PCCP

Charge patching method for the calculation of Electronic Structure of Polypeptide

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID	CP-ART-03-2018-001803.R2
Article Type:	Paper
Date Submitted by the Author:	13-Aug-2018

Complete List of Authors:	Sun, Chang-liang; Shenyang University of Chemical Technology; Lawrence Berkeley National Laboratory, Berkeley, Materials Science Division Liu, Liping; Lawrence Berkeley National Laboratory, Berkeley, Materials Science Division; Beijing Institute of Technology, School of Physics Tian, Fubo; Jilin University, State Key Laboratory of Superhard Materials, College of Physics; Lawrence Berkeley National Laboratory, Berkeley, Materials Science Division Ding, Fu; Shenyang University of Chemical Technology, Faculty of Chemical Technology Wang, Lin-Wang; Lawrence Berkeley National Laboratory,

Charge patching method for the calculation of Electronic Structure of Polypeptide

Chang-Liang Sun[*1,2], Li-Ping Liu[2,3], Fubo Tian[2,4], Fu Ding[1], Lin-Wang Wang[*2]

(1. Center of Physical Chemistry Test, Shenyang University of Chemical Technology, Shenyang 110142, People's Republic of China)

(2. Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA)

(3. School of Physics, Beijing Institute of Technology, Beijing 100081, People's Republic of China)

(4. College of Physics, Jilin University, Changchun 130012, People's Republic of China)

Abstract: Theoretical study of electronic structures of protein is a fundamental challenge in computational biochemistry due to the large size of the systems. The electronic structure of protein is important for some of the important protein functionalities such as photosynthesis. In this study, we explored charge patching method to calculate the electronic structure of polypeptide. This method generates the charge densities of the systems by patching the charge motifs calculated from small prototype systems. The method was tested on a range of polypetides including glycine polypeptide in 2_7-ribbon, α-helix, 3_{10}-helix, and β-strand structures. After the charge density profiles of these systems are obtained, the electronic structure of these glycine polypeptides are further calculated based on density functional theory (DFT) using a folded-spectrum-method. The highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) are analyzed and compared with conventional direct DFT calculations. The charge patching method results are found to be in good agreement with directed DFT results.

Key Words: electronic structures; charge patching method; glycine polypeptide; band gap

1. Introduction

Proteins are the basic building block for all living organisms, they construct protein-complex for different mechanical and structural properties,[1-4] serve as

enzymes for biological reactions,[5-7] make complex systems enabling the photosynthesis.[8] Proteins are constituted by different polypeptide chains with 20 amino acids and the biological functions of proteins are governed by theirs mutual interactions and interaction with other molecules and solvents. Although the majority of the functionalities are determined by their 3-dimensional structures, the mutual interactions and the related binding energies, some important functionalities are also related to their electronic structures.[9-13] This includes the charge transfer in photosynthesis, the path way of ATP,[8,14-16] as well as the catalytic mechanisms of many enzymes.[5-7] Thus, it is also important to understand the electronic structure of the proteins, including their highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) states, as well as how do these orbitals depend on the protein structural change.

The electronic structure are directly related to many properties of biological molecules, such as the stability of proteins folds, the catalytic activity of enzymes, compatibility with ligands or other proteins, and so on.[9-13] Biological energy conversion in photosynthesis, respiration, or enzyme catalysis is mainly based on electron flow through metalloproteins with several transition metal and other redox centers.[15] For example, biological long-range electron transfer is a fundamental process that enables living cells to channel the flow of energy extracted from sunlight or oxidation of food into adenosine triphosphate (ATP). This energy transport is accomplished by electron transfer between remote redox centers encapsulated in protein matrices of the respiratory chain, ultimately driving the proton pumping mechanisms embedded in ATP synthase.[16] Some studies suggested that electron transfer may be a natural mechanism for redox sensing and signaling in the genome, for example, to localize DNA damage before it is repaired by DNA repair enzymes.[17] Therefore, accurate determination of electronic structure of protein and other biological molecule is of both fundamental importance and great practical utility.

High quality ab initio method can provide an accurate and parameter free approach to calculate the electronic structure of proteins. Density functional theory (DFT) describes the electronic properties based on the electron charge density of the

system, and has been used widely in material simulations studying systems from crystal to organic molecules. However, due to their high order scaling (e.g., $O(N^3)$ scaling for DFT where N is the size of the system), the DFT method can only be applied to relatively small systems, e.g., restrict the system size to about a thousand atoms. However, for the majority of proteins or other large biological molecules, the system size could be a few or tens of thousands of atoms, thus beyond the capability of the straight forward DFT calculations. To calculate the large biological systems, some kind of linear scaling methods is thus necessary.[18-21]

In the past years, some researchers have proposed a series of fragment-based method to surmount the size problem of proteins and other biological molecules.[22-29] Most of these methods are based on utilizing chemical knowledge to divide the system into smaller subsystems, for example, cutting up a protein molecule into small fragments. The difference between these methods is the way they treat the fragment boundaries and the interactions between the subsystems. For example, Zhang et al. have proposed a linear-scaling method for direct calculation of total electron density of protein based on the molecular fractionation with conjugate caps (MFCC) approach.[30-35] In this approach, a protein molecule is cut into amino acid fragments along the backbone and a pair of conjugate caps is inserted at locations of cut to cap the protein fragments. The total electron density of a protein molecule is obtained by summing over all individual electron densities of protein fragments subtracted by electron densities of the cap atoms. In another work, Visscher et al. have developed the 3-partition frozen-density embedding (3-FDE) scheme to calculate the electronic densities of proteins and other biomolecular systems based on DFT.[36] In material science simulation, similar approach also exists. For example, the linear scaling three dimensional fragment (LS3DF)[37-41] method cut off a large system into three dimensional fragments, passivates the fragment surface, and patch the charge densities of the fragments into the charge density of the whole system. In LS3DF, overlapping fragments are used, so only the center part of the fragment charge density is used in the patching process. Besides, the Poisson equation for the electron-electron Coulomb interaction is calculated based on the global system, which ensures the long

range Coulomb interaction is described correctly.[37-41]

The above linear scaling methods still requires the quantum mechanical, or DFT calculations for each fragments, and possibly interaction between them treated with self-consistent field (SCF) iterations. As a result, these methods are still computationally expensive. Sometime it will be beneficial to have even faster methods for protein calculations, based on non-SCF patching of fragment charge densities. Mezey et al. proposed the Molecular Electron Density Lego Approach (MEDLA)[42,43] for the construction of ab initio quality electronic charge distributions for large molecules from charge densities of small molecular fragments. The method is based on the assumption that contribution of a given molecular fragment to the complete molecular electron distribution should be quite similar in different molecules or in different locations of the same molecule, provided that the molecular environments are similar. That is to say, the distribution of electron density of molecular fragments is "transferable". In order to get accurately results for large molecules, the molecular environment of the fragment in the parent molecule should be the same as that in the desired molecule for a distance of at least six atomic units.[42,43] This makes this method not very flexible, and might require the recalculations of the fragment densities after the large molecule has changed its shape. In another extreme, the neutral spherical atomic charge densities have been summed up to generate the so called promolecular density.[44,45] It has been shown, such simple promolecular density can reproduce many qualitative features for dispersion energy analysis between molecules. Based on this theory, Yang et al. constructed the approximate promolecular densities or many biological molecules, and have used them to reveal noncovalent interaction features, distinguishing van der Waals interactions, hydrogen bonds, and steric clashes.[46] In another work, Koritsanszky et al. has built pseudoatom databank,[47] expressing the atomic motif charge densities in analytical forms with the parameters derived from fitting with ab initio densities of tripeptides. This procedure parallels that used in the experimental databank,[48,49] except it is fitted to ab initio charge density. Following this approach, Dominiak et.al have built an extended database for C, H, N, and O pseudoatoms and have applied the

databank to calculate the electrostatic interaction of different amino acids and their dimmers, as well as between different domains of the scaffolding protein syntenin and peptides.[50-52] One common feature of these previous non-selfconsistent patching scheme in biological study is that, they are all used to analyze or calculate the nonbonding electrostatic and dispersion interactions. However, the concerns for the energy calculation and electronic structure calculations are rather different. In a way, the electronic structure calculations (e.g., the band gap) are more sensitive to the electron charge density, and the long range electric field. In this work, we are more interested in electronic structure calculation. One of the purposes of the current work is to test how good can such non-selfconsistent patching scheme can work for the electronic structure calculation. Thus the main point of our work is not to show our current charge patching scheme is better than the previous schemes. As a matter of fact, many of these schemes (including our scheme) are rather similar in spirit, and only differ in details. Our point here is to show, such charge patching method, perhaps including some of the previous methods, can be used to calculate the electronic structure of the biological systems.

The charge patching method (CPM) is one non-selfconsistent fragment patching scheme developed by our group specifically for electronic structure calculations.[53-60] However, previously we have applied CPM for inorganic systems and organic polymers, but never applied to biological system. In CPM, instead of calculating each fragment quantum mechanically on-the-flight, it pre-calculates the charge motifs of atoms (not fragments) in each bonding environments. The global charge density of the whole system is patched together from the atomic charge motifs without on-the-flight calculations. Compared with the work of pseudoatom databank[47], we used numerical representation and storage of the charge density modifs, instead of fitted analytical expression. This ensures the accuracy and the ability to exactly reproduce the original ab initio molecule charge density. Besides, we have considered larger molecules and larger area environment in many cases, not just deriving the motif from small molecule. On the other hand, compared with the fragment based approaches as in MEDLA[42,43], the use of atom based motifs provides larger flexibility and

transferability. The atomic configuration for each fragment (e.g., an amino acid unit) might change a little bit for different proteins and at different position of the protein, which makes it necessary to recalculate the rigid fragment charge density in a fragment based approach. On the other hand, even though there is a small variation in the bonding configuration, as will be shown later, the pre-calculated atomic charge motifs can still be used since the center and the orientation of the motif can follow the shift of the atom and its local bonding orientation. As a result, a fixed set of pre-calculated atomic charge motifs can be used to describe the charge density of different proteins and for different configurations. This will be particularly useful if one desires to study the dynamics property of the system and the carrier transport. For example, one can carry out a molecular dynamics simulation using classical force field, and for any given time step, use the atomic configuration at that time step to construct the charge density of the system based on CPM, then use the folded spectrum method (FSM)[61] to calculate the electronic structure, including the HOMO and LUMO states. One can also use the variation of these eigen states as the basis to carry out non-adiabatic molecular dynamics simulations following the time-dependent Schrodinger's equation, hence to study carrier transport in such systems[62].

The CPM has been successfully applied to inorganic semiconductor systems, such as diluted nitrogen alloys, quantum dots and wires, impurities, as well as carbon nanotubes and fullerenes.[53-56] It have also been tested on a range of organic systems including alkane and alkene chains, polyacenes, polythiophenes, polypyrroles, polyfuranes, polyphenylene vinylene, and poly(amidoamine) dendrimers.[57-60] In this work, we test its application to protein and other biomolecular systems. In contrast to inorganic semiconductors where all atoms of the same element usually have the same environment, in biomolecular systems, there are more bonding types and local environments for a given atom. In addition, the non-covalent interactions, such as hydrogen bond interactions, electrostatic interaction, polarization effect, play an important role in determine both the 3D structure of protein and possible also its electronic structures. Therefore, there is a need to develop the appropriate schemes for the classification of atoms, taking into account their bonding environment. There is

also a need to test the CPM, to see how well it will work for proteins.

In this paper, a systematic approach will be developed for performing CPM calculations on the electronic structures of polypeptide molecules. The results will be tested by comparing the results from direct DFT calculations. As the first step in this direction, we will only focus on relatively small polypeptide, instead of fully folded proteins. The structure of this work is organized as follows: In section 2, the computational methodology is briefly outlined: the specification of CPM method, the classification of atoms for glycine polypeptide, and computational details. In section 3 for Results and Discussion, the performance of CPM on four kinds of glycine polypeptide structures will be presented. In section 4, the conclusion remarks will be provided.

2. Computational Methodology

The charge patching method exploits the fact that in systems with covalent bonds, the distribution of charge density around a given atom depends mainly on its local environment. One can, therefore, define a charge density motif corresponding to a given atom in a particular bonding environment. To find the motif, one performs the DFT calculation on a small system where a given atom has a similar bonding environment and extracts the charge motif from the small system result. The charge density of the large system can then be found by adding together the motifs corresponding to each of the atoms in the system, rather than performing a self-consistent DFT calculation. Here we chose the simplest glycine peptide as the test system to develop the CPM strategy for biological molecules.

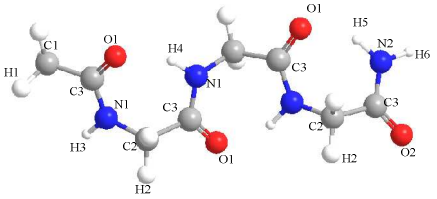


Fig. 1 Defined the atomtypes in Gly-polypeptide structure.

To classify the atoms and the motifs, we follow the idea present in the

classification of atoms for force field calculations that the atoms of the same physical element are defined as different types if their bonding environments differ. As shown Fig. 1 and Table 1, in glycine peptide chain, there are six types of hydrogen, the first one is H atom in the methyl; the second one is in the methylene, which we can lable as H1 and H2, respectively. H3 represent the H atom in NH with no hydrogen bond formation. H4 represent the H atom in NH which has hydrogen bond formation. Similar definitions exist for H5 and H6 in the terminal NH2. The atom types for carbon, nitrogen and oxygen are shown in Table 1.

Then the charge density motif is defined by the type of the central atom, plus the types of atoms connected to the central atom (its nearest neighbors), and the corresponding bonds. In the cases of atoms that are connected to only one atom (for example, H and O atoms), the definition is extended to include the next nearest neighbors. To illustrate the definition, we take glycine polypeptide in 2_7-ribbon structures as example. As shown in Table 2, according to our definition, there are nineteen charge motifs in 2_7-ribbon structures. In the motif_C1_C3H1H1H1, C1 is the central atom of the motif, C3, H1, H1, H1 are the atoms connected to the central atom. In motif_O1_C3_N1C1, O1 is the central atom of the motif, C3 is the neighbor of O1, and the N1 and C2 are the next nearest neighbors. The use of the second nearest neighbors in this case can help us to re-orient the motif when we patch the charge density.

Table 1. Defined and description the atomtypes in Gly-polypeptide structure

atom	types	Description
H	H1	H in CH3
	H2	H in CH2
	H3	H in NH and with no HB formation
	H4	H in NH and with HB formation
	H5	H in terminal NH2 with HB formation
	H6	H in terminal NH2 with no HB formation
C	C1	C in CH3
	C2	C in CH2
	C3	C in CO
O	O1	O in CO with HB formation
	O2	O in CO with no HB formation

N	N1	N in NH
	N2	N in terminal NH2

In our method, the Hirshfeld partitioning scheme was used to generate the charge motifs. The charge density of the motif corresponding to the central atom A is extracted from the small system using the following formula

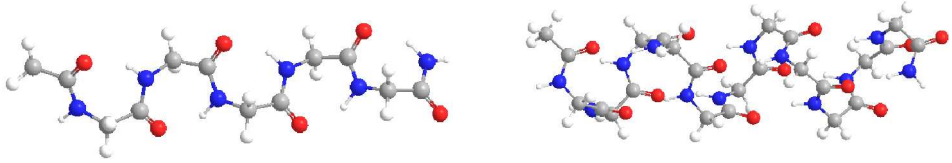
$$m_A(r-R_A)=\frac{w_A(r-R_A)}{\sum_B w_B(r-R_B)}\rho(r), \quad (1)$$

where ρ(r) is the charge density of the small system calculated by using DFT and R_A is the position of atom A. The atomic spherical charge density of isolated neutral atom was used as w_A. The motif charge density $m_A(r-R_A)$ is numerically stored in a 3D numerical grid defined within a box centered at R_A. After all these motifs are generated, for a polypeptide with a given atomic configuration {R}, the charge density of this polypeptide will be calculated as:

$$\rho_{\text{patch}}(r)=\sum_A m_A(r-r_A). \quad (2)$$

Most importantly, not only the motif will be centered at the atomic position R_A, its orientation will also be rotated so its nearest neighbor orientation aligns with the ones when the motif $m_A(r-R_A)$ was generated from the small prototype system. Notice that, in most situations, there could be some small difference between the nearest neighbor orientation/configuration and the original bonding environment when the motif is generated. In such case, we just rotate the motif to the best aligned orientation in average and ignore the small difference. One could generate the derivative motifs when represents the change of the motif when the local bond length and angle have changes. In inorganic systems, it was found such derivative motifs can further improve the charge density accuracy by a factor of 3 to 5.[53,54,56] However, in the current work, we found that relatively good charge density can be obtained without the use of such derivative motifs. In this work, the charge density of the small model molecule is obtained by the DFT calculation with General Gradient Approximate (GGA) of the exchange-correlation energy, and by using plane-wave pseudopotential

code PWmat. Troullier–Martins norm-conserving pseudopotentials with a kinetic energy cutoff of 60 Ry were used. Once the charge density is obtained by using the charge patching procedure, the single-particle Hamiltonian can be generated by solving the Poisson equation for the Hartree potential and using the GGA formula for the exchange-correlation potential. They will give us the Kohn-Sham single particle Hamiltonian H. The single-particle states ψ_i, especially for states at the band gap edge can be calculated using the folded spectrum method (FSM)[61] as implemented in the ESCAN code [http://cmsn.lbl.gov/html/Escan/DOE-nano/pescan.htm]. In FSM, instead of solving $H\psi_i=E_i\psi_i$, one solves the folded spectrum eigen states: $(H-E_{ref})^2\psi_i=(E_i-E_{ref})^2\psi_i$, here E_{ref} is a reference energy placed inside the band gap. The exactly value of E_{ref} will not affect the final results as long as it is located inside the gap. The position of the gap can be found through a density of state (DOS) calculation using the linear scaling generalized moments method (GMM)[63] also implemented in the ESCAN code. Thus the overall procedure is: using CPM, we found the charge density of a molecule; solving the Poisson equation to obtain the total potential, hence the Hamiltonian H; using the GMM to calculate the DOS, locate the band gap, and choose E_{ref}; using FSM and E_{ref} to calculate HOMO and LUMO states. This approach allows the direct solution of the HOMO and LUMO states without solving the thousands of valence states below HOMO, thus makes the whole calculation scales linearly to the size of the system.



a b

Fig. 2 Gly-6-peptide in ribbon structure and Gly-12-peptide in α-helix structures were chosen as model structure to get the charge motifs.

Table 2. The nineteen charge motifs including in 2_7-ribbon structures.

	charge motif		charge motif
1	motif_C1_C3H1H1H1	11	motif_H5_N2_C3H6
2	motif_C2_N1C3H2H2	12	motif_H6_N2_C3H5
3	motif_C3_O1N1C1	13	motif_N1_C3C2H3

4	motif_C3_O1N1C2	14	motif_N1_C3C2H4
5	motif_C3_O2N1C2	15	motif_N2_C3H6H5
6	motif_C3_O2N2C2	16	motif_O1_C3_N1C1
7	motif_H1_C1_C3H1H1	17	motif_O1_C3_N1C2
8	motif_H2_C2_N1C3H2	18	motif_O2_C3_N1C2
9	motif_H3_N1_C3C2	19	motif_O2_C3_N2C2
10	motif_H4_N1_C3C2		

3. Results and Discussion

3.1. The 2_7-ribbon structure. The charge patching method was firstly applied to the glycine polypeptide in 2_7-ribbon structure. As shown in Fig. 2a, we chose Gly-6-peptide ribbon structure as model molecule to get the charge motifs. Firstly, the charge density of the Gly-6-peptide was calculated by DFT method with PBE formula. Then we applied the program to obtain the nineteen charge density motifs as listed in Table 2. In Fig. 3, we illustrated some of the charge motifs as examples. Based on the charge density motifs, the electron density of other glycine polypeptide in 2_7-ribbon structure can be patched with the help of the patching program. The electron densities of Gly-8-peptide obtained from direct DFT calculations and patched by the CPM method are compared in Fig. 4. To further judge the effects of this small difference, we have calculated the band gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) based on the CPM electron density and compared the direct DFT calculation results.

Table 3 and Fig. 5a list the band gap of nineteen 2_7-ribbon structures from Gly-6-peptide to Gly-24-peptide. The band gaps of these ribbon structures calculated from CPM are in good agreement with the direct DFT results. We can also get the wave function of HOMO and LUMO from the CPM method. As shown in Fig. 6, CPM wave functions of HOMO and LUMO are in good agreement with the DFT results for Gly-8-peptide. In the 2_7-ribbon structures, the HOMO located in the N-terminal of the peptide chain and the LUMO in the C-terminal. All the DFT wave function characters are well reproduced by the CPM method.

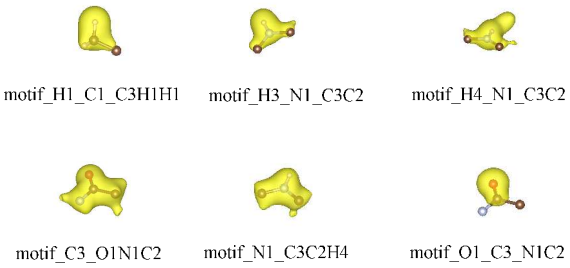


Fig. 3 The images of charge motifs obtained from Gly-6-peptide in 2_7-ribbon structure.

Fig. 4 The electronic structures of Gly-8-peptide obtained from DFT calculation and patched by our CPM method.

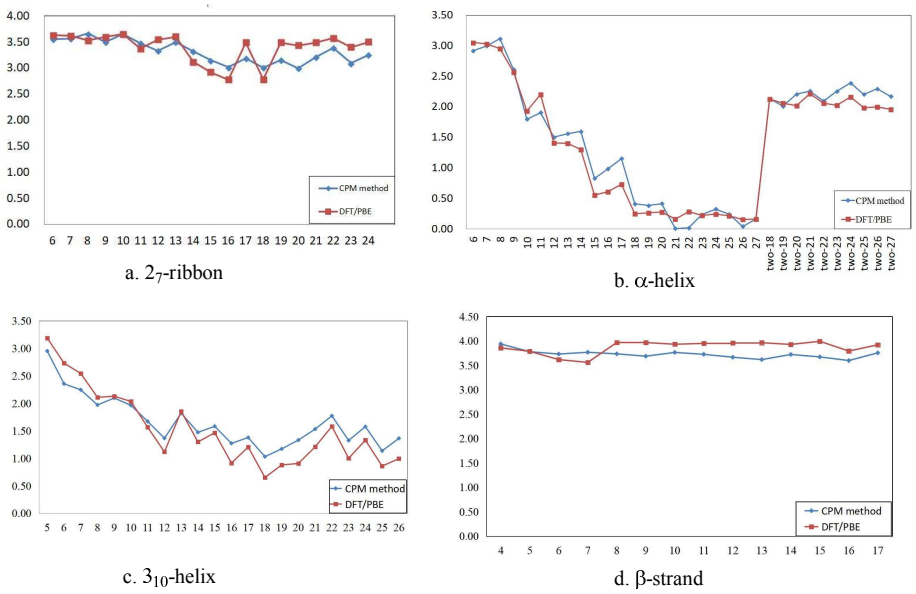


Fig. 5 The band gaps of Gly-polypeptide in 2_7-ribbon, α-helix, 3_{10}-helix, and β-strand structures calculated by our CPM method and DFT with PBE calculations.

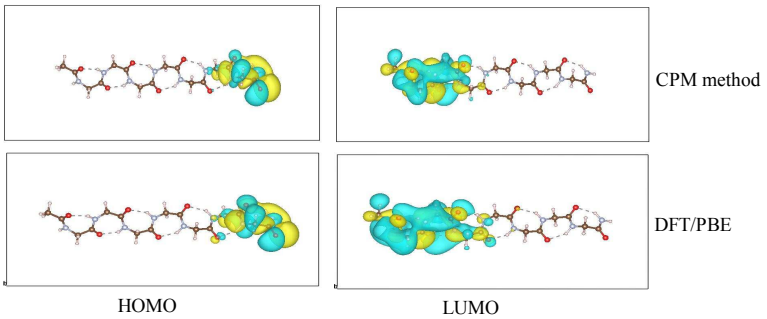


Fig. 6 The wave function of HOMO and LUMO calculated by CPM method and DFT with PBE calculations for Gly-8-peptide in 2_7-ribbon structures.

Table 3. The band gap of Gly-polypeptide in ribbon structures calculated by the CPM method and DFT with PBE calculations.

Gly-n-peptide	CPM method(eV)			DFT/PBE(eV)		
	HOMO	LUMO	GAP	HOMO	LUMO	GAP
6	-5.16	-1.60	3.56	-5.11	-1.48	3.63
7	-5.14	-1.57	3.57	-5.06	-1.44	3.62
8	-5.17	-1.50	3.66	-4.98	-1.45	3.53
9	-5.12	-1.61	3.50	-5.02	-1.42	3.60
10	-5.18	-1.53	3.65	-5.02	-1.37	3.66
11	-5.11	-1.64	3.48	-4.88	-1.51	3.37
12	-5.06	-1.73	3.33	-4.99	-1.44	3.55
13	-5.14	-1.64	3.50	-4.99	-1.39	3.60
14	-5.07	-1.75	3.32	-4.76	-1.64	3.12
15	-4.99	-1.84	3.15	-4.67	-1.74	2.92
16	-4.95	-1.93	3.02	-4.60	-1.83	2.78
17	-5.03	-1.85	3.19	-4.94	-1.45	3.49
18	-4.96	-1.95	3.01	-4.59	-1.81	2.78
19	-5.03	-1.87	3.15	-4.94	-1.44	3.49
20	-4.96	-1.97	3.00	-4.92	-1.48	3.44
21	-5.08	-1.87	3.21	-4.93	-1.44	3.49
22	-5.17	-1.78	3.39	-4.96	-1.39	3.57

23	-4.93	-1.83	3.09	-4.91	-1.51	3.40
24	-5.01	-1.75	3.25	-4.96	-1.46	3.50

3.2. The α-helix structure. The CPM method was further applied to the glycine polypeptide in α-helix structures. It is well known that the α-helix is the most important secondary structure in protein 3D structures. In this structure, the C=O... H-N hydrogen-bond ring in the backbone have thirteen members, which is different from the seven members in 2_7-ribbon structures. According to the studies of Wu et al., the hydrogen-bond cooperativity exists in α-helix structure.[65] Furthermore, there is polarization interaction in α-helix structure because of the dipole moments of C=O... H-N hydrogen bond. This is thus a more stringent test for the CPM method.

It can be seen from Table 4 and Fig. 5b, the band gap of α-helix structures calculated by DFT with PBE calculations decrease gradually as the length of Gly-polypeptide chains added. After it added to 18-peptide, the band gaps no longer change substantially. The reason for this band gap trend is that: there are many C=O... H-N hydrogen bonds around the α-helix structure and these hydrogen bonds can be considered as a series of small dipole moments. Furthermore, our calculation indicates that the HOMO and LUMO of the α-helix structure are respectively located in N-terminal and C-terminal, at the two ends of the chain. When the length is short, the total dipole moment of these C=O... H-N hydrogen bonds point in the perpendicular direction of the helix chain. However, as the length increases, the total dipole moment gradually become parallel to peptide chain and it can induce a static electric field along the α-helix chain. As a result, the band gap of the system reduces with the chain length as the electric potential is tilted towards one direction. To test this assumption, we design some special α-helix structure in our research, in which there are antiparallel two peptide chains. In these systems, the electric fields from the two chains cancel each other. It can be seen from Table 4 and Fig. 5b (the two-18 etc), in these antiparallel α-helix structures, the band gaps of the systems are significantly larger than the corresponding ones with only one chain.

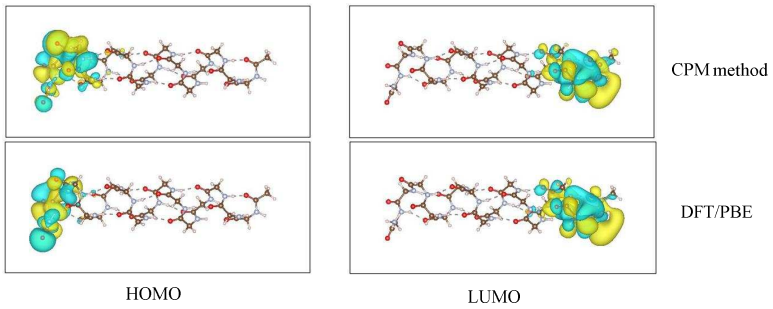


Fig. 7 The wave function of HOMO and LUMO calculated by CPM method and DFT with PBE calculations for Gly-17-peptide chains in α-helix structures.

We have tried applied the same charge density motifs as those used in 2_7-ribbon structures, to patch the electronic structures of α-helix structures. However, the results were not good in compared with the DFT result. Part of the reason is the overall difference of these two configurations, especially for the hydrogen bond difference, another reason is the existence of the electric field as discussed above. Both the hydrogen-bonding cooperativity and polarization interaction greatly affect the charge density of α-helix structure. So we have chosen Gly-12-peptide, shown in Fig. 2b, as our model molecule to extract charge density motifs in α-helix structures. During extraction of charge density motifs, we use the atoms in the middle of the chain and nineteen motifs are generated. Based on these new charge motifs, we patched the electron densities of 32 α-helix structures containing one peptide chain and two antiparallel chains. As shown in Table 4 and Fig. 5b, the band gap calculated from CPM are in good agreement with the results of DFT calculations. The wave functions of HOMO and LUMO of Gly-17-peptide are shown in Fig. 7 as an illustration. Thus, for the α-helix structures, our CPM method also obtained the wave functions similar to the DFT calculations.

Table 4. The band gap of Gly-polypeptide in α-helix structure calculated by our CPM method and DFT with PBE calculations.

Gly-n-peptide	CPM method(eV)			DFT/PBE(eV)		
	HOMO	LUMO	GAP	HOMO	LUMO	GAP
6	-5.09	-2.17	2.92	-4.60	-1.55	3.05
7	-5.08	-2.08	3.00	-4.51	-1.48	3.03
8	-5.28	-2.17	3.12	-4.45	-1.49	2.96
9	-4.89	-2.28	2.61	-4.27	-1.70	2.57
10	-4.34	-2.54	1.80	-3.87	-1.94	1.93
11	-4.24	-2.34	1.91	-3.88	-1.68	2.20
12	-4.22	-2.72	1.50	-3.56	-2.15	1.41

13	-4.16	-2.60	1.56	-3.49	-2.08	1.40
14	-4.12	-2.53	1.60	-3.43	-2.13	1.30
15	-3.83	-3.00	0.83	-3.10	-2.54	0.56
16	-3.85	-2.86	0.98	-3.09	-2.49	0.61
17	-3.85	-2.70	1.15	-3.10	-2.37	0.73
18	-3.57	-3.15	0.41	-2.91	-2.66	0.25
19	-3.63	-3.25	0.38	-2.88	-2.61	0.26
20	-3.62	-3.21	0.42	-2.86	-2.59	0.28
21	-3.32	-3.31	0.01	-2.86	-2.69	0.16
22	-3.39	-3.37	0.02	-2.84	-2.55	0.29
23	-3.34	-3.10	0.24	-2.79	-2.57	0.22
24	-3.44	-3.11	0.33	-2.79	-2.54	0.24
25	-3.34	-3.10	0.24	-2.81	-2.60	0.21
26	-3.20	-3.16	0.04	-2.75	-2.59	0.16
27	-3.24	-3.08	0.17	-2.73	-2.57	0.17
two-18	-4.55	-2.42	2.13	-3.92	-1.79	2.13
two-19	-4.47	-2.46	2.01	-3.82	-1.76	2.06
two-20	-4.42	-2.21	2.21	-3.75	-1.73	2.02
two-21	-4.36	-2.10	2.26	-3.82	-1.60	2.22
two-22	-4.33	-2.24	2.09	-3.67	-1.61	2.06
two-23	-4.23	-1.97	2.26	-3.61	-1.59	2.03
two-24	-4.23	-1.84	2.39	-3.64	-1.47	2.16
two-25	-4.23	-2.03	2.20	-3.63	-1.64	1.98
two-26	-4.19	-1.89	2.30	-3.58	-1.58	2.00
two-27	-4.14	-1.97	2.17	-3.45	-1.49	1.96

3.3. The 3_{10}-helix and β-strand structures. We further applied the CPM method to other two structures of glycine polypeptide, 3_{10}-helix and β-strand structures. The 3_{10}-helix structure is very rare in natural protein and peptide. There are only ten members in hydrogen bond ring. Comparing to the thirteen members in α-helix, the small hydrogen bond ring is believed to be unstable. Wu et al. suggested there is hydrogen-bond cooperativity in 3_{10}-helix structure too.[65] It can be seen from Table 5 and Fig. 5c, the band gap also decrease gradually as the Gly-polypeptide chain increases. Then have used the same charge motifs as those in α-helix structures to patch the electron charge of the 3_{10}-helix structures. It can be seen from Table 5 and Fig. 5c, the results is reasonably good. So we did not extract new charge density motifs for the 3_{10}-helix structures, although that can be done if necessary.

β-strand structure is usually to form β-sheet structure, which is very important in protein folding. There are five-membered hydrogen bond rings in β-strand structure. The angle of the hydrogen bond is significantly smaller than that in normal hydrogen bond, but the length of hydrogen bond is much smaller than those in α-helix and 3_{10}-helix. Despite these differences, we like to test the transferability of the motifs by applying the charge motif obtained from α-helix structure to the β-strand structure. Nevertheless, it is worth to note that we need to add six new charge motifs for the β-strand structures: motif_C3_O2N1C1, motif_H6_N2_C3H6, motif_N2_C3H6H6, motif_O2_C3_N1C1, motif_O2_C3_N1C2, motif_O1_C3_N2C2. The bonding situations of these new motifs do not exist in α-helix, but exist in the β-strand structures. The charge densities of them were obtained from the simplest Gly-4-peptide. In Fig. 8, the wave functions of HOMO and LUMO of Gly-24-peptide in 3_{10}-helix structure and Gly-17-peptide in β-strand structure are illustrated as examples. As shown in Table 6 and Fig. 5d, the band gap calculated from the CPM patching are in good agreement with the DFT results. For most of the 3_{10}-helix and β-strand structures discussed in this work, the CPM method can obtain the wave functions of HOMO and LUMO similar to the DFT calculations.

Table 5. The band gap of Gly-polypeptide in 3_{10}-helix structure calculated by our CPM method and DFT with PBE calculations.

Gly-n-polypeptide	CPM method(eV)			DFT/PBE(eV)		
	HOMO	LUMO	GAP	HOMO	LUMO	GAP
5	-5.46	-2.51	2.95	-4.74	-1.56	3.19
6	-5.43	-3.07	2.35	-4.53	-1.80	2.73
7	-5.39	-3.14	2.25	-4.38	-1.84	2.54
8	-5.26	-3.29	1.97	-4.19	-2.08	2.11
9	-5.50	-3.41	2.10	-3.92	-1.79	2.13
10	-5.28	-3.31	1.96	-4.05	-2.01	2.03
11	-5.13	-3.46	1.67	-3.84	-2.27	1.56
12	-4.99	-3.63	1.37	-3.64	-2.52	1.12
13	-5.23	-3.40	1.82	-3.89	-2.03	1.85
14	-5.05	-3.58	1.47	-3.64	-2.34	1.30
15	-5.11	-3.53	1.58	-3.69	-2.23	1.46
16	-4.96	-3.69	1.27	-3.45	-2.54	0.91
17	-5.01	-3.64	1.38	-3.57	-2.36	1.21
18	-4.85	-3.82	1.03	-3.31	-2.66	0.65
19	-4.92	-3.75	1.17	-3.52	-2.64	0.88
20	-5.01	-3.68	1.33	-3.38	-2.47	0.91

21	-5.12	-3.59	1.53	-3.51	-2.30	1.21
22	-5.27	-3.49	1.77	-3.67	-2.09	1.58
23	-5.03	-3.71	1.33	-3.41	-2.41	1.00
24	-5.18	-3.60	1.58	-3.54	-2.21	1.33
25	-4.94	-3.81	1.14	-3.45	-2.59	0.86
26	-5.07	-3.71	1.36	-3.38	-2.39	0.99

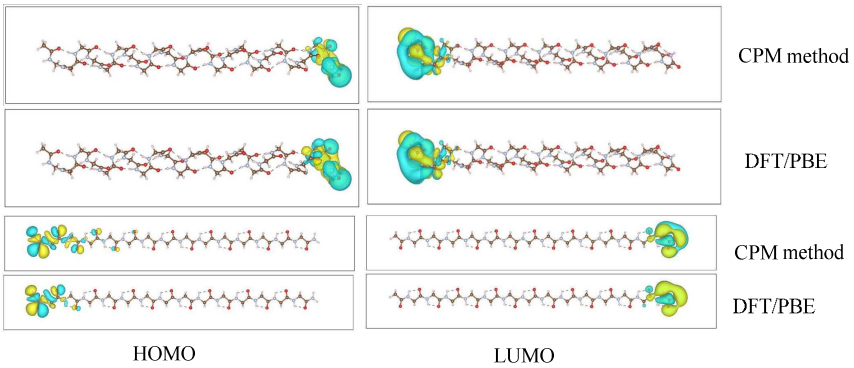


Fig. 8 The wave function of HOMO and LUMO calculated by CPM method and DFT with PBE calculations for Gly-24-peptide in 3_{10}-helix structure and Gly-17-peptide in β-strand structure.

Table 6. The band gap of Gly-polypeptide in β-strand structure calculated by our CPM method and DFT with PBE calculations.

Gly-n-polypeptide	CPM method(eV)			DFT/PBE(eV)		
	HOMO	LUMO	GAP	HOMO	LUMO	GAP
4	-6.64	-2.71	3.94	-5.18	-1.32	3.86
5	-6.67	-2.89	3.78	-5.13	-1.35	3.79
6	-6.71	-2.98	3.73	-5.05	-1.44	3.62
7	-6.79	-3.02	3.77	-5.00	-1.44	3.56
8	-6.81	-3.08	3.73	-5.20	-1.23	3.97
9	-6.82	-3.13	3.69	-5.20	-1.23	3.97
10	-6.92	-3.15	3.76	-5.15	-1.22	3.93
11	-6.93	-3.21	3.73	-5.16	-1.21	3.95
12	-6.93	-3.27	3.67	-5.17	-1.21	3.96
13	-6.94	-3.33	3.62	-5.18	-1.22	3.96
14	-7.05	-3.33	3.72	-5.14	-1.22	3.93
15	-7.06	-3.38	3.67	-5.21	-1.21	3.99
16	-7.04	-3.44	3.60	-5.08	-1.29	3.79
17	-7.18	-3.42	3.75	-5.14	-1.22	3.92

3.4. Comparison with other methods. It is now worth to compare our methods to other existing methods in the literature. As we stated in the introduction, the main difference between our work and the other motif based patching scheme in biology is that we are concerned about the electronic structure while the other works are focused on nonbonding interaction energies. Nevertheless, it will be helpful to compare how the motifs are generated.

The simplest method to generate the molecular density is to take the sum of the spherical neutral atom charge densities. This so called promolecular charge density has been used to extract features to distinguish van der Waals, hydrogen and steric clash interactions[46]. It is found such promolecular charge density can already be useful in such energy analysis. To test how good is this promolecular charge density for electronic structure prediction, we have calculated the band gap and band edge state energies using its charge density. In our ab initio calculation, the first step in the self-consistent field (SCF) iteration is to use the sum of the atomic charge density as the initial molecule charge density. So, the promolecular charge density results are reported in the SI as the NONSCF results. The NONSCF results significantly overestimate the band gap, and their HOMO and LUMO levels are also much deeper than the SCF results. Besides, some of the trends (oscillation and disappearing of band gap) are completely missing in NONSCF results.

We have discussed about the other approaches for density patching schemes. The work by Dominiak et al.[50-52] uses analytical expression to represent the aspherical atomic charge motifs for biological systems. In their approach, the spherical harmonic analytical scheme of Hansen and Coppens are used to describe the atomic charge with parameters. On the other hand, in our approach, Hirshfeld numerical scheme are used to store the motif numerical. One advantage of the numerical representation is that it provides guarantee to reproduce the original molecule total charge density without fitting approximation. In their databank work[51], mostly small molecules are used. In comparison, for many cases, as we discussed previously, it is necessary to use relatively large molecules and environments to exact the charge density motifs.

In another extreme, work of Mezey et al.[42] use finite size molecular motifs to generate the charge density. Mullikan population analysis method is used in their motif charge density partition. Such partition method is only available in atomic basis set quantum chemistry computations, not in plane wave approach like ours. Besides, a finite size fragment approach might encounter difficulty if the atomic positions in a

fragment have changed a little bit, which make it necessary to recalculate the fragment. In an atom motif based approach, such small changes can be accommodated by shifting the motif positions and orientations following the atomic structure deformation.

Finally, there are many discussions for the advantages and problems for Hirshfeld style charge partitioning scheme. Bultinck et al.[64] have listed some problems when Hirshfeld method is used to analyze the charge density of each atom inside a molecule and to provide chemical insights from such analysis. In particular, the simple Hirshfeld method might not satisfy the maximum entropy partitioning criterion. They have proposed an iterative scheme to restore the maximum entropy requirement[64]. It is worth to note that there are two distinct usages for motif base charge density decomposition. One is to use it to analyze the chemistry (e.g., charge transfer from one atom to another). That is the main concern in the work of Bultinck et al. Another is to use it as a tool to reconstruct the charge density of a large system. In this usage, the physical meaning of the atomic motif is secondary, but more essential is the accuracy of the reconstructed charge density. In this sense, the iterative method might not be a good choice. This is because the iteration carried out in different molecules might render the final motifs different even for the same type of atoms, hence makes the motifs less transferable across different molecules. In this case, we believe it might be better to stay in the original simple scheme where the locality of the motif to its environment is preserved.

4. Conclusion

In summary, the charge-density patching method for the calculation of electronic structure in protein systems was developed. Test calculations for a range of glycine polypeptide systems including 2_7-ribbon, α-helix, 3_{10}-helix, and β-strand structures show that this charge patching method can generate good charge density, and electronic structures. The HOMO-LUMO band gaps calculated from the patching electronic structure are in good agreement with the results of DFT calculations. For all the systems tested, the CPM can also generate wave functions of HOMO and LUMO similar to that of DFT calculations. The transferability of the motifs is tested. It is

found that, when the binding configurations of the structures are similar, the motifs generated from one type of polypeptide structure (e.g., α-helix) can be used to other polypeptide structures (e.g., 3_{10}-helix, and β-strand). On the other hand, when the binding topology or the internal electric field caused by the dipole moment is very different, different motifs should be generated from their own structures (e.g., the 2_7-ribbon structure). But if needed, different sets of motifs can be generated from different polypeptide structure. These different sets are mostly used to describe different internal electric field effects, which is not described by the local bonding topology. One can include polarization motifs[66] to describe such long range polarization effects. Future study is needed to fully explore that topic.

With the help of accurately electronic structure information, one can investigate many biological processes involving electron transfer, such as catalytic activity of enzymes, DNA repair enzymes, electrical conductivity of DNA, and so on. However, since the purpose of this work is to test the charge-density patching method, we have only chosen Glycine polypeptide as the constitutional unit to form 2_7-ribbon, α-helix, 3_{10}-helix, and β-strand structures. More work need to be done to extend the method to proteins consisted with all the twenty types of amino acid, and complex folding and interaction exist. Nevertheless, we hope the current work opens a new approach for large scale electronic structure calculations of proteins.

Acknowledgements

This work was supported by the Director, Office of Science (SC), Basic Energy Science (BES), Materials Science and Engineering Division (MSED), of the US Department of Energy (DOE) under Contract No.DE-AC02-05CH11231 through the Materials Theory program (KC2301). We use the resource of National Energy Research Scientific Computing center (NERSC) located in Lawrence Berkeley National Laboratory. This work was also supported by the National Natural Science Foundation of China (No.11574109).

Supporting Information Available. We also applied the sum of neutral atoms to calculate the band gaps of these glycine polypeptides. The additional results are shown in the supporting information (SI).This material is available free of charge via the internet.

References

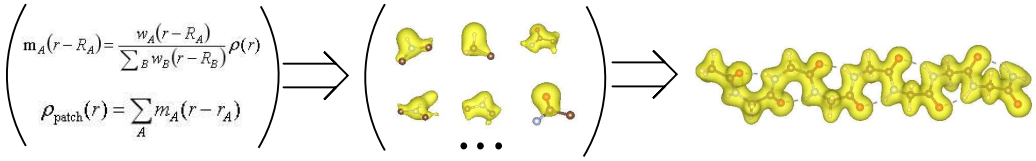
1 K. Fukushima, M. Wada, M. Sakurai, *Proteins: Struct., Funct., Genet.,* 2008, **71**, 1940.
2 S. K. Burley, J. B. Bonanno, *Curr. Opin. Struct. Boil.,* 2002, **12**, 383.
3 J.-M. Chandonia, E. S. Brenner, *Science,* 2006, **311,** 347.
4 C. Zhang, S.-H. Kim, *Curr. Opin. Chem. Boil.,* 2003, **7**, 28.
5 T. Ishida, *J. Am. Chem. Soc.,* 2010, **132**, 7104.
6 A. Fersht, Structure and Mechanism in Protein Science. A Guide to Enzyme Catalysis and Protein Folding, 2nd ed.; W. H. Freeman and Company: New York, **1999**.
7 Frey, P. A.; Hegeman, A. D. Enzymatic Reaction Mechanism. Oxford University Press: New York, **2007**.
8 Blankenship, R. E. Molecular Mechanisms of Photosynthesis. World Scientific, London, **2002**.
9 A. R. Milosavljević, C. Nicolas, M. Lj. Ranković, F. Canon, C. Miron, A. Giuliani, *J. Phys. Chem. Lett.,* 2015, **6**, 3132
10 M. G. Rossmann, P. Argos, A*nnu. Rev. Biochem.,* 1981, **50**, 497.
11 N. Jones, *Nature,* 2014, **505**, 602.
12 K. Lindorff-Larsen, R. B. Best, M. A. Depristo, C. M. Dobson, M. Vendruscolo, *Nature,* 2005, **433**, 128.
13 F. Pichierri, *Chem. Phys. Lett.,* 2005, **410**, 462.
14 M. E. Siwko, S. Corni, *Phys. Chem. Chem. Phys.*, 2013, **15**, 5945.
15 R. R. Nazmutdinov, M. D. Bronshtein, T. T. Zinkicheva, Q. J. Chi, J. D. Zhang, J. Ulstrup *Phys. Chem. Chem. Phys.,* 2012, **14**, 5953.
16 A. De la Lande, N. S. Babcock, J. Řezáč, B. Lévy, B. C.Sanders, D. R. Salahub, *Phys. Chem. Chem. Phys.,* 2012, **14**, 5902.
17 J. C. Genereux, A. K. Boal, J. K. Barton, *J. Am. Chem. Soc.,* 2010, **132**, 891.
18 Jensen, F. *Introduction to Computational Chemistry*, 2nd ed. Wiley, Chichester, **2007**.
19 S. Goedecker, *Rev. Mod. Phys.,* 1999, **71**, 1085.
20 C. Ochsenfeld, J. Kussmann, D. S. Lambrecht, Linear-Scaling Methods in Quantum Chemistry, *Rev. Comp. Chem.*, Vol. **23**, K. Lipkowitz and D. B. Boyd (Eds.), pp. 1-82, Wiley-VCH, New York, **2007**.
21 S. J. Fox, C. Pittock, T. Fox, C. S. Tautermann, N. Malcolm, C.-K. Skylaris, *J. Chem. Phys.,* 2011, **135**, 224107.
22 M. S. Gordon, D. G. Fedorov, S. R. Pruitt, L. V. Slipchenko, *Chem. Rev.,* 2012, **112**, 632
23 A. S. P. Gomes, C. R. Jacob, *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.,* 2012, **108**, 222.
24 F. D. Fedorov, K. Kitaura, *J. Phys. Chem. A,* 2007, **111**, 6904.
25 D. G. Fedorov, K. Ishimura, T. Ishida, K. Kitaura, P. Pulay, S. Nagase, *J. Comput. Chem.,* 2007, **28**, 1476.
26 S. R. Gadre, V. Ganesh, *J. Theor. Comput. Chem.*, 2006, ***5***, 835.
27 M. Elango, V. Subramanian, A. P. Rahalkar, S. R. Gadre, N. Sathyamurthy, *J. Phys. Chem. A*, 2008, **112,** 7699.
28 W. Li, S. H. Li, Y. S. Jiang, *J. Phys. Chem. A,* 2007, **111**, 2193.
29 S. G. Hua, W. J. Hua, S. H. Li, *J. Phys. Chem. A,* 2010, **114**, 8126.
30 D. W. Zhang, J. Z. H. Zhang, *J. Chem. Phys.,* 2003, **119**, 3599.

31 A. M. Gao, D. W. Zhang, J. Z. H. Zhang, Y. Zhang, *Chem. Phys. Lett.* 2004, **394**, 293.
32 X. He, J. Z. H. Zhang, *J. Chem. Phys.,* 2006, **124**, 184703.
33 Y. Mei, D. W. Zhang, J. Z. H. Zhang, *J. Phys. Chem. A,* 2005, **109**, 2.
34 X. H. Chen, J. Z. H. Zhang, *J. Chem. Phys.* 2016, **125**, 044903.
35 X. W. Wang, J. F. Liu, J. Z. H. Zhang, X. He, *J. Phys. Chem. A,* 2013, **117**, 7149.
36 K. Kiewisch, C. R. Jacob, L. Visscher, *J. Chem. Theory Comput.,* 2013, **9**, 2425.
37 S. Dag, S. Z. Wang, L.-W. Wang, *Nano Lett.,* 2011, **11**, 2348.
38 L.-W. Wang, A*nnu. Rev. Phys. Chem.,* 2010, **61**, 19.
39 L.-W. Wang, Z. J. Zhao, J. Meza, *Phys. Rev. B,* 2008, **77**, 165113.
40 Z. J. Zhao, J. Meza, L.-W. Wang, *J. Phys: Condens. Matter,* 2008, **20**, 294203.
41 J. Ma, L.-W. Wang, *Nano Lett.,* 2015, **15**, 248.
42 P. D. Walker, P. G. Mezey, *J. Am. Chem. Soc.,* 1993, **115**, 12423.
43 Zs. Szekeres, T. E. Exner, P. G. Mezey, *Int. J. Quantum Chem.* 2005, **104**, 847.
44 M. A. Spackman, E. N. Maslen, *J. Phys. Chem.* 1986, **90**, 2020.
45 A. M. Pendás, V. Luaña, L. Pueyo, E. Francisco, P. Mori-Sánchez, *J. Chem. Phys.,* 2002, **117**, 1017.
46 E. R. Johnson, S. Keinan, P.Mori-Sánchez, J. Contreras-García, A. J. Cohen, W.-T. Yang, *J. Am. Chem. Soc.,* 2010, **132**, 6498.
47 T. Koritsanszky, A. Volkov, P. Coppens, *Acta Crystallogr., Sect. A,* 2002, **58**, 464.
48 R. Wiest, V. Pichon-Pseme, M. Benard, C. Lecomte, *J. Phys. Chem.* 1994, **98**, 1351.
49 V. Pichon-Pesme, C. Lecomte, H. Lachkar, *J. Phys. Chem.* 1995, **99**, 6242.
50 A. Volkov, X. Li, T. Loritasnszky, P. Coppens, *J. Phys. Chem. A,* 2004**, 108***,* 4283.
51 P. M. Dominiak, A. Volkov, X. Li, M. Messerschmidt, P. Coppens, *J. Chem. Theory Comput.,* 2007, **3**, 232.
52 P. Kumar, S. A. Bojarowski, K.N. Jarzembska, S. Domagala, K. Vanommeslaeghe, A. D. MachKerell, Jr., P. M. Dominiak, *J. Chem. Theory Comput.,* 2014, **10**, 1652.
53 L.-W. Wang, *Phys. Rev. Lett.,* 2002, **88**, 256402.
54 J. B. Li, L.-W. Wang, *Phys. Rev. B,* 2005, **72**, 125325.
55 J. B. Li, L.-W. Wang, *Phys. Rev. B,* 2003, **67**, 033102.
56 L.-W. Wang, *Phys. Rev. B,* 2002, **65**, 153410.
57 N. Vukmirović, L.-W. Wang, *J. Chem. Phys.,* 2008, **128**, 121102.
58 N. Vukmirović, L.-W. Wang, *J. Phys. Chem. B,* 2009, **113**, 409.
59 N. Vukmirović, L.-W. Wang, *Nano Lett*., 2009, **9**, 3996.
60 J. M. Granadino-Roldán, N. Vukmirović, M. Fernández-Gómez, L.-W. Wang, *Phys. Chem. Chem. Phys.,* 2011, **13**, 14500.
61 L.W. Wang; A. Zunger, *J. Chem. Phys.,* 1994,**100**, 2394.
62 J. Ren, N. Vukmirovic, L.-W. Wang, *Phys. Rev. B*, 2013, **87**, 205117.
63 L.W. Wang, *Phy. Rev. B.,* 1994, **49**, 10154.
64 P. Bultinck, C. V. Alsenoy, P. W. Ayers, R. Carbó-Dorca, *J. Chem. Phys.,* 2007, **126**, 144111.
65 Y.-D. Wu, Y.-L. Zhao, *J. Am. Chem. Soc.,* 2001, **123**, 5313.
66 L.W. Wang, X. Cartoixa, *Phys. Rev. B* 2007,**75**, 205334.

For Table of Contents use only:

Charge patching method for the calculation of Electronic Structure of Polypeptide

Chang-Liang Sun*, Li-Ping Liu, Fubo Tian, Fu Ding, Lin-Wang Wang*



Based on CPM method, the charge densities of polypeptide can be generated and the electronic structure can be further calculated.