# PepGrow

## *Protocol*
ver. 14-06-2023

Balázs Zoltán Zsidó[1], Bayartsetseg Bayarsaikhan[1], Rita Börzsei[1], Csaba Hetényi[1]

*[1]Pharmacoinformatics Unit, Department of Pharmacology and Pharmacotherapy, Medical School, University of Pécs, Szigeti út 12, 7624 Pécs, Hungary.*

PepGrow is a protocol for calculation of target-ligand complexes with peptide ligands. The protocol produces the complexes using the structures of unbound target, and in silico generated dipeptide ligand fragments. PepGrow is based on open source and freely available software packages.

This Tutorial provides a practical help with the details of the usage of PepGrow following the steps described in the original publication (Zsidó et al., xxx). Input and output files **marked with bold** are available as Supplementary Files (Protocol.tgz) available online at https://zenodo.org/record/8036490.

**Table of Contents**

*Step 1.1 Target and ligand preparation for docking (Folder: seed_selection > docking_example)*
The names referring to systems (2ke1) and dipeptide fragment seeds (RT) should be modified upon switching systems and fragment seeds to their respective names.
For the docking, the target and the ligand dipeptide fragment structures were needed. Their respective *.pdbqt formats were created in AutoDock Tools using the following menu commands.

File>Read Molecule > **2ke1_target.pdb**
Edit>Hydrogens>Add
Save>Write pdb: **2ke1_target_h.pdb**
Grid>Macromolecule>open> **2ke1_target_h.pdb**
Save as **2ke1_target_h.pdbqt**

Ligand>Input > **RT_orig.pdb**
Torsion tree > Choose torsions
Output>Save as pdbqt> **rt.pdbqt**

*Step 1.2 Creating grid maps for docking (Folder: seed_selection > docking_example)*
The grid maps were created with the following command line, after setting the coordinates of the grid box in AutoDock Tools:

```
autogrid4 -p rt.gpf -l rt.glg
```

*Step 1.3 Fragment seed docking (Folder: seed_selection > docking_example)*
After setting the docking parameters in AutoDock Tools, the following command line was used for the docking:

```
autodock4 -p rt.dpf -l rt.dlg
```

The docked ligand file: **O_rt_rank_1-7.pdb**
The complex file: **2ke1_m.pdb**

*Step 1.4 Growing of whole ligand copies from docked fragment copies in the binding site (Folder: seed_selection > rt >1)*
The folder name rt refers to the used dipeptide fragment seed, subfolder 1 refers to the first ranked RT fragment seed from the previous step (subfolder 2 to the second ranked, etc…).

Other folders contain the same files for other dipeptide fragments, the nomenclature is changed to the respective dipeptide fragment name.

The complex structure of the target and the first rank of the docked RT fragment seed is **2ke1_m.pdb** and this file was used as the input of this step.

2ke1.ali file contains the sequence of the to-be-built target and whole ligand complex.

Using the following command line the sequence alignment is performed using the 2**ke1_m.pdb** and **2ke1.ali** files as inputs.

```
mod9.22 align2d.py
```

This step generates **komplex.ali** file. This file and **2ke1_m.pdb** file are the inputs for the following command:

```
mod9.22 model.py
```

The command generates 100 target-whole ligand complex files: **2ke1.B99990001-100.pdb**.

*Step 1.5 Evaluation of the generated target-whole ligand complexes (Folder: seed_selection > rt >1)*

For the evaluation of these target-whole ligand complexes the **energy_loop.sh** script is used. For different systems the file names should be modified in the script to match the name of the new system (2ke1_ref.pdb to xxxx_ref.pdb).

```
./energy_loop.sh
```

The input files for the script are the target-whole ligand complexes, the **target.pdb** file for alignment, the **2ke1_ref.pdb** and **2ke1_ref_5.pdb** files for RMSD calculation and **energy_types.prm** and **general_2012.prm** files for the pantherg program. The script will calculate all heavy atom RMSD values for the whole ligand, the first 5 amino acids of the ligand and Coulomb (Mehler-Solmajer) and Lennard-Jones interaction energies.

The output results are listed in **rmsd.log, first5_rmsd.log, Coulom_ms.log, and Lennard.log** files.

*Step 2.1 Target and ligand preparation for docking (Folder: fragment_docking > 2)*
Folder 2 is mentioned here also as an example, the files of other fragments (named 1-9, as seen in Fig. 4 of main text) are also found in subfolders of Folder fragment_docking.
For the docking, the target and the ligand dipeptide fragment structures were needed. Their respective *.pdbqt formats were created in AutoDock Tools using the following menu commands.

File>Read Molecule > **2ke1_target.pdb**
Edit>Hydrogens>Add
Save>Write pdb: **2ke1_target_h.pdb**
Grid>Macromolecule>open> **2ke1_target_h.pdb**
Save as **2ke1_target_h.pdbqt**

Ligand>Input > **RT_orig.pdb**
Torsion tree > Choose torsions
Output>Save as pdbqt> **rt.pdbqt**

*Step 2.2 Creating grid maps for docking (Folder: fragment_docking > 2)*
The grid maps were created with the following command line, after setting the coordinates of the grid box in AutoDock Tools:

```
autogrid4 -p rt.gpf -l rt.glg
```

*Step 2.3 Fragment seed docking (Folder: fragment_docking > 2)*
After setting the docking parameters in AutoDock Tools, the following command line was used for the docking:

```
autodock4 -p rt.dpf -l rt.dlg
```

The docked ligand file: **O_rt_rank_1-7.pdb**
The complex file: **2ke1_m.pdb**

Names of complex files follow the PDB ID name of the systems, and names of fragment seeds follow the names seen in Fig. 4 of the main text.

Description: In this Step, whole ligand copies are grown from docked fragment copies (from Step 2) in the binding site of the target, using Modeller for homology modeling.
Software used:  Modeller (https://salilab.org/modeller/download_installation.html)
**Available files: Protocol.tgz > seed_selection Folder**

*Step 3.1 Growing of whole ligand copies from docked fragment copies in the binding site (Folder: seed_selection > ar_selected_fragment >2)*
Every numbered folder corresponds to the number of rank of the docked AR fragment from Step 2.
The complex structure of the target and the second rank of the docked AR fragment seed is **2ke1_m.pdb** and this file was used as the input of this step.
2ke1.ali file contains the sequence of the to-be-built target and whole ligand complex.
Using the following command line the sequence alignment is performed using the 2**ke1_m.pdb** and **2ke1.ali** files as inputs.

```
mod9.22 align2d.py
```

This step generates **komplex.ali** file. This file and **2ke1_m.pdb** file are the inputs for the following command:

```
mod9.22 model.py
```

 The command generates 100 target-whole ligand complex files: **2ke1.B99990001-100.pdb**.

**Step 4 – Ranking of the whole ligand copies by E_inter**
Description: In this Step, the grown whole ligand copies from docked fragment copies in the binding pocket are ranked by $E_{inter}$ (see Methods of the main text for formula).
Software used:  lsqman (http://www.csb.yale.edu/userguides/
                         datamanip/uppsala/manuals/dejavu.html)
                 pantherg A to-be-published program, the binary is made available.
                 rmsd A to-be-published program, the binary is made available.
**Available files: Protocol.tgz > seed_selection Folder**

The command from Step 3 generates 100 target-whole ligand complex files: **2ke1.B99990001-100.pdb**. The names of the files change when other systems are investigated to the respective PDB ID.B99990001-100.pdb.

*Step 4.1 Evaluation of the generated target-whole ligand complexes (Folder: seed_selection > ar_selected_fragment >2)*
For the evaluation of these target-whole ligand complexes the **energy_loop.sh** script is used. For different systems the file names should be modified in the script to match the name of the new system (2ke1_ref.pdb to xxxx_ref.pdb).

```
./energy_loop.sh
```

This is repeated for every docked fragment seed rank (the numbering of the folders in the ar_selected_fragment folder correspond to these ranks), generating 500 models in total.

The input files for the script are the target-whole ligand complexes, the **target.pdb** file for alignment, the **2ke1_ref.pdb** and **2ke1_ref_5.pdb** files for RMSD calculation and **energy_types.prm** and **general_2012.prm** files for the pantherg program. The script will calculate all heavy atom RMSD values for the whole ligand, the first 5 amino acids of the ligand and Coulomb (Mehler-Solmajer) and Lennard-Jones interaction energies. The output results are listed in **rmsd.log, first5_rmsd.log, Coulom_ms.log, and Lennard.log** files. For other systems the filenames containing the PDB ID of the systems should be changed to the PDB ID of the new system.

The sum of these output files for all 500 models are found in the root of Folder ar_selected_fragment.

Average structure calculation examples for the top ranked 1% of the complexes are found in Folder top for whole ligands, Folder top_e5 for the top 1% selection based on the $E_{inter}$ calculated for the first 5 amino acids.

*Step 4.2 Representative complex selection (Folder: seed_selection > ar_selected_fragment > top_e5)*

In Folders top and top_e5 the top ranked 5 structures (based on whole ligand $E_{inter}$ and $E_{inter}$ for the first 5 amino acids, respectively) of the 2ke1 system (top 1%, 5 out of 500) are present as the input for **average.sh** script, producing **avg.txt** file.

```
./average.sh
```

The **avg.txt** text file has to be edited to have the PDB format → **average_ligand.pdb**. Then the **calc.sh** script produces **rmsd_100-104.log** files.

```
./calc.sh
```

The file corresponding to the lowest RMSD is selected as the closest to average, representative structure of the top 1% ranked structures.