

The 2nd ISV pre-conference Computational Vaccinology Workshop (ICoVax 2012 2012/10/13)



Protein-ligand Binding Region Prediction based on Geometric Features and CUDA Acceleration

Ying-Tsang Lo¹, Hsin-Wei Wang¹, Tun-Wen Pai^{1,3*},
Wen-Shoung Tzou^{2,3}, Hui-Huang Hsu⁴, Hao-Teng Chang^{5,6}

¹Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan, R.O.C.

²Department of Life Sciences, National Taiwan Ocean University, Keelung, Taiwan, R.O.C.

³Center of Excellence for Marine Bioenvironment and Biotechnology, National Taiwan Ocean University, Keelung, Taiwan, R.O.C.

⁴Department of Computer Science and Information Engineering, Tamkang University, Taipei, Taiwan, R.O.C.

⁵Graduate Institute of Molecular Systems Biomedicine, China Medical University, Taichung, Taiwan, R.O.C.

⁶China Medical University Hospital, Taichung, Taiwan, R.O.C..



PLB-SAVE

Protein-Ligand Binding region prediction based on features of Solid Angle, Volume, and dEpth

Outlines:

- Introduction
- Methods and System Configuration
- Materials and Experimental Results
- System Demonstration
- Conclusions



PLB-SAVE

Protein-Ligand Binding region prediction based on features of Solid Angle, Volume, and dEpth

Protein binding region and binding site prediction:

- the first *in silico* step to study protein functions regarding to structure-based drug design and vaccine development
- one of the best ways to understand the mechanisms, principles and specificities in [Molecular Recognition](#)
- providing deterministic information for
 - protein function annotation
 - construction of protein-protein interaction networks
 - high-through virtual screening for drug design and discovery
 - vaccine design and development



Related Research

- Traditional Way for protein-ligand binding analysis :
pockets/cavities
- **Two major categories** : geometry based / energy based

Geometry based: grid based, sphere based, and α -shape based

- LIGSITE (Hendlich et al., 1997), LIGSITE^{CS} (Huang and Schroeder, 2006), PocketPicker (Weisel et al., 2007), GHECOM (Kawabata, 2010) and ConCavity (Capra et al., 2009)
- SURFNET (Laskowski, 1995), PASS (Brady and Stouten, 2000), PHECOM (Kawabata and Go, 2007) and POCASA (Yu et al., 2010).
- CAST (Binkowski et al., 2003; Dundas et al., 2006) and Fpocket (Le Guilloux et al., 2009).

Energy based:

- Q-SiteFinder (Laurie and Jackson, 2005) SiteHound (Gherzi and Sanchez, 2009; Hernandez et al., 2009)

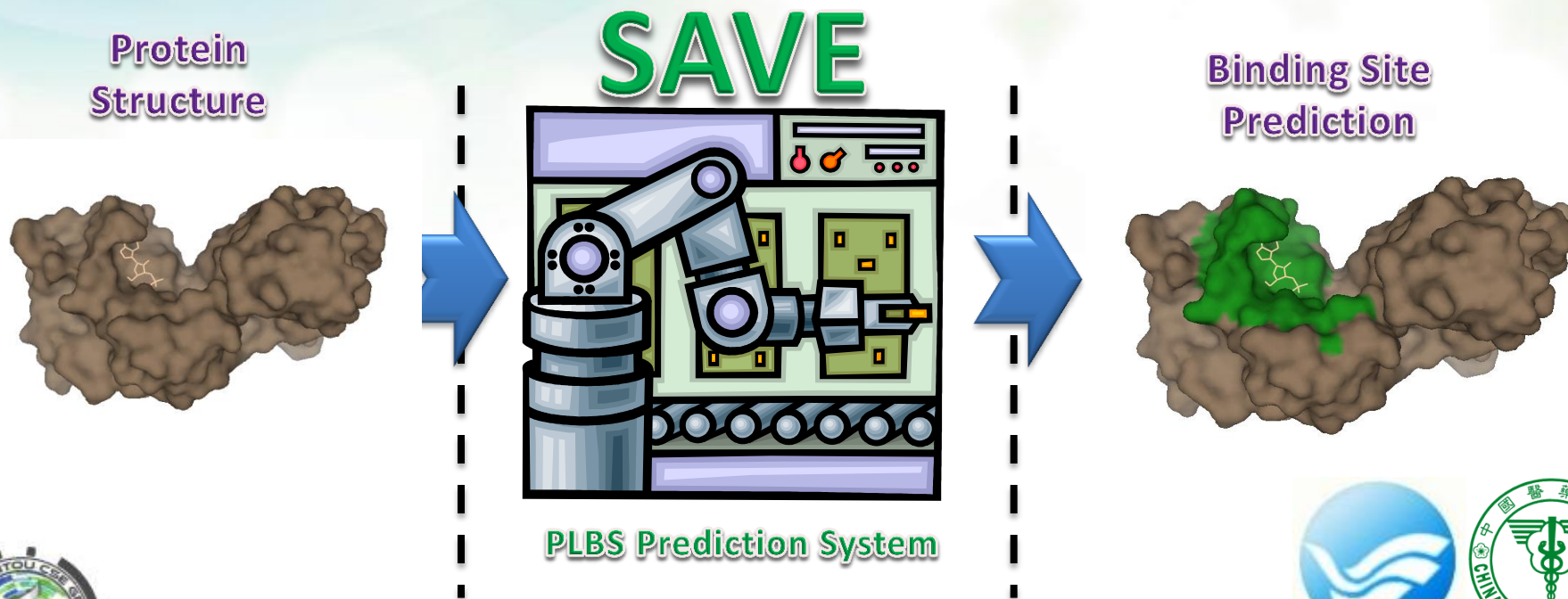
- **MetaPocket 2.0 (MPK2)**

- (LIGSITE^{CS}, SURFNET, PASS, Q-SiteFinder, Fpocket, GHECOM, ConCavity and POCASA) (Huang, 2009/2011)
- **MPK2 achieved >12% success rate over the best single method**

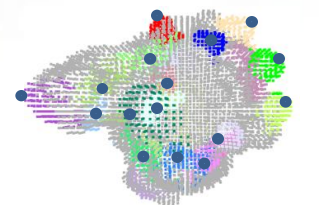
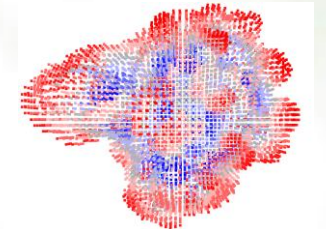
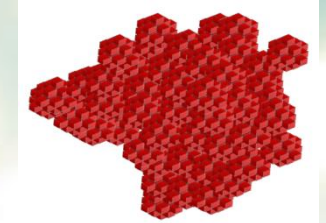
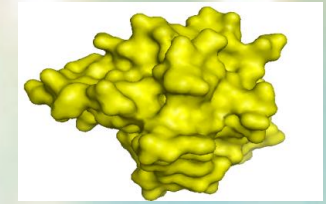
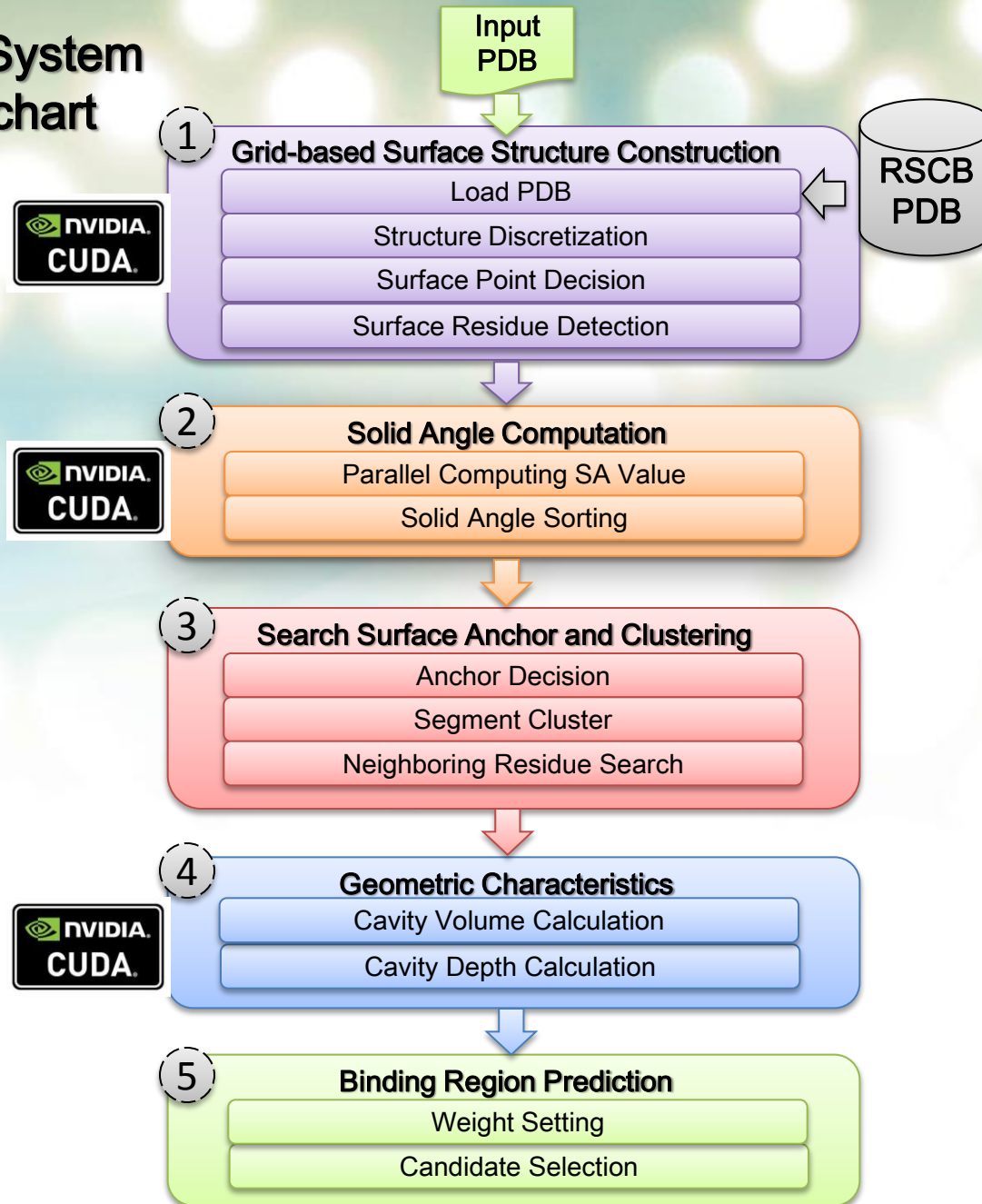


Our GOAL

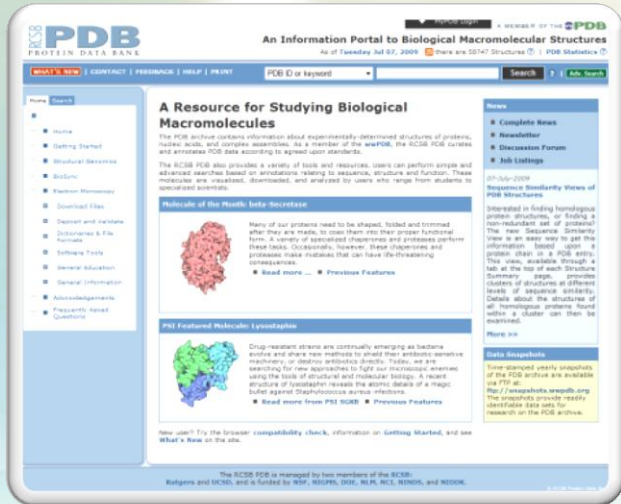
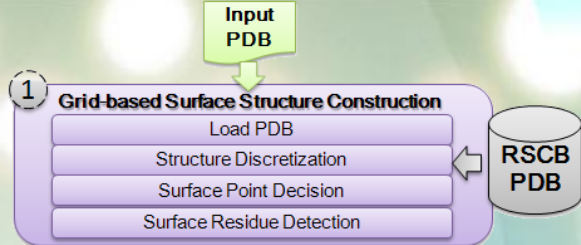
- To design an effective and efficient system to predict **protein-ligand binding regions/sites**
- Effective aspect: using simple/straight forward geometric features
- Efficient aspect: employing CUDA acceleration
- Better performance than previous existing systems



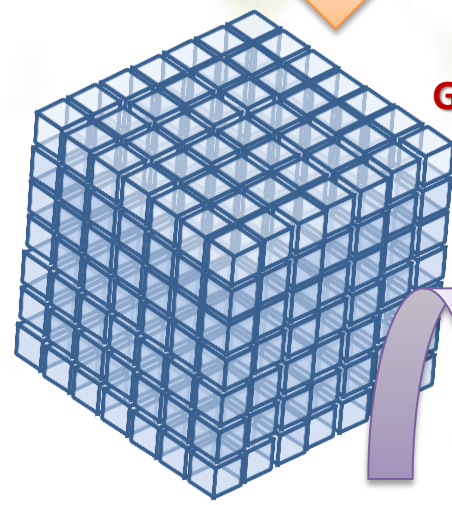
SAVE System Flowchart



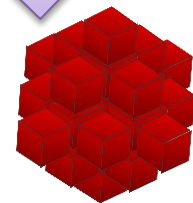
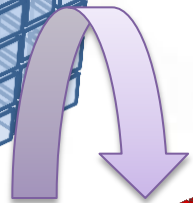
System Initialization



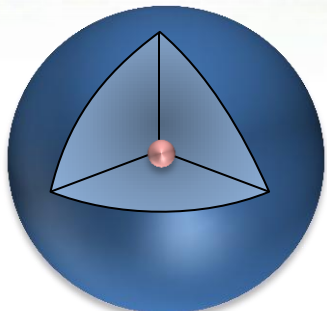
Protein atom



Grid Discretization



A protein atom in grid mode

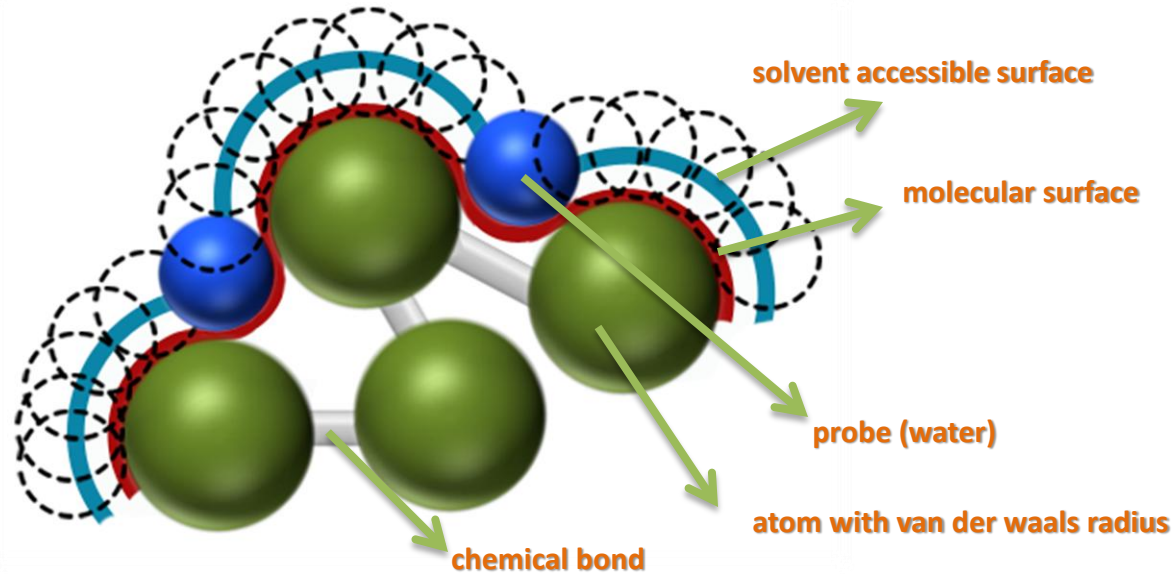
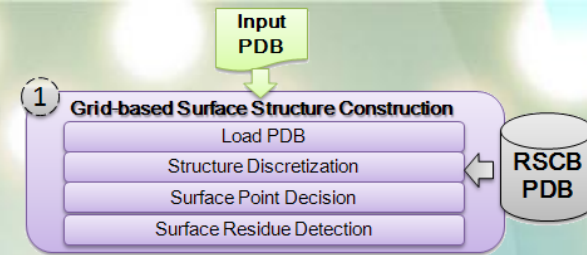


Obtained coordinates of an atom and corresponding radii



Surface Extraction(1/3)

Protein Surface Definition (Connolly, 1983)

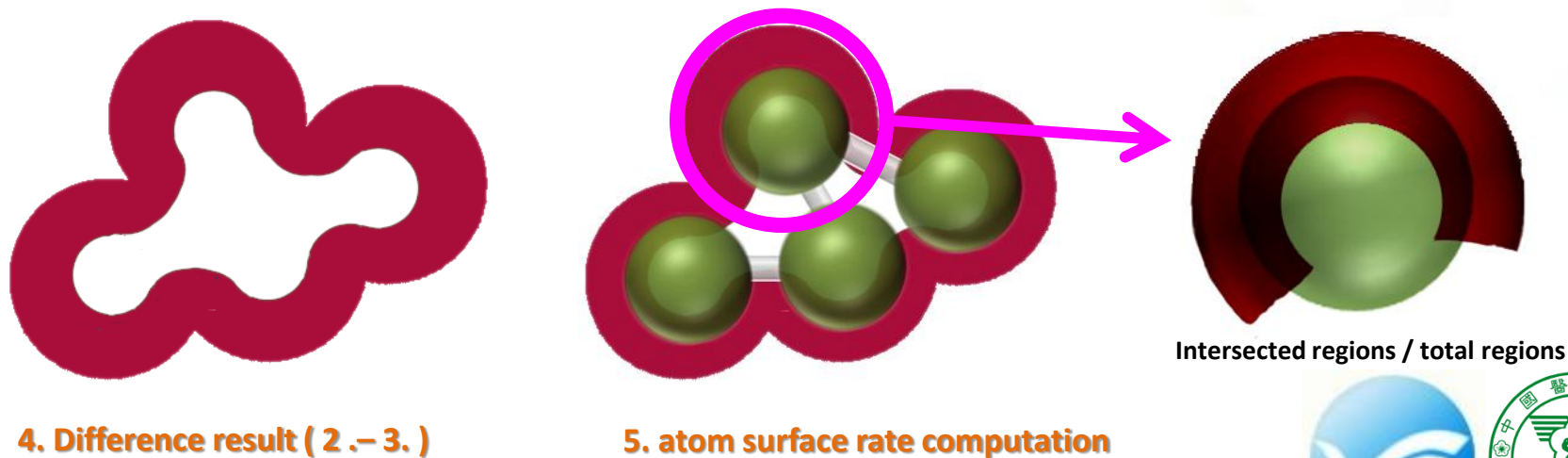
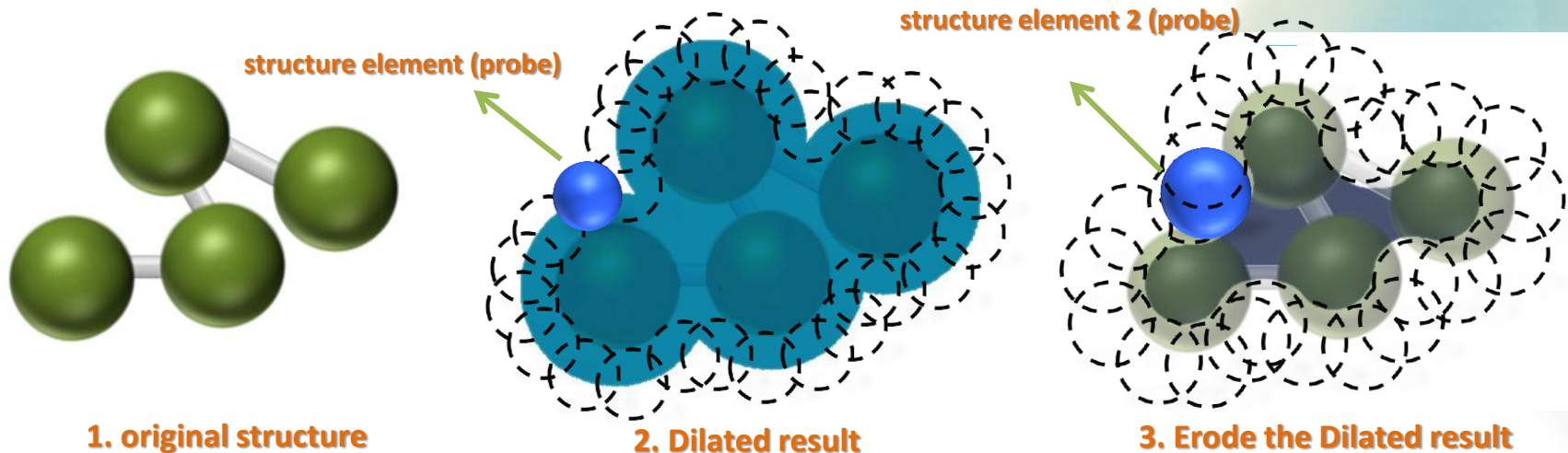
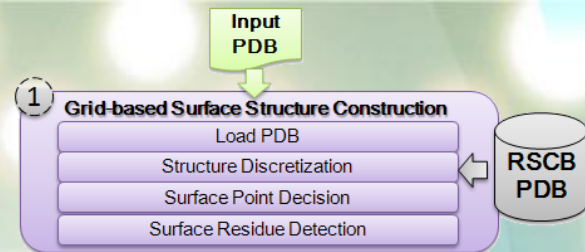


→ The proposed system utilized 3-D mathematical morphology to extract protein surface.

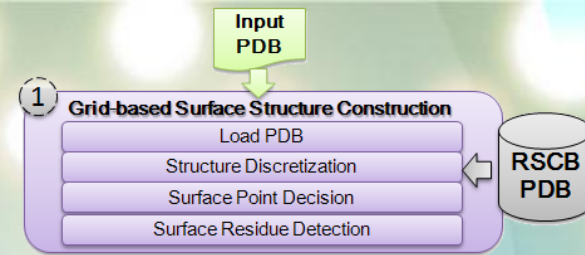


Surface Extraction(2/3)

3-D mathematical morphology operators and flowchart

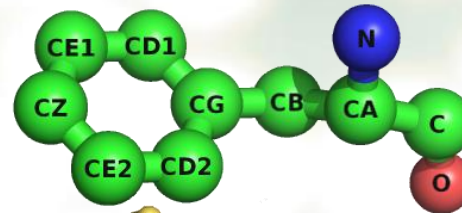


Surface Extraction(3/3)

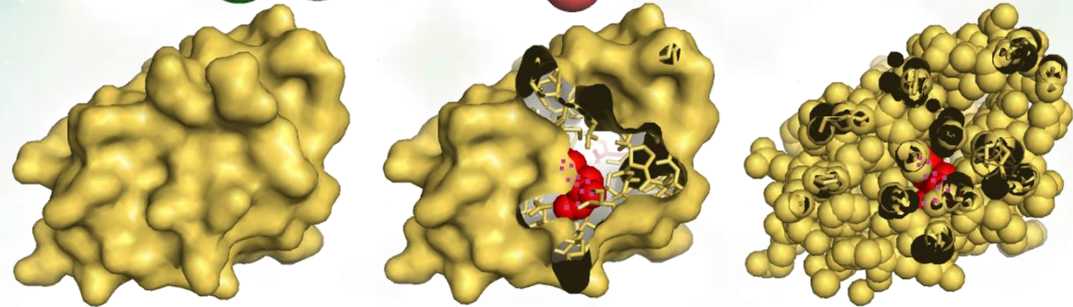


- After calculated all side-chain atoms of each residue in the query protein, the residue surface rate was computed by this formula:

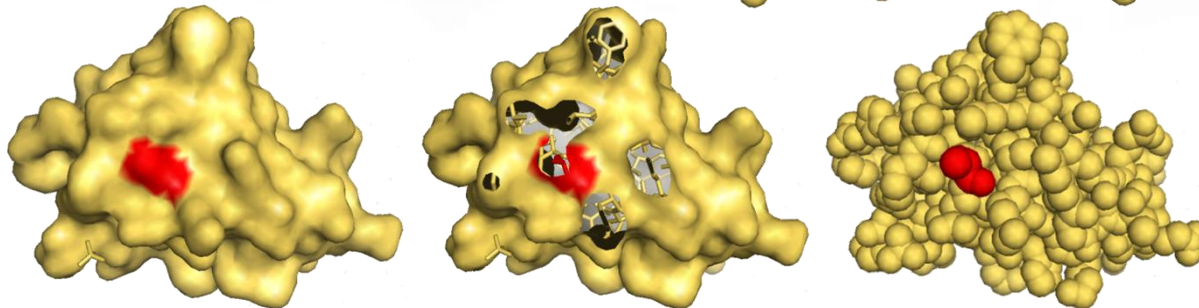
$$SR(r) = \left\{ i \in R : \frac{1}{N} \sum_{i=1}^N AR(i) \right\}$$



→ When $SR(r) = 0.0$



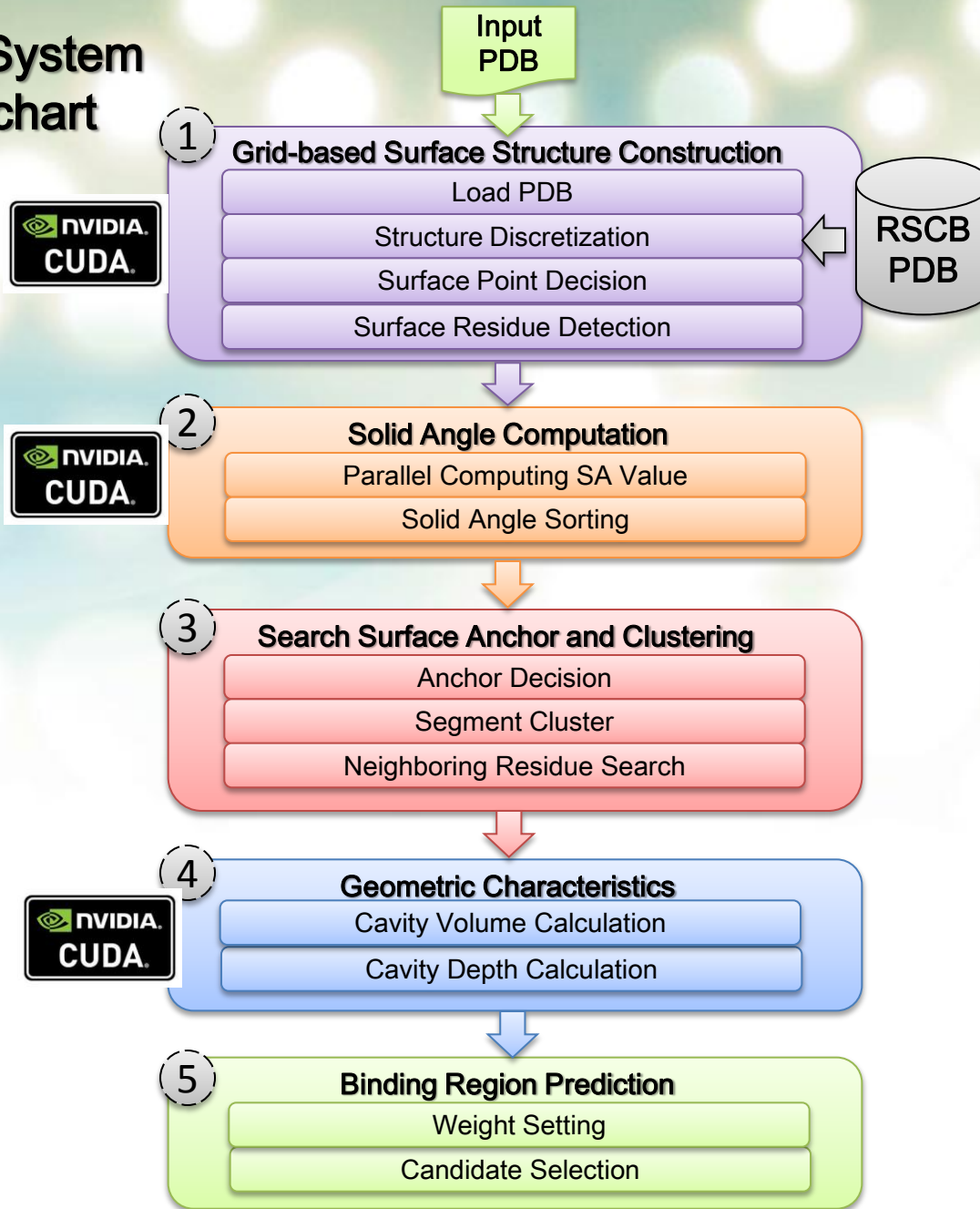
← When $SR(r) = 75.0$

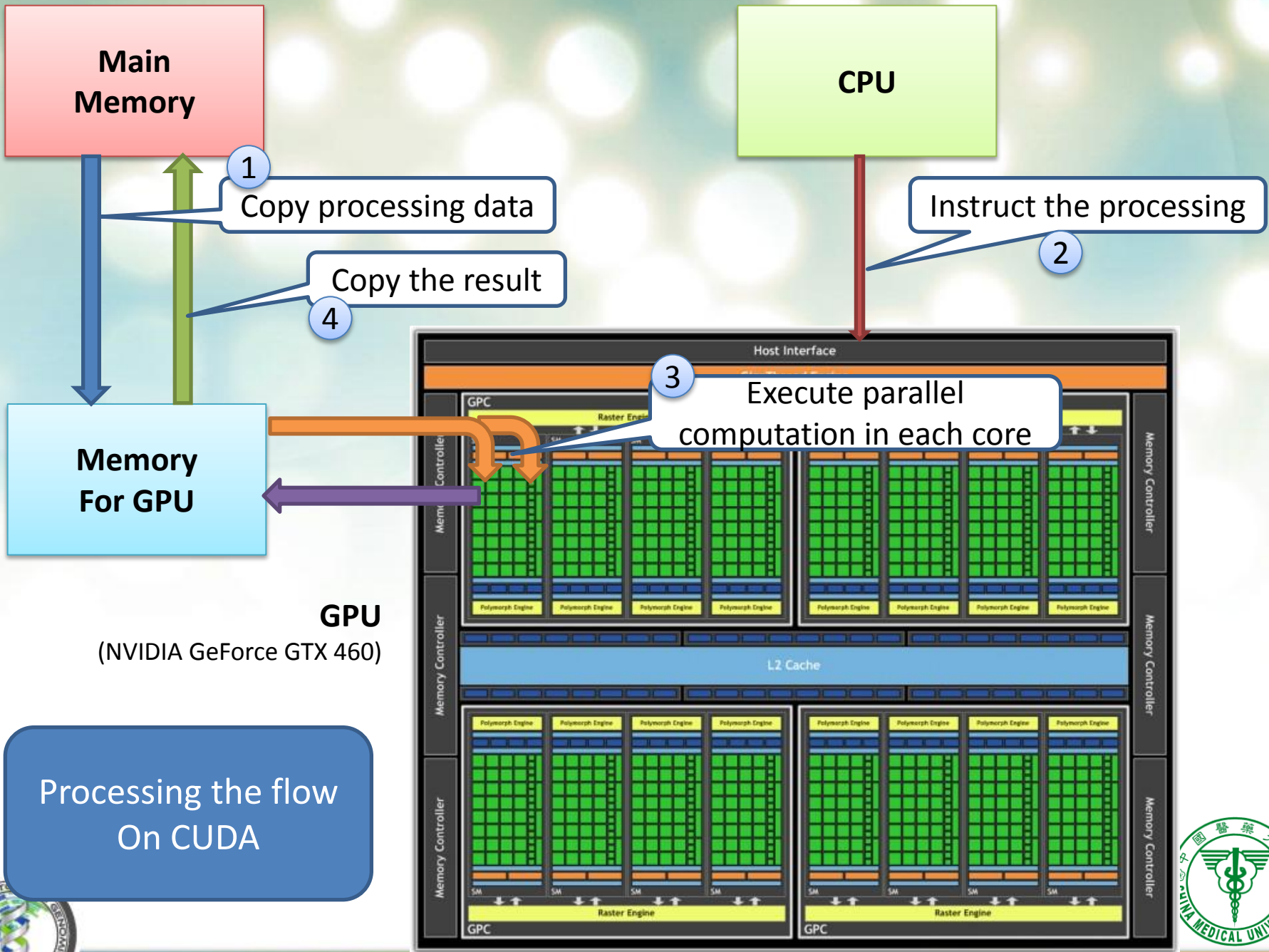


- In SAVE system, each detected voxel on the surface were applied to calculate its corresponding solid angle feature.



SAVE System Flowchart





What is solid angle?

2

Solid Angle Computation

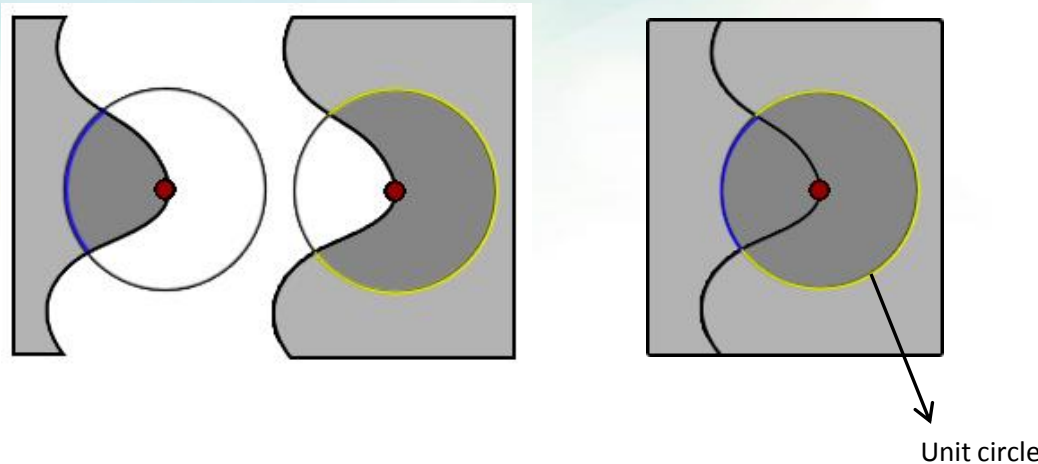
Parallel Computing SA Value

Solid Angle Sorting

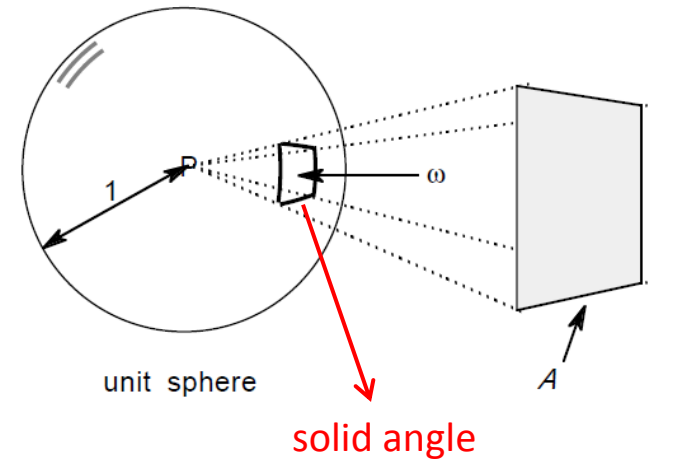


- The **solid angle**, Ω , is a measure of how large an object appears to an observer looking from a point.

2D



3D



- The compact matched two solid-angles is 2π in 2D space, and 4π in 3D space.



Formula for the solid angle of protein surface

2

Solid Angle Computation

Parallel Computing SA Value

Solid Angle Sorting



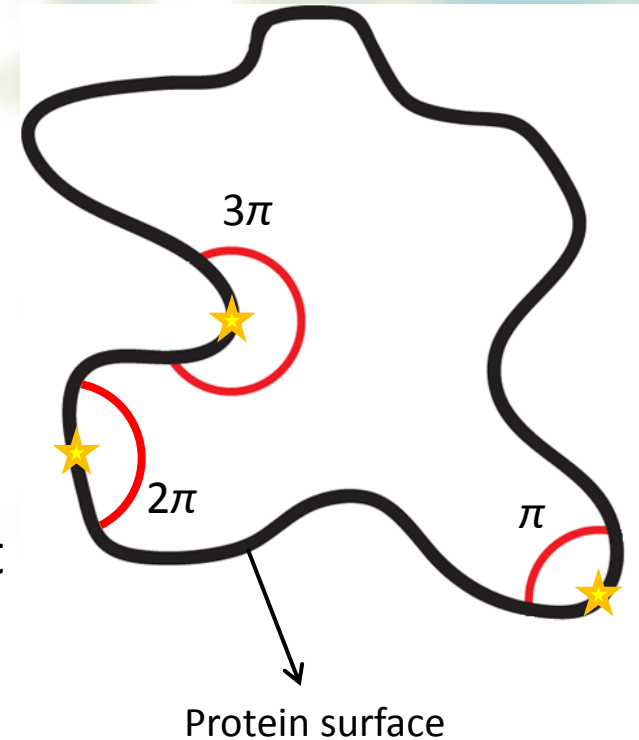
For each surface point in 3D:

- $SA = (inSP / nP) * 4\pi$
 - SA : the value of solid-angle
 - $inSP$: the number of voxels within the unit sphere which were located inside the protein (red arc)
 - nP : the entire voxels within unit sphere

$SA < 2\pi \rightarrow$ convex surface

$SA = 2\pi \rightarrow$ flat surface

$SA > 2\pi \rightarrow$ concave surface



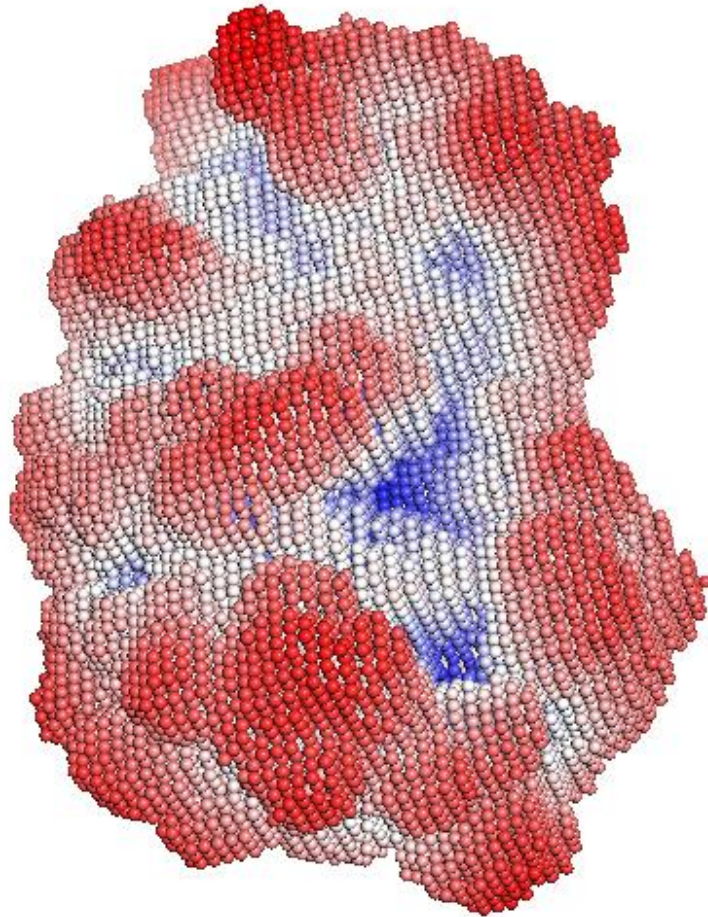
Solid Angle Computation

2

Solid Angle Computation

Parallel Computing SA Value

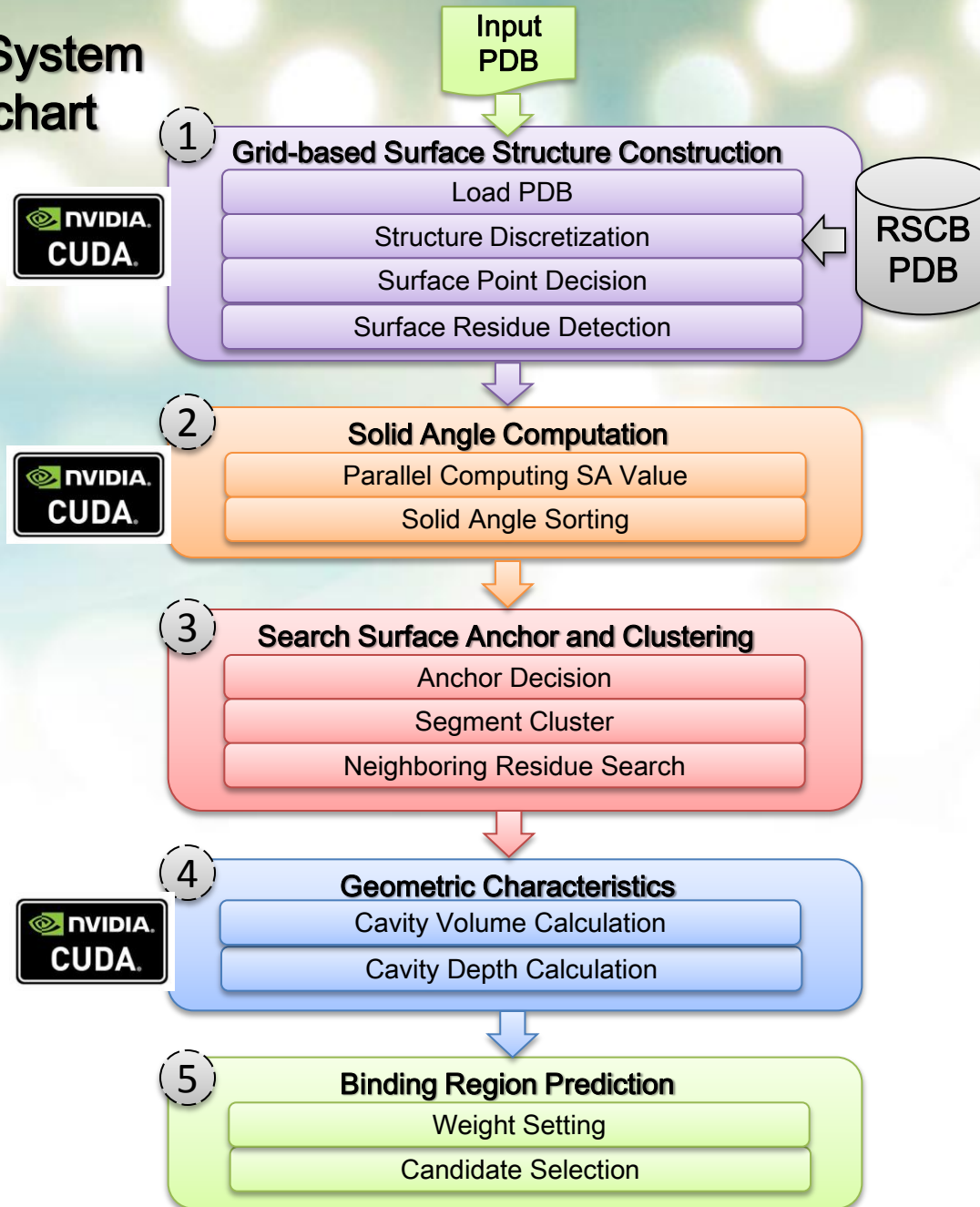
Solid Angle Sorting



- An example of calculated solid angles for all surface voxels of PDB:1GOY protein
- **Red spheres** represented the solid angles with small values
=> Located on **convex regions**
- **Blue spheres** represented relatively large values of solid angles
=> Located on **concave regions**
- White spheres represented surface voxels located on flat regions



SAVE System Flowchart

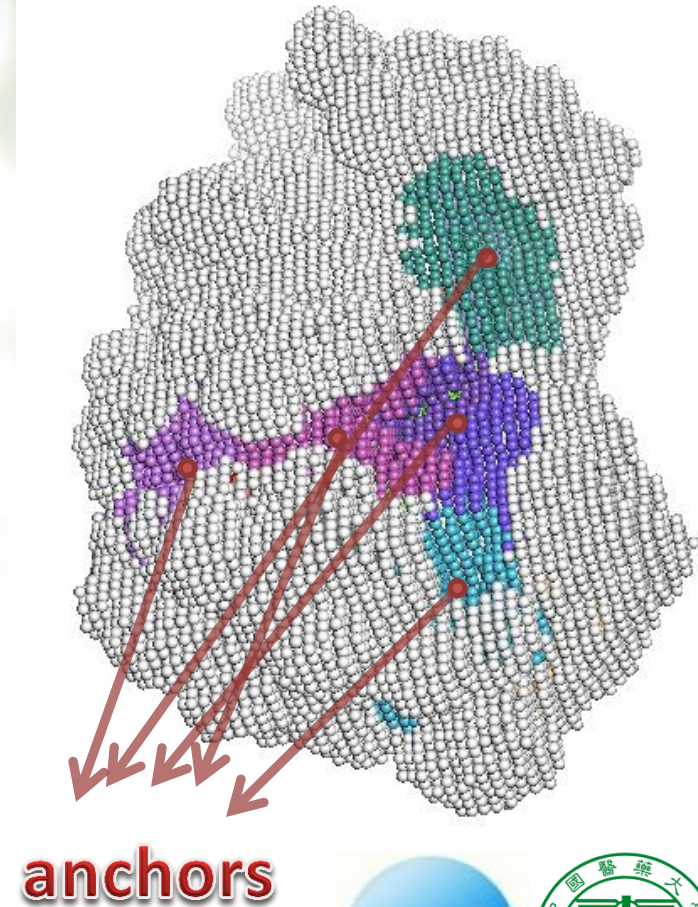


Search Surface Anchor and Clustering

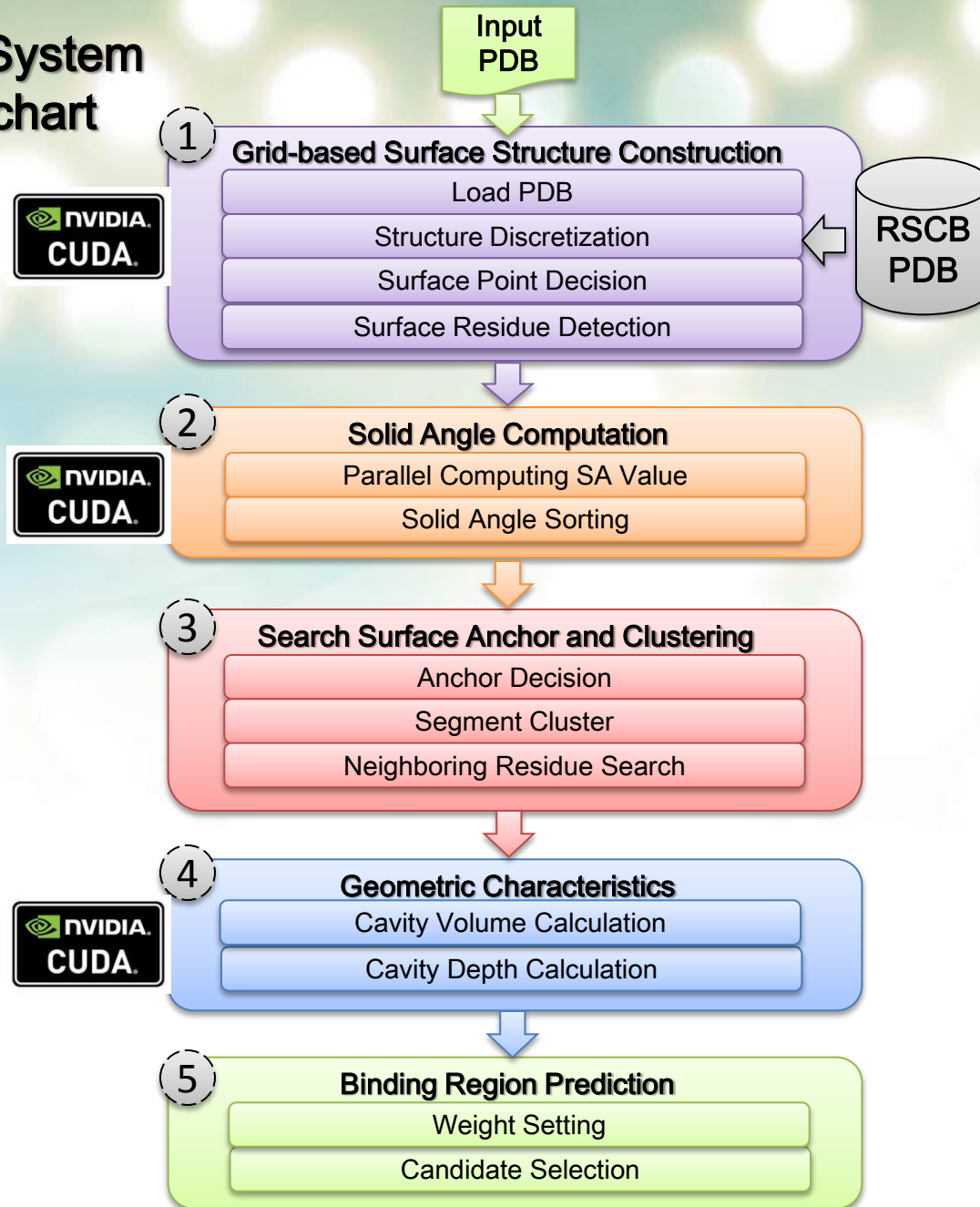
3 Search Surface Anchor and Clustering

- Anchor Decision
- Segment Cluster
- Neighboring Residue Search

- Surface voxels with solid angles ranked in **top 20%** were selected and clustered into representative groups
- Two neighboring surface voxels would be clustered into an identical group with a **threshold distance (8 Å)** and **solid angle value at similar level**
- The largest solid angle in a group was considered as the representative **anchor**



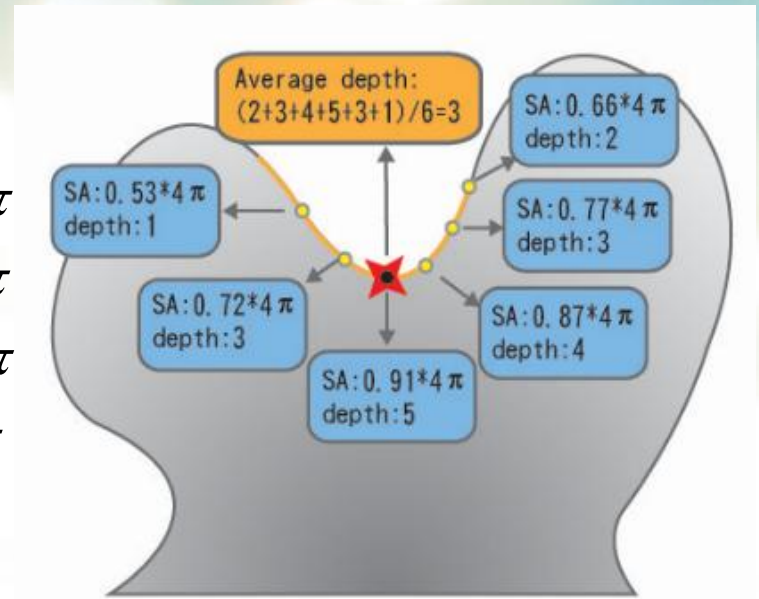
SAVE System Flowchart



Average depth of a cavity

Depth index transformation formula

$$Depth = \begin{cases} 5 & \text{if } SA > 0.9 * 4\pi \\ 4 & \text{if } 0.8 * 4\pi < SA \leq 0.9 * 4\pi \\ 3 & \text{if } 0.7 * 4\pi < SA \leq 0.8 * 4\pi \\ 2 & \text{if } 0.6 * 4\pi < SA \leq 0.7 * 4\pi \\ 1 & \text{if } 0.5 * 4\pi < SA \leq 0.6 * 4\pi \\ 0 & \text{else} \end{cases}$$



- The average depth indicator of a cavity candidate was obtained by taking an average of transformed depths in the cluster.

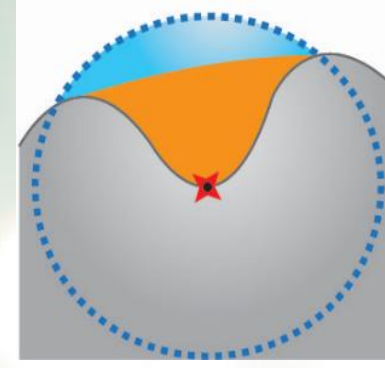
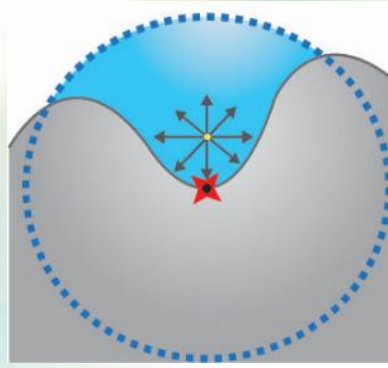
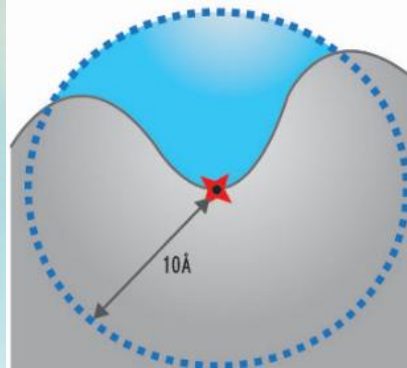
Volume of a potential cavity

4

Geometric Characteristics

Cavity Volume Calculation

Cavity Depth Calculation



The Idea by:2006_LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation

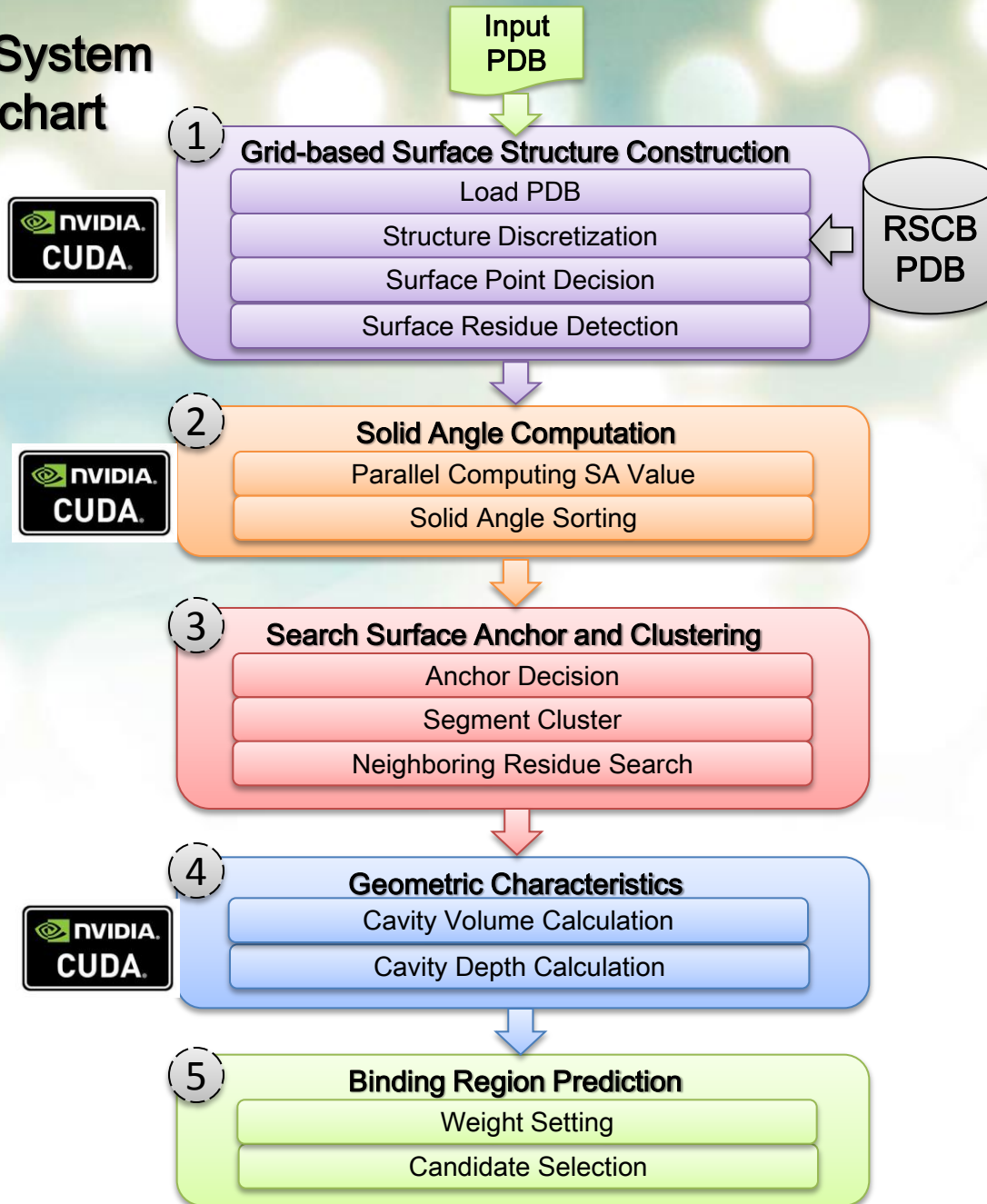
Step1: Taking the anchor surface residue as a center and formulating a **virtual sphere** with a radius of 10 Å

Step2: Each voxel taking 7 directional vectors to extend
If > 3 **vectors intersecting with the query protein**
→ this voxel is an inner voxel within the cavity

Step3: Examining all voxels in the virtual sphere, **total interior voxel counts** → **volume** of the cavity



SAVE System Flowchart



Binding Region Prediction

5

Binding Region Prediction

Weight Setting

Candidate Selection

$$RV(p) = \frac{CD(p)_{avg}}{CD_{max}} \times w_1 + \frac{CV(p)}{CV_{max}} \times w_2$$

- $RV(p)$ is the ranking value for anchor residue p cluster
- $CD(p)_{avg}$ is the **average depth** value of p cluster
- CD_{max} is the maximum depth of the query protein
- $CV(p)$ is the **volume** of p cluster
- CV_{max} is the maximum volume of the query protein

– P.S. sum of w_1 and w_2 is equal to 1



Experimental Results



LigASite dataset



- LigASite version 9.5 released July 2011
- **388** non-redundant **unbound protein structures** from LigASite dataset (APO)
- **388** non-redundant **bound protein structures** from LigASite dataset(HOLO)
- LigASite dataset provide residue numbers of binding site and PDB IDs
- Website: <http://www.bigre.ulb.ac.be/Users/benoit/LigASite/index.php?home>



Performance measurement

pre \ data	T	F
P	TP	FP
N	FN	TN

- TP is the number of true positive
- FP is the number of false positive
- TN is the number of true negative
- FN is the number of false negative



Performance measurement

- Sensitivity (SE) = $TP \div [TP + FN]$
- Specificity (SP) = $TN \div [TN + FP]$
- Positive Prediction Value (PPV) = $TP \div [TP + FP]$
- Accuracy (ACC) = $[TP + TN] \div [TP + TN + FN + FP]$
- Matthews correlation coefficient (MCC) =
$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$



Ten-fold Cross Validation

PLB-SAVE (10-fold cross-validation)	APO-388 Proteins (unbound)	HOLO-388 Proteins (bound)
Sensitivity	0.579043	0.642564
Specificity	0.972336	0.976363
Accuracy	0.942588	0.955269
PPV	0.634765	0.651935
MCC	0.566041	0.613089

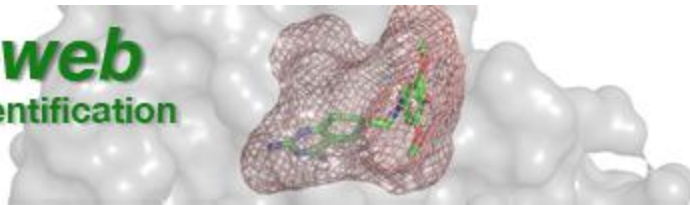
Prediction system evaluated under a ten-fold cross-validation mechanism



System Comparison

System	Year	Nation	Laboratory / University	Journal / conference	Method
SITEHOUND	2009	USA	Mount Sinai School of Medicine	Nucleic Acids Research	interaction energy and cavity volume
MegaPocket2.0 (MPK2)	2011	Germany	Technical University of Dresden	Bioinformatics	Consensus method (combine with LIGSITE ^{CS} , PASS, QSiteFinder, SURFNET, Fpocket , GHECOM, ConCavity and POCASA)

SiteHound-web
Ligand Binding Site Identification
in Protein Structures



metaPocket 2.0



Comparison with other systems on LigASite dataset

Comparison of performance of system for **388 APO** (unbound protein structures)

System	Fail number	Complete Rate
SAVE	0	100%
MegaPocket2.0	207	53.4%
SITEHOUND	15	3.9%

System	Execution time
SAVE	2sec ~ 60sec
MegaPocket2.0	20sec ~ 120sec
SITEHOUND	60sec ~ 600sec

System	Top1	Top2	Top3	Total	success rates
SAVE	146	87	79	312	80.4%
MegaPocket2.0	113	31	13	157	40.5%
SITEHOUND	109	72	43	224	57.7%

If the forecasting results with sensitivity value were less than 25% is considered as an error prediction.



Comparison with Comparative System on LigASite

Comparison of performance of system for **388 HOLO**(bound protein structures).

System	Fail number	Complete Rate
SAVE	0	100%
MegaPocket2.0	240	61.9%
SITEHOUND	14	3.6%

System	Execution time
SAVE	2sec ~ 60sec
MegaPocket2.0	20sec ~120sec
SITEHOUND	60sec ~ 600sec

System	Top1	Top2	Top3	Total	success rates
SAVE	146	91	80	317	81.7%
MegaPocket2.0	92	27	21	140	36.1%
SITEHOUND	159	87	46	292	75.2%

If the forecasting results with sensitivity value less than 25% is considered as an error prediction.



SAVE v.s SiteHound

APO (unbound structures)	PLB-SAVE (373 proteins)	SiteHound (373 proteins)
Sensitivity	<u>0.527</u>	0.379
Specificity	<u>0.968</u>	0.955
Accuracy	<u>0.934</u>	0.912
PPV	<u>0.583</u>	0.399
MCC	<u>0.509</u>	0.332

HOLO (bound structures)	PLB-SAVE (374 proteins)	SiteHound (374 proteins)
Sensitivity	<u>0.623</u>	0.538
Specificity	<u>0.975</u>	0.975
Accuracy	<u>0.953</u>	0.952
PPV	<u>0.629</u>	0.625
MCC	<u>0.589</u>	0.585



SAVE v.S MPK2

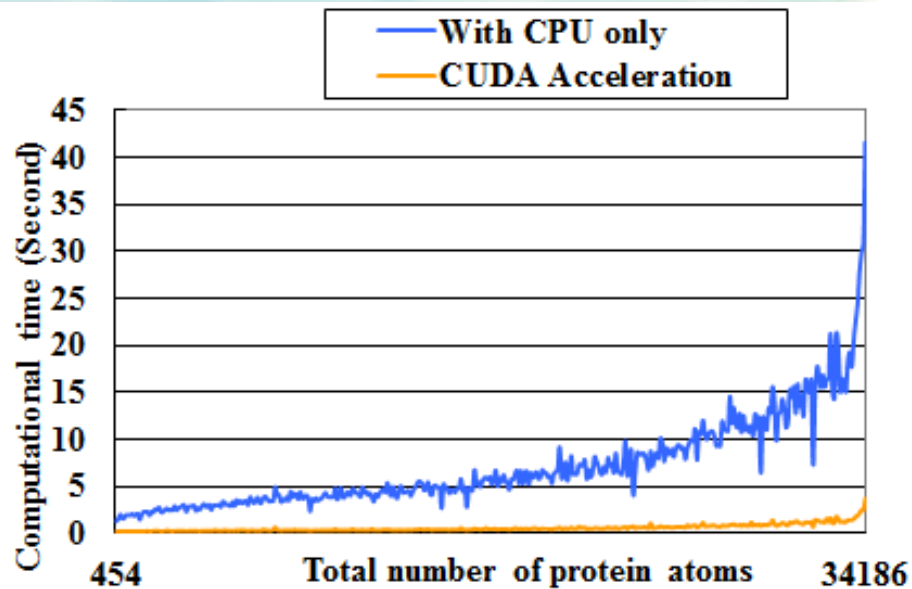
APO (unbound structures)	PLB-SAVE (171 proteins)	MPK2 (171 proteins)
Sensitivity	0.567	<u>0.710</u>
Specificity	<u>0.953</u>	0.904
Accuracy	<u>0.905</u>	0.878
PPV	<u>0.609</u>	0.478
MCC	<u>0.524</u>	0.500

HOLO (bound structures)	PLB-SAVE (148 proteins)	MPK2 (148 proteins)
Sensitivity	0.673	<u>0.861</u>
Specificity	<u>0.959</u>	0.912
Accuracy	<u>0.927</u>	0.905
PPV	<u>0.654</u>	0.556
MCC	0.615	<u>0.634</u>

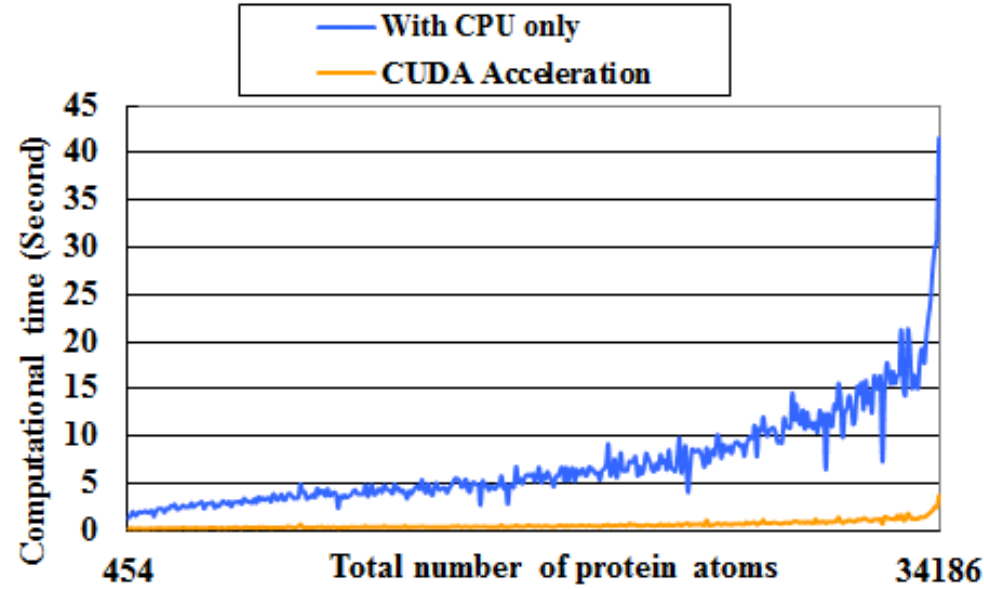


Running Time Comparison

388 unbound dataset (APO)



388 bound dataset (HOL0)



System Demonstration

PLB-SAVE

(<http://save.cs.ntou.edu.tw/>)



PLB-SAVE
Protein-Ligand Binding region prediction based on features
of Solid Angle, Volume, and dEpth

Main

Result

Method

Contact

Welcome to our system(Use [Firefox](#) to get the best performance!)

Enter a PDB ID and its chain ID:

PDB ID:

or upload a pdbfile:

sample_PDB:[1NOA]



PLB-SAVE

Protein-Ligand Binding region prediction based on features of Solid Angle, Volume, and dEpth

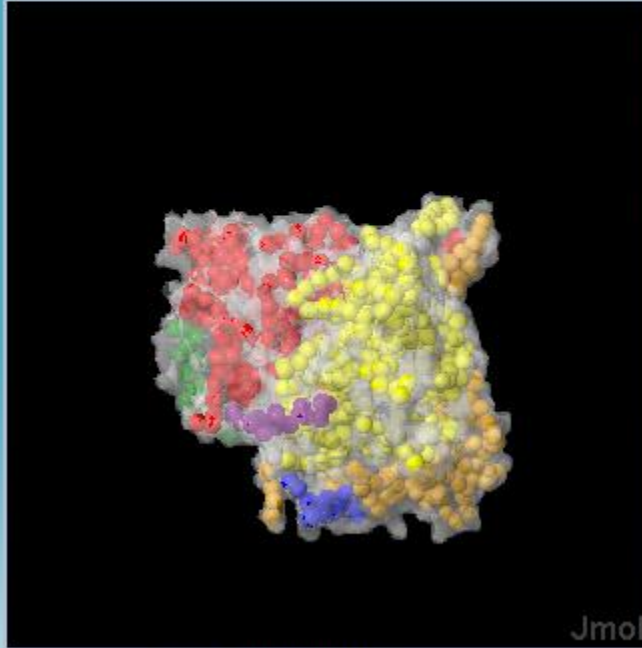
Main

Result

Method

Contact

~ Welcome to our system! Use [Linkbox](#) to get the best performance! ~



Jmol

Select 1NOV its chain :

ALL A B C D E F

Ok



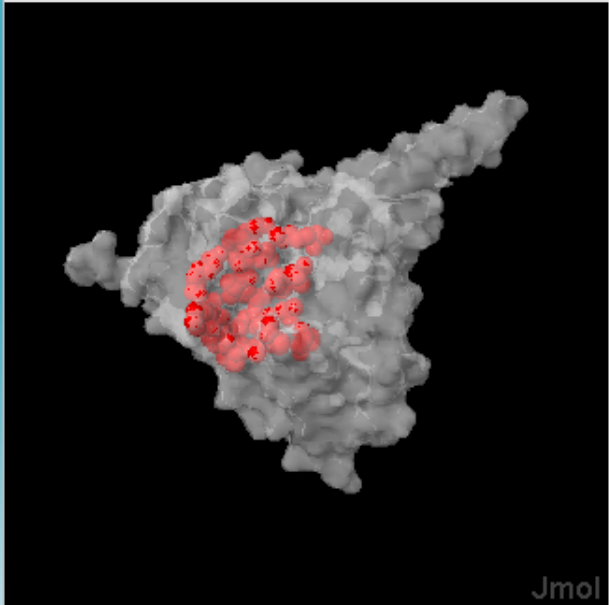
PLB-SAVE

Protein-Ligand Binding region prediction based on features of Solid Angle, Volume, and dEpth

[Main](#)
[Result](#)
[Method](#)
[Contact](#)

Prediction Result (Use [Pilotfox](#) to get the best performance)

PDB ID: 1nov | chain ID: A | Residue #: 309



1 2 3 4 5 6 7 8 9 10

Depth: 1.943069
 Volume: 61
 Residues: 192:A, 194:A, 196:A, 227:A, 229:A, 242:A, 243:A, 244:A, 245:A, 246:A, 247:A, 295:A, 348:A, 351:A, 352:A, 354:A, 355:A, 356:A
 Sequence: KPKPNQSVCNEDRARIP

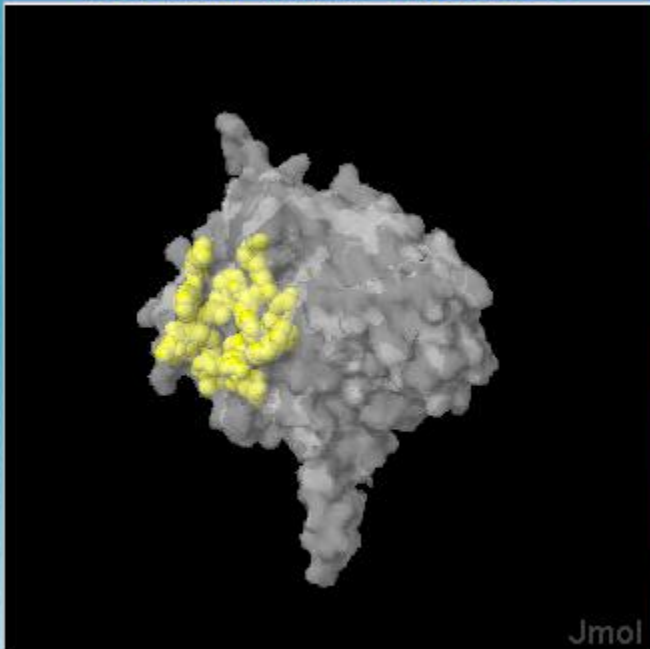
Chain	0	10	20	30	40	50	60	70	80	90	
A0							NMLKM	SAPGLDFLKC	AFASPDFSTD	PGKGIQDKFQ	GLVLPKKHCL
A100	IQSITFTFGK	QIMLLVAPIF	GIACLKAEAN	VGASFSQVBL	ASVEFPGFDQ	LFQTSATDTA	ANVTIAPRYAS	MAAGVYPTSN	LMQFAGSIQV	YRPFKQVILN	
A200	SYSQIVATVP	PTNLAQNTIA	IDGLEALNA	FNNNYSQSFY	EGYSQSVYCN	EPEPEFHPIM	EGYASVPPAN	VTNAQASMPY	NLTFSGARYT	GLGDMIAIAI	
A300	LVITPTGAVN	TAVLKWACV	EYRPNFNSIL	YEPARESPAN	DEYALAAAYK	I	ND	DA	AVA	CKDN	

Rank# #all #1 #2 #3 #4 #5 #6 #7 #8 #9 #10



Prediction Result (Use [Info](#) to get the best performance)

PDB ID: 1nov | chain ID: A | Residue #: 309



Jmol

Rank# #all #1 #2 #3 #4 #5 #6 #7 #8 #9 #10

1 2 3 4 5 6 7 8 9 10

Depth: 1.631313

Volume: 32

Residues: 81:A, 93:A, 94:A, 95:A, 96:A, 97:A, 146:A, 147:A, 149:A, 150:A, 151:A, 161:A, 162:A, 327:A, 328:A, 329:A, 332:A, 334:A,

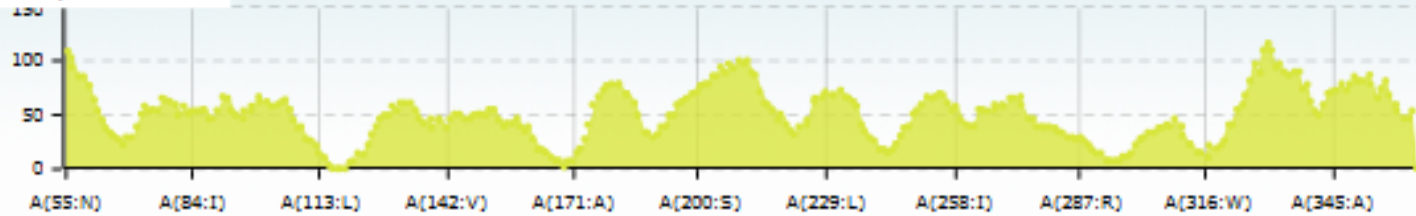
Sequence GLPKKHFGDQLANSTLFR

Chain	0	10	20	30	40	50	60	70	80	90	
A0							NMLKM	SAPGLDFLKC	AFASPDFSTD	FGKGIKDFQ	GLVLPKKHCL
A100	TQSIITFPGK	QTMLLVAPIP	GIACLKAEAN	VGASFSGVPL	ASVEFFGFDQ	LFQTSATDIA	ANVTAFRIAS	MAAGVYPTSN	LMQFAGSIQV	YKIPLKQVLM	
A200	SYSQIVATVP	PINLAQNTIA	IDGLEALDAL	PNNNYSQSP	EGCYSQSVCM	EPEFEPHPIM	EGVASVPPAN	VTNAQASMT	NLTFSGARYT	GLGDMDAIAI	
A300	LVTITPGAVN	TAVLKVWACV	EYRPNPSTL	YEFARESPAN	DEVALAAYRK	IARDIPIAVA	CKDN				



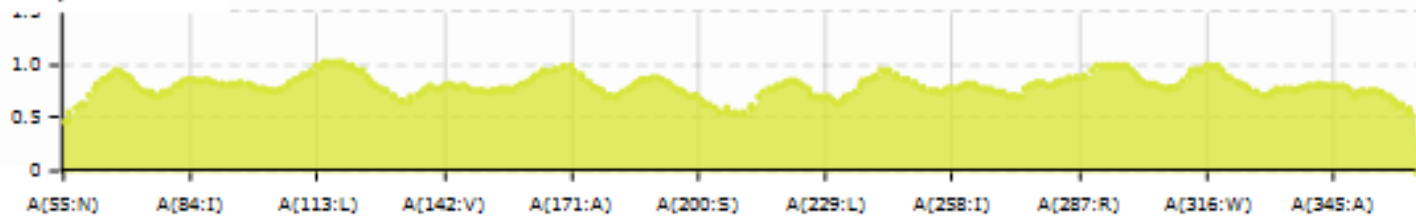
accessible surface area (ASA)

chart by amcharts.com



SolidAngle

chart by amcharts.com



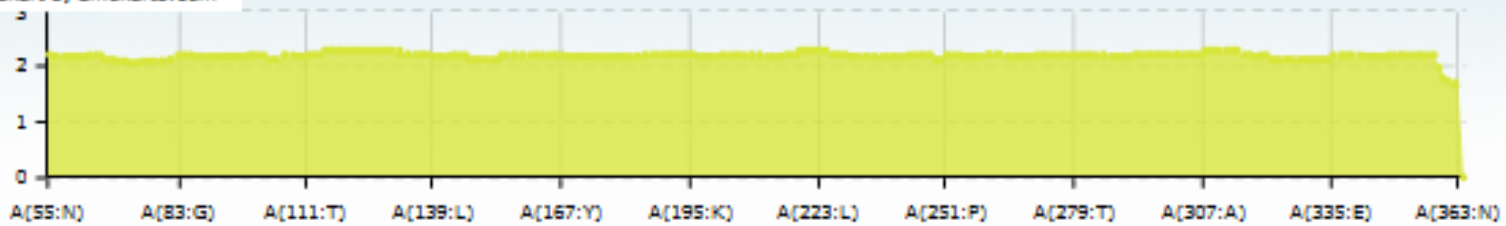
Hydrophobicity

chart by amcharts.com



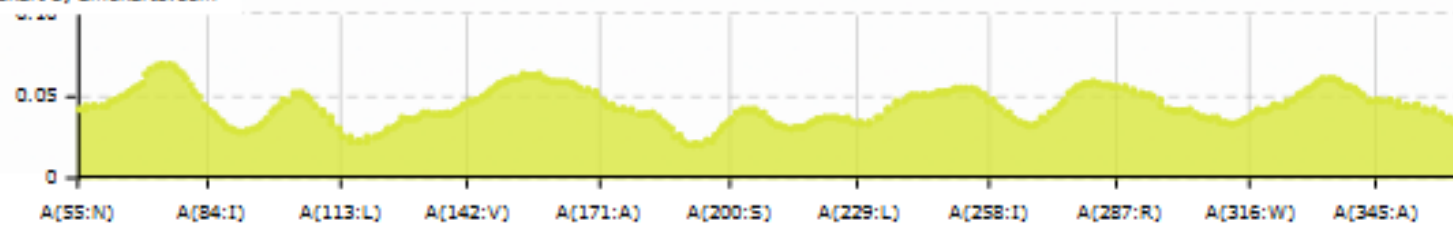
Ionization

chart by amcharts.com



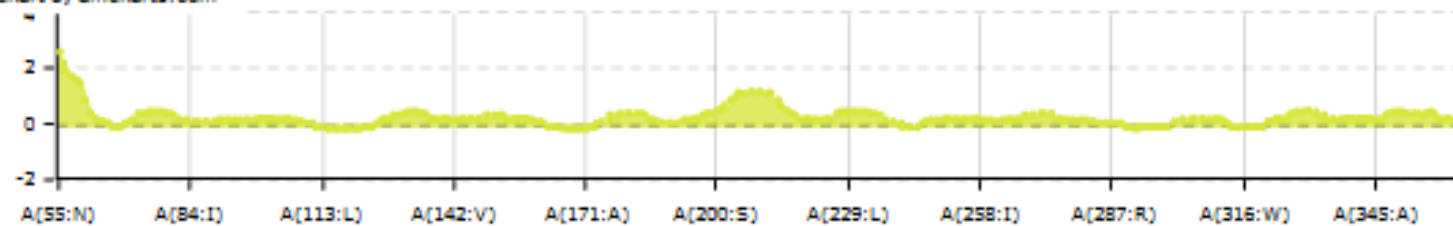
electron-ion interaction potential (EIIP)

chart by amcharts.com



CX value

chart by amcharts.com



Conclusions



Conclusions

- To design an effective and efficient system to predict protein-ligand binding regions/sites
 - PLB - SAVE (<http://save.cs.ntou.edu.tw/>)
- Effective aspect: using simple/straight forward geometric features
 - ranked solid angles / volume / depth features
- Efficient aspect: employing CUDA acceleration
 - an average of 11-fold faster by employing GPU acceleration
- Performance should be better than previously existing systems
 - Robust performance compared to SiteHound and MPK2 systems
 - the proposed parallel algorithms achieved an average accuracy rate of 94.9%
- Carbohydrate-based vaccine design could be applied for pathogen infection and cancer diseases



Acknowledgements



Mr. Ying-Tsang Lo (Ph.D. candidate)
Mr. Hsin-Wei Wang (Ph.D. candidate)
Dr. Wen-Shoung Tzou (NTOU, Taiwan)
Dr. Hui-Huang Hsu (Tamkang Univ., Taiwan)
Dr. Hao-Teng Chang (CMU, Taiwan)

Project Funding supported by:
Center of Excellence for Marine
Bioenvironment and Biotechnology , NTOU
National Science Council, Taiwan
(NSC 101-2321-B-019-001 and
NSC 100-2627-B-019-006)



thanks for your attention

