

## Accepted Manuscript

Title: Molecular modeling and identification of novel  
Glucokinase activators through stepwise virtual screening

Author: Pabitra Mohan Behera Deepak Kumar Behera Suresh  
Satpati Geetanjali Agnihotri Sanghamitra Nayak Payodhar  
Padhi Anshuman Dixit



PII: S1093-3263(15)00031-5  
DOI: <http://dx.doi.org/doi:10.1016/j.jmgm.2015.01.012>  
Reference: JMG 6513

To appear in: *Journal of Molecular Graphics and Modelling*

Received date: 15-9-2014  
Revised date: 28-1-2015  
Accepted date: 29-1-2015

Please cite this article as: P.M. Behera, D.K. Behera, S. Satpati, G. Agnihotri, S.N. Payodhar Padhi, A. Dixit, Molecular modeling and identification of novel Glucokinase activators through stepwise virtual screening, *Journal of Molecular Graphics and Modelling* (2015), <http://dx.doi.org/10.1016/j.jmgm.2015.01.012>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

**Research Highlights**

1. Homology models of three natural variants of human Glucokinase were made.
2. A 10ns MD simulation was done on models to assess flexibility of the active site.
3. Virtual screening was done to identify potential GK activators for the 3 variants.
4. A frequency analysis was done to identify single and multiple variant binders.
5. Analysis for presence of different scaffolds was done in identified molecules.

## Molecular modeling and identification of novel Glucokinase activators through stepwise virtual screening

Pabitra Mohan Behera <sup>1</sup>, Deepak Kumar Behera <sup>1</sup>, Suresh Satpati <sup>2</sup>, Geetanjali Agnihotri <sup>2</sup>,  
Sanghamitra Nayak <sup>1</sup> Payodhar Padhi <sup>3</sup> and Anshuman Dixit <sup>2</sup>

<sup>1</sup> Centre of Biotechnology, Siksha O Anusandhan University, Bhubaneswar, Odisha, India-751030

<sup>2</sup> Institute of Life Sciences, Nalco Square, Bhubaneswar, Odisha, India-751023

<sup>3</sup> Research and Development Centre, Hi-Tech Medical College and Hospital, Bhubaneswar, Odisha, India-751025

*Corresponding Author: Dr. Anshuman Dixit, Department of Translational Research and Technology Development, Institute of Life Sciences, Nalco Square, Bhubaneswar, Odisha, India-751023*

*E-mail: anshumandixit@gmail.com*

*Phone No.: 09777257570*

### Abstract

The glucose phosphorylating enzyme glucokinase (GK) is a 50kD monomeric protein having 465 amino acids. It maintains glucose homeostasis inside cells, acts as a glucose sensor in pancreatic  $\beta$ -cells and as a rate controlling enzyme for hepatic glucose clearance and glycogen synthesis. It has two binding sites, one for binding D-glucose and the other for a putative allosteric activator named glucokinase activator (GKA). The GKAs interact with the same region of the GK enzyme that is commonly affected by naturally occurring mutations in humans. However, many GKAs don't bind to GK in the absence of glucose. Recently, it has been reported that GKAs are highly effective in patients with type 2 diabetes mellitus.

In this milieu a molecular modeling study has been carried out on three natural variants of GK that lie in the GKA binding site and are known to cause maturity onset diabetes of young (MODY). Additionally, a 10ns molecular dynamics simulation was done on each of the modeled variant in order to explore the flexibility of this site. Subsequently, a systematic virtual screening study was done to identify compounds which can bind with high affinity at GKA binding site of mutant GK.

**Keywords:** Diabetes, Glucokinase, Glucokinase activator, Molecular dynamics simulation, Virtual screening, Molecular docking

## 1. Introduction

Diabetes is a metabolic disorder characterized by malfunction of glucose metabolism. It leads to other complications like cardiovascular, peripheral vascular, ocular, neurologic and renal abnormalities etc. The rising problem of diabetes has led to integrated research activities globally for development of preventive and therapeutic strategies. The World Health Organization (WHO) has estimated that ~1.6 and 2.5 million people may die from diabetes in 2015 and 2030 respectively. It will be the 5<sup>th</sup> leading cause of death worldwide by 2030. [1] Majority of deaths from diabetes occur in low and middle income countries leading to their socio-economic loss. [2] Two most common forms of diabetes, type 1 or insulin dependent and type 2 or non-insulin dependent, occurs due to synergistic association of both genetic and environmental risk factors. More than 20 regions of the human genome are susceptible to type 1 diabetes, majority of which belong to genes of HLA class II present in chromosome 6 and contribute approximately 40-50% of the cases. [3] Similarly more than 50 candidate genes have been reported to be involved in type 2 diabetes with variations among populations worldwide. These genes have significant involvement in pancreatic  $\beta$  cell function, glycolysis and other metabolic pathways which increase the risk of type 2 diabetes. The Maturity-Onset Diabetes of the Young (MODY) is an uncommon variety of type 2 diabetes, characterized by a slow onset of diabetic symptoms with no evidence of obesity, ketosis and  $\beta$  cell autoimmunity. Advance molecular genetics studies reveals that there are six forms of MODY (1-6) and each of which is caused by mutations in different genes which are directly involved with  $\beta$  cell function. [4] MODY2 is caused by ~200 mutations in the GCK gene that is located on chromosome 7. GCK codes for Glucokinase (GK), an important enzyme occurring in cells of liver, pancreas, gut and brain. GK plays an important role in carbohydrate metabolism by facilitating phosphorylation of glucose to glucose-6-phosphate. This single autosomal gene GCK has 10 exons [5, 6] and it begins with two promoter regions. [7] The first promoter or the neuroendocrine promoter is active in pancreatic islet cells, neural tissues and enterocytes to produce neuroendocrine isoforms of glucokinase. The second promoter or the liver promoter is active in hepatocytes and produce liver isoforms of glucokinase. [8] In liver, GK acts as the gateway for bulk processing of available glucose but in neuroendocrine cells it acts as a sensor and triggers the cell responses affecting carbohydrate metabolism throughout body.

Most of the GK is found in liver. It provides approximately 95% of the hexokinase activity in hepatocytes. The GK activity can be regulated by the actions of a regulatory protein, Glucokinase Regulatory Protein (GKRP). The GKRP maintains an inactive reserve of GK

inside hepatocytes and recruits active GKs in response to rising levels of glucose in portal vein. The GKRPs move between nucleus and cytoplasm of the hepatocytes and form reversible complexes with GK. It acts as a competitive inhibitor with glucose, such that the enzyme activity is reduced to non-zero while bound. The crystal structures of GK are solved as a tertiary complex with one D-glucose and an activator (GKA) forming a closed conformation. [9-11] The GKAs increase the affinity of recombinant GK for glucose by 10 fold. Therefore, intensive research efforts are being made globally for identification of novel GKAs that can activate glucokinase for the treatment of type 2 diabetes. [12, 13]

In the current study, we present a molecular modeling study of 3 natural variants of GK that lie in the GKA binding site and are known to cause MODY. In order to explore the flexibility of this site, a 10ns molecular dynamics simulation was done on each of the modeled variant. Subsequently, a systematic virtual screening study was done to identify the compounds which can bind with high affinity at GKA binding site of mutant GK.

## **2. Materials and Methods**

### **2.1 Materials**

#### **2.1.1 Important natural variants of Glucokinase**

The Human Glucokinase (UniProtKB ID: P35557) is a 465 amino acids long protein with hexokinase type-1 and hexokinase type-2 domains. It has three isoforms isoform1 (P35557-1), isoform2 (P35557-2) and isoform3 (P35557-3) produced by alternate splicing. There are 4 nucleotide binding regions, 3 substrate binding regions and 5 binding sites in GK. It has 47 natural variants having point mutations at various positions in the main sequence. At the time of this investigation there were 23 reported crystal structures of GK in PDB [14] and majority of them were complexed with small molecule activator.

#### **2.1.2 The Glucokinase activator**

Reported glucokinase activators are small compounds with little variance in their chemical structure. Generally, the parent molecule is either a central carbon atom or an aromatic ring with three attachments to it. Out of the three attachments two are hydrophobic and at least one of the two is aromatic in nature. The third one is either a 2-aminoheterocycle or N-acyl urea moiety which forms an electron donor/acceptor interaction with R63 of GK. [15] The 3D representation of a GKA is shown in Figure-1. However, there are a few compounds which significantly differ in structure from the above discussed one.

**Please insert Figure-1 about here.**

## 2.2 Methods

### 2.2.1 Molecular modeling of natural variants of Glucokinase

An analysis of the GKA bound GK crystal structures revealed that residues e.g. V62, R63, S64, T65, G68, S69, G72, V91, W99, M210, I211, Y214, Y215, M235, V452, V455 and K459 line the GKA binding site of GK. Three variants M210K, M210T, and E221K out of the total 47 were found to be located in the GKA binding site (defined as residues within 5Å of the bound GKA). These variants M210K (variant 1), M210T (variant 2), and E221K (variant 3) were considered for further analysis.

The homology models of these variants were made by software Modeller v9.12 [16, 17] using the X-ray crystal structure of Glucokinase with a small molecule activator 1JD at 2Å resolution (PDB: 4IXC) [18]. Ten models were generated for each variant and the best model for each of them was selected on the basis of lowest DOPE score. The modeled structures were checked for stereochemical quality using Procheck\_NT [19].

### 2.2.2 Molecular dynamics simulation

To investigate the effect of mutation on the flexibility/ rearrangement of the GKA binding site, a 10ns molecular dynamics simulation was carried out by the following protocol on each of the modeled structure.

For each of the variant, the structure was prepared for simulation in VMD [20] package while the simulation was done using NAMD [21] software. The structure was solvated in a box of water with buffering distance of 10 Å and sodium ( $\text{Na}^+$ ) & chloride ( $\text{Cl}^-$ ) counter-ions were added to achieve charge neutrality. All  $\text{Na}^+$  and  $\text{Cl}^-$  ions were placed at least 8 Å away from any protein atoms and from each other.

The system was subjected to initial minimization for 20000 steps (40ps) with protein backbone fixed. The constraints were then removed and the system was allowed to relax freely by minimizing for another 20000 steps (40ps). These steps were done to eliminate any bad steric contacts and regions of higher energy. The minimization was followed by equilibration of the modeled variants by gradually increasing the system temperature in increments of 20K starting from 0K until 300K and at each step 15000 steps (30 ps) equilibration was run keeping a restraint of 10 Kcal mol<sup>-1</sup> Å<sup>-2</sup> on protein alpha carbons ( $\text{C}_\alpha$ ). The system was then equilibrated for 150000 steps (300ps) at 310K (NVT) and then for further 150000 steps (300ps) at 310K using Langevin piston (NPT) algorithm as implemented in NAMD. Finally the restrains were removed and the system was equilibrated for 500000 steps (1ns) to prepare the system for simulation.

An NPT simulation was run on the equilibrated structure for 10 ns keeping the temperature at

310 K and pressure at 1 bar using Langevin piston coupling algorithm. The integration time step of the simulations was set to 2.0 fs, the SHAKE algorithm [22] was used to constrain the lengths of all chemical bonds involving hydrogen atoms at their equilibrium values and the water geometry was restrained rigid by using the SETTLE algorithm. [23] Nonbonded van der Waals interactions were treated by using a switching function at 10Å and reaching zero at a distance of 12Å. The particle-mesh Ewald algorithm (PME) as implied in NAMD was used to handle long range electrostatic forces. The refined structures obtained by molecular dynamics were used for further analysis.

### **Active Site clustering**

MMTSB toolset (<http://blue11.bch.msu.edu/mmts/>) was used to generate representative clusters from 100 equally spaced snapshots from the molecular dynamics run. The clustering was done to explore the conformational flexibility of the binding site using root mean square deviation in the residues lining the active site. Since, it has been reported that including conformational flexibility of the binding site increase the robustness of protein centric virtual screening. [24] The derived clusters were used in the subsequent virtual screening procedure in order to get compounds that can fit multiple conformations of the binding site. Such compounds can be expected to be more promising than those selected using single protein conformation. As a result of clustering, three conformations corresponding to three most populated clusters for each variant were selected for further studies.

### **2.2.3 Selection of suitable compounds from PubChem**

The PubChem [25] database provides valuable information on the biological activity of small compounds shared by depositors across the globe. It includes substance information, compound structures and bioactivity data stored in three primary databases, Pcsubstance, Pccompound and PCBioAssay respectively. The PubChem database has 461937 compounds (release 3) available in sdf format. Each compound has been given a unique compound identification (CID). These compounds were downloaded and used in the current study for the identification of probable GKA ligands.

### **2.2.4 Preparation of compound database**

The compounds were imported in OpenEye FILTER program ([www.eyesopen.com/filter](http://www.eyesopen.com/filter)) to filter out non-drug like compounds from the database. The FILTER program was run with default settings except for the molecular weight limit which was set to 500. The remaining compounds after filtering were taken to OpenEye OMEGA [26] software for conformer generation using default settings except for maximum number of conformers that was set at 400. The resulting conformer database was used for shape based screening using OpenEye

ROCS [27] program with 1JD molecule (PDB: 4IXC) as query. The ROCS program screens the compounds by producing overlays of them over query molecule(s). The compounds having appreciable similarity to the query molecule(s), based on different metrics, are then selected for further processing. In the current study, “TanimotoCombo” as implied in ROCS was used for ranking of the compounds. It measures the shape and electrostatic similarity between query and database compounds. All the compounds having TanimotoCombo>0.9 (Total 11272 compounds) were selected for the next docking step.

### 2.2.5 Molecular docking studies

The selected compounds were docked in the different receptor conformations generated by active site clustering. Prior to docking the compounds were prepared using Ligprep module of Schrodinger [28] with default settings. Three conformations for each of the three variants were used for docking calculations. The protein structures were first preprocessed by assigning correct bond orders and adding hydrogens. The H-bonds were optimized and restrained minimization of hydrogen atoms was carried out. The protein heavy atoms were minimized to a gradient of 0.3Å prior to docking. A receptor grid was generated separately for each of the receptor conformation by specifying residues which were found within 5Å of the co-crystallized compound in GK structure (PDB: 4IXC). The selected compounds were then docked in the binding site conformations using standard precision (SP) docking mode of Glide [29]. For each receptor conformation top 1000 compounds, as ranked by the Glide Score, were retained for next step. In the next step extra precision (XP) mode of Glide program was used to dock the previously selected compounds in respective conformation. Since the residue R63 is known to be pivotal in binding of ligands to the GK receptor [30], two hydrogen bonding (HB) constraints with R63 were applied during the docking process. The compounds which could not make these HB interactions with R63 were filtered out. The XP descriptor information was recorded and Epik state penalties were added to the docking score. The ligands were sampled as flexible.

### 2.2.6 Identification of diverse active scaffolds:

The diverse active scaffolds were identified using Scaffold hunter, which is a tool for the analysis of scaffolds in chemical datasets. [31] The compounds can be clustered based on the common scaffolds and can be depicted graphically in the form of a tree or dendrogram. The visual representation of a dataset in this manner can help medicinal chemists in making decisions to select compounds for synthesis. To generate scaffold tree for the compounds, smiles and different fingerprints using DaylightBitFingerprinter, EStateBitFingerprinter and EstateNumericalFingerprinter, were calculated using the largest fragment. A scaffold tree was



generated using default rule set and a radial layout was prepared for analysis. Prior to analysis the tree compounds were sorted by GScore and the scaffolds were sorted by number of aromatic rings.

The detailed methodology followed in this study is shown in Figure-2.

**Please insert Figure-2 about here.**

### **3. Results and Discussions**

#### **3.1 Models of important natural variants**

The stereochemical analysis of generated models revealed that >90% residues were in the allowed region of the Ramachandran plot. The statistical information of this analysis is presented in Table-1. A structural alignment of the best models (V1, V2 and V3) with the template 4IXC was done in PyMOL. [32] The RMSD values of these alignments were found to be 0.203 Å, 0.182 Å and 0.190 Å respectively, suggesting high structural similarity with the template.

**Please insert Table-1 about here.**

An analysis of available GK crystal structures revealed that there are two domains. The domain 1 ranges from 12 to 217 (206 amino acids long) and somewhat open in structure. The domain 2 ranges from 219 to 458 (240 amino acids long) and shows a compact structure. The compact 3D structure is formed by these two domains in which an  $\alpha$ -helix of 16 residues (441-456) from the larger domain 2 is buried into the crevice formed by the smaller domain1. The arrangement of these two domains forms the allosteric site where GKA binds. (Figure-3)

**Please insert Figure-3 about here.**

#### **3.2 Simulation of natural variant models**

The simulations were checked for root mean square deviation (RMSD) for stability of the simulations. RMSD measures the change in overall conformation of a molecule with respect to simulation time. Variant 2 showed least RMS deviation while Variant 1 and 3 showed higher fluctuation. The variant V3 (E221K) showed largest relative fluctuation as compared to others. The higher RMSD in case of V1 and V3 can be explained because of the substitution of M210 and E221 with lysine respectively. Lysine has quite different properties and size as compared to these amino acids. Therefore, it may cause some local rearrangements in the active site to get accommodated. However, it can be seen that the systems become stable quickly i.e. RMSD moving within 2 Å. (Figure-4) The simulations were later extended to 20ns, details of which are given in supplementary data.

**Please insert Figure-4 about here.**

#### **3.3 Virtual screening**

As stated earlier the 11272 ROCS hits were docked into the different conformations of the variants of the GK using standard precision docking. Top 1000 molecules for each conformation were selected and redocked using Glide extra precision (XP) dock. The XP docking resulted in generation of three hit lists per variant (total 9 hit lists). A total of 183, 467 and 283 different compounds for the variants V1, V2 and V3 respectively were able to make two HB interactions with R63 in at least one or more variant conformation(s). Since the Glide score (GScore) can be a rough estimate of the binding free energy, all the compounds having average GScore < -4.00 for individual variants were selected to identify better fitting compounds. Additionally, a frequency analysis was done by analyzing occurrence of each molecule in different hit lists. It is assumed that the ligands showing appreciable binding in different binding site geometries (i.e. high frequency) have better chances of being true hits. A frequency table was generated in which each ligand was given 1 mark if it appear in one list, thus a ligand can have a maximum and minimum score of 3 and 1 respectively for each variant. Those compounds with a frequency of three for an individual variant were selected. Thus a total of 17, 59 and 27 compounds were retained for variant V1, V2 and V3 respectively. (Table-2) These compounds can be proposed as potential ligands for respective variants.

**Please insert Table-2 about here.**

The top 10 compounds as per the docking score for each variant are listed with GScore and average GScore for 3 different conformations. (Table-3) Detailed results are given in supplementary data.

**Please insert Table-3 about here.**

A comparison of the final hit lists showed that six compounds are common between variant 1 and 2, seven compounds are common between variant 2 and 3, and three compounds are common between variant 1 and 3. It is interesting to note that two compounds (PubChem CID: 2815231, 2815230) are common for all variants. (Figure-5)

**Please insert Figure-5 about here.**

Since there are only two compounds that bind to all conformations of each variant (total frequency nine), we were interested to see if we can select some more compounds that can be proposed as potential hits for all variants by slight relaxation in criteria. Out of the three criterion i.e. GScore, HB with R63 and Frequency, we made a slight relaxation in frequency. Therefore compounds having minimum frequency of two in each of the three variants (total frequency six) were selected. There are eleven compounds which satisfy this criterion. (Table-4) The 2D views of the compounds along with their PubChem CID are shown in

Figure-6.

**Please insert Table-4 about here.**

**Please insert Figure-6 about here.**

The docking of top scoring compounds in different variants is shown in Figure-7. The panel A has a compound (CID: 16673125) docked in variant 1, panel B has a compound (CID: 5312112) docked in variant 2 and panel C has a compound (CID: 5342515) docked in variant 3. It can be seen that all compounds are making at least two hydrogen bonds with R63. The compound in Figure-7(A) makes additional hydrogen bond with K459. In general the phenyl ring of the compounds is ensconced in a hydrophobic cavity lined by W99, I211, Y214, Y215 and L451. The other part goes into a partial polar cavity lined by D158, I159 and K459. A closer inspection of the active site shows relative flexibility of the amino acid side chains (especially R63, I211, Y214 and Y215) forming the binding cavity in different variants. This may be the reason why we got different compounds binding to different structures. It is expected that the molecules, which can bind to most of the binding site conformations in all or most structures i.e. having high structural compatibility with the binding site should show a better spectrum of activity.

**Please insert Figure-7 about here.**

### 3.4 Identification of important common scaffold

The idea of 'privileged substructures' that preferentially bind to a given target class has been a source of motivation for the analysis of scaffolds. They may also indicate about the pharmacodynamic underpinnings of a new molecule's activity. In medicinal chemistry identification of diverse active scaffolds is an important task prior to synthesis. The identification of diverse active scaffolds minimizes the risk of costly late stage failures. A chemist can also generate new scaffolds by incorporating mutation in or hybridization of existing ones. We have analyzed the compounds (total 11) selected in earlier step for identification of diverse scaffolds. It is important to note that our main objective here was to cluster the virtual screening hits as per the scaffolds present in them.

All the 11 compounds with their corresponding GScore were imported into Scaffold Hunter. A scaffold tree was generated as described earlier. This depiction is also annotated with the effects of structural changes on binding profile. The nodes are represented in different background color based on Glide Score of the corresponding molecule. The nodes in blue and in magenta represent molecules in GScore range of -8 to -6 (high affinity) and -6 to -4 (moderate affinity) respectively.

There were 7 different scaffolds identified from the scaffold tree. We have isolated three

important scaffolds as subsets from the main tree with 3 (subset 1), 2 (subset 2) and 2 (subset 3) compounds respectively. (Figure-8) The structures of basic scaffolds of these compounds, and PubChem CID are shown in Table-5.

**Please insert Figure-8 about here.**

The scaffold generation process, where a scaffold is extracted and successively deconstructed to a single ring, is shown in Figure-9 taking subset 1 as an example. Successive deconstruction of two molecules (CID 3205404 and CID 3205407) resulted in a two ring scaffold which further gave a single ring. Similarly, third molecule (CID 661821) of this subset gave another two ring scaffold and finally gave a single ring scaffold.

**Please insert Table-5 about here.**

**Please insert Figure-9 about here.**

#### **4. Conclusion**

The glucokinase activators are emerging as new class of therapeutics for treatment of diabetes. In the present study we carried out systematic virtual screening of PubChem database against three mutant forms of glucokinase receptor to identify compounds which can bind to single or multiple mutant forms of GK. The compounds were further classified by scaffolds and three major subsets were found. The identified compounds can be proposed as novel glucokinase activators.

#### **5. Acknowledgement**

Sincere acknowledgements for the Vice Chancellor of Siksha O Anusandhan University and Chairman of Hi-Tech Group of Institutions for providing essential facilities in successful research and bringing the paper to the level of publication. S. Suresh and G. Agnihotri are thankful to DST for fellowship. The award of Rajiv Gandhi National Fellowship to Pabitra Mohan Behera by University Grants Commission, India is duly acknowledged.

#### **6. References**

6. Global status report on non communicable diseases 2010. Geneva, World Health Organization, 2011.
7. C. D. Mathers, D. Loncar, Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med., 3(11) (2006) e442.
8. J. N. Hirschhorn, Genetic epidemiology of type 1 diabetes. Pediatr Diabetes, 4 (2003) 87-100.
9. W. E. Winter, Newly defined genetic diabetes syndromes: maturity onset diabetes of the young. Rev Endocr Metab Disord, 4 (2003) 43-51.

10. A. Matsutani, R. Janssen, H. Donis-Keller, M. A. Permutt, A polymorphic (CA)<sub>n</sub> repeat element maps the human glucokinase gene (GCK) to chromosome 7p". *Genomics* 12 (2) (1992) 319–325.
11. M. Stoffel, P. Froguel, J. Takeda, H. Zouali, N. Vionnet, S. Nishi, I. T. Weber, R. W. Harrison, S. J. Pilkis, S. Lesage, Human glucokinase gene: isolation, characterization, and identification of two missense mutations linked to early-onset non-insulin-dependent (type 2) diabetes mellitus. *Proc. Natl. Acad. Sci. U.S.A.* 89 (16) (1992) 7698–702.
12. P. B. Iynedjian, P. R. Pilot, T. Nospikel, Differential expression and regulation of the glucokinase gene in liver and islets of Langerhans. *Proc. Natl. Acad. Sci. U.S.A.* 86 (20) (1989) 7838–7842.
13. P. B. Iynedjian, D. Jotterand, T. Nospikel, M. Asfari, P. R. Pilot Transcriptional induction of glucokinase gene by insulin in cultured liver cells and its repression by the glucagon-cAMP system. *J. Biol. Chem.* 264 (36) (1989) 21824–21829.
14. K. Kamata, M. Mitsuya, T. Nishimura, J. Eiki, Y. Nagata, Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure* 12 (2004) 429–438.
15. P. Dunten, A. Swain, U. Kammlott, Crystal structure of human liver glucokinase bound to a small molecule allosteric activator in glucokinase and glycemic disease: from basics to novel therapeutics. *Front Diabetes* 16 (2004) 145–154.
16. A. M. Efanov, D. G. Barrett, M. B. Brenner, A novel glucokinase activator modulates pancreatic islet and hepatocyte function. *Endocrinology* 146 (2005) 3696–3701.
17. F. Matschinsky, Assessing the potential of glucokinase activators in diabetes therapy 8 (5) (2009) 399–419.
18. G. E. Meininger, R. Scott, M. Alba, Y. Shentu, E. Luo, H. Amin, M. J. Davies, K. D. Kaufman, B. J. Goldstein, Effects of MK-0941, a novel glucokinase activator, on glycemic control in insulin-treated patients with type 2 diabetes. *Diabetes Care* 34 (12) (2011) 2560–2566.
19. F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures, *J. of Mol. Biol.*, 112 (1977) 535.
20. F. M. Matschinsky, D. Porte Jr, Glucokinase activators (GKAs) promise a new pharmacotherapy for diabetics, *Medicine Reports*, 2:43 (2010) (doi:10.3410/M2-43)

21. N. Eswar, M. A. Marti-Renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali. Comparative Protein Structure Modeling With MODELLER. *Current Protocols in Bioinformatics*, John Wiley & Sons, Inc., Supplement 15 (2006) 5.6.1-5.6.30.
22. M. A. Marti-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Sali. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29 (2000) 291-325.
23. M. J. Waring, S. N. L. Bennett, S. Boyd, L. Campbell, R. D. M. Davies, S. Gerhardt, D. Hargreaves, N. G. Martin, G. R. Robb, G. Wilkinson, Matched triplicate design sets in the optimisation of glucokinase activators maximising medicinal chemistry information content (to be published).
24. R. A. Laskowski, M. W. MacArthur, J. M. ōfornton, PROCHECK: validation of protein structure coordinates," in *International Tables of Crystallography, Volume F: Crystallography of Biological Macromolecules*, M. G. Grossmann and E. Arnold, Eds., Kluwer Academic, Dordrecht, The Netherlands (2001) pp. 722-725.
25. W. Humphrey, A. Dalke, K. Schulten, VMD - Visual Molecular Dynamics, *J. Molec. Graphics*, 14.1 (1996) 33-38.
26. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, K. Schulten, Scalable molecular dynamics with NAMD, *Journal of Computational Chemistry*, 26 (2005) 1781-1802.
27. J. P. Ryckaert, G. Ciccotti, H.J.C. Berendsen, Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *Journal of Computational Physics*, 1977, 23 (3): 327-341.
28. S. Miyamoto, P. A. Kollman, SETTLE: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models, *Journal of Computational Chemistry*, 1992, 13 (8): 952-962.
29. A. Dixit, G. M. Verkhivker, Integrating Ligand-Based and Protein-Centric Virtual Screening of Kinase Inhibitors Using Ensembles of Multiple Protein Kinase Genes and Conformations, *J. Chem. Inf. Model.*, 52 (2012), 2501-2515.
30. E. Bolton, Y. Wang, P. A. Thiessen, S. H. Bryant, PubChem: Integrated Platform of Small compounds and Biological Activities. Chapter 12 IN *Annual Reports in Computational Chemistry*, Elsevier: Oxford, UK, 4 (2008) 217-240.
31. P.C.D. Hawkins, A.G. Skillman, G.L. Warren, B.A. Ellingson and M.T. Stahl, Conformer Generation with OMEGA: Algorithm and Validation Using High Quality

- Structures from the Protein Databank and Cambridge Structural Database, *J. Chem. Inf. Model.*, 50 (2010) 572.
32. P.C.D. Hawkins, A.G. Skillman, A. Nicholls, Comparison of Shape-Matching and Docking as Virtual Screening Tools *J. Med. Chem.*, 50 (2007) 74.
33. Ligprep: Schrödinger Suite 2012: LigPrep, version 2.5, Schrödinger, LLC, New York, NY, 2012.
34. R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin, D. T. Mainz, Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes, *J. Med. Chem.*, 2006, 49, 6177–6196.
35. F. M. Matschinsky, D. Porte, Jr, Glucokinase activators (GKAs) promise a new pharmacotherapy for diabetics, *F1000 Med Rep.*, 2010, 2, 43.
36. Karsten Klein, Oliver Koch, Nils Kriege, Petra Mutzel, Till Schäfer, Visual Analysis of Biological Activity Data with Scaffold Hunter Molecular Informatics, WILEY-VCH Verlag, 2013, 32, 964-975
37. The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.

**Figure captions:**

**Figure-1:** The 3D representation of GKA.

**Figure-2:** Schematic representation of the virtual screening methodology.

**Figure-3:** The compact arrangement of two domains forming the allosteric site.

**Figure-4:** The RMSD analysis of the simulations.

**Figure-5:** Venn diagram showing number of unique and common compounds among variants.

**Figure-6:** The structures of selected compounds with their PubChem compound ID (CID).

**Figure-7:** Binding of top scoring compounds in (A) Variant 1 (B) Variant 2 (C) Variant 3.

**Figure-8:** The clustering of compounds by Scaffold Hunter.

**Figure-9:** Scaffold generation in Scaffold Hunter.



**Table-1:** Model evaluation statistics of natural variants by Procheck\_NT standalone version.

Sl. No.	Model	DOPE score	Regions of Ramachandran plot			
			A, B, L	a, b, l, p	~a, ~b, ~l, ~p	Disallowed
1	Variant 1	-55817.51953	94.2%	5.8%	0.0%	0.0%
2	Variant 2	-55858.74609	93.2%	6.3%	0.2%	0.2%
3	Variant 3	-55991.16797	94.2%	5.6%	0.2%	0.0%

A=Core alpha, B=Core beta, L=Core left-handed alpha, a=Allowed alpha, b=Allowed beta, l=Allowed left-handed alpha, ~a=Generous alpha, ~b=Generous beta, ~l=Generous left-handed alpha, p=Allowed epsilon, ~p=Generous epsilon.

**Table-2:** The summary of docking results

Sl. No.	Variant	Conformation	No. of compounds	Total retained compounds*
1	V1	1	71	
2		2	79	
3		3	112	17
4	V2	1	320	
5		2	325	
6		3	153	59
7	V3	1	46	
8		2	227	
9		3	189	26

\*The compounds appearing in all 3 hit list for individual variants with average docking score <-4.0.

**Table-3:** Docking result of top 10 compounds

Comp. <sup>@</sup>	GScore_C1 <sup>*</sup>	GScore_C2 <sup>*</sup>	GScore_C3 <sup>*</sup>	Avg. GScore <sup>#</sup>
<b>Variant 1</b>				
16673125	-6.869	-8.161	-5.698	-6.909
750996	-7.696	-6.491	-6.420	-6.869
661821	-7.793	-6.274	-5.399	-6.489
5342515	-6.019	-5.445	-6.968	-6.144
1112997	-3.458	-7.692	-6.751	-5.967
2815230	-5.543	-6.087	-5.870	-5.833
1250330	-6.029	-6.804	-4.446	-5.760
2725422	-4.460	-6.152	-6.498	-5.703
3101296	-4.621	-5.001	-6.474	-5.365
2815231	-5.888	-5.102	-4.966	-5.319
<b>Variant 2</b>				
5312112	-10.151	-8.804	-10.666	-9.874
845433	-9.345	-9.624	-7.376	-8.782
1094212	-8.721	-9.280	-8.254	-8.752
2099678	-9.315	-9.975	-6.756	-8.682
5161259	-8.496	-8.543	-8.818	-8.619
2078915	-9.084	-8.445	-8.324	-8.618
7205478	-8.336	-8.038	-9.460	-8.611
1245295	-8.467	-8.669	-8.622	-8.586
2815231	-8.026	-9.180	-8.537	-8.581
3205407	-7.163	-10.746	-7.026	-8.312
<b>Variant 3</b>				
5342515	-8.840	-7.009	-10.078	-8.642
2583317	-7.571	-8.888	-9.456	-8.638
5345544	-7.425	-9.107	-8.303	-8.278
2789377	-7.619	-8.349	-8.611	-8.193
4315981	-8.062	-7.089	-7.842	-7.664
5080694	-6.507	-7.847	-8.474	-7.609
5345858	-8.248	-6.998	-7.362	-7.536
761743	-8.248	-6.998	-7.362	-7.536

4271676	-6.857	-7.731	-7.867	-7.485
899219	-5.801	-7.718	-8.842	-7.454

---

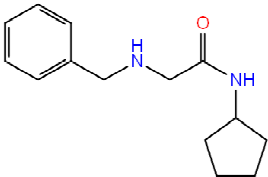
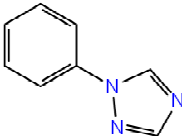
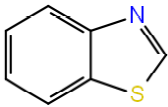
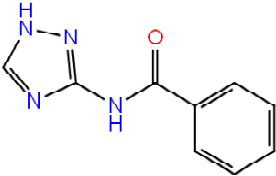

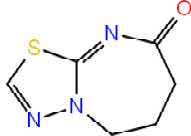
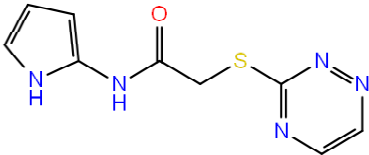
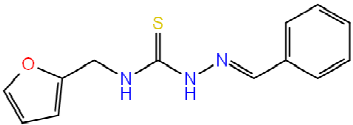
@The PubChem CID of compound. \*GScore\_C1, GScore\_C2, GScore\_C3 is GScore in conformation 1, 2 and 3 respectively. #Avg. GScore=Average GScore.

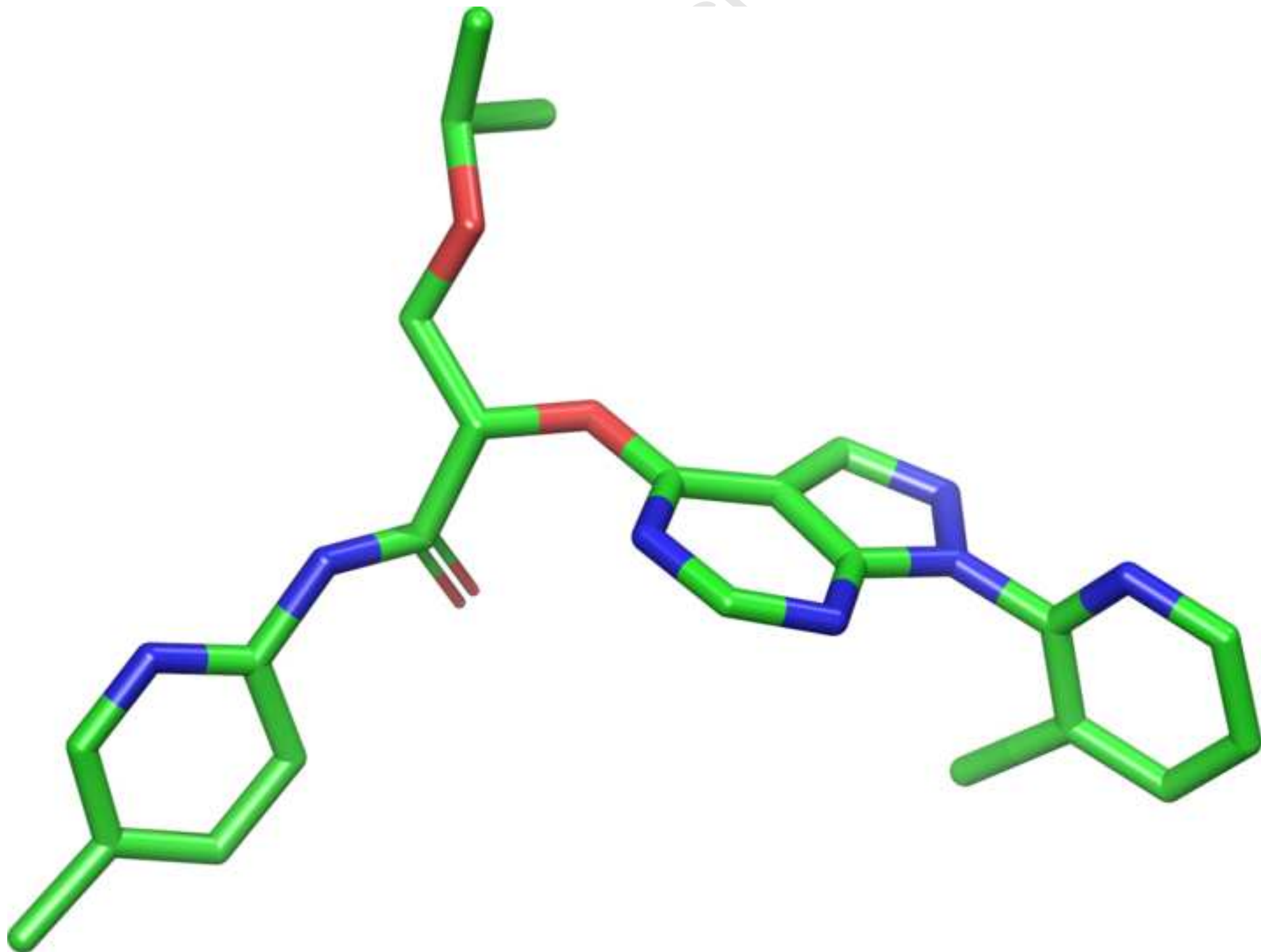
**Table-4:** The docking poses frequency of compounds in different variant models

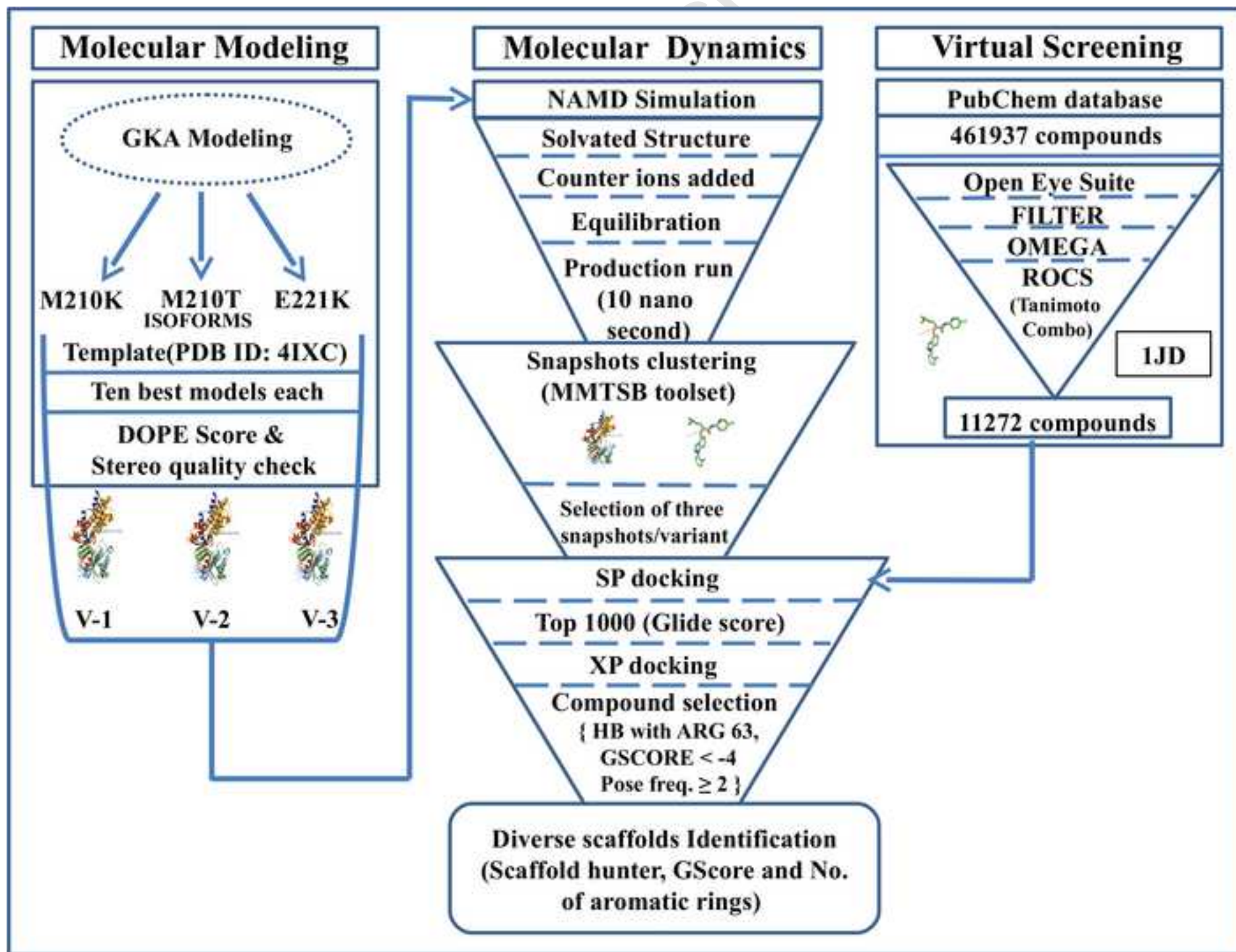
Comp.	V1_Freq	V1_Avg	V2_Freq	V2_Avg	V3_Freq	V3_Avg	Total_Freq	Total_Avg
2815231	3	-5.319	3	-8.581	3	-6.369	9	-6.756
2815230	3	-5.833	3	-6.468	3	-6.565	9	-6.289
3205407	3	-4.338	3	-8.312	2	-9.393	8	-7.348
661821	3	-6.489	3	-7.112	2	-6.979	8	-6.86
1035395	2	-4.728	3	-7.633	2	-8.003	7	-6.788
1637725	2	-5.832	2	-8.062	3	-6.188	7	-6.694
3205404	3	-4.678	2	-5.951	2	-9.248	7	-6.626
749956	2	-5.284	3	-6.905	2	-7.626	7	-6.605
2725422	3	-5.703	2	-6.259	2	-7.123	7	-6.362
677049	2	-4.129	2	-6.036	3	-6.399	7	-5.521
4051685	2	-4.464	3	-5.484	2	-6.403	7	-5.45

Comp.=PubChem CID of a compound. V1\_Freq, V2\_Freq and V3\_Freq are the frequency of compound in three variants. V1\_Avg, V2\_Avg and V3\_Avg are average GScore of compound in three variants. Total\_freq is the cumulative frequency of a compound. Total\_Avg is average GScore of a compound.

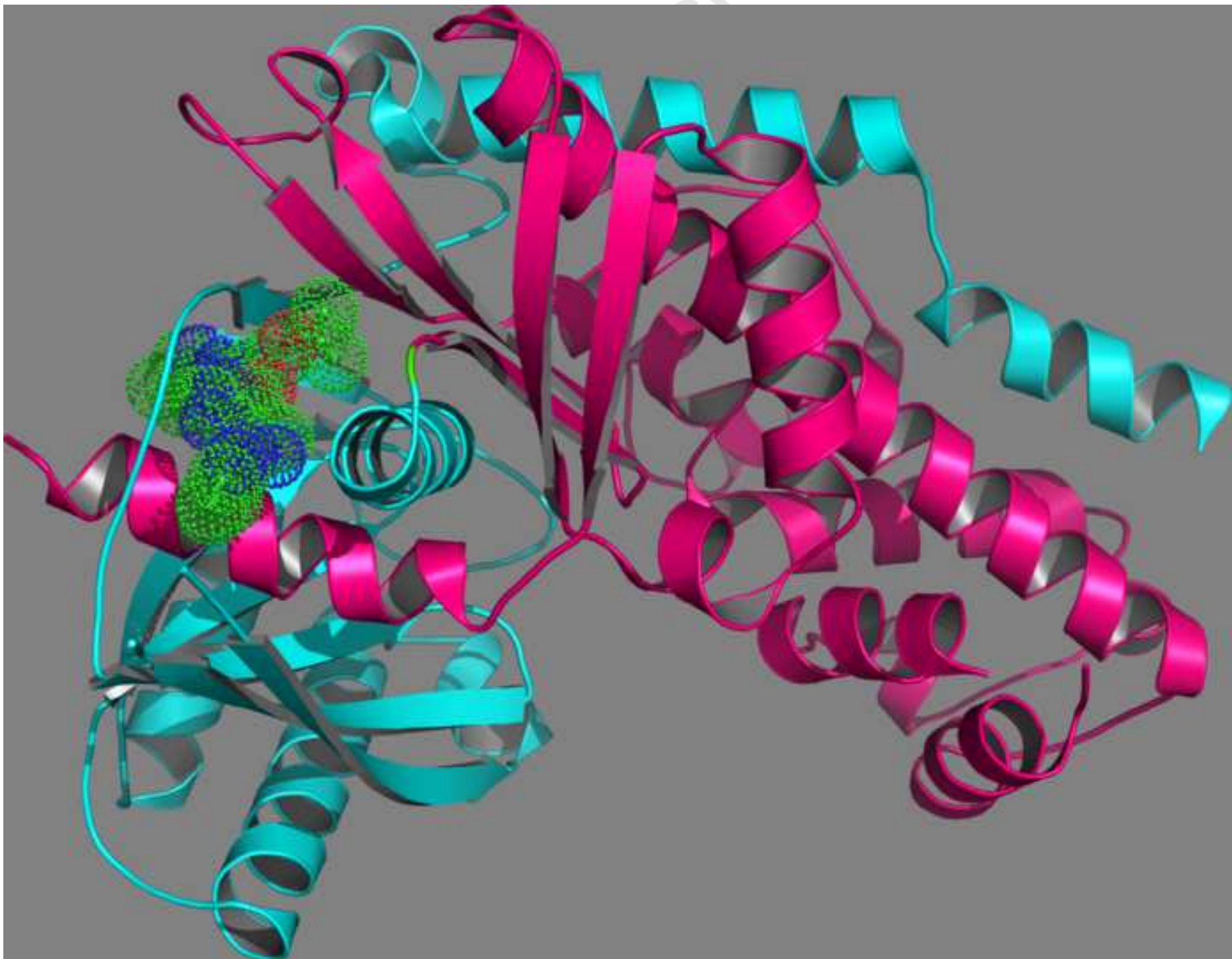
**Table-5:** Scaffold Hunter tree analysis featuring scaffolds.

Subset No.	Scaffold name	Scaffold structure	No. of compounds	PubChem CIDs
1	2-(benzylamino)-N-cyclopentylacetamide		3	3205404, 3205407
	1-phenyl-1H-1,2,4-triazole			661821
2	Benzo[d]thiazole		2	1637725, 677049
3	N-(4H-1,2,4-triazol-3-yl)benzamide		2	2815230, 2815231
4	2H-pyrazolo[3,4-d]pyrimidine		1	749956
5	6,7-dihydro-[1,3,4]thiadiazolo[3,2-a][1,3]diazepin-8(5H)-one		1	4051685
6	2-((1,2,4-triazin-3-yl)thio)-N-(1H-pyrrol-2-yl)acetamide		1	1035395
7	(E)-2-benzylidene-N-(furan-2-ylmethyl)hydrazinecarbotioamide		1	2725422

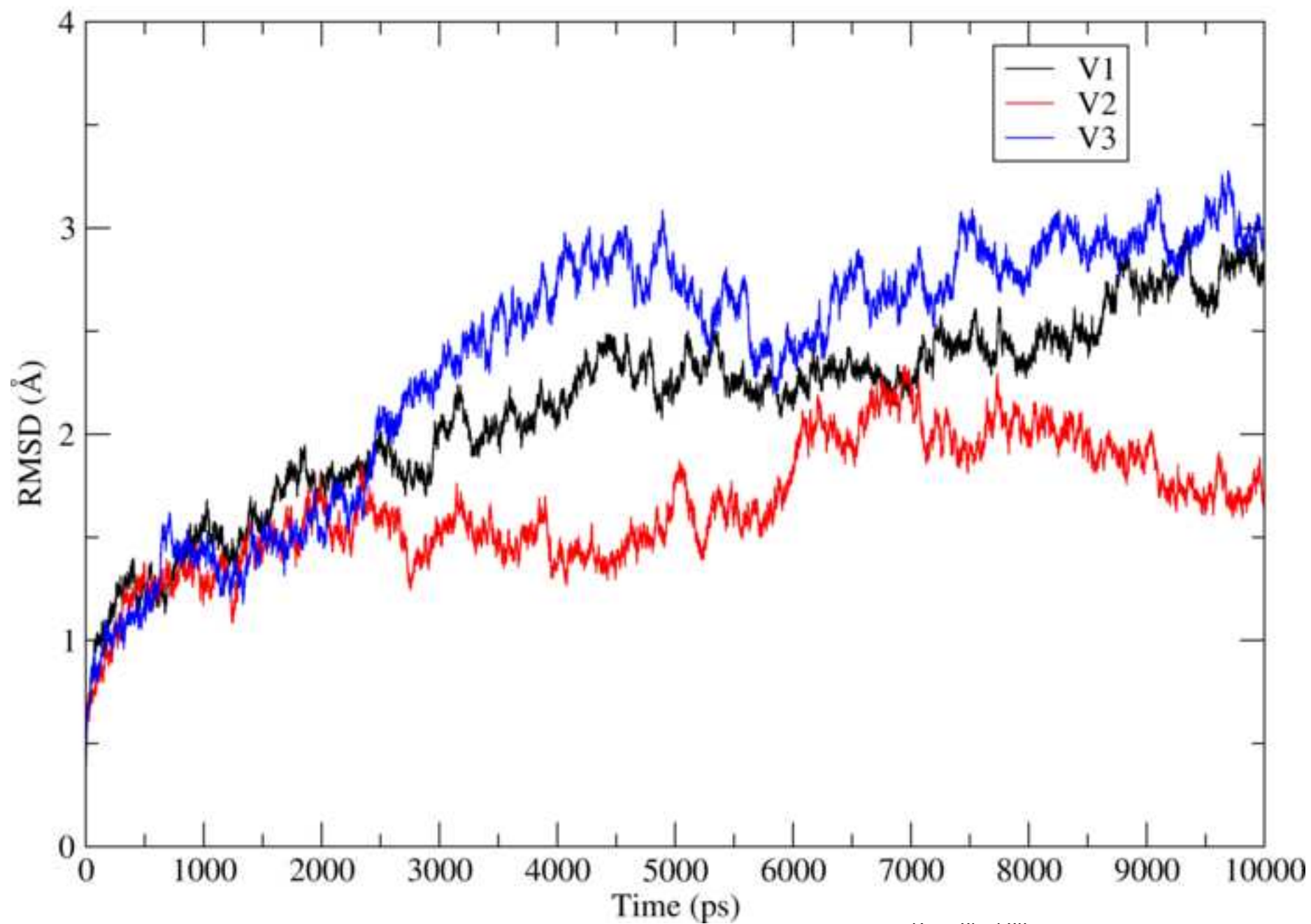


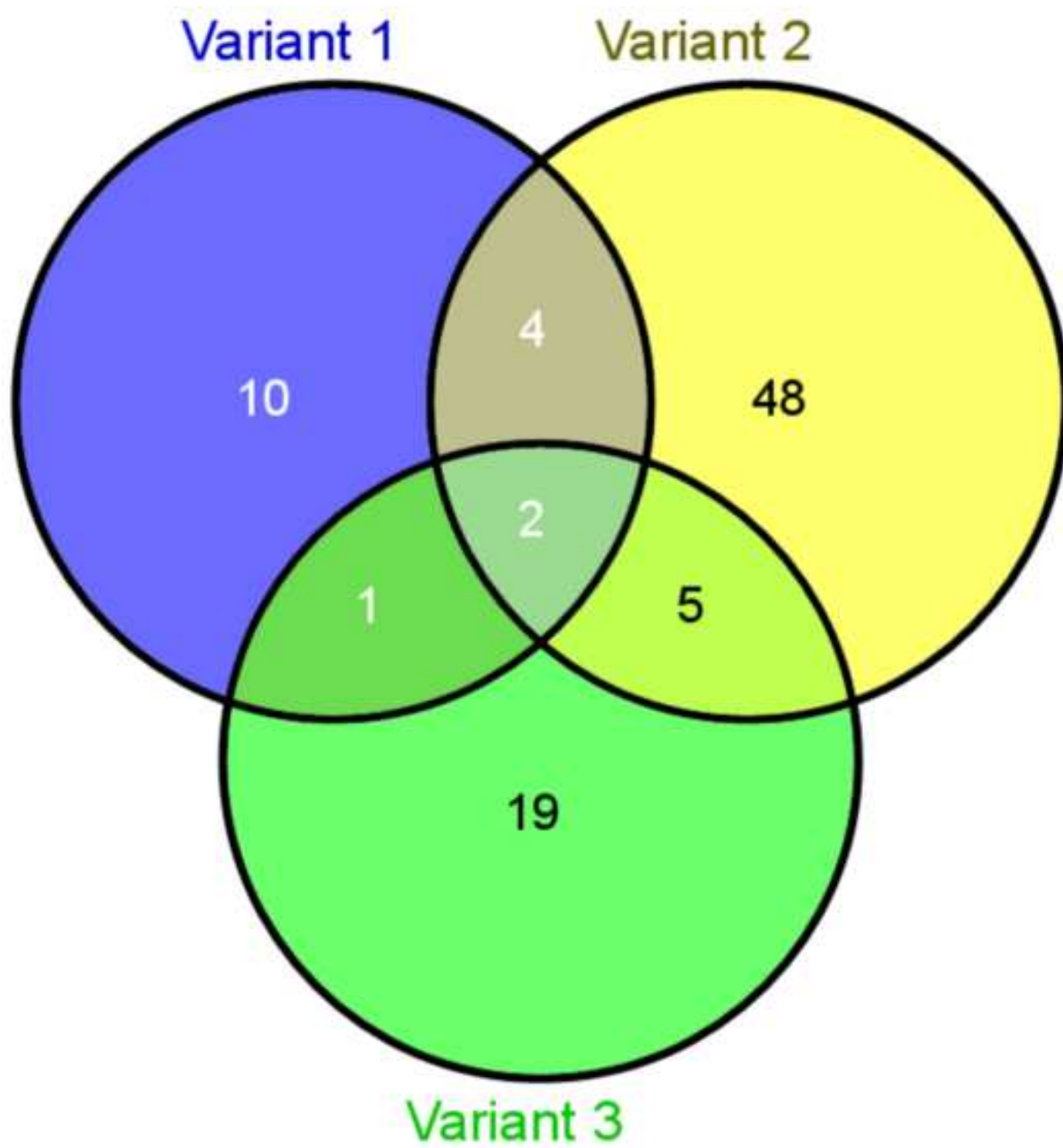






## RMSD vs Time



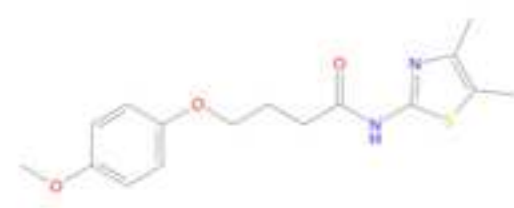




CID: 2815231



CID: 2815230



CID: 1637725



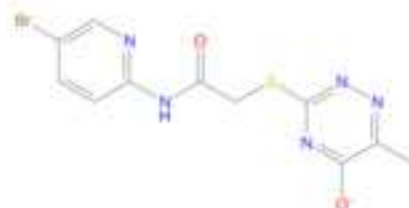
CID: 749956



CID: 2725422



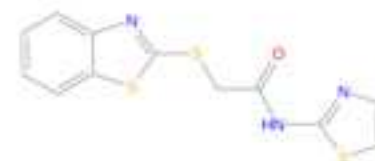
CID: 661821



CID: 1035395



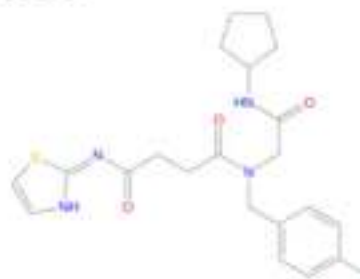
CID: 3205404



CID: 677049



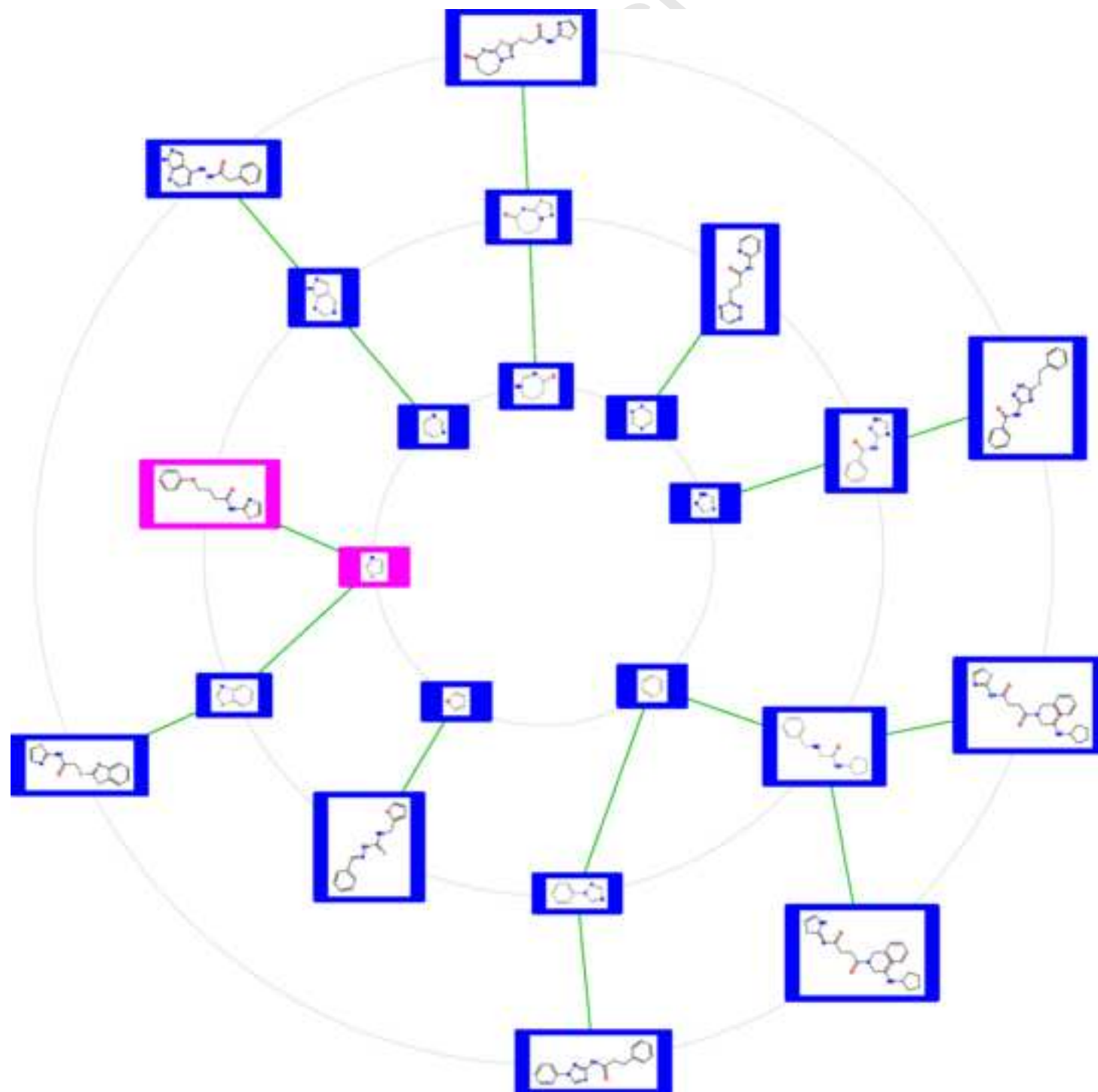
CID: 4051685



CID: 3205407







Figure(s)

