FANTASTIC ALLOSTERIC BINDING SITES AND WHY DEEP LEARNING CANNOT FIND THEM

Dhvani S. Vora*
Department of Biochemistry and Molecular Biology
Mayo Clinic
Rochester, MN 55905, USA
{vora.dhvani}@mayo.edu

Shashank Yadav*
Department of Biomedical Engineering
University of Arizona
Tucson, AZ 85721, USA
{shashank}@arizona.edu

ABSTRACT

The discovery of druggable and structurally distinct allosteric sites across various protein classes has introduced new avenues for small molecules to modulate protein activity and, hence, cellular functions. Ligands that target allosteric sites may provide advantages like enhanced selectivity and often exhibit the possibility of targeting existing drug-resistant mutations. However, recent deep learning approaches show limited effectiveness in predicting allosteric sites, as demonstrated in the present study. We compare the performance of two deep learning methods, PUResNetV2.0 and VNEGNN, with Fpocket, a traditional geometry-based method and P2Rank, a geometry and machine learning ensemble approach.

1 INTRODUCTION

Proteins contain different types of functional sites that play critical roles in their biological activities- two key types being orthosteric and allosteric sites. Orthosteric sites are the primary binding sites where substrates or inhibitors directly interact with the protein, often leading to a functional or catalytic response. Allosteric sites, in contrast, are located distally from the orthosteric site and regulate protein function by binding to effectors that induce conformational changes (Christopoulos, 2002; Lu et al., 2019). Identifying these sites is essential for drug discovery. Multiple in silico approaches to predict ligand-binding sites (LBS) on three-dimensional protein structures have emerged in the past years, employing various approaches from geometry-based techniques like usage of grids, spheres, or tessellations; energy-based methods, conservation-based methods, template-based as well as ensemble methods (Xia et al., 2024; Zhao et al., 2020). While existing computational methods predict multiple LBS, predicting allosteric sites is still faced with significant challenges. In this study, we compare the performance of four popular methods for LBS prediction- Fpocket (Le Guilloux et al., 2009), P2Rank (Krivák & Hoksza, 2018), VNEGNN (Sestak et al., 2024) and PUResNetV2.0 (Jeevan et al., 2024). Fpocket is a geometry-based method, P2Rank a geometry and machine learning-based method, while VNEGNN and PUResNetV2.0 are deep learning (DL) methods. While all methods perform comparably on orthosteric sites of our dataset, the DL methods perform worse on allosteric sites. Our study provides compelling evidence to question if DL approaches truly outperform traditional geometry-based methods, especially on allosteric LBS.

2 METHODS

All human proteins from the Allosteric Database (ASD) with binding sites of known allosteric function were compiled (He et al., 2024). An additional 22 proteins with experimentally reported ortho- and allosteric sites were added to the dataset after a manual search of recent PDB entries, which were not included in the ASD. The protein structures with complexed orthosteric and allosteric ligands were downloaded from RCSB PDB (Berman et al., 2000), stripped of water molecules and ions. The ligands were then removed to generate a "clean" PDB structure used for pocket prediction. The Fpocket, P2Rank and VNEGNN methods were installed and run according to instructions on the GitHub repositories. PUResNetV2.0 predictions were obtained from the webserver. The predicted sites were manually verified with the ground truth. The results for each protein with all

*Equal contribution

binding sites and predictions are saved as PyMOL session files (DeLano et al., 2002) accessible at https://doi.org/10.5281/zenodo.14977798.

3 RESULTS

A manually curated dataset of 60 proteins with 152 LBS was compiled to evaluate the different prediction methods. Out of 152, 58 sites were orthosteric, while 94 were allosteric in nature. The predictions were counted as true positives if the predicted pocket completely or partially overlaps the ligand in the PDB structure. An example of the ligand-bound protein structure overlayed with prediction visualizations is shown in Figure 1, and additional examples are provided in Appendix A1. A comparison of the models' predictions is summarized in Table 1 based on recall, considering it is impossible to obtain true negatives and false negatives without experimental support. A detailed site-wise comparison is included in Appendix A2. The geometry-based Fpocket ranks highest in predicting ligandable cavities on the protein across all types of pockets, as well as for predicting allosteric sites. However, P2Rank proves to be the superior method in predicting orthosteric sites. The more recent deep learning based PUResNetV2.0 and VNEGNN are not state-of-the-art for LBS prediction.

Table 1: Model Recall Comparison

	Ligand Binding Site Type		
Model	**All**	**Orthosteric**	**Allosteric**
PUResNet	0.460	0.845	0.223
VNEGNN	0.328	0.707	0.096
Fpocket	**0.954**	0.983	**0.936**
P2Rank	0.789	**1.000**	0.659

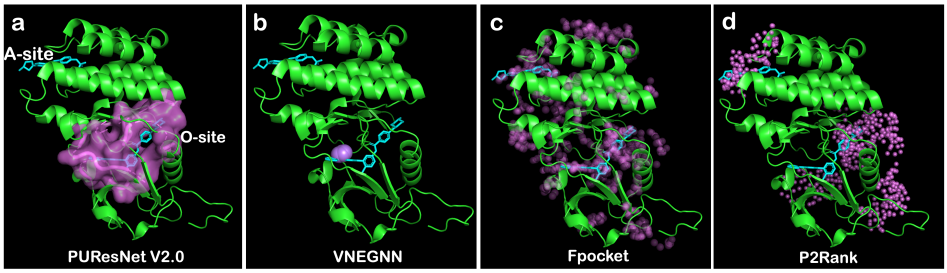

Figure 1: The tyrosine protein kinase ABL1, represented in green, complexed with asciminib (at the allosteric or A-site) and nilotinib (orthosteric or O-site), with the reference PDB structure 5MO4, overlayed with predictions of (a) PUResNetV2.0, (b) VNEGNN, (c) Fpocket and (d) P2Rank predicted pockets. The ligands are shown in blue, and the predicted pockets are shown in violet.

4 DISCUSSION

Despite the advanced capabilities of DL, which is quoted for its ability to model complex patterns, the results from this study suggest that DL techniques struggle to accurately predict allosteric LBS. PUResNetV2.0 may often predict allosteric sites that are present in close proximity to the orthosteric site, while distant sites are missed. The poor performance of VNEGNN and PUResNetV2.0 may be attributed to the absence of extensive, high-quality training data. In contrast, methods that rely on surface topology and shape analysis show more robust performance possibly due to their reliance on inherent structural features. DL methods typically require large amounts of annotated examples to generalize effectively, and the relatively small number of experimentally validated allosteric sites available for training further exacerbates this issue. Allosteric regulation often involves conformational changes that are difficult for models to capture, as these changes may not be directly observable from static protein structures alone. Without sufficient diversity in training examples, DL models may fail to learn the subtle patterns necessary for accurate allosteric LBS prediction. The high structural and functional group complexity inherent in the allosteric mechanisms may also be inadequately represented, which leads to DL methods being unable to fully capture such intricate biological processes. Geometry-based methods excel by detecting physical pocket characteristics and conformational flexibility without being overly dependent on training data. DL methods may improve on allosteric prediction by integrating molecular dynamics features, expanding allosteric datasets and explicitly modeling structural flexibility. The findings from this study highlight the need for investigation into the limitations of deep learning approaches and emphasize that, in certain contexts, traditional methods may still offer competitive performance in allosteric site prediction.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of the Tiny Papers Track at the "I Can't Believe It's Not Better" Workshop @ ICLR 2025.

REFERENCES

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Arthur Christopoulos. Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nature reviews Drug discovery*, 1(3):198–210, 2002.

Warren L DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, 40(1):82–92, 2002.

Jixiao He, Xinyi Liu, Chunhao Zhu, Jinyin Zha, Qian Li, Mingzhu Zhao, Jiacheng Wei, Mingyu Li, Chengwei Wu, Junyuan Wang, et al. Asd2023: towards the integrating landscapes of allosteric knowledgebase. *Nucleic Acids Research*, 52(D1):D376–D383, 2024.

Kandel Jeevan, Shrestha Palistha, Hilal Tayara, and Kil T Chong. Puresnetv2. 0: a deep learning model leveraging sparse representation for improved ligand binding site prediction. *Journal of Cheminformatics*, 16(1):1–16, 2024.

Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10:1–12, 2018.

Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10:1–11, 2009.

Shaoyong Lu, Qiancheng Shen, and Jian Zhang. Allosteric methods and their applications: facilitating the discovery of allosteric drugs and the investigation of allosteric mechanisms. *Accounts of Chemical Research*, 52(2):492–500, 2019.

Florian Sestak, Lisa Schneckenreiter, Johannes Brandstetter, Sepp Hochreiter, Andreas Mayr, and Günter Klambauer. Vn-egnn: E (3)-equivariant graph neural networks with virtual nodes enhance protein binding site identification. *arXiv preprint arXiv:2404.07194*, 2024.

Ying Xia, Xiaoyong Pan, and Hong-Bin Shen. A comprehensive survey on protein-ligand binding site prediction. *Current Opinion in Structural Biology*, 86:102793, 2024.

Jingtian Zhao, Yang Cao, and Le Zhang. Exploring the computational methods for protein-ligand binding site prediction. *Computational and structural biotechnology journal*, 18:417–426, 2020.

A Appendix

A.1 Extended Figures

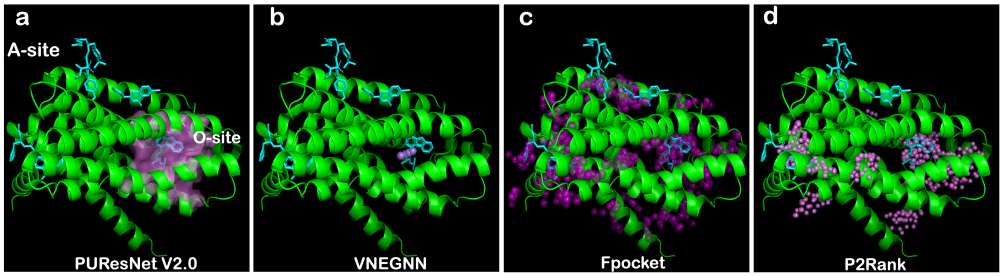

Figure 2: The beta2 adrenoceptor, represented in green, complexed with allosteric (A-site) and orthosteric (O-site) ligands shown in blue, with the reference PDB structure 5X7D, overlayed with predictions of (a) PUResNetV2.0, (b) VNEGNN, (c) Fpocket and (d) P2Rank predicted pockets. The predicted pockets are shown in violet.

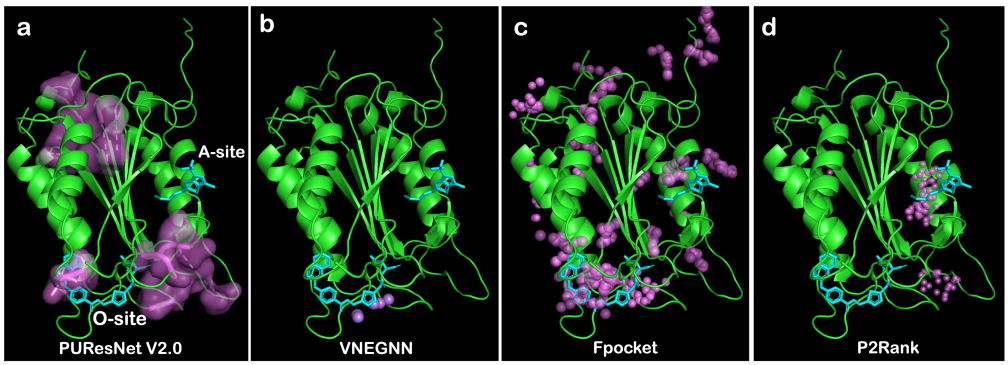

Figure 3: The caspase-1, represented in green, complexed with allosteric (A-site) and orthosteric (O-site) ligands shown in blue, with the reference PDB structure 2FQQ, overlayed with predictions of (a) PUResNetV2.0, (b) VNEGNN, (c) Fpocket and (d) P2Rank predicted pockets. The predicted pockets are shown in violet.

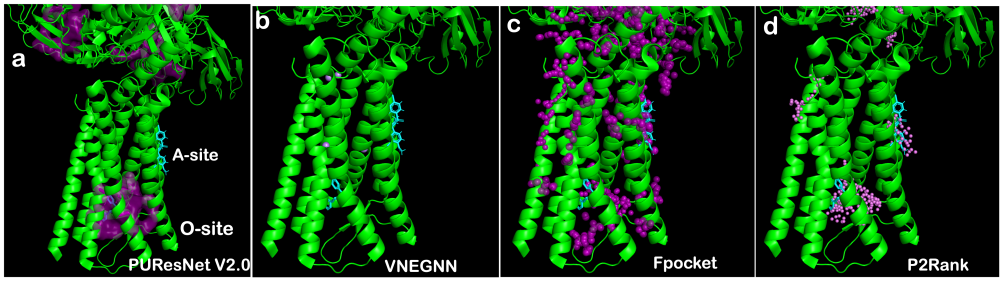

Figure 4: The mu-type opioid receptor-G protein complex, represented in green, complexed with allosteric (A-site) and orthosteric (O-site) ligands shown in blue, with the reference PDB structure 8K9L, overlayed with predictions of (a) PUResNetV2.0, (b) VNEGNN, (c) Fpocket and (d) P2Rank predicted pockets. The predicted pockets are shown in violet.

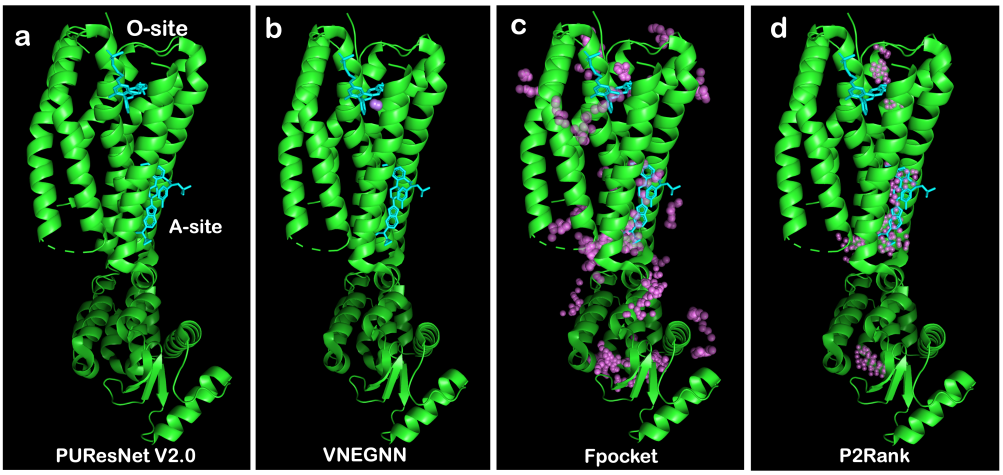

Figure 5: The G-protein-coupled receptor 40 (GPR40), represented in green, complexed with allosteric (A-site) and orthosteric (O-site) ligands shown in blue, with the reference PDB structure 5KW2, overlayed with predictions of (a) PUResNetV2.0 (no pocket predicted), (b) VNEGNN, (c) Fpocket and (d) P2Rank predicted pockets. The predicted pockets are shown in violet.

A.2 Detailed comparison of predictions on each Ligand binding site

Table 2: Summary of proteins, ligand types, and binding information. Abbreviations: PUR—PUResNetV2.0, VNN—VNEGNN, FPK—FPOCKET, P2R—P2RANK, O—Orthosteric, A—Allosteric, 1-Correct, 0-Incorrect.

#	Gene ID	PDB Base	Multi-mer	LBS Index	LBS Type	Ligand Bound	PUR	VNN	FPK	P2R
1	CDK2	8VQ4	No	1	O	JWS	1	1	1	1
	-CCNE1			2	A	A1AC5	0	0	1	1
2	GBA1	8P3E	Yes	1	O	WYC	0	1	1	1
				2	O	WYC	0	0	0	1
				3	O	PN8	1	1	1	1
				4	O	PN8	1	1	1	1
				5	A	WSI	0	0	0	0
3	PTPN2	9C56	No	1	O	527	1	1	1	1
				2	A	FRJ	0	0	1	0
4	MALT1	8V4X	Yes	1	A	A1A	1	0	1	1
				2	A	A1A	1	0	1	1
				3	A	A1A	1	0	1	1
				4	A	A1A	1	0	1	1
5	MOR	8K9L	No	1	O	7V7	1	1	1	1
				2	A	VV9	0	0	1	1
6	MAT2A	8XAM	Yes	1	O	SAM	1	1	1	1
				2	O	SAM	0	1	1	1
				3	A	XRH	0	0	1	1
				4	A	VUO	0	0	1	0
7	KCNC1	8QUC	Yes	1	O	PCF	0	0	1	1
				2	A	WY9	0	0	1	1
8	ABL1	5MO4	No	1	O	STI	1	1	1	1
				2	A	AY7	0	0	1	1
9	CNR1	8IKH	No	1	O	FMN	1	1	1	1
				2	O	8D0	1	0	1	1

#	Gene ID	PDB Base	Multi-mer	LBS Index	LBS Type	Ligand Bound	PUR	VNN	FPK	P2R
				3	A	Q2L	0	0	1	0
10	ADRB2	5X7D	No	1	O	CAU	1	1	1	1
				2	A	8VS	0	0	1	1
				3	A	M3J	0	0	1	0
				4	A	KBY	0	0	1	0
11	AKT1	4GV1	No	1	O	UCB	0	1	1	1
				2	A	0XZ	1	0	1	1
12	AMD1	3EPA	No	1	O	MAO	0	1	1	1
				2	A	PUT	0	0	1	1
13	AD	2PIQ	No	1	O	TES	1	1	1	1
				2	A	17W	0	0	1	0
				3	A	RB1	0	0	1	0
14	ALB	2BX8	No	1	A	MYR	0	0	1	1
				2	A	MYR	0	0	1	1
				3	A	MYR	0	0	1	1
				4	A	MYR	0	0	1	1
				5	A	RWF	1	1	1	1
				6	A	AZQ	0	1	1	1
				7	A	MYR	0	0	1	1
				8	A	MYR	0	0	1	0
15	AURKA	5DN3	No	1	O	ATP	1	1	1	1
				2	A	5DN	0	0	1	0
16	BRAF	1UWH	No	1	O	ACP	1	1	1	1
				2	A	0LI	1	1	1	1
17	CASP1	2FQQ	No	1	O	Q2Y	1	1	1	1
				2	A	F1G	0	0	1	1
18	CASP7	1SHJ	Yes	1	A	NXN	0	1	1	1
				2	A	8YM	0	0	1	1
				3	A	NXN	0	1	1	1
				4	A	SE1	0	0	1	1
19	CPS1	5DOU	Yes	1	O	ADP	1	1	1	1
				2	O	ADP	1	0	1	1
				3	A	NLG	0	0	1	1
				4	A	374	0	0	1	1
20	CRF1R	4K5Y	Yes	1	A	1Q5	1	0	1	1
				2	A	1Q5	1	0	1	1
21	CTSK	5J94	No	1	O	INA	1	1	1	1
				2	A	6HM	0	0	0	0
22	DHPS	6PGR	Yes	1	A	8XY	1	0	1	1
				2	A	8XY	1	0	1	1
23	EP300	6PF1	No	1	O	COA	1	1	1	1
				2	A	OJ7	1	0	1	1
24	FDPS	5DGM	No	1	O	ZOL	1	1	1	1
				2	A	7AM	1	0	1	0
				3	A	7AM	0	0	1	0
25	FFAR1	5KW2	No	1	O	MK6	0	1	1	1
				2	A	6XQ	0	0	1	1
26	FLT3	4RT7	No	1	O	P30	1	1	1	1
				2	A	CXS	0	0	0	0
27	GLS	5JYO	Yes	1	O	GLU	1	0	1	1
				2	A	ZBS	0	1	1	1
				3	A	63J	0	1	1	1
28	GMDS	5IN4	Yes	1	O	GDP	1	0	1	1
				2	O	NAP	1	1	1	1

#	Gene ID	PDB Base	Multi-mer	LBS Index	LBS Type	Ligand Bound	PUR	VNN	FPK	P2R
				3	A	6CK	0	0	1	1
29	GRM5	5CGC	No	1	O	MES	0	0	1	1
				2	A	51D	1	1	1	1
30	H3C1	6IQ4	No	1	A	XIS	0	0	1	0
				2	A	D0X	0	0	1	0
31	HBB	4ROL	Yes	1	O	HEM	1	0	1	1
				2	A	KOH	0	0	1	1
32	HRAS	5ZC6	No	1	O	GNP	1	1	1	1
				2	A	KBF	0	0	1	0
33	IDH1	4UMX	Yes	1	O	NAP	1	1	1	1
				2	A	59D	0	0	1	1
34	IDH3	6L57	Yes	1	O	NAI	1	1	1	1
				2	A	CIT	1	0	1	1
35	KDM4A	5D6W	No	1	A	92Y	1	0	1	1
36	KIF11	6HKX	Yes	1	O	ADP	1	1	1	1
				2	A	4A2	0	0	1	1
				3	A	GCE	1	0	1	1
37	KIF18A	5OCU	No	1	O	ATP	1	0	1	1
				2	A	9V5	0	0	1	0
38	TUBB	5OCU	No	1	O	GDP	1	0	1	1
				2	A	TA1	0	0	1	0
39	ALOX5	6NCF	No	1	O	ACD	1	1	1	1
				2	A	AF7	0	0	1	1
40	MYH7	4PA0	Yes	1	O	CRO	0	0	1	1
				2	A	2OW	0	0	1	1
41	NRAS	8VM2	Yes	1	O	GTP	1	1	1	1
				2	A	EZZ	0	0	1	0
42	P2RX3	9IK1	Yes	1	O	ATP	1	0	1	1
				2	A	128	0	0	1	1
				3	A	A1L	1	0	1	1
				4	A	PG4	0	0	1	1
43	PAK1	4ZJI	Yes	1	O	4OQ	1	0	1	1
				2	A	59U	0	0	0	1
44	PANK3	3MK6	Yes	1	O	AN2	1	0	1	1
				2	A	PKZ	0	0	1	1
45	PARP1	6BHV	Yes	1	O	DQV	1	0	1	1
				2	A	09L	0	0	1	0
46	PCSK9	6U3X	No	1	A	63	1	0	1	1
47	MAPK14	8X3M	No	1	O	LBE	1	1	1	1
				2	A	B8Z	0	0	1	1
48	PGYB	5IKP	No	1	O	PLP	1	1	1	1
				2	A	AMP	0	0	1	1
49	PIK3CA	9C15	No	1	O	GNP	1	0	1	1
				2	A	70S	0	0	1	0
				3	A	A1A	0	0	1	1
				4	A	71K	0	0	1	1
				5	A	FB1	0	0	1	0
50	PGC1A	3GN8	No	1	O	MOF	1	1	1	1
				2	A	EPE	0	0	1	0
				3	A	TLA	0	0	1	1
51	PTPN1	5T19	No	1	O	IX1	1	1	1	1
				2	A	DOP	0	0	0	0
52	PTPN5	6H8R	No	1	O	CIT	1	1	1	1

#	Gene ID	PDB Base	Multi-mer	LBS Index	LBS Type	Ligand Bound	PUR	VNN	FPK	P2R
				2	A	FWB	0	0	0	1
53	PTPN11	5XZR	No	1	O	JZG	1	1	1	1
				2	A	50D	1	0	1	0
				3	A	DZV	0	0	1	0
54	DUSP3	8TK3	No	1	O	PO4	1	1	1	1
				2	A	I2X	0	0	1	0
				3	A	I2X	0	0	1	0
55	SIRT1	5BTR	Yes	1	O	4I5	1	1	1	1
				2	A	STL	0	0	1	1
				3	A	8QF	0	0	1	0
56	SIRT6	5MFP	Yes	1	O	AR6	1	1	1	1
				2	A	8L9	0	0	1	1
57	SMO	5L7I	Yes	1	A	VIS	1	1	1	1
58	THRB	2PIN	Yes	1	O	4HY	1	1	1	1
				2	A	LEG	0	0	1	1
59	TTR	5EZP	Yes	1	O	IPJ	1	0	1	1
				2	A	AJU	0	0	1	1
				3	A	04B	0	0	1	0
60	SMYD3	6YUH	No	1	O	SAM	1	1	1	1
				2	A	POW	0	0	1	0