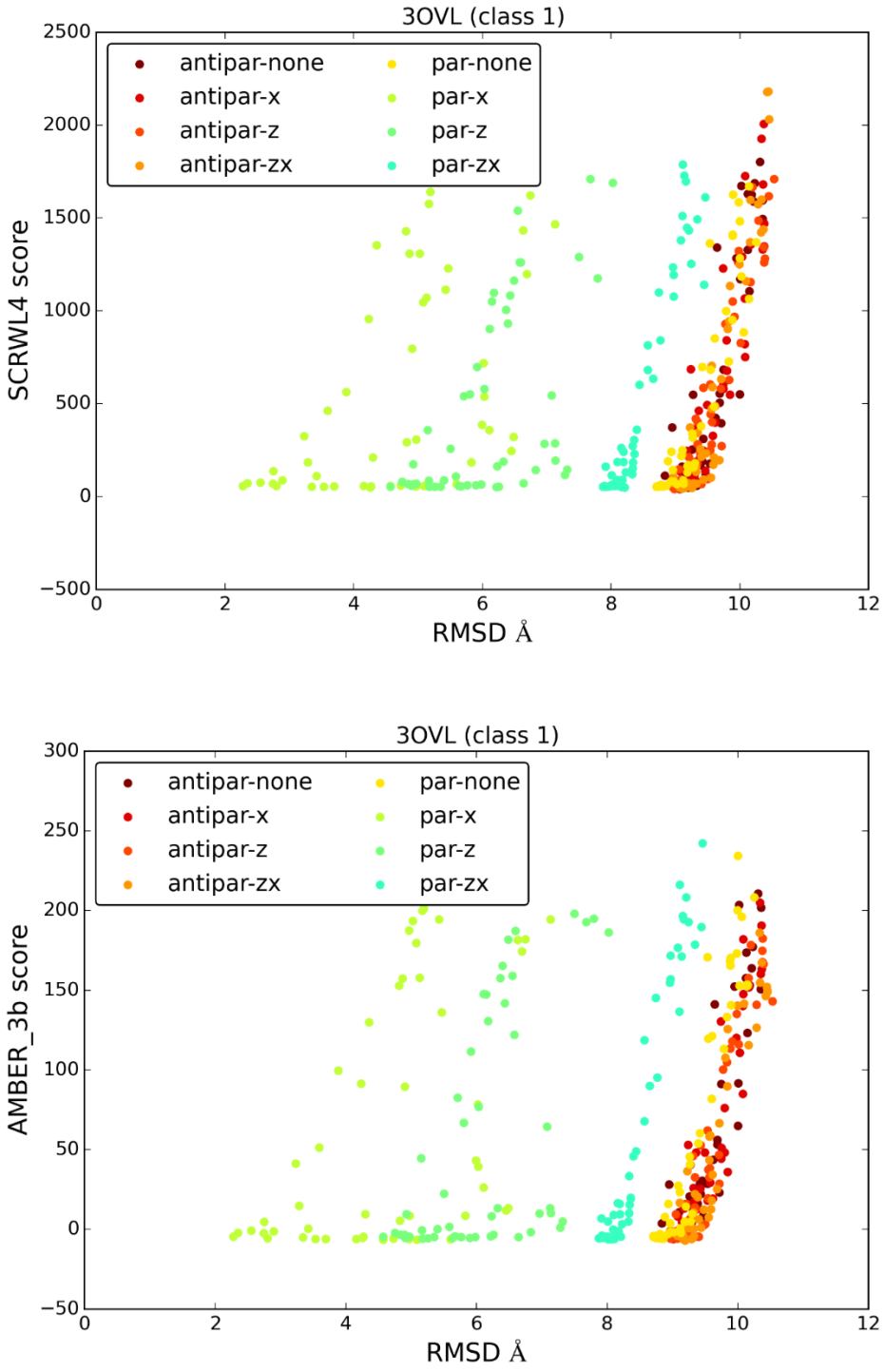

Figure 25 continued

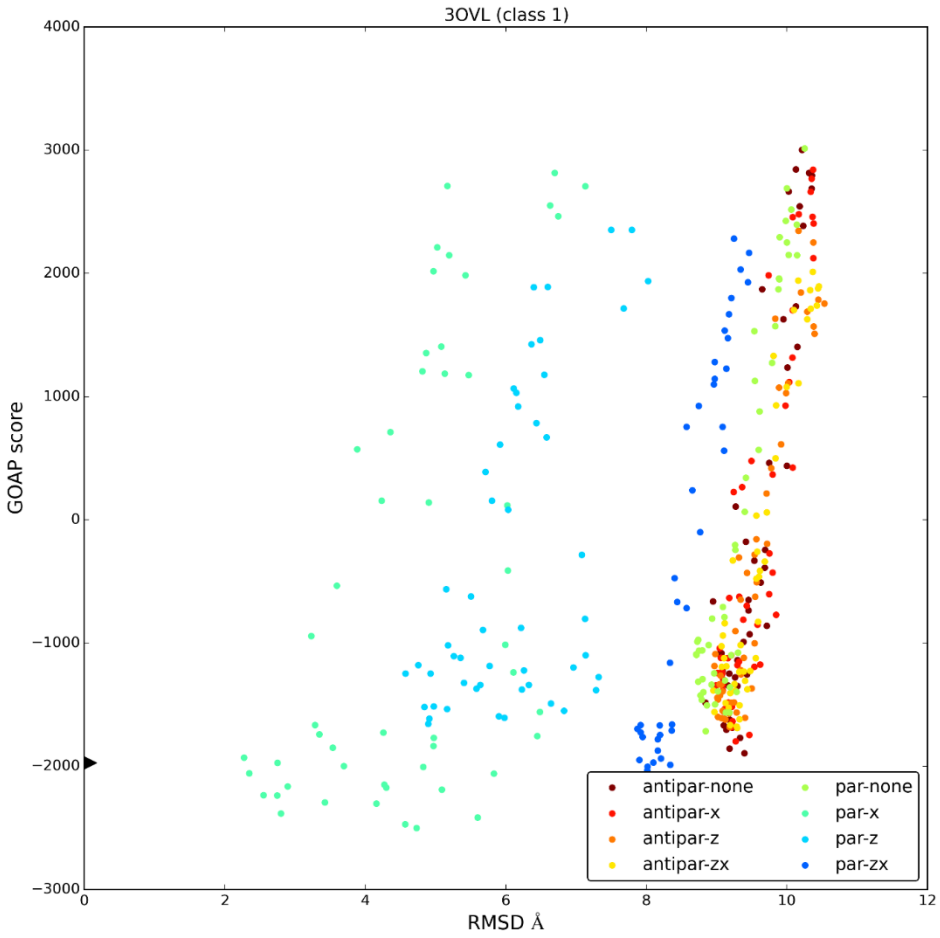

Figure 25 continued

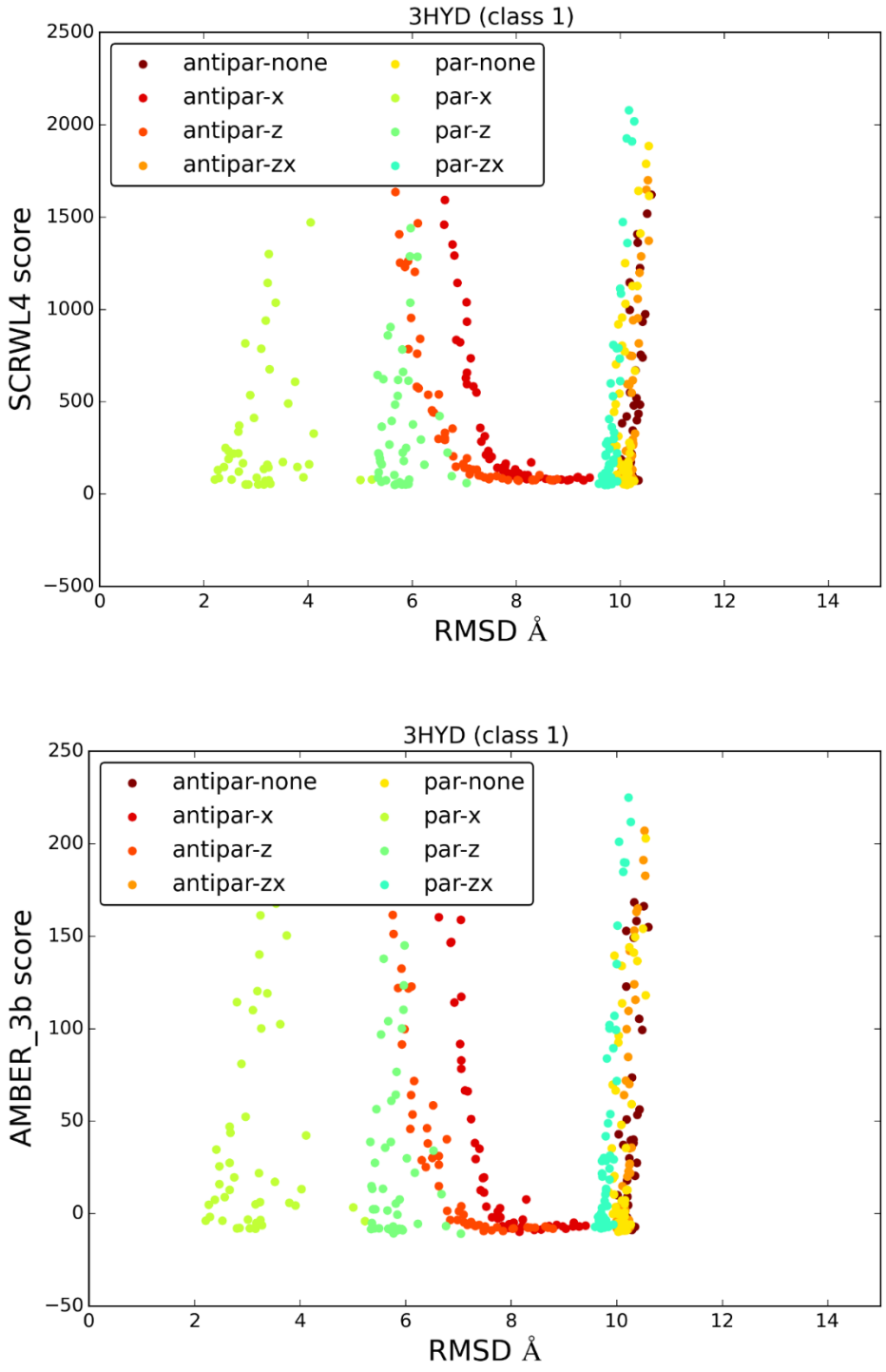

Figure 25 continued

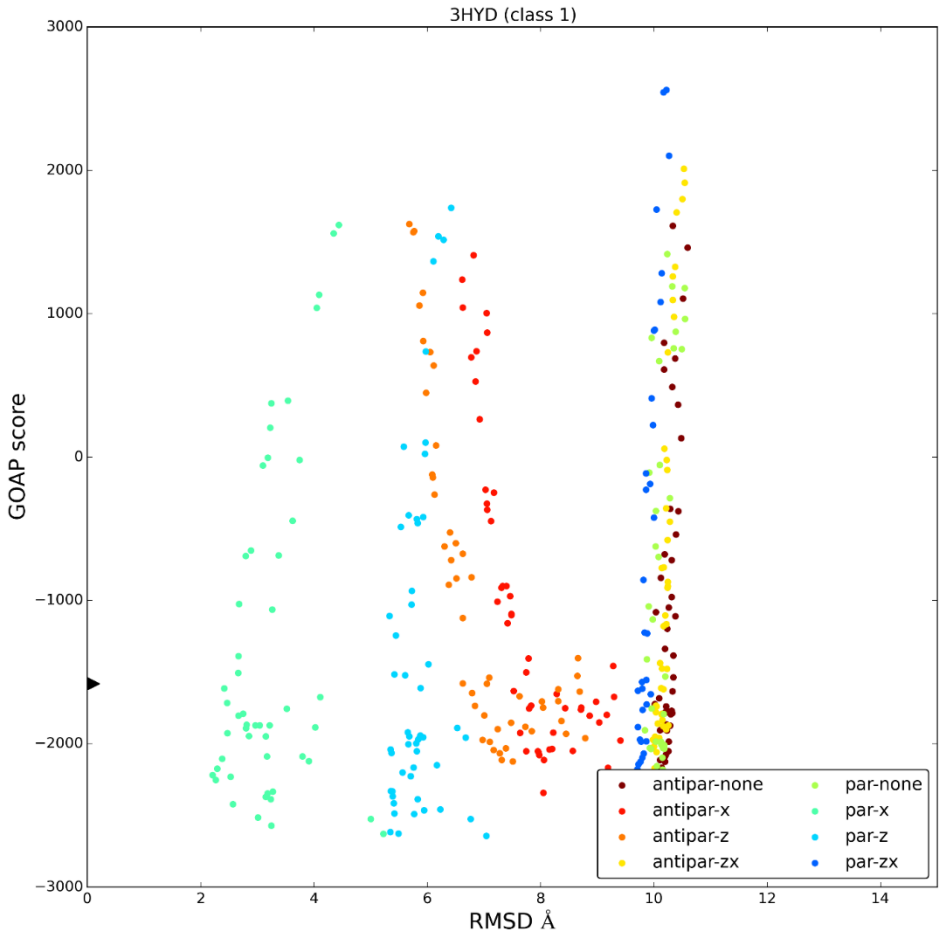

Figure 25 continued

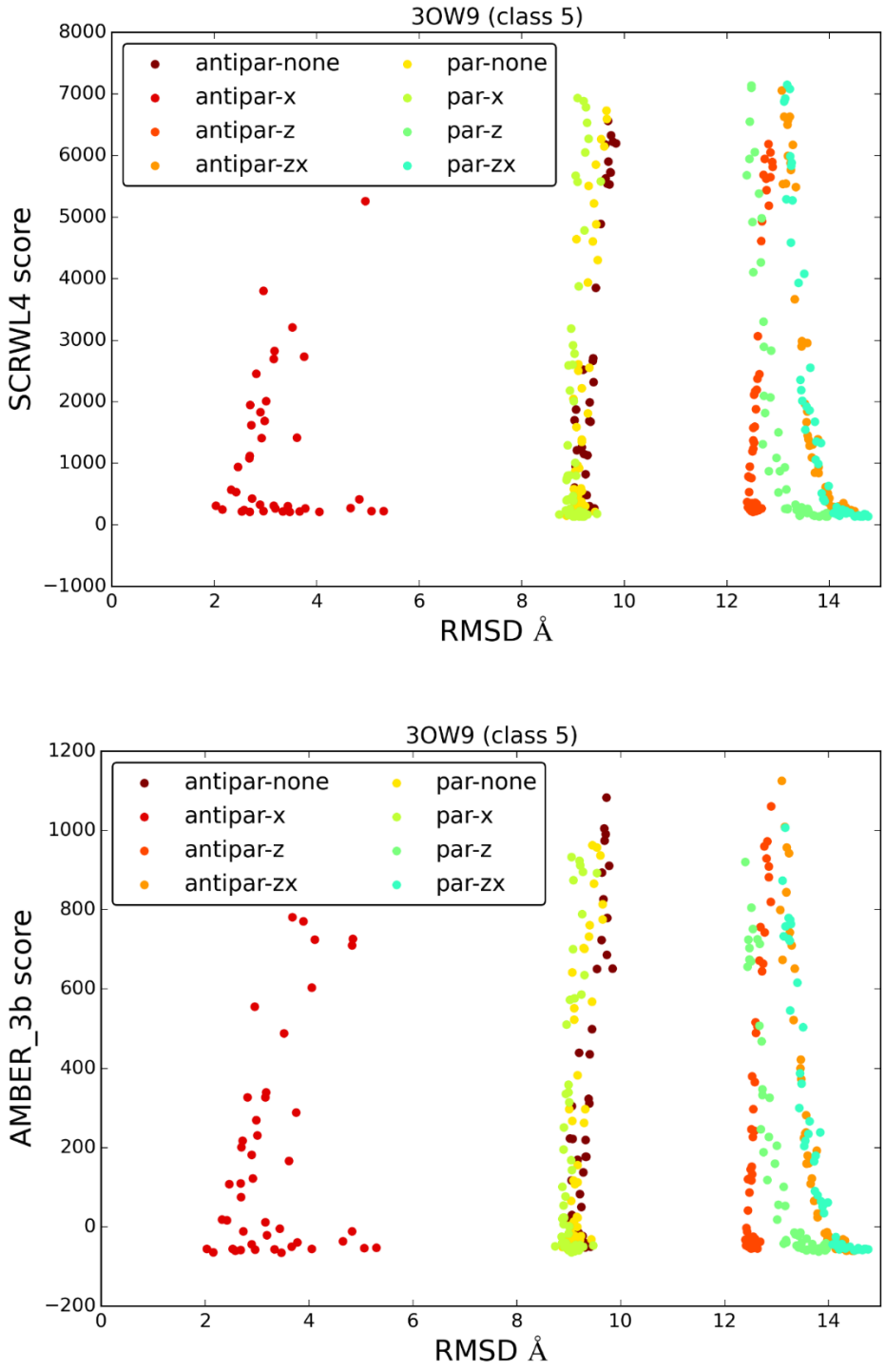

Figure 25 continued

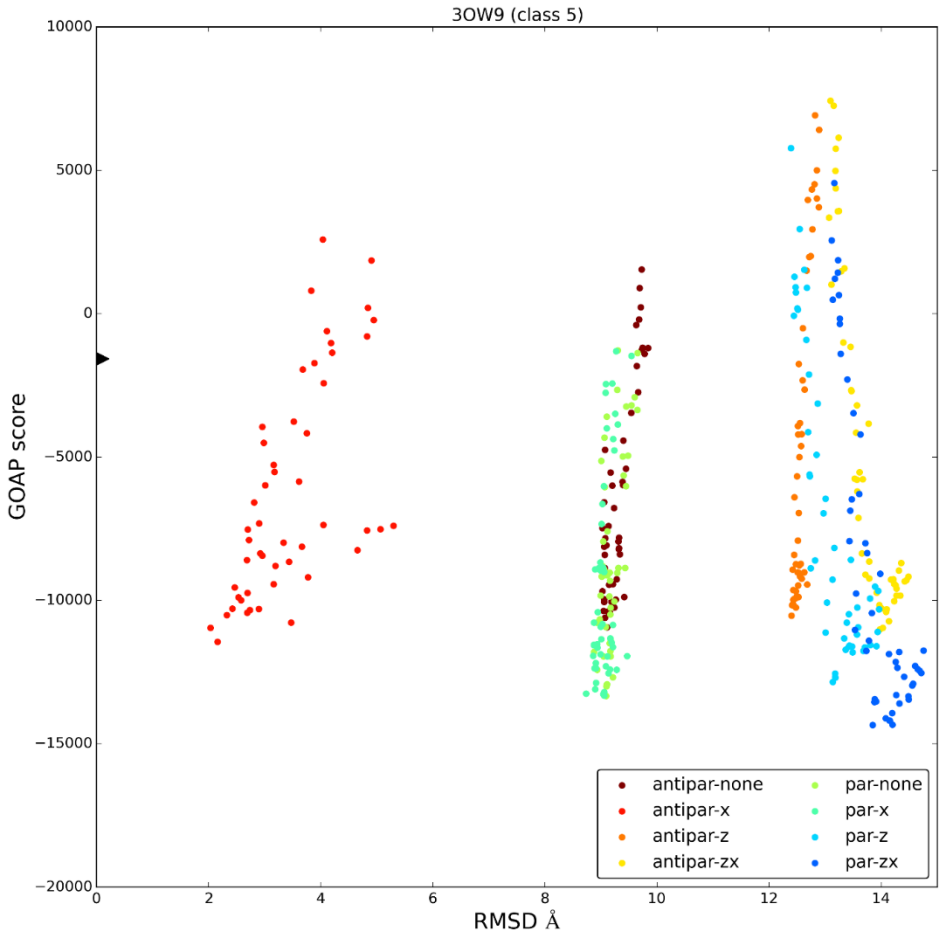

Figure 25 continued

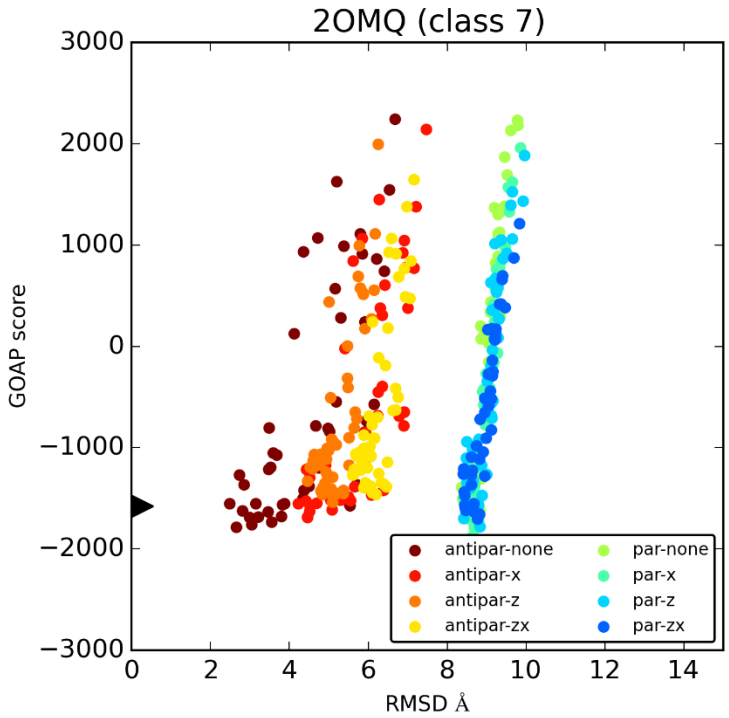

Figure 26 GOAP score vs. RMSD for 2OMQ fibril. The triangle displays the score of the reference PDB structure.

4.4 Conclusions

In this paper, we have reported a general, computationally efficient method for structure prediction of amyloid fibrils. We have demonstrated that native-like amyloid fibril structures can be generated based on the sequence alone. Currently, due to the lack of knowledge about potential structural polymorphism of amyloid fibrils, it is unclear whether the method identifies the most energetically favorable class of amyloid fibril for a peptide sequence, or whether equally favorable amyloid fibril structures for the same sequence exist. Thus, FibPredictor results should be combined with experimental data to determine the sense of the amyloid β-spine [78], to reduce the analysis to a small subset of fibril classes. In such cases, FibPredictor demonstrated the ability to identify the correct amyloid fibril structures among the top-ranked conformations.

The structures generated by our program can also be useful in interpreting experimental data, e.g., by fitting them to SAXS spectra or for interpreting residue interactions observed by NMR. Fibpredictor results can also be combined with more sophisticated but computationally demanding simulation methods to further refine the initial predicted structures, identify potentially important interactions in amyloid fibrils, study mechanical properties of amyloid fibrils[79,80] and quantify the free energies of amyloid fibril stability. Finally, analysis of ensembles of energetically favorable structures generated by FibPredictor can be used to identify important interactions in the steric zipper.

Table 7 Eight classes of amyloid β-fibrils [16] and the rotation operations used by FibPredictor to generate each amyloid class. Figure 19 presents visualization of the different fibril classes.

Class	Sense of β-Sheet	Directions of the two β-sheets	Steric Zipper	Rotation operation
1	Parallel	Up-up	Face-to-face	X
2	Parallel	Up-up	Face-to-back	Z
3	Parallel	Up-down	Face-to-face	ZX
4	Parallel	Up-down	Face-to-back	No rotation
5	Anti-Parallel	Up = down	Face-to-face	X
6	Anti-Parallel	Up = down	Face-to-back	Z
7	Anti-Parallel	Up-up	Face = back	No rotation
8	Anti-Parallel	Up-down	Face = back	ZX

CHAPTER 5. PHOSPHATE ESTER DERIVATIVES OF GLUCAGON

5.1 Introduction

In this chapter stable phosphorylated glucagon derivatives are introduced as glucagon pro-drugs which are soluble in neutral pH. Phosphate groups which can be removed upon administration by serum phosphatases have been successfully used in past to enhance small molecule and peptidomimetic drug solubility and delivery [4–6,81,82]. Also, phosphorylation has been shown to be able to affect fibril formation of small peptides [5,6,83–88]. Based on the idea of phosphate derivate prodrugs and phosphate-mediated fibrillation modulation, we designed stable and soluble phospho-glucagon prodrugs. This design was based on a rigorous computational analysis which suggested that phosphorylation at certain rationally-picked residues can effectively prevent fibrillation. The enhanced solubility and chemical and physical stability of these prodrugs are shown by various methods. Also results show the phosphate group can be removed enzymatically in phosphatase enzyme concentrations close to serum conditions, resulting in free native glucagon.

5.2 Materials and Methods

5.2.1 Phosphorylation Sites and Possible Phospho-glucagon Prodrugs

There are 10 readily phosporhylatable sites on glucagon (i.e., His1, Ser2, Thr5, Thr7, Ser8, Tyr10, Ser11, Tyr13, Ser16, Thr29), which means there are hypothetically 10

singly phosphorylated, 45 doubly phosphorylated and 120 triply phosphorylated possible glucagon prodrugs carrying between one and three phosphate groups, a total of 175 distinct molecules. Allowing for up to ten sites of phosphorylation, the number of distinct phospho-glucagon derivatives increases to 1023. This study is focused on phospho-glucagon derivatives containing only one phosphate group, since these are the simplest to produce and serve to demonstrate the approach.

5.2.2 Computational Modelling of Glucagon Fibrils

Crystal structures of a glucagon fibril have not been resolved yet. Therefore, computational modeling was used to rationally identify strategies for inhibiting glucagon fibrillation. Small angle X-ray scattering (SAXS) [30] and Fourier transform infrared (FTIR) spectroscopy data of glucagon fibril structures[21] were used to limit the possible geometries for glucagon fibrils. FTIR data[21] shows that the glucagon fibril is formed by antiparallel β-sheets. SAXS data shows that the glucagon fibril has a diameter of 45 Å, which is half of the length of a fully extended glucagon. This suggests that glucagon folds onto itself and is not fully extended in its fibril form. Combining the SAXS data with FTIR, only two different fibril classes remain possible for glucagon. Each of these classes can form steric zipper by entanglements of side chains on two sides of the β-sheet which results in four formations in total (Figure 28).

Due to impreciseness inherent to the SAXS data, however, it is not clear how many of the amino acids engage in forming the steric zipper, how many form the loop and how many terminal residues are free and unstructured. With a loop length ranging from 3 to 17 amino acids and allowing zero, one or two free terminal amino acids yields 64 possible fibril formations (Figure 29). Moreover, each of these fibril formations can fold in two

different directions (Figure 28) doubling the number of possible formations. FibPredictor, a program developed in the investigators' lab for computational modelling of steric zipper regions of amyloid fibrils[89], was then used to generate 100 candidate structures for each of 128 formations, amounting to 12,800 structural models covering a comprehensive set of hypothetically possible structures for the geometries compatible with the SAX and FTIR data. Fibpredictor models the steric zipper by first placing the backbone atoms of the two sheets within a user defined minimum and maximum distance (and a certain range of tilting angles). Then the side chains are optimized for each relative position of the two sheets using SCWRL4 [70] and the energy of the final structure model is calculated. Fibpredictor has two options for scoring, GOAP[71] and Amb_3b[73]. For this study, we used the GOAP score to identify energetically favorable models.

From the 128,000 steric zipper models, the top 500 most energetically favorable were investigated by an in-house program for the most frequent inter-residue contacts. To overcome the preference for larger models, the energy was normalized by the number of residues. A pair of residues with any two heavy atom closer than 5 Å to each other were considered as a contact.

5.2.3 MD Simulations

The model of glucagon steric zipper generated by FibPredictor with the lowest average energy per residue among all models (NOP) (Figure 29A) and three phosphorylated analogues (Figure 29B) were simulated to investigate the effect of phosphorylation on the stability of the steric zipper. The phosphorylated analogues represented the phosphorylated steric zipper in three different protonation states: doubly protonated (SEN)[90], singly protonated (S1P) and not protonated phosphate group (SEP)[91]. NOP,

SEN and S1P were simulated in pH 2.5, and SEP was simulated in pH 7.4 to reproduce different experimental conditions of fluorescence studies described below. The proteins were solvated in a pre-equilibrated octahedron of TIP3P water molecules with a minimum distance of 20 Å between the box boundary and any solute atom[37]. Simulations were performed using the AMBER constant pH force field [92].

The shake algorithm was used to constrain hydrogen containing bonds [39]. The simulations were performed in an NPT ensemble. The temperature was maintained at 298 K with a Langevin thermostat[40] with collision frequency of 1 ps^{-1}. Isotropic position scaling with pressure relaxation time of 2 ps was used to maintain pressure at 1 atm. The electrostatic interactions in periodic boundary conditions were treated using the particle mesh Ewald method [38]. The cut-off for van der Waals interactions was set to 10 Å. The integration time step was 2 fs.

The water molecules and peptide were energy minimized first with and then without restraints. The system was then heated from 0 K to 298 K gradually over a 20 ps. The system was then equilibrated in an NPT ensemble for 100 ps. The main production MD runs were performed for 60 ns. 1200 snapshots were saved for each production simulation. Contacts were defined as two residues on the two sides of the steric zipper with closest distance to each other. The terminal amino acids were excluded from this study due to their flexibility. Initial contacts were defined as contacts that were identified in the initial equilibrated structure. Contacts were defined between two closest amino acids on the two sides of the steric zipper. The contacts were tracked over the full simulation length using an in-house python code.

5.2.4 Peptides and their solubility

Research grade human glucagon was purchased from ProSpec (East Brunswick, NJ). Phosphorylated-glucagon derivatives were purchased from GenScript (Piscataway, NJ). Solubility of glucagon derivatives were reported by GenScript.

5.2.5 Stability study (24 h)

Glucagon, phospho-Ser2-, phospho-Thr5- and phospho-Ser8-glucagon were prepared at 1.6 mg/mL in 3.2 mM HCl, 0.9% NaCl (w/v) (pH 2.5) and phospho-Thr5-glucagon and phospho-Ser8-glucagon were prepared at 1.6 mg/ml in 1X phosphate buffer saline (PBS), pH 7.4. Samples were centrifuged at 14,000 rpm for 5 min and filtered through 0.1 μm filters to remove any insoluble material. 100 μL of the filtered samples were quickly transferred to a 96-well black flat bottom microtiter plate in triplicate and incubated with 50 μM ThT final concentration. The final volume was adjusted to 200 μL using the corresponding buffer as mentioned above. The plate was sealed with a crystal clear sealing tape. Fluorescence measurements were carried out in a BioTek Synergy 4 Multi-Detection microplate reader as described below.

5.2.6 Initial stability study (31 days)

Phospho-Thr5- and phospho-Ser8-glucagon were prepared at 1 mg/mL in 50 mM sodium phosphate, pH 7.4. Samples were centrifuged at 14,000 rpm for 5 min and filtered through 0.1 μm filters to remove any insoluble material. 100 μL of the filtered samples were quickly transferred to a 96-well black flat bottom microtiter plate in triplicate and incubated with 50 μM ThT final concentration. The final volume was adjusted to 200 μL using 50 mM sodium phosphate, pH 7.4. For monitoring fibrillation under different temperature conditions, all the samples were prepared in three separate plates. The plates

were sealed with a crystal clear sealing tape and incubated at 5 ˚C, 23 ˚C and 37 ˚C. Fluorescence measurements were carried out at regular intervals for 31 days as described below.

For turbidity measurement, 100 µL of the above filtered samples were quickly transferred to a 96-well crystal-clear microtiter plates in triplicate and the final volume was made up to 200 µL using 50 mM sodium phosphate, pH 7.4. For monitoring aggregation under different temperature conditions, all the samples were prepared in three separate plates. The plates were sealed with a crystal clear sealing tape and incubated at 5 ˚C, 23 ˚C and 37 ˚C. Measurements were carried out as described below.

To determine the chemical stability, 1 mL of the above filtered samples were transferred to 2 mL glass vials which were stored at three temperatures (5 ˚C, 23 ˚C and 37 ˚C).

5.2.7 Satibility study (35 days)

Phospho-Thr5- and phospho-Ser8-glucagon were prepared at 1 mg/mL in 50 mM sodium phosphate, pH 7.4 and 1 mg/mL in 50 mM sodium phosphate with 10^{-4} M EDTA, pH 7.4. Both with EDTA and without EDTA samples were centrifuged at 14,000 rpm for 5 min and filtered through 0.1 µm filters to remove any insoluble material. The samples were aliquoted to vials and sealed under nitrogen gas and stored away from light in room temperature. At regulars intervals, sample vials were taken out to for the measurements described below. Used sample vials were then discarded.

For fluorescence measurements, 100 µL of the filtered samples were quickly transferred to a 96-well black flat bottom microtiter plate in triplicate and incubated with 50 µM ThT final concentration. The final volume was adjusted to 200 µL using 50 mM sodium phosphate, pH 7.4. Plates were also prepared at half of this concentration by transferring

50 μL of the filtered samples microtiter plate and following the same procedure as above. The plates were sealed with a crystal clear sealing tape. Fluorescence measurements were performed as described below.

Same vials were used for turbidity measurement. 100 μL samples were quickly transferred to a 96-well crystal-clear microtiter plates in triplicate and the final volume was made up to 200 μL using 50 mM sodium phosphate, pH 7.4. Measurements were carried out as described below.

5.2.8 ThT fluorescence measurements

Fibrillation was followed by measuring the fluorescence intensity of ThT with the excitation and emission wavelengths set to 440 nm and 482 nm, respectively. For the 24-hour studies, measurements were carried out at 15-min intervals for 24 h at 23°C with 5 s automixing before each reading. For the initial 31 day studies, measurements were carried out every other day for 31 days with 5 s automixing before each reading. For the second 35 days studies measurements were carried out every week for 31 days with 5 s automixing before each reading. Fluorescence signals of over 100,000 (overflow) were set to 100,000 for visualization purposes.

5.2.9 Intrinsic fluorescence measurements

The excitation and emission wavelengths were set to 295 nm and 355 nm, respectively, corresponding to the fluorescence of Trp25. For the 24-hour study, measurement was carried out for 24 h at 23°C at 15-min intervals preceded by 5 s automixing before each reading. For the initial 31 day study, measurement was carried out every other day for 31 days preceded by 5 s automixing before each reading. For the second 35 day study, measurement was carried out every week for 35 days preceded by 5 s automixing before

each reading. Very high fluorescence signals of over 100,000 (overflow) were set to 100,000 for visualization purposes.

5.2.10 Turbidity measurements

The turbidity of the peptide solutions was measured by UV absorbance at 405 nm and 340 nm using a BioTek Synergy 4 Multi-Detection microplate reader (BioTek Instruments, Winooski, VT). UV absorbance at 280 nm and 450 nm (Eq.14) and UV absorbance at 280 nm and 450 nm (Eq.15) were used to calculate the aggregation index-1 (AI1) and aggregation index-2 (AI2) respectively. Measurement was carried out every other day for initial 31 day stability study and every week for the second 35 stability study preceded by 5 s automixing before each reading. The aggregation index was calculated using Eq. 14 and/or Eq. 15.

$$AI1 = 100 \times \left(\frac{Abs\ 450nm}{Abs\ 280nm - Abs\ 450nm}\right) \quad \text{Eq. 14}$$

$$AI2 = 100 \times \left(\frac{Abs\ 340nm}{Abs\ 280nm - Abs\ 340nm}\right) \quad \text{Eq. 15}$$

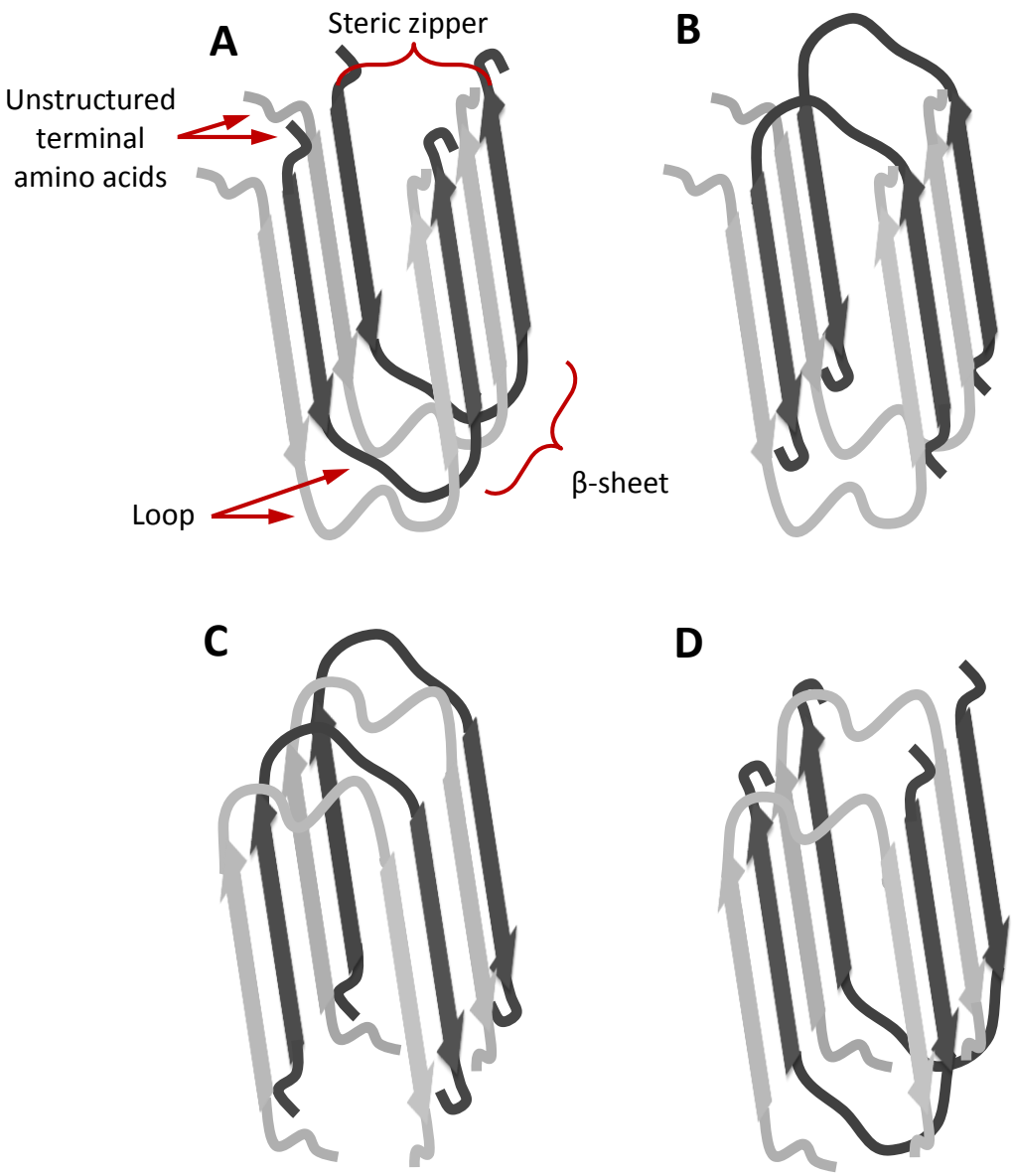

Figure 27: Possible conformations for glucagon fibril according to SAXS and FTIR data. A and B show the two classes of possible formations. Each of these classes can form the steric zipper also on the other side of the sheet resulting in formations shown in C and D.

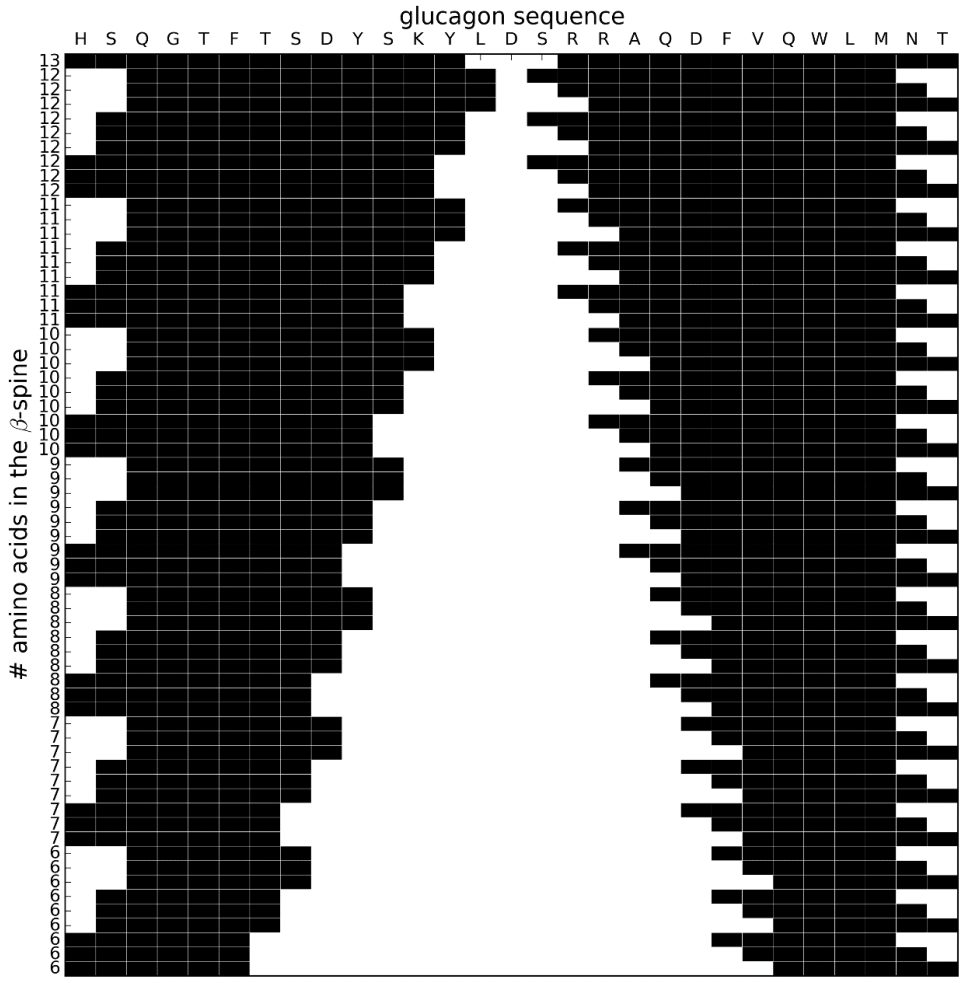

Figure 28: All formations of glucagon steric zipper modelled by FibPredictor. The black blocks show the sequence engaged in the steric zipper.

5.3 Results

5.3.1 Computational analysis

An example of an energetically favorable model of the steric zipper of glucagon fibril is in shown in Figure 29A. The model shown has the lowest average energy per residue among all models. The top 10 most frequent inter-residue contacts in the top 500 most energetically favorable models of the steric zipper region of the glucagon fibril are shown in Table 8. The top three most frequent contacts, Trp25-Phe6, Val23-Phe6 and Trp25-Gly4 are of hydrophobic nature. In addition, hydrophobic residues such as Phe6, Val23 and Trp25 are involved in seven out of the ten most frequent contacts, which confirms the importance of hydrophobic interactions within the steric zipper. It is also observed that four residues (Ser2, Thr5, Ser8 and Tyr10) which are involved in the top-10 most frequent contacts can be phosphorylated. Based on this contact analysis, out hypothesis was that the addition of a phosphate group on these four residues will insert a charged and highly hydrophilic group into the core of a highly hydrophobic steric zipper, thus "opening" the zipper and inhibiting fibril formation. Moreover, the charged phosphate groups are expected to increase the solubility of the peptide.

5.3.2 MD Simulations of the Steric Zipper Model with and without Phosphorylation

Figure 30 shows the percentage of the initial contacts lost over the course of the simulation of a model of the steric zipper (NOP) and its doubly protonated (SEN), singly protonated (S1P) and doubly charged (SEP) phosphorylated analogues. FibPredictor is designed to generate energetically favorable steric zippers and therefore, native contacts are supposed to contribute to fibril formation. Loss of initial contact therefore, suggest a potential instability of the steric zipper. NOP loses less than 10% of its initial contacts

over the course of simulation and the steric zipper maintains its original formation. SEN and S1P, which reproduce different protonation species of the phosphorylated analogues in pH 2.5, lose around 10% and 15% of initial contacts. It should be noted that according to the pKa of phosphoserine[93]the dominant species in pH=2.5 is the doubly protonated analogue and therefore, the steric zipper of phosphorylated analogues at this pH is nearly as stable as native glucagon. This observation is in line of fibril formation of phos-Ser8-glucagon in pH=2.5. Nevertheless, SEP in pH=7.4 loses over 20% of its initial contacts suggesting that the steric zipper is less stable at this pH value. This observation is in agreement with the experimental results showing no fibrillation for phos-Ser8-glucagon at pH=7.5.

5.3.3 Solubility

While glucagon is not soluble in pH 7.4, two of the glucagon derivatives, phospho-Thr5- and phospho-Ser8-glucagon are soluble (10 mg/ml and 8 mg/ml respectively) in neutral pH (Table 9). The solubility values presented here are according to reports by GenScript. More accurate solubility measurements are underway in Dr. Elizabeth Topp's lab.

5.3.4 Fluorescence measurements over 24 hours

Fluorescence measurements over 24 hours are shown in Figure 31A-D. Interaction of ThT with amyloid β-fibrils results in an increase in the ThT fluorescence signal and allows amyloid β-fibril formation to be probed. In pH 2.5 (Figure 31A), native glucagon begins to fibrillate after a lag time of approximately 8 hours. Glucagon rapidly goes to complete fibrillation after this lag time and the ThT signal reaches a plateau after approximately 16 hours. The phosphorylated prodrugs also fibrillate at this pH but with a longer lag time of approximately 15 hours. However, at pH 7.4 (Figure 31B) phospho-

Thr5- and phospho-Ser8-glucagon show no fibrillation over 24 hours and the ThT signal remains low for the period of study. Native glucagon and phospho-Ser2-glucagon cannot be tested for fibrillation at pH 7.4 since they are not soluble at this pH. This demonstrates that while glucagon and the phosphorylated prodrugs studied fibrillate under acidic conditions, phospho-Thr-5- and phospho-Ser8-glucagon do not fibrillate in neutral pH over 24-hours. A decrease in the Trp intrinsic fluorescence signal indicates oligomerization of the peptide. In pH 2.5 (Figure 31C), glucagon intrinsic fluorescence shows a sudden decrease after a lag time of nearly 9 hours. Similar behavior is observed for phospho-glucagon prodrugs at pH 2.5, but with longer lag times of approximately 18 hours (phospho-Ser8-glucagon) and 21 hours (phospho-Ser2- and phospho-Thr5-glucagon). Nonetheless, at 7.4 (Figure 31D) phospho-Ser8- and phospho-Thr5-glucagon intrinsic fluorescence signals remain high with no decreasing trend, which indicates a lack of oligomerization for these peptides at pH 7.4.

5.3.5 ThT fluorescence measurements over the initial 31-day stability study

Figure. 32A-C show the results for ThT assays for 31 days. As mentioned above, upon interaction of ThT with amyloid fibrils, the ThT fluorescence signal increases and allows identification of amyloid fibril formation. The ThT fluorescence remained low for samples stored at 5˚C (Figure 32A), 23˚C (Figure 32B) and 37˚C (Figure 32C) for 31 days. This indicates lack of fibrillation in these samples over the extended time period and the three temperatures studied.

5.3.6 ThT fluorescence measurements over the 35-day stability study

Figure 33 shows the results for ThT assays for 35 days. As mentioned above, upon interaction of ThT with amyloid fibrils, the ThT fluorescence signal increases and allows

identification of amyloid fibril formation. The ThT fluorescence remained low for samples stored with and without EDTA for 35 days. This indicates lack of fibrillation in these samples over the extended time period and the three temperatures studied.

5.3.7 Intrinsic fluorescence measurement over the initial 31-day stability study

Figure 34A-C show results from Trp intrinsic fluorescence measurements over 31 days. A decrease in the Trp intrinsic fluorescence signal indicates oligomerization of the peptide. No such decrease was observed in the Trp fluorescence signal for samples stored at 5˚C (FIG. 4A), 23˚C (FIG. 4B) and 37˚C (FIG. 4C) for 31 days. This indicates no oligomerization at any of the incubation temperatures over 31 days.

5.3.8 Intrinsic fluorescence measurement over the 35-day stability study

Figure 35A-C show results from Trp intrinsic fluorescence measurements over 35 days. A decrease in the Trp intrinsic fluorescence signal indicates oligomerization of the peptide. No such decrease was observed in the Trp fluorescence signal for samples stored with or without EDTA for 35 days. This indicates no oligomerization over 35 days in either of the formulations.

5.3.9 Turbidity measurement over the initial 31-day stability study

Figure 36A-C show the results of aggregation index measurements over 31 days. Proteins do not absorb UV light at 450 nm. Any absorbance observed in this wavelength is generally due the light scattering by particles resulting from aggregation, and the aggregation index-1 (AI-1) helps quantify this. AI-1 values remained below 5 for samples stored at 5˚C (Figure 36A), 23˚C (Figure 36B) and 37˚C (Figure 36C) for 31 days. This indicates that no significant turbidity was observed for either of the two phospho-glucagon peptides at any of the incubation temperatures.

5.3.10 Turbidity measurement over the 35-day stability study

Figures. 37 and 38 shows the results of AI-1 and AI-2 respectively over 35 days. Proteins do not absorb UV light at 450 nm and 340 nm. Any absorbance observed in these wavelength is generally due the light scattering by particles resulting from aggregation, and the aggregation indices (AI-1 and AI-2) helps quantify this. AI values remained below 5% for samples stored with or without EDTA for 35 days. This indicates that no significant turbidity was observed for either of the two phospho-glucagon peptides in any of the two formulations.

5.3.11 Visual Inspection of Vials in the Second Stability Study

FIGs 10A-D show photographs of sample of the second stability study up to 28 days. No turbidity or visible particles were observed in the vials of phospho-Ser8-Glucagon and phospho-Thr5-Glucagon with or without EDTA.

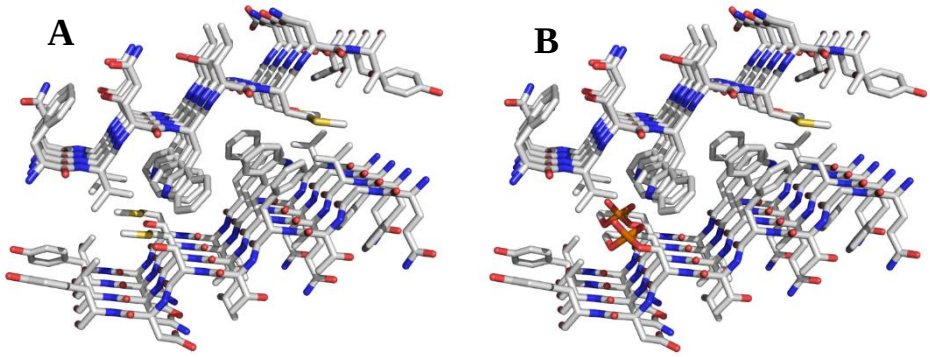

Figure 29: A) An example of energetically favorable steric zipper models generated by FibPredictor. B) Same model phosphorylated at Ser-8.

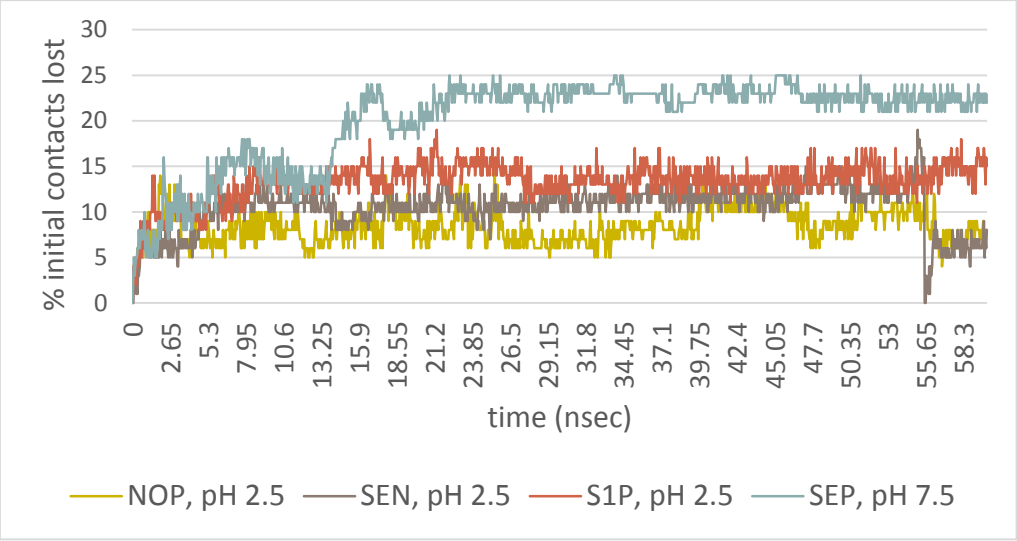

Figure 30: percentage of the native contacts lost over the course of the simulation of a model of steric zipper (NOP) and its doubly protonated (SEN), singly protonated (S1P) and doubly charged (SEP) phosphorylated analogues in different pH conditions

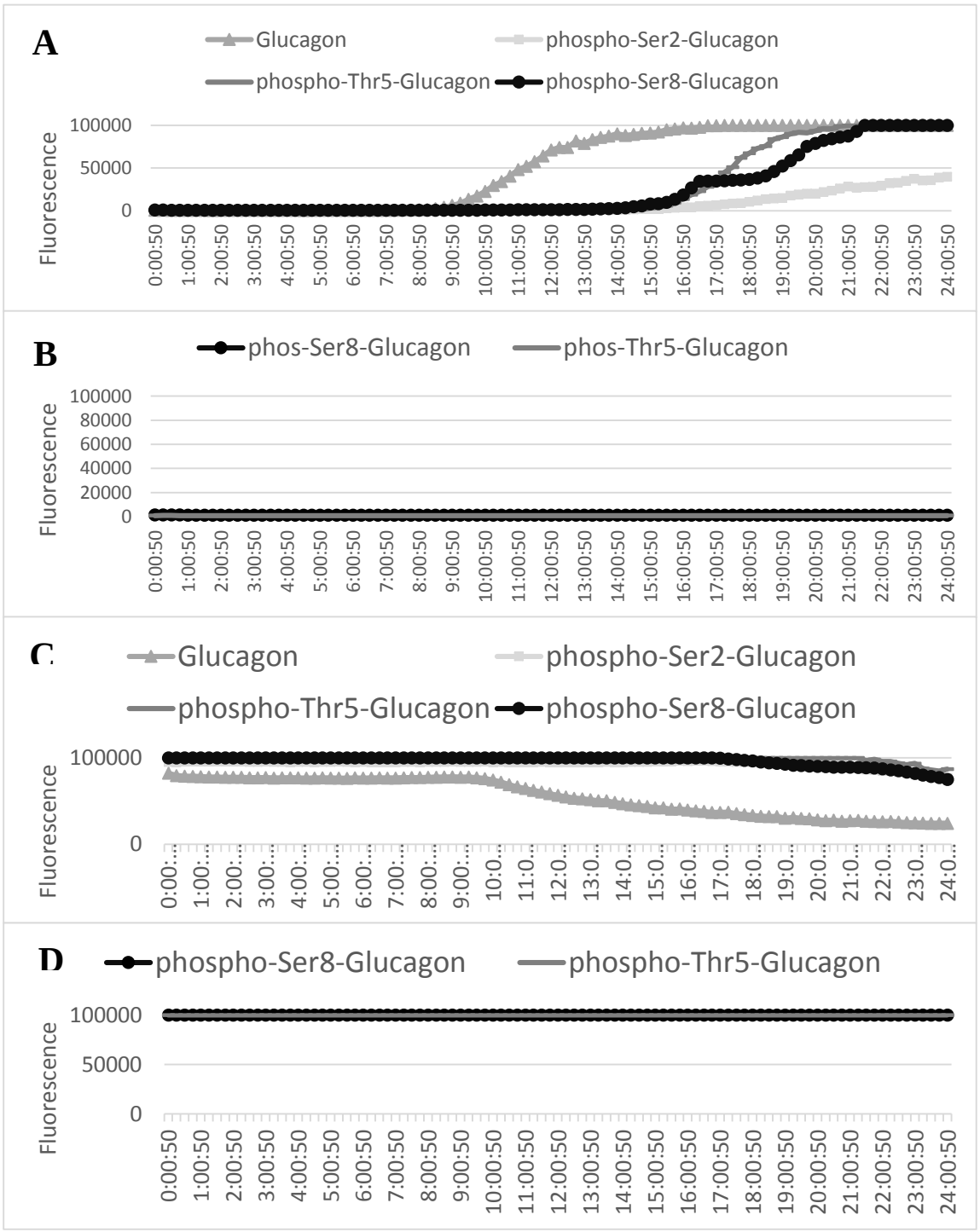

Figure 31: Fluorescence measurements over 24 hours

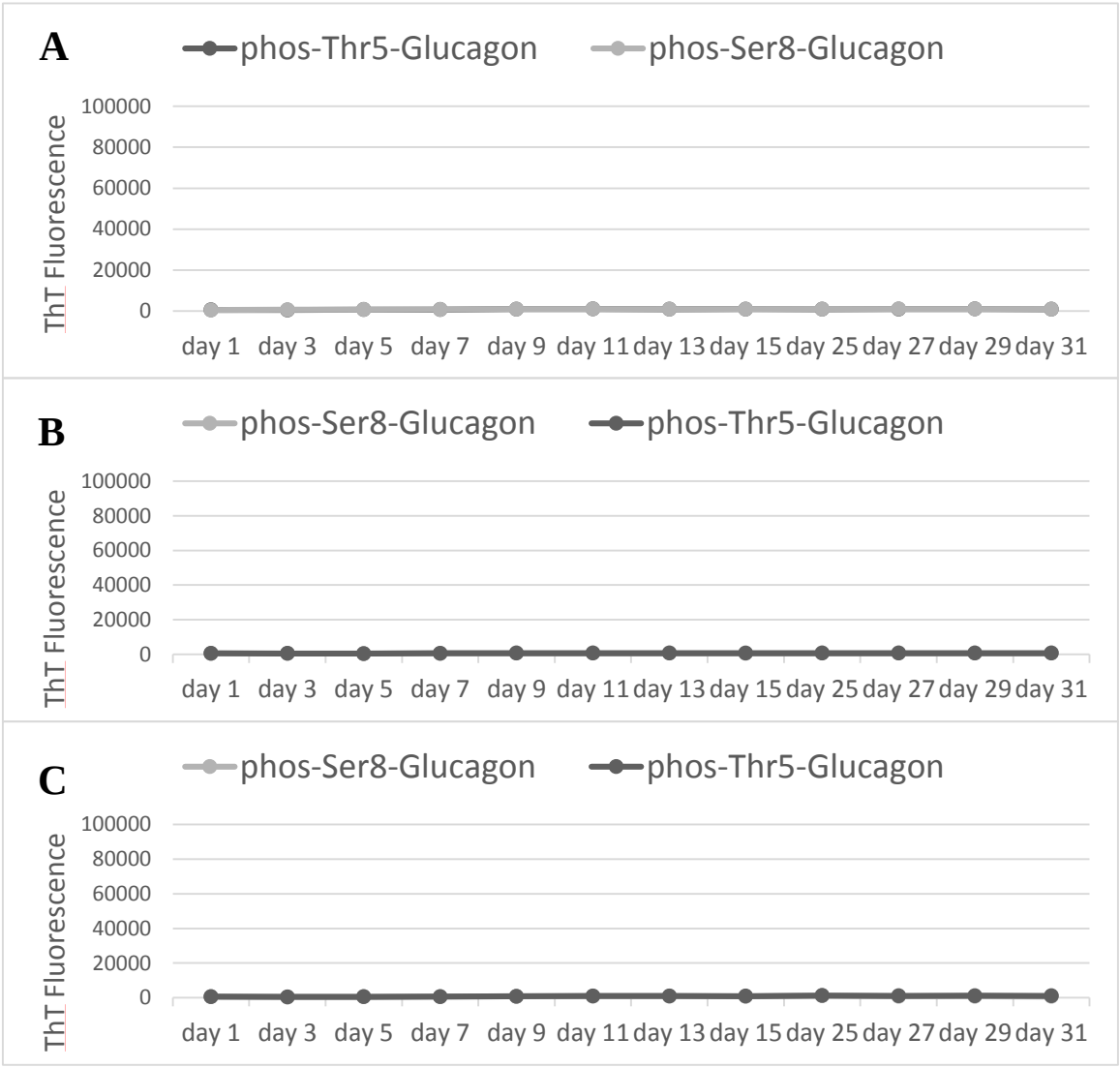

Figure 32: ThT assay of the initial 31-day stability study in A) 5℃, B) 23℃ and C) 37℃

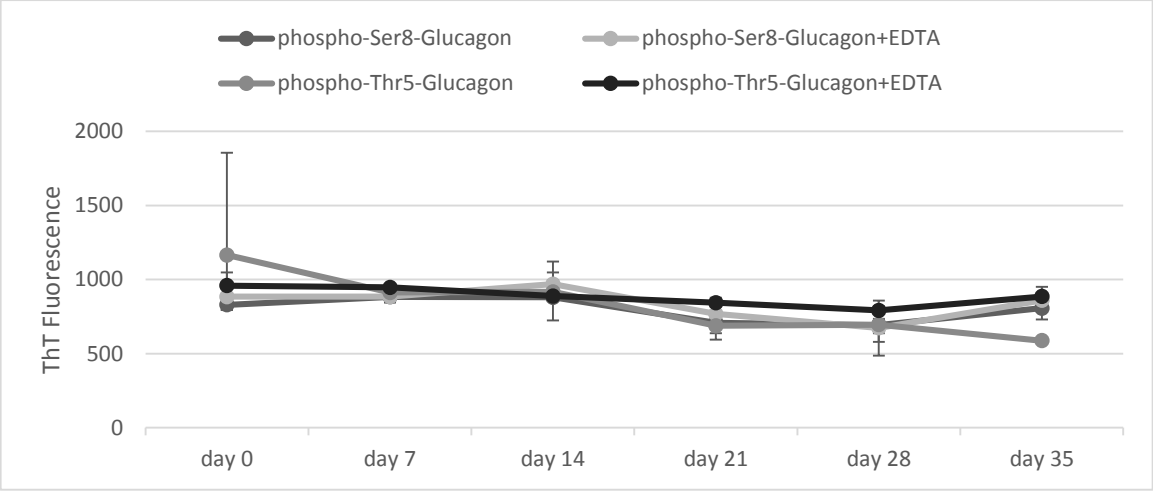

Figure 33: ThT assay of the 35-day stability study

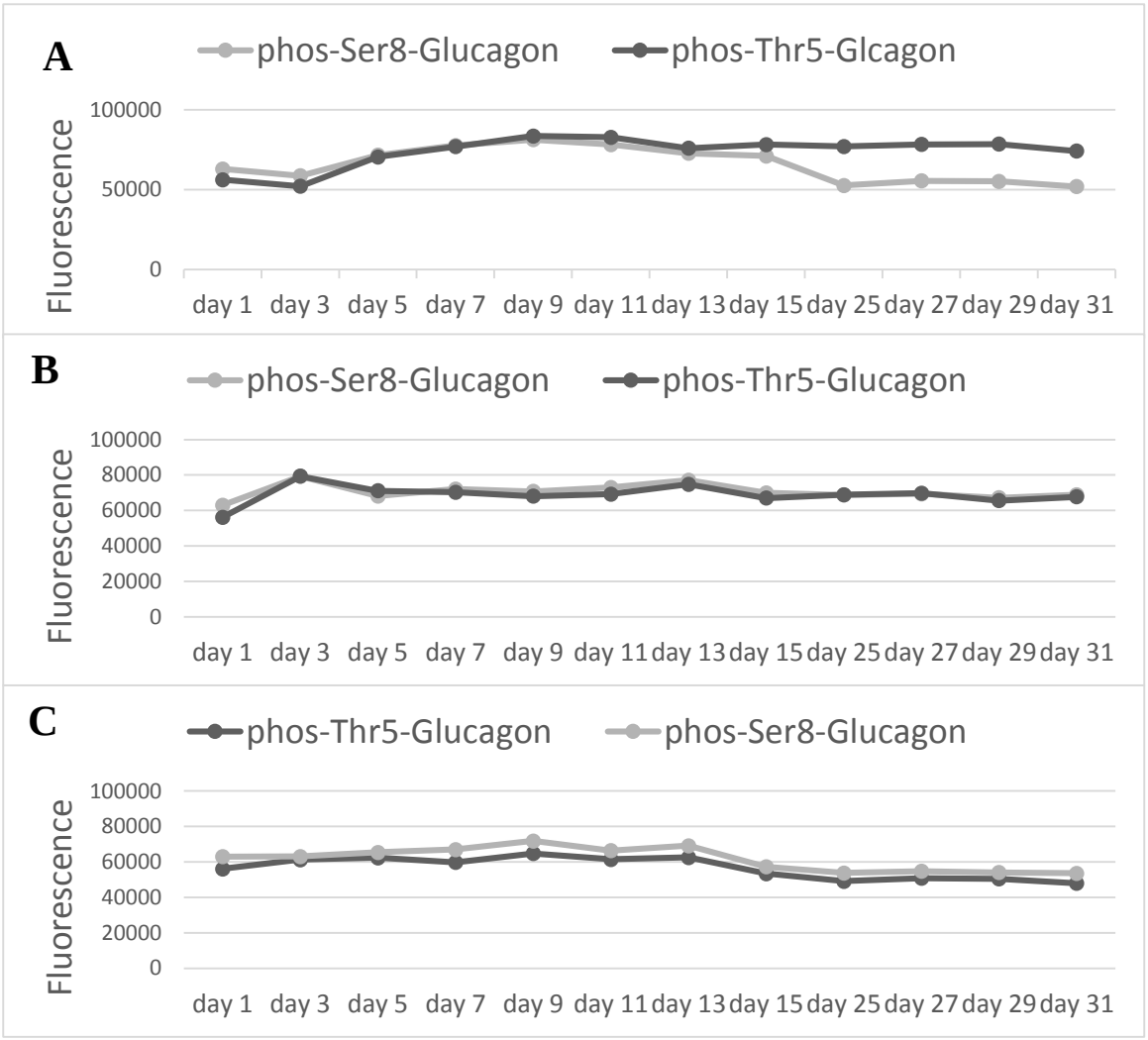

Figure 34: Intrinsic fluorescence assay of the initial 31-day stability study in A) 5℃, B) 23℃ and C) 37℃.

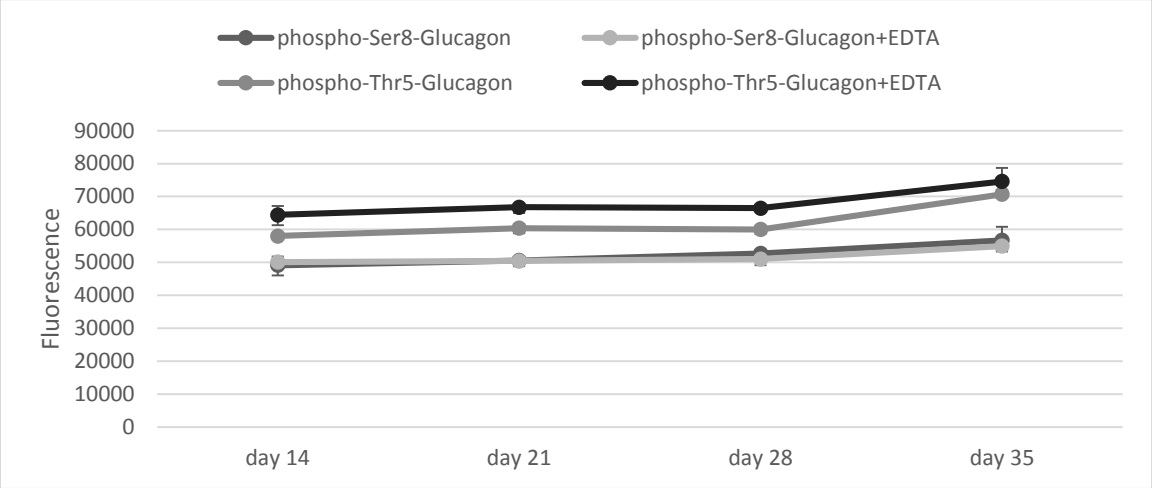

Figure 35: Intrinsic fluorescence assay of the 35-day stability study

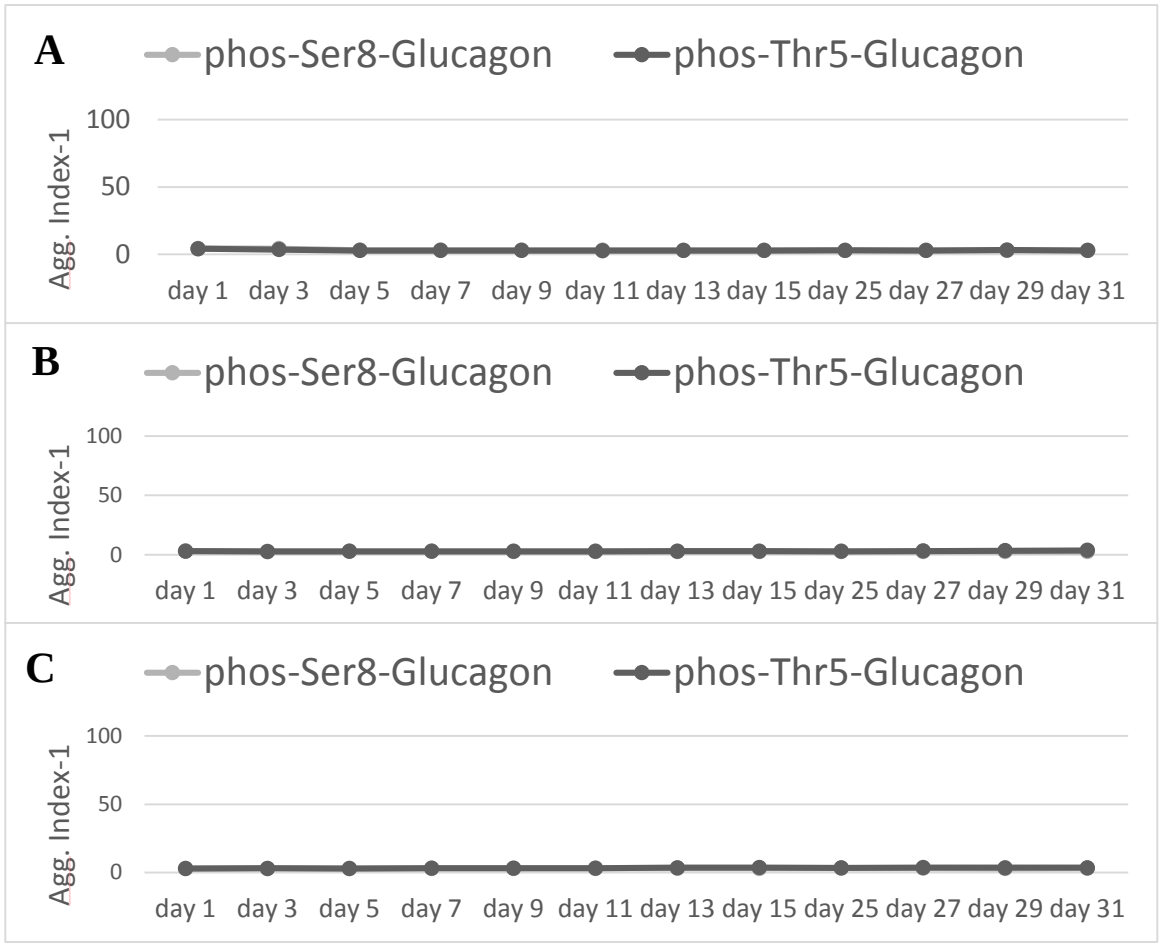

Figure 36: Aggregation index over 31-day initial stability study study in A) 5℃, B) 23℃ and C) 37℃.

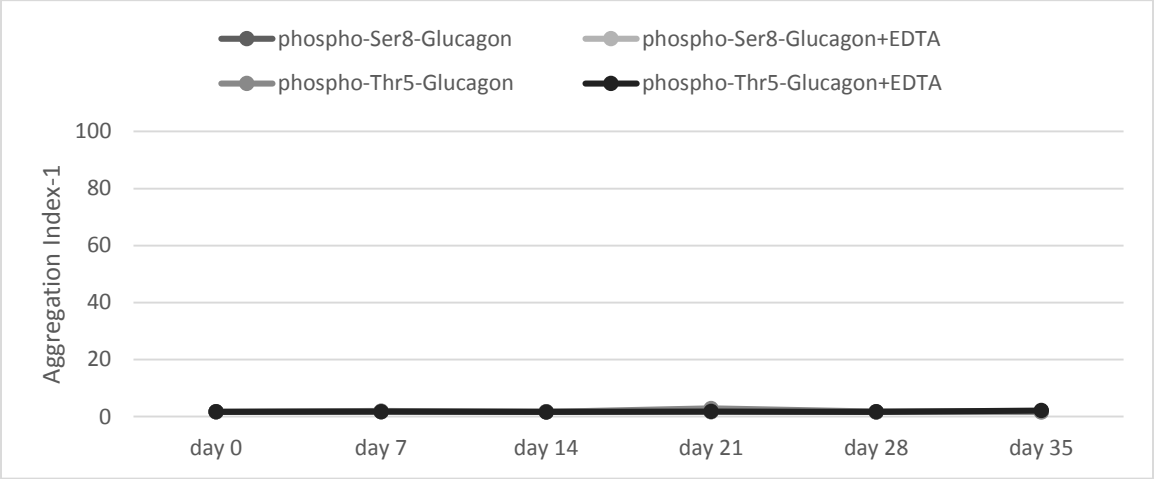

Figure 37: Aggregation index-1 over 35 days

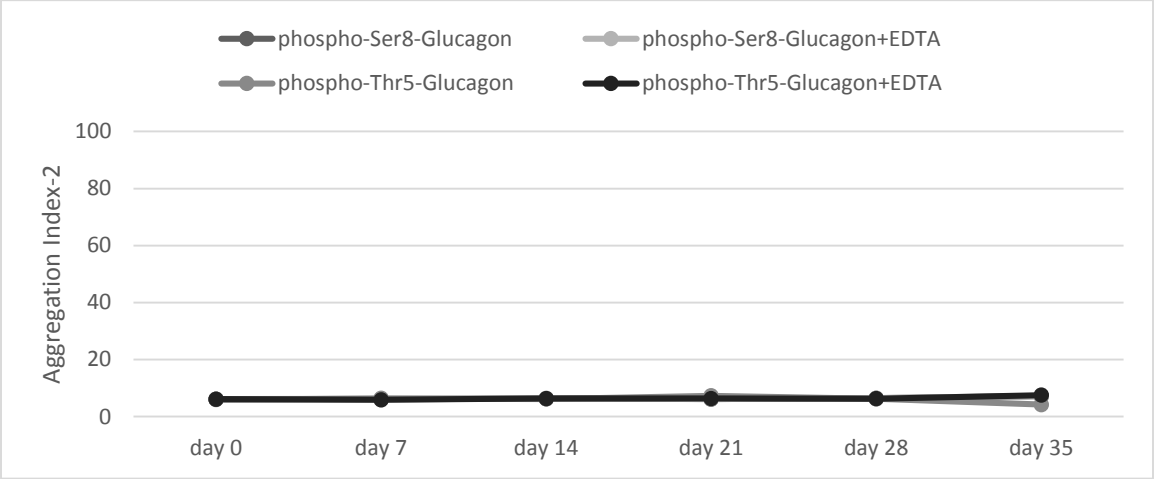

Figure 38: Aggregation index-2 over 35 days

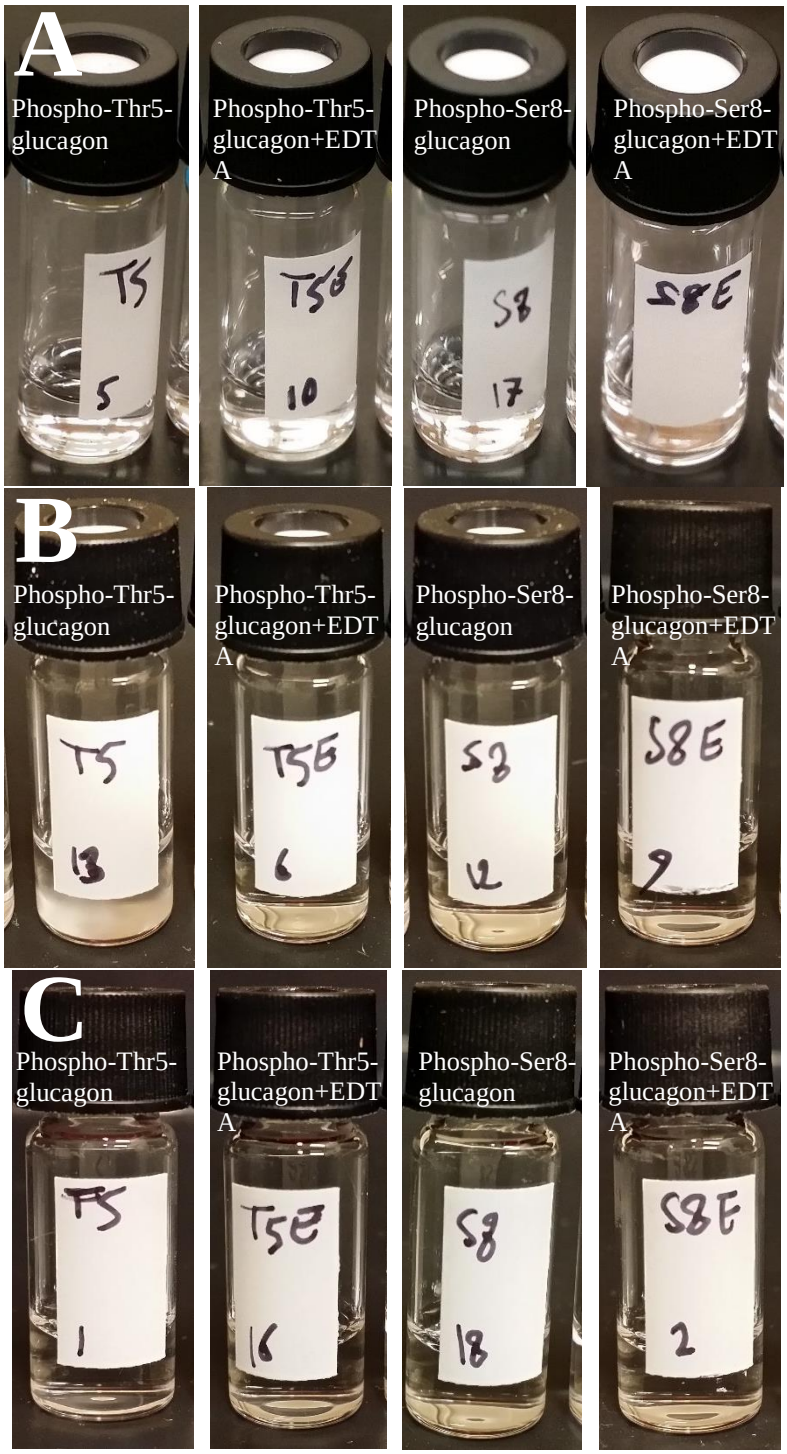

Figure 39: photographs of sample of the second stability study on day A) 7, B) 14, C) 21 and D) 28 and E) 35. Samples remain clear with no visible particle.

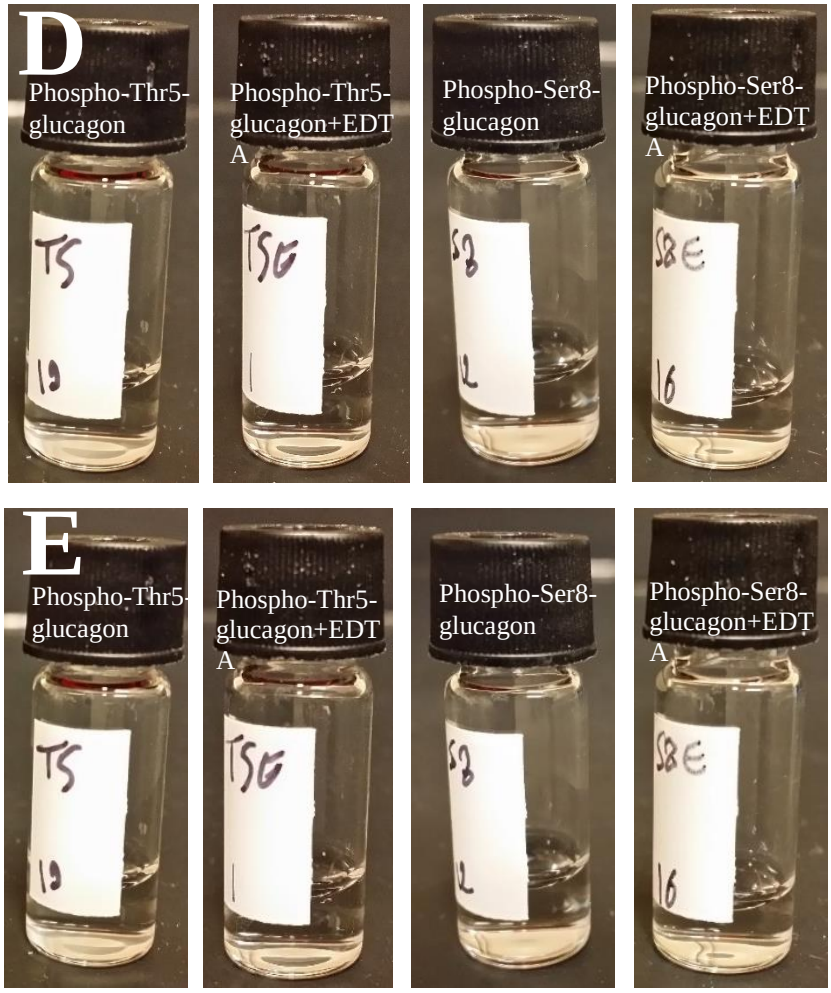

Figure 39 continued

5.4 Discussions

Computational modelling of the glucagon fibril steric zipper suggests that phosphorylation on Ser8 and Thr5 can effectively inhibit fibrillation by introducing an anionic charged group into the hydrophobic entanglements of sidechains between the two β-sheets. This computational prediction is verified by experiments that show phospho-Ser8- and phospho-Thr5-glucagon are soluble and stable. Neither phospho-Thr5-Glucagon nor phospho-Ser8-Glucagon shows fibrillation and neutral pH solution of both remain clear with no turbidity for more than one month. This indicates the potential of these molecules to be formulated as injection pen or for use in artificial pancreas devices. The fact that phospho-Ser2-glucagon fibrillates in both acidic and neutral pH shows that phosphorylation inhibits fibrillation in a site-specific way and merely hanging a charged group on glucagon is not enough for preventing its fibrillation. The charged group should be placed on the correct residue.

Phosphate derivatization increases the net charge of glucagon and, consequently, increases the solubility of glucagon at neutral pH. As a result, while glucagon is not soluble at neutral pH and should be solubilized in acidic pH, all three of the phosphorylated glucagon analogues tested in this study are soluble at both acidic and neutral pH.

ThT assays and intrinsic fluorescence assays over 24 hours show that phospho-Thr5-Glucagon and phospho-Ser8-Glucagon are both stable and do not fibrillation at pH 7.4. Both of these molecules, however, fibrillate at pH 2.5. This observation can be explained by different charge states of the phosphate group in the acidic and neutral pH. MD simulations of glucagon and its phosphorylated analogues in different pHs confirm and

clarify the effect of the charge state of the phosphate group on its fibrillation inhibition effects. MD simulations show that the steric zipper of phospho- glucagon is stable in absence of the phosphate group, and in singly and doubly protonated states in neutral pH simulations. The steric zipper, however, does not remain stable in acidic pH simulations where the phosphate group is not protonated and is doubly charged. The steric effect of the additional volume introduced by the phosphate group therefore, is not enough to destabilize the steric zipper and the electric charge of the phosphate moieties plays an important role in their fibrillation inhibition effects.

5.5 Conclusions

In this study phosphate-ester derivatives of glucagon were computationally designed and tested as soluble and stable prodrugs or active derivatives of glucagon. The phosphorylated glucagons showed significantly improved solubility in neutral pH compared to glucagon. Also, contrary to glucagon which fibrillates in few hours, the phosphorylated glucagons did not fibrillate and were stable for weeks.

Our research group has applied for a patent on all phosphate ester derivatives of glucagon (patent application number 62/195,537).

Table 8: The 10 most frequent inter-residue contacts in the 500 most energetically favorable models of the steric zipper region of glucagon fibril.

Contact	Frequency
Trp25-Phe6	327
Val23-Phe6	256
Trp25-Gly4	249
Trp25-Thr5	206
Met27-Ser2	183
Asp21-Tyr10	183
Trp25-Gln3	159
Val23-Ser8	158
Gln24-Phe6	145
Asp21-Ser8	142

Table 9: Solubility in neutral pH

Peptide	pH 7.4 (PBS)
Glucagon	Not Soluble
phospho-Ser2-Gluc.	Not Soluble
phospho-Thr5-Gluc.	8 mg/ml
phospho-Ser8-Gluc.	10 mg/ml

CHAPTER 6. CONCLUSIONS

Two strategies were tested for stabilization of glucagon formulation and preventing its fibrillation, penta-peptide chaperon excipients and derivatization of glucagon itself. Penta-peptides were shown to delay glucagon fibrillation for a few hundred minutes. However, after this lag time, glucagon entered a log phase and fibrillated rapidly. This delay, therefore, was not enough for stable formulation of glucagon.

Derivatization of glucagon was shown to be more effective for inhibiting glucagon fibrillation. Two phosphate ester derivatives of glucagon, phospho-Ser8- and phospho-Thr5-glucagon designed in this study, were stable and stayed in solution at neutral pH for at least one month. Currently dephosphorization studies, chemical stability studies, cell-based assays and animal studies are underway to test the activity of these molecules and their mechanism of actions and find ways to further improve their formulation.

The phosphate ester derivatives of glucagon were designed based on a computational method (FibPredictor) developed as a part of this project to model the steric zipper of amyloid fibrils. This computational method is not limited to a specific protein and can be applied to generate models of steric zippers starting from any user-defined sequence. The generated models can be used in combination with experimental data or as input for further computational studies. This computational method is now publicly available on anoHub.org.

Another developments in this dissertation, were a number of quasi-three body statistical potentials for protein structure predictions. The most successful of these potentials (Amb-3b) has been implemented in FibPredictor. However, the application of these potentials is not limited to fibrils and they can be broadly used for any type of protein structure prediction. Moreover, the theoretical framework of these quasi-three body potentials can be further expanded for information-theoric studies on protein structure[94].

REFERENCES

REFERENCES

(1) Rambaran, R. N.; Serpell, L. C. *PRION* **2008**, *2*, 112–117.

(2) Zhang, S. *Nature biotechnology* **2003**, *21*, 1171–1178.

(3) Brummitt, R. K.; Nesta, D. P.; Chang, L.; Chase, S. F.; Laue, T. M.; Roberts, C. J. *Journal of pharmaceutical sciences* **2011**, *100*, 2087–2103.

(4) Oliyai, R. *Advanced drug delivery reviews* **1996**, 275–286.

(5) Kühnle, H.; Börner, H. G. *Angewandte Chemie (International ed. in English)* **2009**, *48*, 6431–4.

(6) Zhang, J.; Gao, J.; Chen, M.; Yang, Z. *Antioxidants & redox signaling* **2014**, *20*, 2179–90.

(7) Nelson, R.; Eisenberg, D. *Current opinion in structural biology* **2006**, *16*, 260–265.

(8) Agrawal, N. J.; Kumar, S.; Wang, X.; Helk, B.; Singh, S. K.; Trout, B. L. *Journal of pharmaceutical sciences* **2011**, *100*, 5081–95.

(9) André, I.; Bradley, P. *Proceedings of the National Academy of Science* **2007**, *104*, 17656–17661.

(10) Zhang, J.; Gao, D. Y.; Yearwood, J. *Journal of theoretical biology* **2011**, *284*, 149–57.

(11) Zhang, J. *Journal of molecular modeling* **2011**, *17*, 173–179.

(12) Buck, P. M.; Kumar, S.; Wang, X.; Agrawal, N. J.; Trout, B. L.; Singh, S. K. *Methods in molecular biology (Clifton, N.J.)* **2012**, *899*, 425–51.

(13) Thompson, M. J.; Sievers, S. A.; Karanicolas, J.; Ivanova, M. I.; Baker, D.; Eisenberg, D. *Proceedings of the National Academy of Science* **2006**, *103*, 4074–4078.

(14) Frousios, K. K.; Iconomidou, V. a; Karletidi, C.-M.; Hamodrakas, S. J. *BMC structural biology* **2009**, *9*, 44.

(15) Pedersen, J. *Journal of diabetes science and technology* **2010**, *4*, 1357–1367.

(16) Sawaya, M. R.; Sambashivan, S.; Nelson, R.; Ivanova, M. I.; Sievers, S. A.; Apostol, M. I.; Thompson, M. J.; Balbirnie, M.; Wiltzius, J. J. W.; McFarlane, H. T.; others *Nature* **2007**, *447*, 453–457.

(17) Fitzpatrick, A. W. P.; Debelouchina, G. T.; Bayro, M. J.; Clare, D. K.; Caporini, M. A. **2013**, *110*, 5468–5473.

(18) Pearson, T. *The Diabetes educator* **2008**, *34*, 128–34.

(19) Pedersen, J.; Dikov, D.; Otzen, D. *Biochemistry* **2006**, *45*, 14503–14512.

(20) Stigsnaes, P.; Frokjaer, S.; Bjerregaard, S.; Van de Weert, M.; Kingshott, P.; Moeller, E. H. *International journal of pharmaceutics* **2007**, *330*, 89–98.

(21) Ghodke, S.; Nielsen, S. *FEBS ...* **2012**, *279*, 752–65.

(22) Pedersen, J. S.; Flink, J. M.; Dikov, D.; Otzen, D. E. *Biophysical journal* **2006**, *90*, 4181–94.

(23) Steiner, S.; Li, M.; Hauser, R.; Pohl, R. *Journal of diabetes science and technology* **2010**, *4*, 1332–1337.

(24) Pedersen, J. S.; Dikov, D.; Flink, J. L.; Hjuler, H. A.; Christiansen, G.; Otzen, D. E. *Journal of molecular biology* **2006**, *355*, 501–523.

(25) Matilainen, L.; Maunu, S. L.; Pajander, J.; Auriola, S.; Jääskeläinen, I.; Larsen, K. L.; Järvinen, T.; Jarho, P. *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences* **2009**, *36*, 412–20.

(26) Soto, C.; Sigurdsson, E. M.; Morelli, L.; Kumar, R. A.; Castaño, E. M.; Frangione, B. *Nature medicine* **1998**, *4*, 822–826.

(27) Soto, C.; Kascsak, R. J.; Saborío, G. P.; Aucouturier, P.; Wisniewski, T.; Prelli, F.; Kascsak, R.; Mendez, E.; Harris, D. A.; Ironside, J.; others *The Lancet* **2000**, *355*, 192–197.

(28) Soto, C.; Kindy, M. S.; Baumann, M.; Frangione, B. *Biochemical and biophysical research communications* **1996**, *226*, 672–680.

(29) Sievers, S. A.; Karanicolas, J.; Chang, H. W.; Zhao, A.; Jiang, L.; Zirafi, O.; Stevens, J. T.; Münch, J.; Baker, D.; Eisenberg, D. *Nature* **2011**, *475*, 96–100.

(30) Oliveira, C. L. P.; Behrens, M. A.; Pedersen, J. S.; Erlacher, K.; Otzen, D.; Pedersen, J. S. *Journal of Molecular Biology* **2009**, *387*, 147–161.

(31) Hellberg, S.; Eriksson, L.; Jonsoon, J.; Lindgren, F.; Sjostrom, M.; Skagerberg, B.; Worl, S.; Andrews, P. *International journal of peptide and protein research* **1991**, *37*, 414–424.

(32) Muthas, D.; Lek, P. M.; Nurbo, J.; Karlén, A.; Lundstedt, T. *Journal of Chemometrics* **2007**, *21*, 486–495.

(33) NIST Fractional Factorial Design http://www.itl.nist.gov/div898/handbook/pri/section3/pri3347.htm (accessed Feb 10, 2014).

(34) Moorthy, B. S.; Ghomi, H. T.; Lill, M. A.; Topp, E. M. *Biophysical journal* **2015**, *108*, 937–948.

(35) Wold, S.; Sjöström, M.; Eriksson, L. *Chemometrics and intelligent laboratory systems* **2001**, *58*, 109–130.

(36) Braun, W.; Wider, G.; Lee, K. H.; Wüthrich, K. *Journal of molecular biology* **1983**, *169*, 921–948.

(37) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *The Journal of chemical physics* **1983**, *79*, 926–935.

(38) Darden, T.; York, D.; Pedersen, L. *The Journal of chemical physics* **1993**, *98*, 10089–10092.

(39) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *Journal of Computational Physics* **1977**, *23*, 327–341.

(40) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. *Biopolymers* **1992**, *32*, 523–535.

(41) Joosten, R. P.; Te Beek, T. a H.; Krieger, E.; Hekkelman, M. L.; Hooft, R. W. W.; Schneider, R.; Sander, C.; Vriend, G. *Nucleic acids research* **2011**, *39*, D411–9.

(42) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.

(43) Schrödinger, L. The PyMOL Molecular Graphics System, Version~1.3r1 **2010**.

(44) Schrödinger, L. The JyMOL Molecular Graphics Development Component, Version~1.0 **2010**.

(45) Schrödinger, L. The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version~1.0 **2010**.

(46) Hunter, J. D. *Computing In Science & Engineering* **2007**, *9*, 90–95.

(47) Ghodke, S.; Nielsen, S. B.; Christiansen, G.; Hjuler, H. A.; Flink, J.; Otzen, D. *FEBS Journal* **2012**, *279*, 752–765.

(48) Wang, W. *International journal of pharmaceutics* **2005**, *289*, 1–30.

(49) Lorenzo, A.; Yankner, B. A. *Annals of the New York Academy of Sciences* **1996**, *777*, 89–95.

(50) Lu, H.; Skolnick, J. *Proteins* **2001**, *44*, 223–32.

(51) Samudrala, R.; Moult, J. *Journal of molecular biology* **1998**, *275*, 895–916.

(52) Sippl, M. *Journal of molecular biology* **1990**, *213*, 859–883.

(53) Poupon, A. *Current opinion in structural biology* **2004**, *14*, 233–41.

(54) Gan, H.; Tropsha, A.; Schlick, T. *Proteins: Structure, Function, and Bioinformatics* **2001**, *43*, 161–174.

(55) Krishnamoorthy, B.; Tropsha, A. *Bioinformatics* **2003**, *19*, 1540–1548.

(56) Mirzaie, M.; Sadeghi, M. *Journal of Paramedical Sciences* **2010**, *1*, 63–73.

(57) Li, X.; Liang, J. *Proteins* **2005**, *60*, 46–65.

(58) Mayewski, S. *Proteins* **2005**, *59*, 152–69.

(59) Betancourt, M.; Thirumalai, D. *Protein Science* **1999**, *8*, 361–369.

(60) Ejtehadi, M. *Proceedings of the National Academy of Science* **2004**, *101*, 15088–93.

(61) Samudrala, R.; Levitt, M. *Protein Science* **2000**, *9*, 1399–1401.

(62) Rykunov, D.; Fiser, A. *Proteins* **2007**, *67*, 559–68.

(63) Pace, C.; Shirley, B.; McNutt, M.; Gajiwala, K. *The FASEB journal* **1996**, 75–83.

(64) Smirnov, N. *The Annals of Mathematical Statistics* **1948**, *19*, 279–281.

(65) Kolmogorov, A. N. *Giornale dell'Istituto Italiano degli Attuari* **1933**, *4*, 83–91.

(66) Munson, P. J.; Singh, R. K. *Protein science : a publication of the Protein Society* **1997**, *6*, 1467–81.

(67) Feng, Y.; Kloczkowski, A.; Jernigan, R. L. *Proteins: Structure, Function, and Bioinformatics* **2007**, *66*, 57–66.

(68) Gniewek, P.; Leelananda, S. P.; Kolinski, A.; Jernigan, R. L.; Kloczkowski, A. *Proteins* **2011**, *79*, 1923–9.

(69) Ghomi, H. T.; Thompson, J. J.; Lill, M. A. *Journal of bioinformatics and computational biology* **2014**, *12*.

(70) Krivov, G. G.; Shapovalov, M. V; Dunbrack, R. L. *Proteins* **2009**, *77*, 778–95.

(71) Zhou, H.; Skolnick, J. *Biophysical journal* **2011**, *101*, 2043–2052.

(72) Webb, B.; Sali, A. *Current protocols in bioinformatics* **2014**, 5–6.

(73) Ghomi, H. T.; Thompson, J. J.; Lill, M. a *Journal of bioinformatics and computational biology* **2014**, *12*, 1450022.

(74) Landau, M.; Sawaya, M. R.; Faull, K. F.; Laganowsky, A.; Jiang, L.; Sievers, S. A.; Liu, J.; Barrio, J. R.; Eisenberg, D. *PLoS biology* **2011**, *9*, e1001080.

(75) Meinhardt, J.; Sachse, C.; Hortschansky, P.; Grigorieff, N.; Fändrich, M. *Journal of molecular biology* **2009**, *386*, 869–877.

(76) Petkova, A. T.; Leapman, R. D.; Guo, Z.; Yau, W.-M.; Mattson, M. P.; Tycko, R. *Science* **2005**, *307*, 262–265.

(77) Tycko, R. *Current opinion in structural biology* **2004**, *14*, 96–103.

(78) Cerf, E.; Sarroukh, R.; Tamamizu-Kato, S.; Breydo, L.; Derclaye, S.; Dufrêne, Y.; Narayanaswami, V.; Goormaghtigh, E.; Ruysschaert, J.; Raussens, V. *Biochem. J* **2009**, *421*, 415–423.

(79) Choi, B.; Yoon, G.; Lee, S. W.; Eom, K. *Physical chemistry chemical physics : PCCP* **2015**, *17*, 1379–89.

(80) Yoon, G.; Kwak, J.; Kim, J. I.; Na, S.; Eom, K. *Advanced Functional Materials* **2011**, *21*, 3454–3463.

(81) Rooseboom, M.; Commandeur, J. N. M.; Vermeulen, N. P. E. *Pharmacological reviews* **2004**, *56*, 53–102.

(82) Hale, J. J.; Mills, S. G.; MacCoss, M.; Dorn, C. P.; Finke, P. E.; Budhu, R. J.; Reamer, R. A.; Huskey, S.-E. W.; Luffer-Atlas, D.; Dean, B. J.; others *Journal of medicinal chemistry* **2000**, *43*, 1234–1241.

(83) Valette, N. M.; Radford, S. E.; Harris, S. a; Warriner, S. L. *Chembiochem : a European journal of chemical biology* **2012**, *13*, 271–81.

(84) Broncel, M.; Wagner, S. C.; Hackenberger, C. P. R.; Koksch, B. *Chemical communications (Cambridge, England)* **2010**, *46*, 3080–2.

(85) Broncel, M.; Falenski, J. a; Wagner, S. C.; Hackenberger, C. P. R.; Koksch, B. *Chemistry (Weinheim an der Bergstrasse, Germany)* **2010**, *16*, 7881–8.

(86) Gorbatyuk, O. S.; Li, S.; Sullivan, L. F.; Chen, W.; Kondrikova, G.; Manfredsson, F. P.; Mandel, R. J.; Muzyczka, N. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, *105*, 763–8.

(87) Paleologou, K. E.; Schmid, A. W.; Rospigliosi, C. C.; Kim, H.-Y.; Lamberto, G. R.; Fredenburg, R. a; Lansbury, P. T.; Fernandez, C. O.; Eliezer, D.; Zweckstetter, M.; Lashuel, H. a *The Journal of biological chemistry* **2008**, *283*, 16895–905.

(88) Schneider, A.; Biernat, J. *Biochemistry* **1999**, 3549–3558.

(89) Tabatabaei Ghomi, H.; Topp, E. M.; Lill, M. a *unpublished* **2015**.

(90) Khoury, G. A.; Thompson, J. P.; Smadbeck, J.; Kieslich, C. A.; Floudas, C. A. *Journal of chemical theory and computation* **2013**, *9*, 5653–5674.

(91) Homeyer, N.; Horn, A. H. C.; Lanig, H.; Sticht, H. *Journal of molecular modeling* **2006**, *12*, 281–289.

(92) Swails, J. M.; York, D. M.; Roitberg, A. E. *Journal of chemical theory and computation* **2014**, *10*, 1341–1352.

(93) Śmiechowski, M. *Chemical Physics Letters* **2010**, *501*, 123–129.

(94) Thompson, J. J.; Tabatabaei Ghomi, H.; Lill, M. a *Proteins* **2014**, *82*, 3450–65.

VITA

VITA

Education

PhD. Medicinal Chemistry and Molecular Pharmacology; **Purdue University, West Lafayette, IN**; Specialization: ***Computational Biophysics and Computer-aided Drug Design***, December 2015

MA. Philosophy; **Purdue University, West Lafayette, IN**; December 2015

MSc. Mathematics; **Purdue University, West Lafayette, IN**; December 2014

Doctor of Pharmacy (PharmD.); Shahid Beheshti Medical University, Tehran, Iran November 2010

Publications:

FibPredictor: A Computational Method for Rapid Prediction of Amyloid beta-Fibril Structures, **Hamed Tabatabaei Ghomi**, Elizabeth M. Topp, Markus A. Lill; under review (Software publicly available on https://nanohub.org/resources/fibpredictor)

Structural Transitions and Interactions in the Early Stages of Human Glucagon Amyloid Fibrillation, Balakrishnan S. Moorthy, **Hamed Tabatabaei Ghomi**, Markus A. Lill, Elizabeth M. Topp; Biophysical Journal 108.4. (2015): 937-948.

Are distance-dependent statistical potentials considering three interacting bodies superior to two-body statistical potentials for protein structure prediction?, **Hamed Tabatabaei Ghomi**, Jared J. Thompson, and Markus A. Lill; J. Bioinform. Comput. Biol. 12.05 (2014).

An Application of Information Theory to a Three-Body Coarse-Grained Representation of Proteins in the PDB: Insights into the Structural and Evolutionary Roles of Residues in Protein Structure, Jared J. Thompson, **Hamed Tabatabaei Ghomi**, and Markus A. Lill; PROTEINS Struct Func Bioinf. 82.12 (2014): 3450-3465.

QSAR and Pharmacophor Studies of Telomerase Inhibitors, Atefeh Hajiagha Bozorgi, **Hamed Tabatabaie Ghomi**, Abolghasem Jouyban , Med. Chem. Res. 21.6 (2012): 853-866.

Design, Synthesis and Pharmacological Evaluation of Some 2-[2-(2-Chlorophenoxy)phenyl]-1,3,4-oxadiazole Derivatives as Benzodiazepine Receptor Agonists, Mehrdad Faizi, Majid Sheikhha, Nematollah Ahangarb, Hamed Tabatabaei Ghomi, Bijan Shafaghi, Abbas Shafiee and Sayyed Abbas Tabatabai, IJPR. 11.1 (2012): 83.

Synthesis of 2,3-Diphenyl-3-oxo-propanamide Derivatives as Selective COX-2 Inhibitors; Thesis project for pharmacy doctorate degree, Shahid Beheshti Medical University (2010)

Patents:

(Glucagon research; title withheld), **Hamed Tabatabaei Ghomi**, Shenbaga Moorthy Balakrishnan, Markus A. Lill, Elizabeth M. Topp; provisional patent filed July 22, 2015; Patent application number 62/195,537.

Presentations:

Computational Modelling of Amyloid β-fibrils

- Biophysical Society 59th annual meeting, 2015
- Purdue University Office of Interdisciplinary Graduate Programs Spring Reception, 2014.

Early Stages of Glucagon Fibrillation

- American Association of Pharmaceutical Scientists, I/ODG Symposium, 2014

Are three-body distance-dependent statistical potentials superior to two-body statistical potentials for protein structure prediction?"

- Purdue University Office of Interdisciplinary Graduate Programs Spring Reception, 2014.
- 6th Yao Yuan Biotech-Pharma Symposium, 2014.
- Purdue University Chapter for Society of Industrial and Applied Mathematics Conference, 2014.

Computer Proficiency:

Programming languages: **C**, **Python**, **MATLAB**

Software Developed:

- FibPredictor (A computational method for rapid prediction of amyloid β-fibril structures); publicly available on https://nanohub.org/resources/fibpredictor
- Quasi-three body statistical potential using functional groups; results available from http://people.pharmacy.purdue.edu/~mlill/software/

- Numerous small codes for analyzing and manipulating MD simulation trajectories and protein structures.

Biomolecular Simulation Packages: **AMBER** and **GROMACS** (experienced in various advanced biomolecular simulation techniques)

Computer-aided molecular design (experienced in various docking, QSAR and molecular modeling techniques)

Large Database Analysis: developed code for compiling, processing and statistical analysis of large protein databases

Relevant Work Experience:

Research Assistant at Lill's lab-Purdue University; West Lafayette, IN; August 2012-December 2015

- Developed scientific software for modelling protein structure and interactions
- Developed statistical force fields for protein structure prediction
- Performed peptide virtual screening
- Simulated and analyzed biomolecular systems
- Developed multivariate quantitative structure activity relationship models
- Rationally designed peptide therapeutics and excipients

Research Assistant at Topp's lab-Purdue University; West Lafayette, IN; August 2014-December 2015

- Designed and evaluated peptides therapeutics and formulations
- Analyzed peptide formulation stability using various spectroscopic methods
- Designed, performed and analyzed peptide screening experiments

Teaching Assistant Purdue University; West Lafayette, IN; August 2011-December 2012

- Organic Chemistry II Laboratory (MCMP205L), Fall 2012
- Medicinal Chemistry and Pharmacology III: Cardiovascular & Renal Pharmacology (MCMP408), Spring 2012
- Biological Chemistry I (MCMP304), Fall 2011

Exir® Pharmaceutical Co.; Tehran, Iran; January 2010-June 2011

- Contributed to patent document preparation
- Evaluated bioequivalence studies for Capetopril, Cephalexin, Acetaminophen, Co-Amoxiclav
- Prepared the periodic safety update report (PSUR) for Insulin
- Contributed to idea generation sub-committee of Product Development Committee
- Revised and renewed standard operating procedure for pharmacovigilence
- Revised and renewed standard operating procedure for consumer complaint handling based on ISO 10002

Medical consultant at **National Drugs and Poisons Information Centre** (DPIC); Tehran, Iran; March 2009-January 2010

- Provided drug and poison control information to medical professionals and public

Research Assistant at Tabatabai's lab-Shahid Beheshti Medical University; Tehran, Iran; January 2008- November 2010

- Synthesized small organic molecules
- Performed docking and molecular modelling studies

Other Teaching Experiences

- "Molecular Modeling and Docking Workshop"; Shahid Beheshti Medical University & SBMU Center for Pharmaceutical Research; December 2010, February 2011 and March 2011;

Awards and Honors

Certificate of Excellence in Interdisciplinary Research, Purdue Office of Interdisciplinary Graduate Programs, Spring Reception, 2014

Distinguished Student, Shahid Beheshti Medical University, Second educational festival, 2006

Ranked **top student (1st place) among all pharmacy students of Iran**, in the nationwide Basic Sciences for Pharmacy Comprehensive Exam, 2005