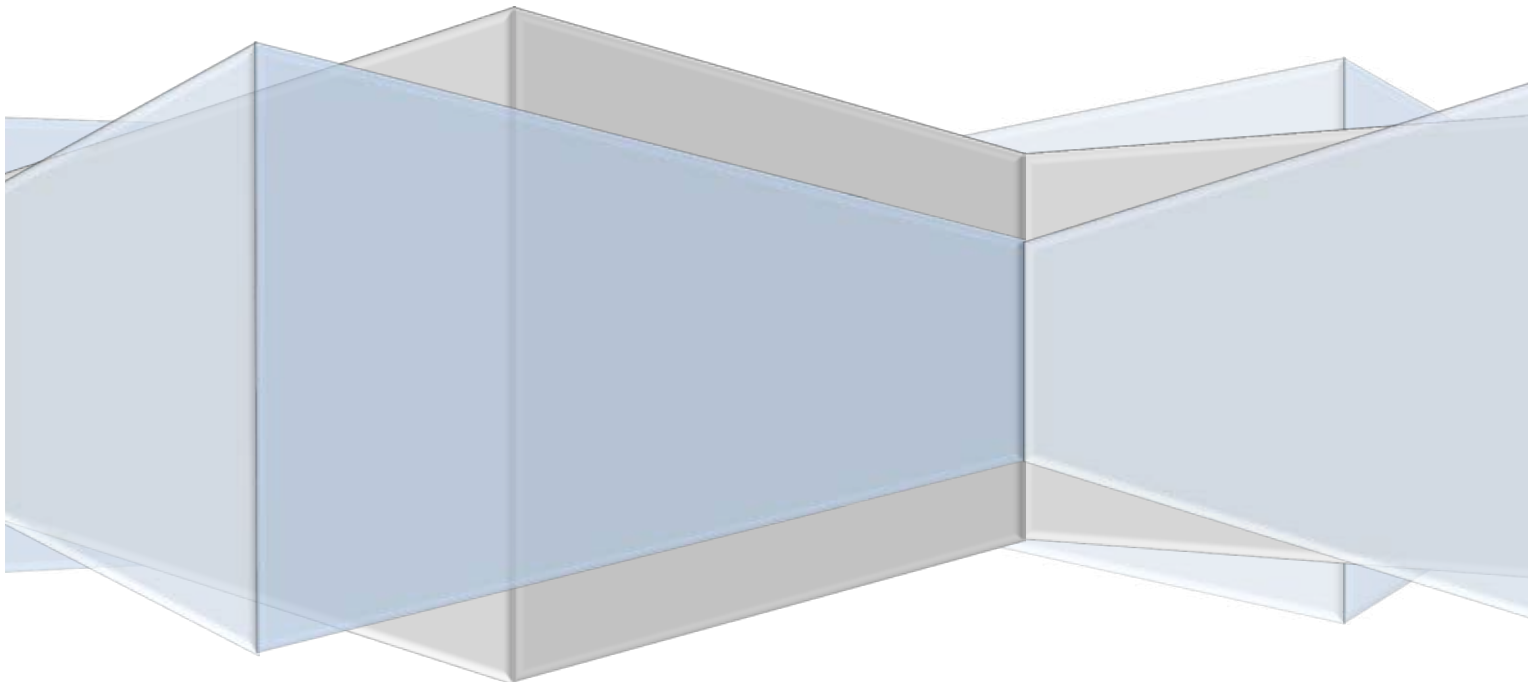


**Kyoto Constella Technologies Co., Ltd.**

# **CzeekD Manual**

**December 2014**



---

## TABLE OF CONTENTS

<b>1. Introduction .....</b>	<b>1</b>
1.1 Regarding use of CzeekD .....	1
1.2 Login and operational policy .....	1
<b>2. Creation of Compounds in CzeekD .....</b>	<b>2</b>
2.1 Compound creation scheme .....	2
2.2 Retrosynthetic Combinatorial Analysis Procedure (RECAP) .....	2
2.3 Particle Swarm Optimization (PSO) Algorithm .....	3
<b>3. Operation – Specifying Calculation Settings and Actual Calculation .....</b>	<b>4</b>
3.1 Target selection .....	5
3.2 Synthesis frame settings .....	7
3.3 Unit settings .....	7
3.4 Bond settings .....	9
3.5 Regarding Combinations (Chemical Space) and Estimated Time .....	11
3.6 Checking the status of a calculation job .....	12
<b>4. Calculation Results Analysis .....</b>	<b>13</b>
4.1 Distribution map .....	13
4.2 Compound list operations .....	14
4.3 Numeric and partial structure search .....	15
4.4 Compound display and fragment display .....	17
4.6 File export .....	21
<b>5. Other Features .....</b>	<b>22</b>
5.1 Modifying user information (changing password) .....	22
5.2 Compound scoring by CGBVS .....	22
5.3 Assigning several core structures .....	24
5.4 Sharing calculation results .....	25
5.5 Expanding the range of compounds that can be generated .....	26
5.6 Using custom-made fragments .....	26
5.7 Advanced calculation parameters .....	27

## Trademarks

All the company and product names appearing in this manual are trademarks or registered trademarks of the respective companies. Furthermore, trademarks are not appended to all the software and product names described in this manual.

## 1. Introduction

### 1.1 Regarding use of CzeekD

This manual explains how a computational chemistry personnel can utilize CzeekD with particular emphasis to medicinal chemists. Starting from the introduction of the calculation method involved in CzeekD, description of operations from calculation up to analysis of results will be described while at the same time showing relevant CzeekD windows. It is assumed that CzeekD will be used in a Windows environment and explanations in this manual are based on that assumption.

CzeekD is a Client-server type of web application offered as a service that designs and analyzes new compound structures based on CGBVS (Chemical Genomics-Based Virtual Screening) prediction score. Computational servers are maintained and managed at Kyoto Constella Technologies and therefore, it does not require any installation on the part of the user. Users can access the service using the web browser installed in their PCs (Internet Explorer, Firefox, Google Chrome). Below are the system requirements to use CzeekD.

- OS: Windows 7, 8, 8.1
- Screen Resolution: HD (1366×768) and above, Full HD (1920×1080) is recommended
- Browser: Firefox, Google Chrome, Internet Explorer
- Java version: 6 or 7
- Flash Player: Adobe Flash Player 11 and above
- Network connection: Broadband

### 1.2 Login and operational policy

Users can reach the CzeekD server by accessing the URL below:

<https://denovo.czeek.jp/czeekd/>

Once you reach the login page, fill in your login name and password then click on the Login button. Please change your password particularly during your first login (Refer to section 5 for details)

Below is our server operation policy.

- Our servers are structured such that, our application server is our frontend and the calculation server is our backend.
- We dedicate one CzeekD server for every customer.
- Data backups are performed automatically every Sunday midnight AM.
- In principle, we retain customer data only for a period of one year.

## 2. Creation of Compounds in CzeekD

The principle behind the creation of virtual compounds using CzeekD will be explained in this section.

### 2.1 Compound creation scheme

Virtual compounds are created according to the conditions that are defined in a synthesis frame (Figure 1). The synthesis frame is generally composed of units and bonds. Units represent areas where fragments are placed, whereas, bonds are parts where binding rules between fragments can be defined. Appropriate fragments are placed into all the units and results to creation of a virtual compound after binding of the fragments involved. Structures that do not conform to the defined conditions of the synthesis are not created. The following should be defined in the synthesis frame:

- ① number and configuration of units (topology)
- ② structures and/or fragment libraries to be used
- ③ rules that would govern binding between units

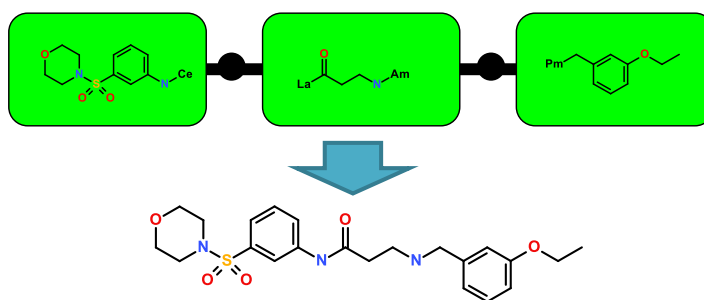


Figure 1. Example of a synthesis frame

There are 5 types of units (Figure 2) that are available in CzeekD, each differing only in the number of bonds possible. In creating the synthesis frame, units are selected and arranged then are bound according to the number of bonds available for each unit. Units with “dangling” bonds are not allowed. Up to 9 units can be set in a synthesis frame.

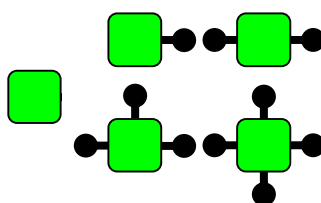
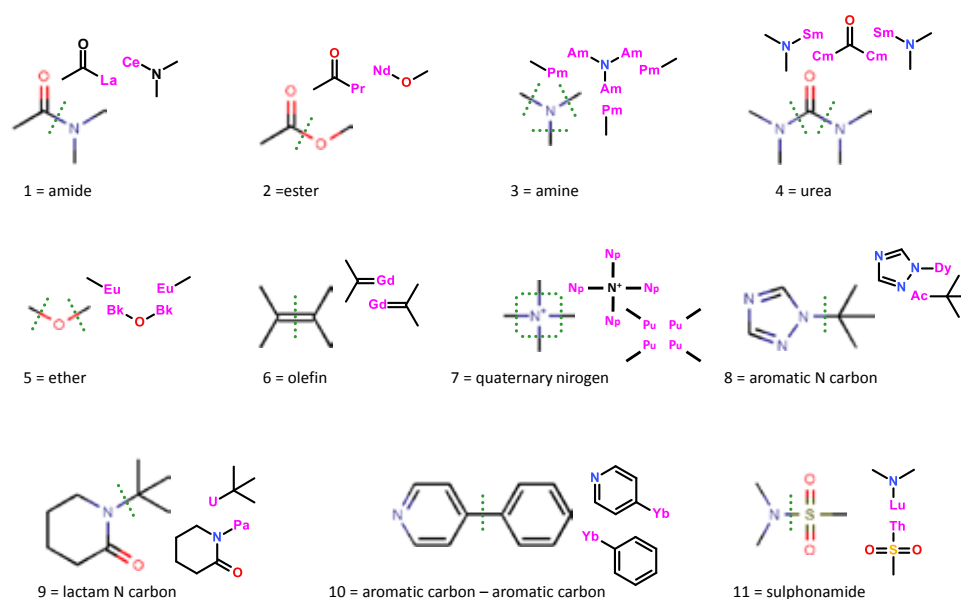


Figure 2. 5 types of units

### 2.2 Retrosynthetic Combinatorial Analysis Procedure (RECAP)

Fragments libraries provided in CzeekD are created based on the RECAP rules. Using the RECAP method, a particular compound is fragmented (cut) based on the rules by which it was originally chemically synthesized. Fragments referred to in CzeekD are basically, structures created after “cutting” compounds using the RECAP method. There are actually 11 basic rules that are followed in the RECAP method and these are indicated in Figure

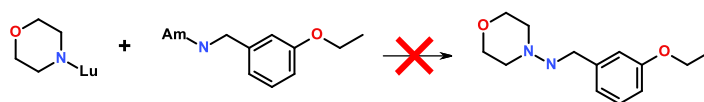
3 below.



**Figure 3. RECAP rules**

In CzeekD, we employ heavy metal symbols, such as rare earth elements, as markers for cutting and binding sites. These metals are not normally used in the creation of drugs, therefore, are used only as markers. Markers differ based on the cutting rule and the type of bond created after cutting. We also refer to these heavy metal symbols as “attachments.” For example, if an amide bond is cut, “La” is attached to the acyl side and “Ce” is attached to the amine side (please refer to Figure 3 for details). In CzeekD, fragment libraries refer to the collection of structures with attachments, and are classified according to the number of attachments.

Fragment binding also follow the RECAP rules unless otherwise specified. That is, binding based only on the above combinations of heavy metal markers is allowed. For example, amide bonding is only allowed with fragments having “La” and “Ce” attachments. In Figure 4 below, combinations that are not allowed in RECAP, like “Lu” and “Am” below, do not result in binding and, therefore, to the formation of new compound structures. Using the RECAP rule to govern the binding of fragments results to creation of virtual compounds with high probability of synthetic accessibility.



**Figure 4. Example of a fragment combination not conforming to the RECAP**

### 2.3 Particle Swarm Optimization (PSO) algorithm

This section describes the process by which appropriate fragments are arranged in each unit. The use of PSO algorithm in CzeekD allows selection of fragments for binding such that the capabilities of the scoring function is maximized.

PSO is an optimization technique that mimics the movement of insect swarms and fish schools that are seeking food or safety. It has been found to have better performance than genetic algorithm (GA). Exchange of information between each particle in a swarm regarding the best solution each has found on its own, leading to the determination of the overall best solution is the unique feature of PSO. The general equation for the PSO algorithm is shown below. Each particle has its own position ( $\vec{X}$ ) and speed ( $\vec{V}$ ) vectors and, based on the equation, searches the chemical space for the best solution.

$$\begin{aligned}\vec{X}_{t+1} &= \vec{X}_t + \vec{V}_t \\ \vec{V}_{t+1} &= w\vec{V}_t + r_1c_1(\vec{X}_{pbest} - \vec{X}_t) + r_2c_2(\vec{X}_{gbest} - \vec{X}_t)\end{aligned}$$

In the equation above,  $\vec{X}_{pbest}$  refers to the best solution found by each particle (personal best),  $\vec{X}_{gbest}$  refers to the best solution found by the entire swarm (global best),  $w$  refers to the inertia constant,  $r_1 r_2$  indicates random digits from 0 to 1, while  $c_1 c_2$  refers to the size of particles attracted to the best solution. The position and speed vectors are constantly recalculated and changes in their values lead the swarm nearer to the best solution. There have been other equations that are proposed to represent PSO, but in our case, we are adopting the equation above.

If you would apply PSO to *de novo* design, a particle represents a compound. On the other hand, a particle's position vector ( $\vec{X}$ ) is a value that represents the properties of the compound represented by the particle. The fitness function score that should be optimized is one that is independent of  $\vec{X}$  and in the case of CzeekD, the CGBVS (Chemical Genomics-Based Virtual Screening) score is adopted. This score is also the same score calculated when using our other product CzeekS (when several proteins are specified, the average score is presented). The position vector  $\vec{X}$  is represented by the following equation:

$$\vec{X} = (\vec{x}_1, \vec{x}_2, \dots, \vec{x}_u)$$

In the above equation,  $\vec{x}_u$  is a vector that represents the properties of the fragment while  $u$  represents the unit. Since a virtual compound is created by combining all the fragment units in the synthesis frame, the position vector  $\vec{X}$  can be represented by all the fragment vectors  $\vec{x}_u$  connected in series. Although the basic equation for PSO indicates that the position vectors are continuously changing, with regards to *de novo* design, the structure of the compound must be obtained from the value of the position vector  $\vec{X}$ . The approach that is used to realize that is to select all the fragments with vectors whose properties are very much alike with those of  $\vec{x}_u$  that constitutes  $\vec{X}$  and subsequently bind them to create the virtual compounds.

If several proteins are selected, an overall score will be calculated. Score calculation is represented by the following equation.

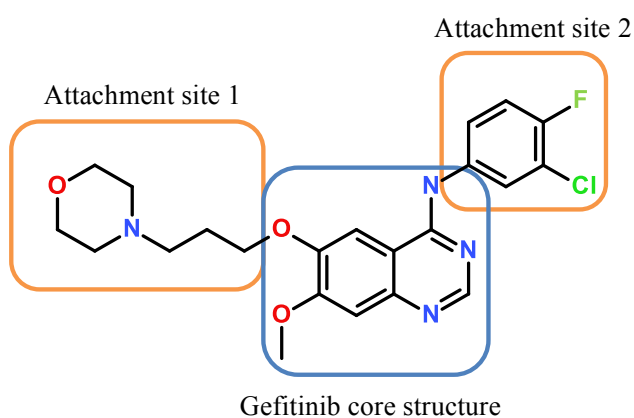
$$\text{Score} = \sqrt[n]{\prod_{i=1}^n S_i} \quad S_i = \begin{cases} P_i & \text{positive} \\ (1 - P_i) & \text{negative} \end{cases}$$

In the equation above,  $n$  represents the number of proteins,  $S_i$  represents the score of the  $i^{\text{th}}$  protein and if negative (that is, off-target protein) is selected,  $S_i$  is calculated using  $(1 - P_i)$ .

### 3. Operation – Specifying Calculation Settings and Actual Calculation

In this section, *de novo* design settings will be explained using Kinase as an example target. The target proteins are EGFR and GAK. Compounds that are active against EGFR while at the same time inactive against GAK will be

calculated (searched) for this particular example. As shown in Figure 5, compounds will be designed having the Quinazoline part of Gefitinib as the core structure with both ends serving as fragment attachment sites.

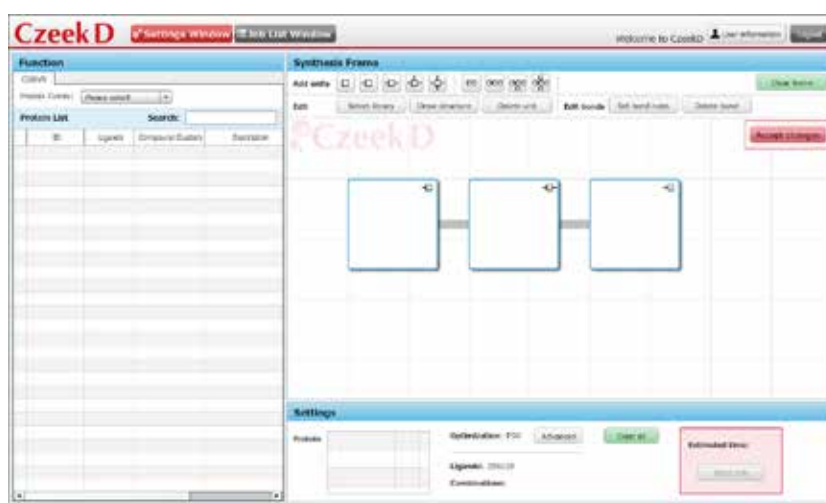


**Figure 5. Compound design scheme**

### 3.1 Target selection

There are 3 general settings that must be performed by the user: a) select target protein (calculation model), b) specify synthesis frame settings, and c) set calculation parameters. The last is optional and default settings will be assumed if not set by the user.

The window below will be shown immediately right after logging in to CzeekD. On the left side of the window is a panel that allows the user to select the target protein family and the actual target proteins itself. The biggest panel located at the right side of the window allows the setting of the synthesis frame. Below is where calculation conditions can be set and is also where the “Start job” button is located.



**Figure 6. Settings window**

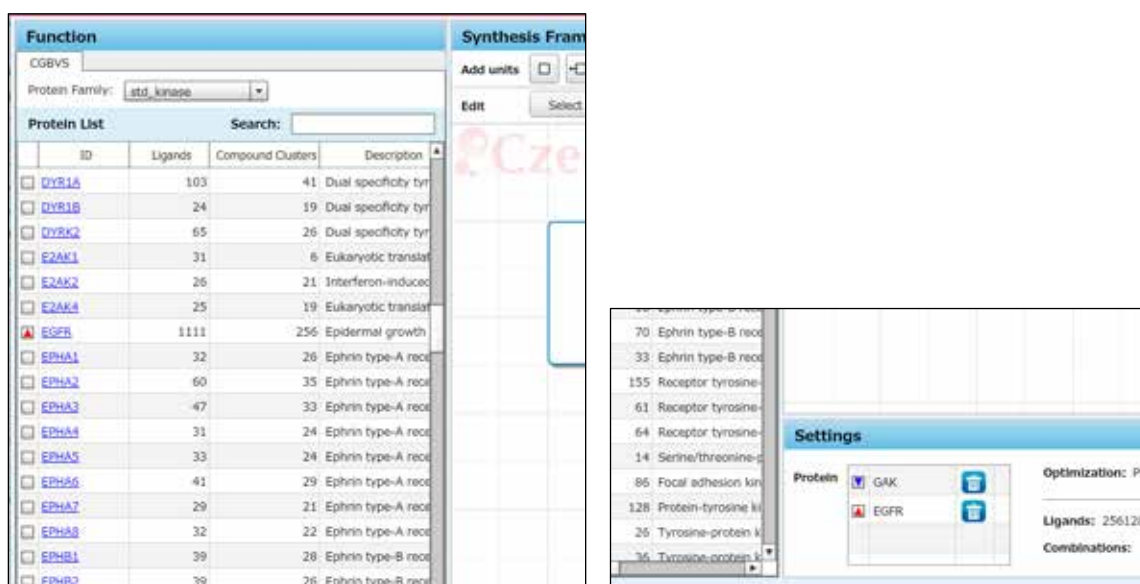
In selecting the target protein, the CGBVS calculation model should be selected first (Figure 7). Models can be selected after clicking on the dropdown list. There are 6 standard models to choose from (Table 1). Once a model is selected by clicking on it, available target proteins will then be shown in a tabulated list (Figure 7). The Ligands



column shows the number of compounds used in machine learning for each target protein and the Compound Clusters column indicates the number of structural variations of those compounds. Please keep in mind that in CGBVS, higher number of structural variations between compound used in machine learning equates to higher prediction accuracy.

Model name	Description	Proteins
std_gpcr	Standard model for GPCRs	233
std_ionch	Standard model for Ion channels	168
std_kinase	Standard model for Kinases	398
std_nuclear	Standard model for Nuclear receptors	28
std_protease	Standard model for Proteases	218
std_trans	Standard model for Transporters	59

**Table 1. List of standard models**



**Figure 7. Protein selection panel**

Figure 7 shows an example in which the Kinase standard was selected. A protein can be selected by clicking on the checkbox located on the left side of the protein ID. Clicking on the box once will show ▲ and a second click will show ▼. A third click will return it to an empty checkbox. ▲ indicates that active compounds against the selected protein will be designed, while ▼ indicates that the compounds to be designed should not be active against the selected protein. In short, ▲ and ▼ indicate the target and off-target, respectively. The protein IDs are hyperlinks to the UniProt database and clicking on them will open the web browser and show the page containing information about the “clicked” protein. Up to 10 proteins can be selected and these proteins will be shown on the lower right hand panel (Figure 7). In the case of Figure 7, EGFR will be the target and GAK will be the off-target.

### 3.2 Synthesis frame settings

Setting the required conditions for the creation of virtual compounds is performed in the synthesis frame panel. Operation for the required settings (shown below) introduced in the section 2 will be explained in this section.

- ① set number and configuration of units (topology)
- ② specify structures and/or fragment libraries to be used
- ③ select rules that would govern binding between units

Immediately after logging in to CzeekD, the settings window will be shown (Figure 6) wherein a synthesis frame with default topology (3 units arranged horizontally) is initially set. In our example, this synthesis frame can be used without changing the topology. But if the topology has to be modified, the default synthesis frame must be deleted first by clicking on the “Clear frame” button.

Creating a new synthesis frame can be performed by first clicking on any of the desired unit located on the right side area of the “Add units” label (upper left hand area of the panel) (Figure 8). Clicking on the frame area will create the selected unit. After placing the desired units, connect the units by clicking and dragging between the connection points of each unit. There are 5 types of units that can be used in CzeekD each having from 0 to 4 bonding points (Figure 9). If a unit having no bonding point is selected, a simple scoring (screening) calculation will be performed (refer to p. 22 for details). There is a limit of 9 units that can be used to compose a synthesis frame. Furthermore, one or more units that have one or more unattached binding points are not allowed.

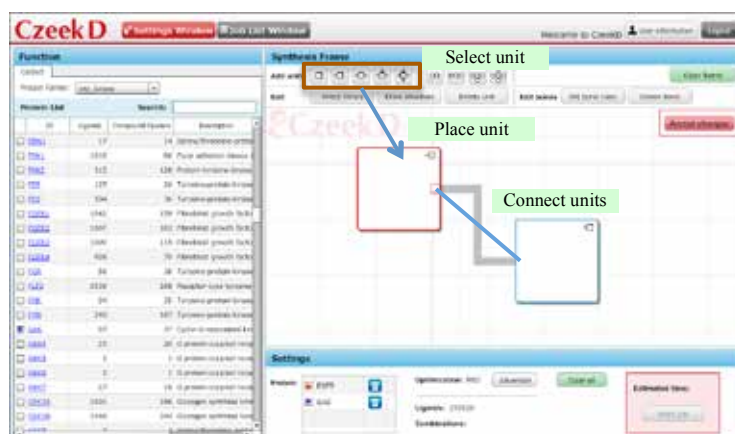


Figure 8. Selecting, arranging and connecting units



Figure 9. 5 types of units according to the number of bonding points

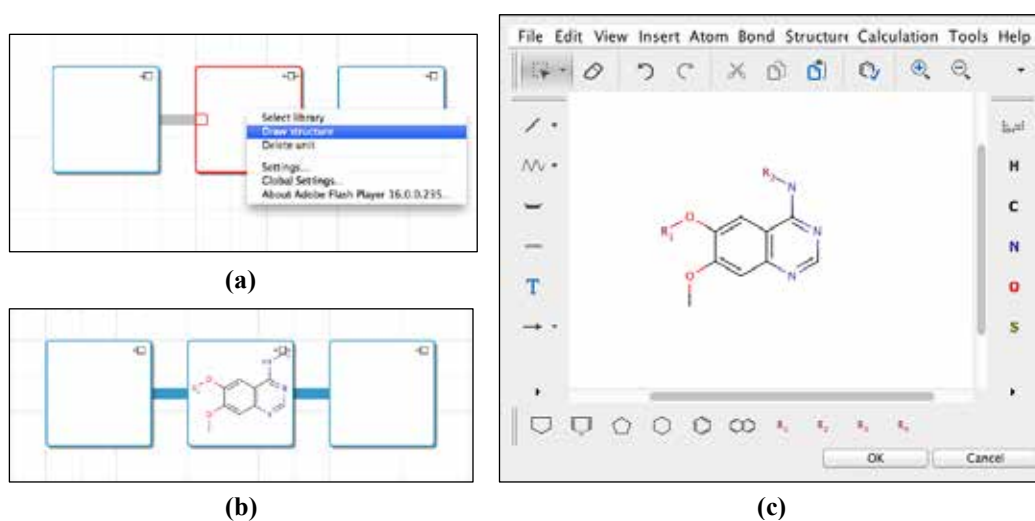
### 3.3 Unit settings

Unit properties can be set in two ways: 1) drawing a structure through the use of a plugin integrated within CzeekD and 2) selecting a fragment library.

#### 3.3.1 Drawing a structure

With this method, a structure (fragment) is set by using a drawing tool integrated within CzeekD. As shown in

Figure 10(a), it can be done by right-clicking on one of the units then clicking on the “Draw structure” item on the context menu. The Marvin Sketch [Figure 10(c)] plugin will appear in a popup window that would allow the user to draw the desired structure. Another way to activate the drawing tool is to select a unit by clicking on it once then clicking on the Draw structure button. Clicking on the OK button after drawing the structure will close the plugin and return to the CzeekD window already showing the structure drawn [Figure 10(b)]. The number of bonding points, as well as, the number of R groups must also be defined. Setting the R groups can be done using the R1 to R4 buttons in Marvin Sketch. Structures that are created using other tools can also imported from files and by copying and pasting from other applications. Please refer to the following link for details regarding the use of Marvin Sketch: <http://www.chemaxon.com/marvin/help/sketch/sketch-index.html>. In our example, the unit at the middle will be set to the structure of Quinazoline as shown in the Figure below.



**Figure 10. Drawing the core structure**

### 3.3.2 Selection of fragment library

There are two ways to select the desired fragment library: 1) by right-clicking on a unit then clicking on “Select library” item on the context menu and 2) by selecting a unit by clicking on it once then clicking on the “Select library” button. This will open a window that will allow the selection of one or more desired fragment library from a list. Several fragment libraries are available depending on the structure and intended usage. Please refer to Table 2 for details.

The number of fragments contained within the selected fragment library can still be narrowed down by using one or more methods shown below:

- According to suppliers: this is possible only with those libraries containing supplier information
- According to quantitative physical properties: number of heavy atoms, molecular weight and ClogP
- According to partial structure: can be done by fragment structure and structure similarity search

After the conditions above have been set, clicking on the “Search” button will start the search for fragments fitting the set conditions. Clicking on the “Show Fragment List” button will reveal the selected fragments on the right side of the window (Figure 11).

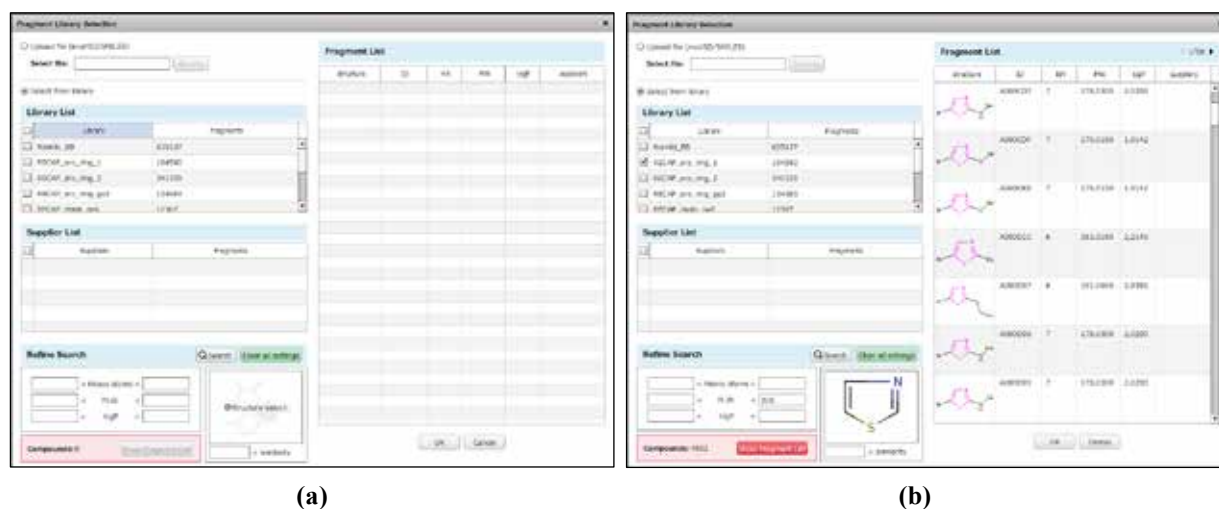


Figure 11. Fragment library selection window

The example in this section sets the left unit to “RECAP\_non-aro\_ring\_1” and the right unit to “RECAP\_aro\_ring\_1” fragment libraries.

Library Name	Description	Supplier Information
Namiki_BB	Building block library being marketed by Namiki Shoji	present
RECAP_aro_ring_1	Has only 1 aromatic ring	absent
RECAP_aro_ring_2	Has 2 ring structures one or more of which is aromatic	absent
RECAP_aro_ring_ge3	Has 3 or more ring structures 1 or more of which is aromatic	absent
RECAP_chain_ge6	Has 6 or more rotatable bonds but no ring structures	absent
RECAP_chain_le5	Has 5 or less rotatable bonds but no ring structures	absent
RECAP_non-aro_ring_1	Has only 1 non-aromatic ring	absent
RECAP_non-aro_ring_2	Has 2 non-aromatic rings	absent
RECAP_non-aro_ring_ge3	Has 3 or more non-aromatic rings	absent

Table 2. Fragment library list

### 3.4 Bond settings

In CzeekD, the RECAP rule is applied to bonding between fragment units and the fragment that conform to these rule are automatically selected. However, if a particular structure is set for a unit, the RECAP rule that must govern the binding must be explicitly defined. The procedure is explained below.

As shown in Figure 12(a), the bond to be set is selected by clicking on it once after which the “Set bond rules” button is clicked. This will reveal a pull down menu from which the appropriate R group is selected. A window similar to that in Figure 12(b) and 13 will then be shown. At the upper part of the window, the previously drawn structure will be shown with the selected R group as shaded. 11 types of bonding rules are shown in Figure 14 and

with the exception of olefins and aromatic carbons, variations are shown in pairs. One or more bond types can be selected.

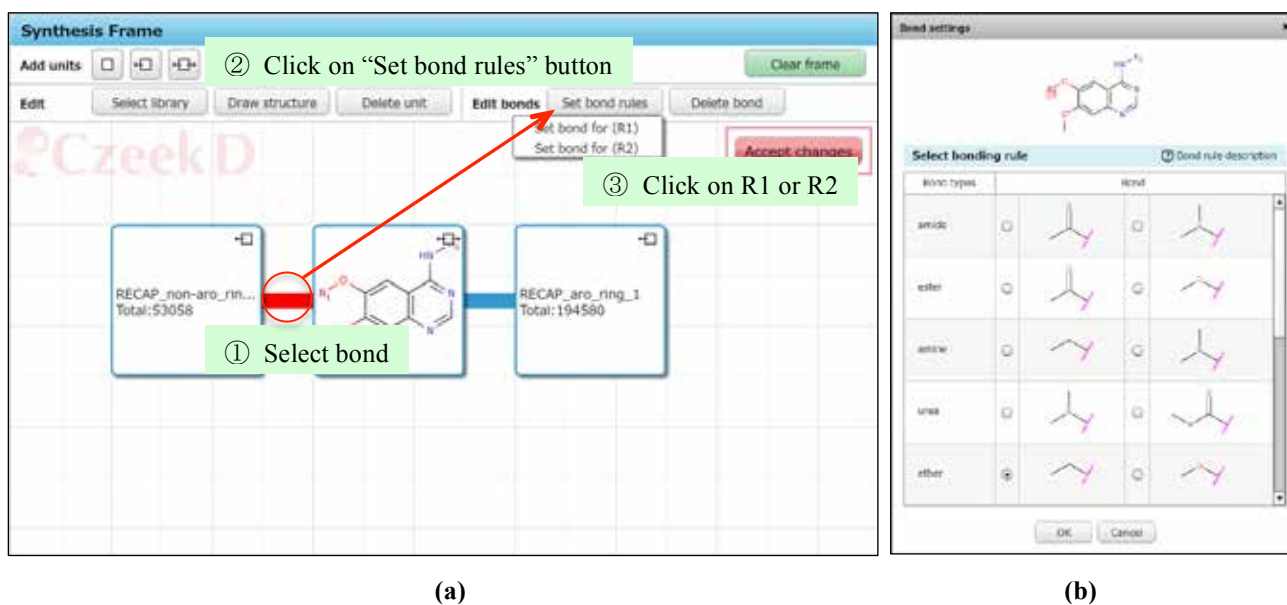


Figure 12. Bond rule settings window

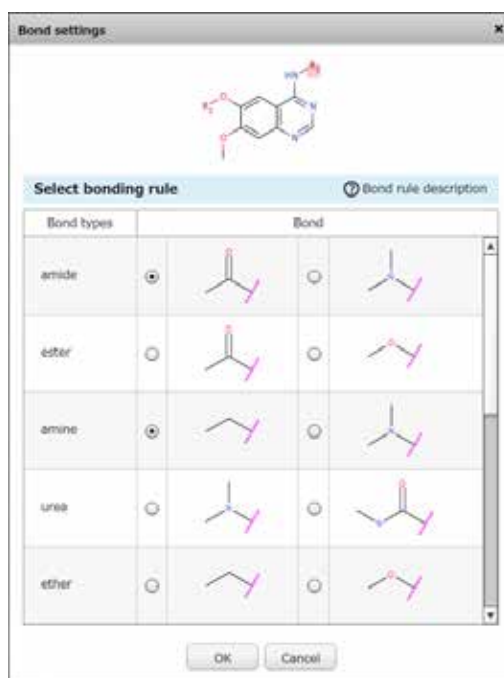


Figure 13. Setting bond rules for R2

Figure 14. Possible settings based on RECAP rules

Bond types	Bond			
amide	<input type="radio"/>		<input type="radio"/>	
ester	<input type="radio"/>		<input type="radio"/>	
amine	<input type="radio"/>		<input type="radio"/>	
urea	<input type="radio"/>		<input type="radio"/>	
ether	<input checked="" type="radio"/>		<input type="radio"/>	
olefin	<input type="radio"/>			
quaternary nitro	<input type="radio"/>		<input type="radio"/>	
aromatic N carb	<input type="radio"/>		<input type="radio"/>	
lactam N carbon	<input type="radio"/>		<input type="radio"/>	
aromatic carbon	<input type="radio"/>			
sulphonamide	<input type="radio"/>		<input type="radio"/>	

### 3.5 Regarding Combinations (Chemical Space) and Estimated Time

The settings required to start a *de novo* design “job” is almost complete at this point. Clicking on the accept changes button will make the settings take effect and if there are no errors encountered will give values for the “Combinations” (equivalent to size of chemical space) and “Estimated Time” (approximate time required to perform *de novo* design calculation).



Figure 15. Summary of calculation settings

“Combinations” refers to the total number of compounds that will be created based on the synthesis frame settings. It is basically shown as a high value exponential notation. “Ligands” refer to the default number of compounds actually included in the calculation. This value depends on the number of particles and number of iterations set within the PSO Advanced settings. “Estimated time ” indicates the approximate time it will take for the calculation to finish and is meant only as a guide.

After carefully checking the settings, the calculation can be commenced by clicking on the “Start job” button. A dialog box just like in Figure 16 will be shown and a job name for the calculation can be entered. As much as possible, use an easy to comprehend job name.

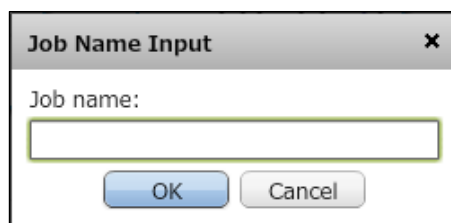


Figure 16. Job name input dialog box

### 3.6 Checking the calculation job

Status of calculation jobs can be checked by clicking on the “Job List Window” button (Figure 17). Check the “Status” column to see whether your job is currently running or have already completed. The status of a currently running job is usually shown as the percentage (%) of the completed calculation. Job status is usually shown in the following order: Queued, Pre-Processing, Running (shown as percentage), Post-Processing, Normal-end. Results of complete jobs can be displayed and further analyzed. A currently running job can be aborted by clicking on the “STOP” button, which is displayed in place of the trash bin icon in case of a currently running job.

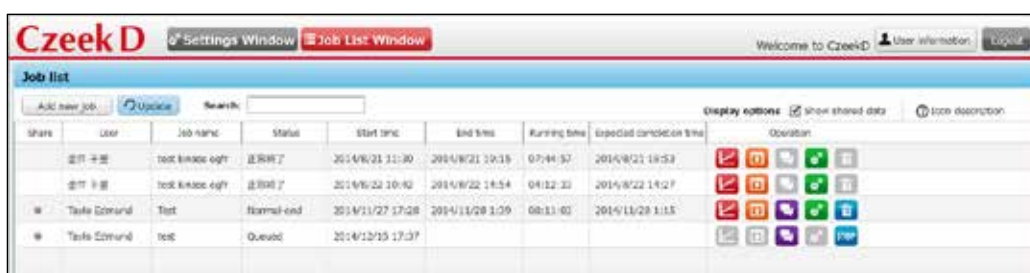


Figure 17. Job list window

The trash bin icon appears in completed jobs, as well as, in aborted jobs. Clicking on the trash bin icon will delete the job. Deleting a job cannot be undone so it is advisable to download your results before doing anything else so as not to accidentally lose important data.



## 4. Calculation Results Analysis

In this section, display of results and analysis feature will be described. We have used EGFR and GAK as example targets since section and we will continue to use the same in this section. Double-clicking on a job or clicking on the red icon at the rightmost column of the job list window will display the results of a calculation job.

### 4.1 Distribution map

A sample of the results analysis window is shown in Figure 18 below. The left hand side shows the data manipulation area while the right hand side displays the selected compound structures with corresponding scores and parameters. Compounds shown in the distribution map range in colors from blue to nearly red indicating the density (number) of compounds. Initially, the distribution map is presented as the plot of molecular weight at the X-axis (“Weight”) against the “score” at the Y-axis. Changing the parameters plotted in any of the axes can be done by clicking on any of the axis labels then selecting the desired parameters in a popup window [Figure 19(a)]. In this example, CGBVS scores of GAK (X-axis) and EGFR (Y-axis) are plotted.

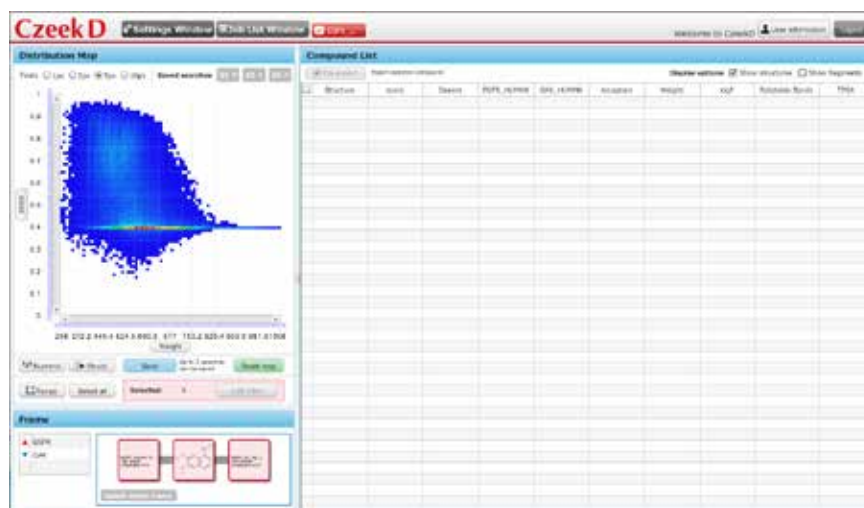


Figure 18. Results analysis window

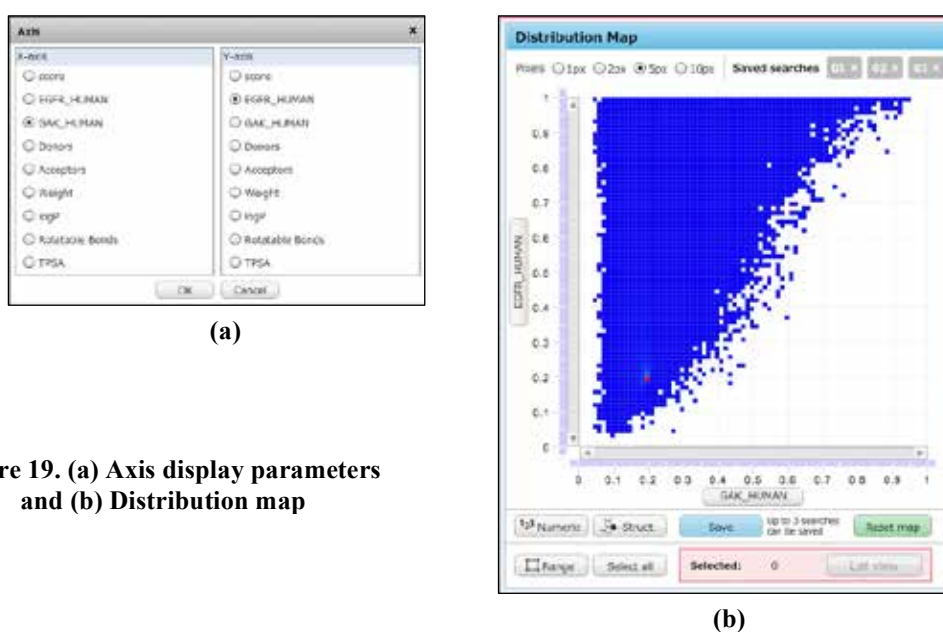


Figure 19. (a) Axis display parameters and (b) Distribution map



Clicking and dragging on an area within the distribution map will select the compound present in that area. Clicking the “List view” button will show the compound structures with its corresponding scores, as well as, other parameters (Figure 20). Selecting a range by specifying values or selecting all the plotted compounds can be performed by clicking on the “Range” or “Select all” buttons, respectively. Canceling the selected range can be done by double clicking within the distribution map. The maximum number of compounds that can be selected is 50,000.

① Click and drag to select an area

Double-clicking within the map will reset selection

② Click List view button

Click on the Numeric button to do numeric search

**Figure 20. Selecting an area within the distribution map and displaying the structures**

## 4.2 Compound list operations

Putting the mouse over a structure shown on the compound list will magnify the structure and clicking on the structure itself will open a structure viewer displaying the clicked structure. Expanding the width of the structure column will also result to the magnification of the displayed structures. Unchecking the “Show structures” checkbox will hide the compound structures while clicking the “Show fragments” checkbox will show the fragment structures. Clicking on the column name of any parameter will reveal a dialog box displaying sorting and filtering options. Clicking and dragging on the column name of any parameter will change the order of the columns. (Figure 21).

Compounds can be selected by clicking on the checkbox to the left of the compound structures. Selected compounds can then be exported in SDF or CSV format by clicking on the “File export” button.

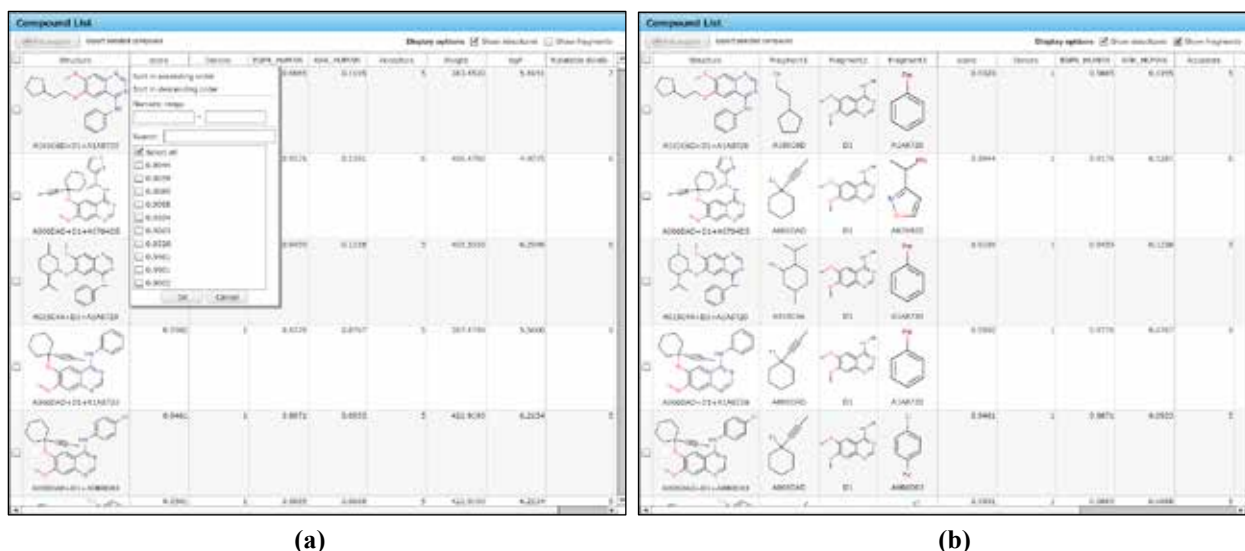


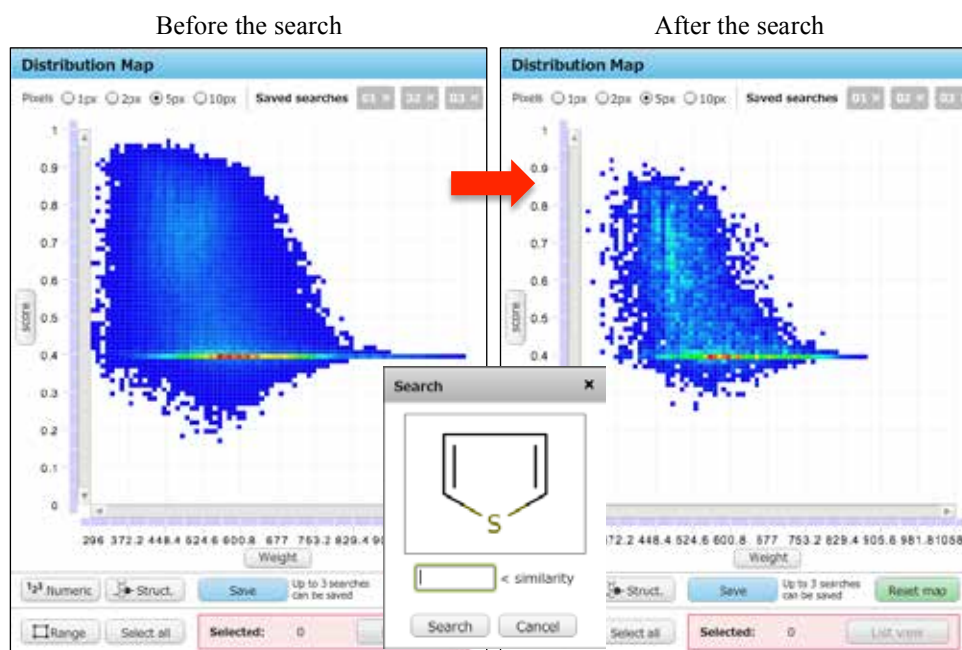
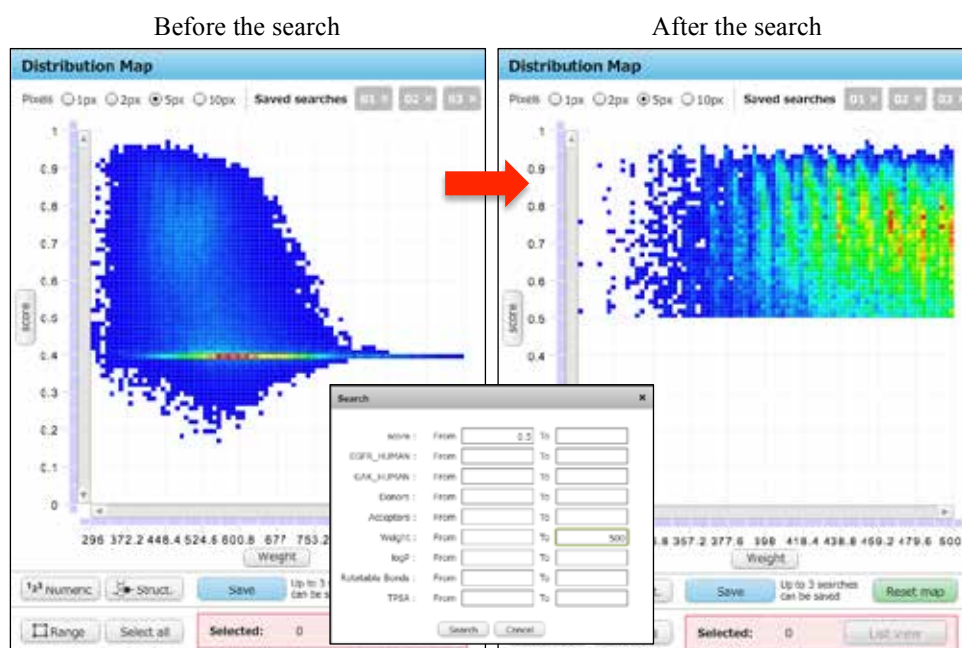
Figure 21. Compound list panel

### 4.3 Numeric and partial structure search

Display filtering has been described in the previous section, but actually, in the first place, only 50,000 compounds can be displayed. However, compounds to be displayed can be narrowed down by filling in search conditions for the distribution map. There are two ways that this can be done: first, by using numerical parameters and second, by partial structure search.

Clicking on the “Numeric” button will open a dialog box that allows entry of numerical values that will be used for the numeric search (Figure 22). In this example, scores of 0.5 and above and molecular weights of 500 or less were used as numeric search parameters. Executing the search changes the display of the distribution map depending on the conditions set in the numeric search parameters.

Compound search based on partial structures can be performed by clicking on the “Struct.” button then clicking within a dialog box that appears afterwards. This opens the Marvin Sketch plugin that allows entry of the desired structure. Clicking OK after drawing the structure in Marvin Sketch returns you to the dialog box that also allows you to enter an optional lower limit for % similarity with the structure previously entered. The example in Figure 23 searches for compounds containing the Thiophene group. Similarly as in quantitative search, the distribution map changes depending on the partial structure search parameters. After partial structure search, compounds can be selected within the distribution map and the structures displayed in the Compound List panel (Figure 24).



Structure	score	Dimers	EDP_HSPAN	CAI_HSPAN	Atomlets	Weight	logP	Rotatable Bonds
	0.9176	1	0.9171	0.1809	5	355.4548	4.7851	6
A181468+D1+AD17238	0.9025	1	0.9476	0.1404	5	430.3230	5.2497	6
	0.8896	1	0.8933	0.1140	3	155.4540	5.1811	6
A181468+D1+AD16582	0.8894	1	0.8850	0.1976	5	430.9500	5.0211	6
A181392+D1+AD16586	0.8872	1	0.9040	0.1321	6	434.5540	4.8831	5
A181392+D1+AD16582	0.8851	1	0.8811	0.1116	6	438.8500	4.8431	5

Figure 24. Resulting compound list after partial structure search

#### 4.4 Compound display and fragment display

The “Frame” panel located at the lower left hand area of the results analysis window displays the synthesis frame (Figure 25). At this area, clicking on any of the fragments then on the “Select entire frame” button toggles between compound list display and fragment list display in the Compound List panel.

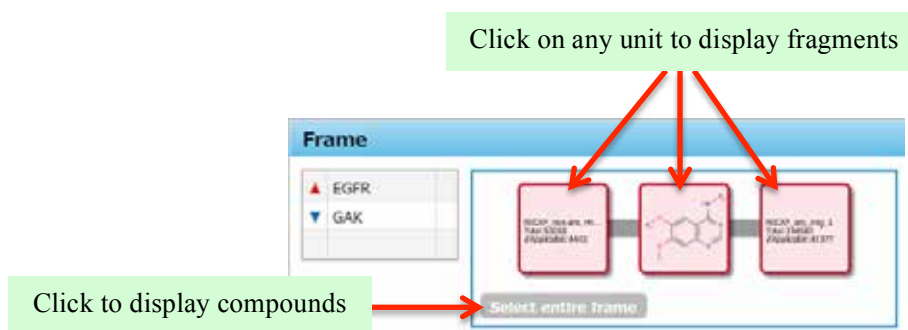


Figure 25. The Frame panel

Figure 26 displays the fragment structures with the corresponding parameters after selecting the rightmost fragment unit within the synthesis frame then selecting an area within the distribution map. In CzeekD, the binding point in a fragment structure is designated by an atom of a heavy element not generally used in the creation of drugs. Parameters displayed are values from which the effects of the heavy atoms were removed.

In the case of fragment display in the distribution map, the X-axis is initially set to “freq” that indicates the frequency in which the fragment is utilized in the generated compounds. The Y-axis is set to “mean” that indicates the average score of compound groups possessing that fragment. In addition, fragments that have available supplier data also show the name and ID of the supplier. That is true, in the case of building block library being offered by Namiki Shoji (Namiki\_BB Library).

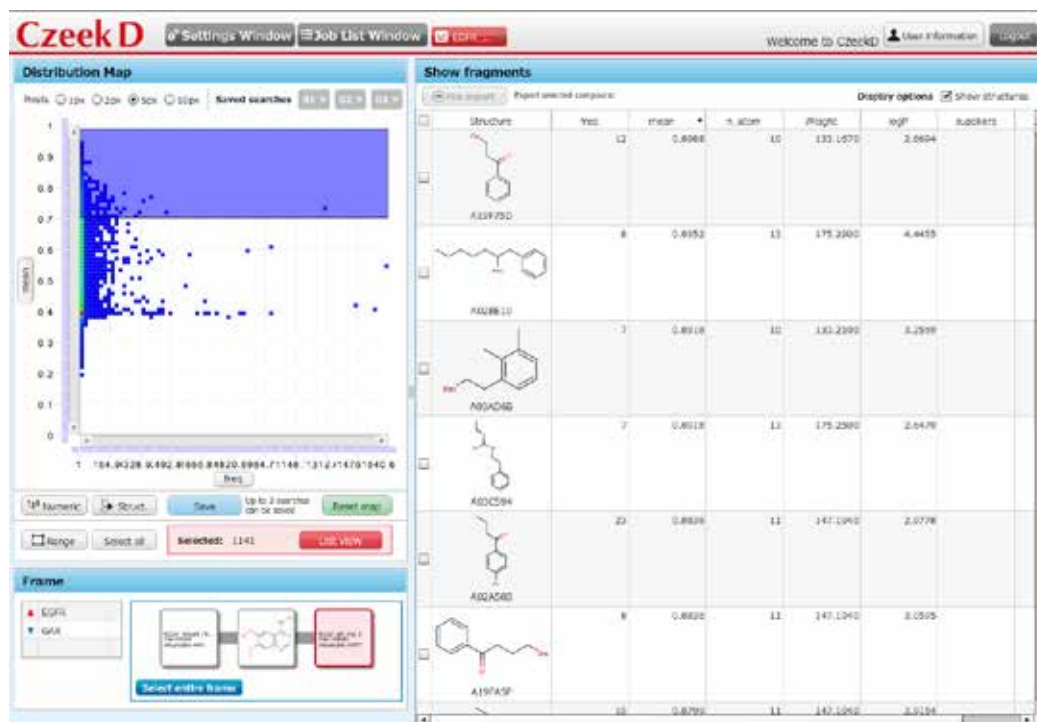


Figure 26. Displaying structures of selected fragments

## 4.5 Refining compound search

Normally, CzeekD generates several hundred thousands of compounds during calculation. While it depends on the selected target and synthesis frame settings, the number of high scoring compounds can reach up to tens of thousands. In our example calculation using EGFR and GAK alone, there are 21256 compounds scoring 0.8 or greater and checking that list for compounds worthy of further examination would be very difficult if not impossible. In cases like these, refining compound search is advisable and is a feature included in CzeekD.

### 4.5.1 Refining fragment-based compound search

This is done by first switching to fragment display mode. As shown in Figure 27, select an area in the distribution map and display the list at the right panel. Select one or more fragments from the list by clicking on the checkbox to the left of each structure. All the fragments in the list can be selected by clicking on the checkbox located at the left of “structure” column heading. Next, go to the Frame panel at the lower left hand side of the window then click on the “Select entire frame” button. The distribution will then change into something similar to that shown in Figure 28. In this example, the fragments having the highest average score (mean > 0.7) are selected, so the center of the distribution map would be composed of those fragments.



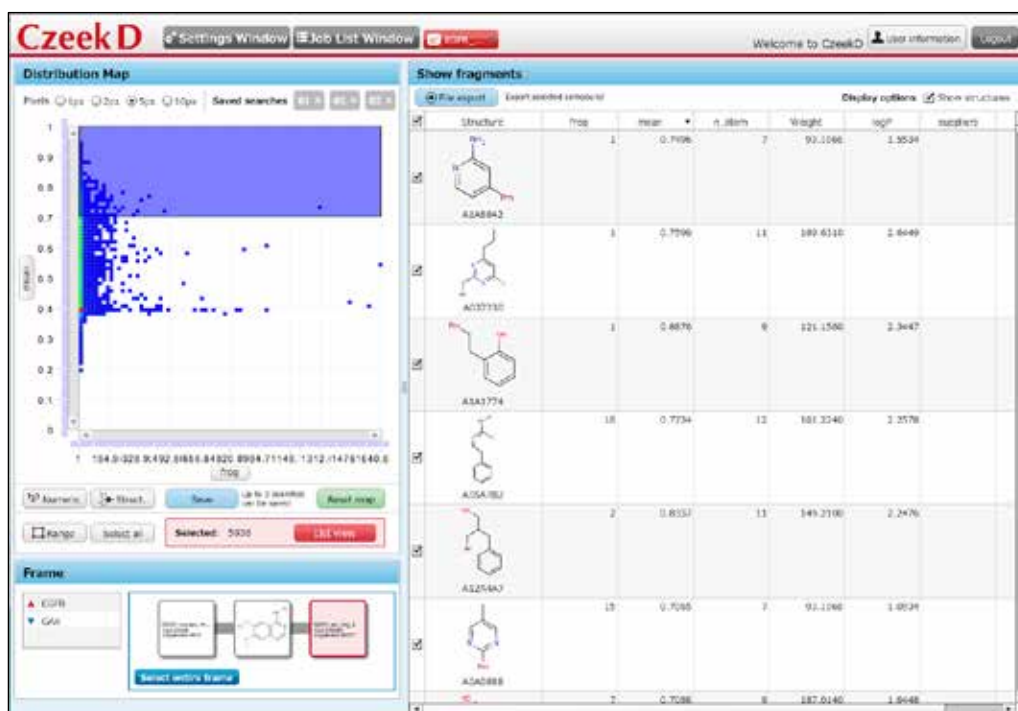


Figure 27. Selecting fragments

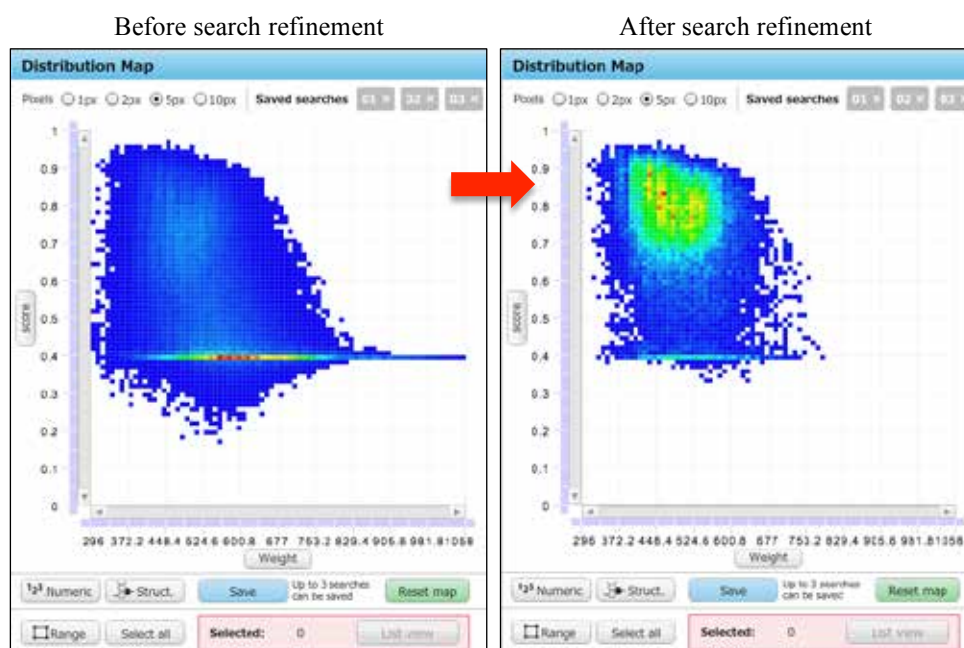


Figure 28. Change in distribution map after fragment-based search refinement

#### 4.5.2 Compound-based fragment search refinement

In compound display mode, select an area by clicking and dragging within the distribution map. Click on the “List view” button to show the compound structures then select one or more compounds by clicking on the checkboxes (Figure 29). On the Frame panel click on the left or right fragment units. The distribution map will change to show only the fragments present in the selected compounds (Figure 30).

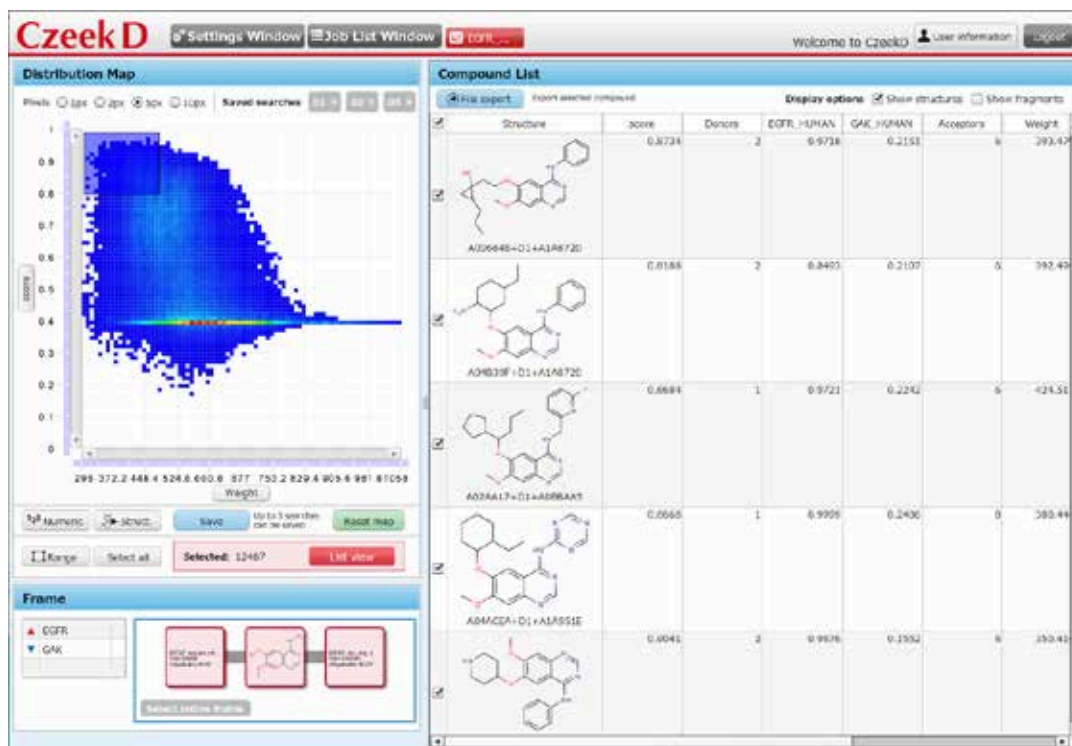


Figure 29. Selecting compounds



Figure 30. Change in distribution map after compound-based search refinement

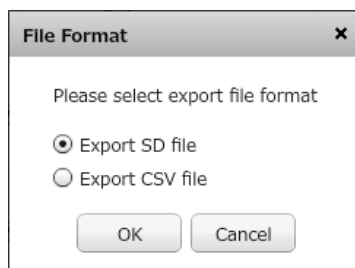
For both cases above, the results of the refined search can still be narrowed down further by setting different conditions and by clicking the checkboxes to select fragments or compounds.

Clicking the “Save” button located below the distribution map will save the conditions and results of the current search. Up to 3 searches can be saved and you can switch between them by clicking on the buttons “x1”, “x2”, or “x3” buttons located above the upper right hand corner of the distribution map. Saving more than 3 searches will

erase the least recent search. Clicking on the “Reset map” button will reset the distribution map to display the default and will erase all the saved searches.

#### 4.6 File export

Data of the selected compounds can be downloaded by clicking on the “File export” button and selecting the appropriate file format within the dialog box that appears (Figure 31). Data can be saved in either SD or CSV file format. Structures saved in CSV files are in SMILES format. Clicking on the OK button allow the user to specify a filename before saving the file.



**Figure 31. Selecting file format before saving data file**



## 5. Other Features

This section describes the other functions of the CzeekD system not described in the previous sections.

### 5.1 Modifying user information (changing password)

Registered name, e-mail address, and passwords can be changed by first clicking on the “User information” button [Figure 32(a)]. The display will switch to a new window showing the information for the current user [Figure 32(b)]. Information can be modified by clicking the “Edit account details” button. This will open a dialog box that allows the user to change the user name, e-mail address and password (Figure 33). Click on “OK” to close the window and effect the changes or “Cancel” to cancel current operation.

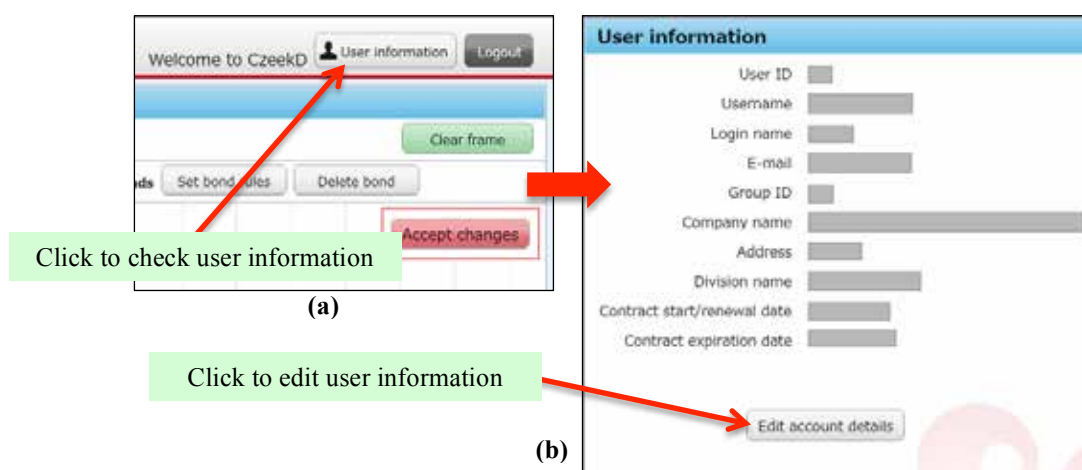


Figure 32. Displaying user information

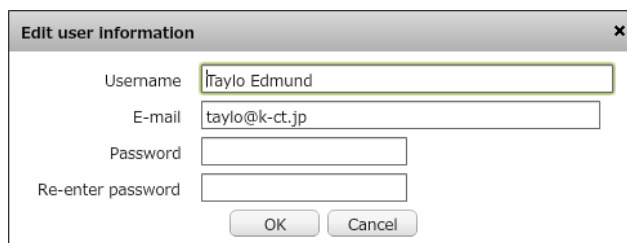


Figure 33. Editing user information

### 5.2 Compound scoring by CGBVS

With CzeekD, scoring of a predefined structure against one or more protein targets can be performed and the structure can be specified in 2 ways: 1) drawing or uploading the desired structure using the Marvin Sketch plugin or 2) by uploading SD or SMILES file within the fragment library selection window.

#### 5.2.1 Using the Marvin Sketch plugin

First, remove the default frame by clicking on the “Clear frame” button. From the selection of units, select the one without any bonding points then click on the frame area to place the unit. Select the unit by clicking on it then click on the “Draw structure” button to activate the Marvin Sketch plugin. Draw the desired structure then click “OK” to return to frame setting window. In this example, the structure of Gefitinib is drawn while EGFR and GAK were set as the target and off-target proteins, respectively. (Figure 34)

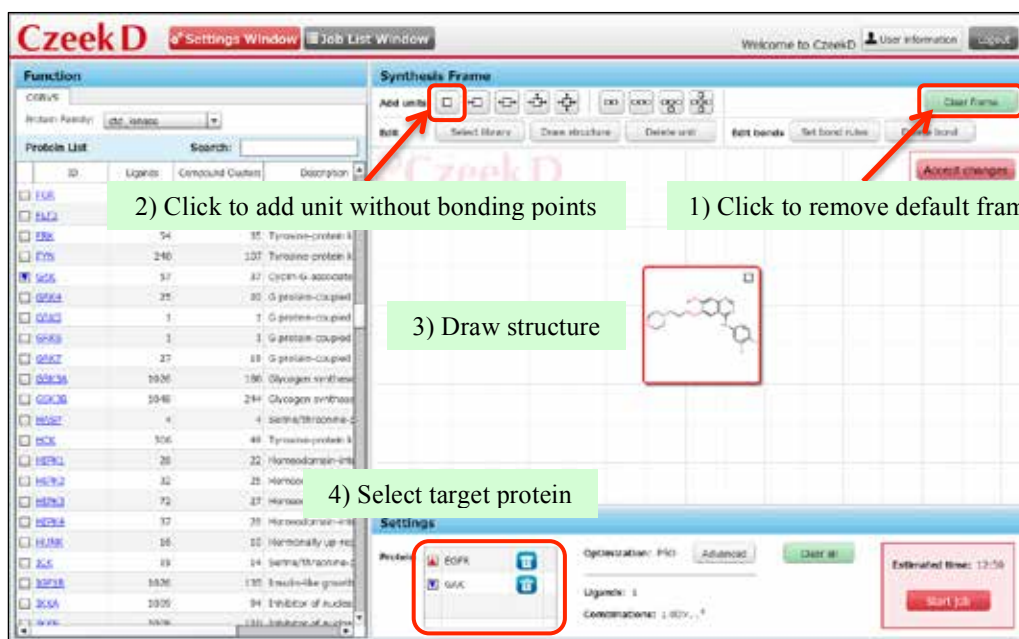


Figure 34. Calculation based on structure entered using the Marvin Sketch plugin

The result of the above calculation is shown in Figure 35.

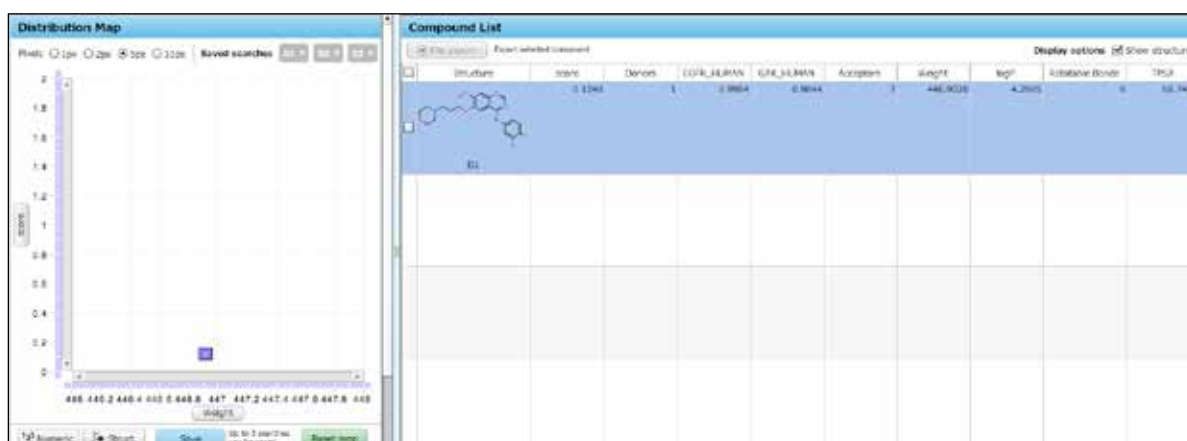


Figure 35. Scoring result for Gefitinib

## 5.2.2 Uploading structure files

If calculation of a set of compounds is desired, the compound structures must be put in a file (SD, SMILES, or mol format) and then uploaded into the system. Up to 100 compound structures can be uploaded. The first line of the descriptors for all compounds must start with the compound name and unexpected results will be obtained if that information is missing.

A unit without any bonding point must be placed first in the synthesis frame area. The unit is then selected before clicking on the “Select library” button. A window similar to that in Figure 36 will open. Click on the “Browse” button to select the file for uploading. The number of compound structures contained in the file will be shown at the lower left hand area of the window. Clicking the “Show Fragment List” button to show the structures at the right

side of the window. Click the OK button to set the compounds for calculation.

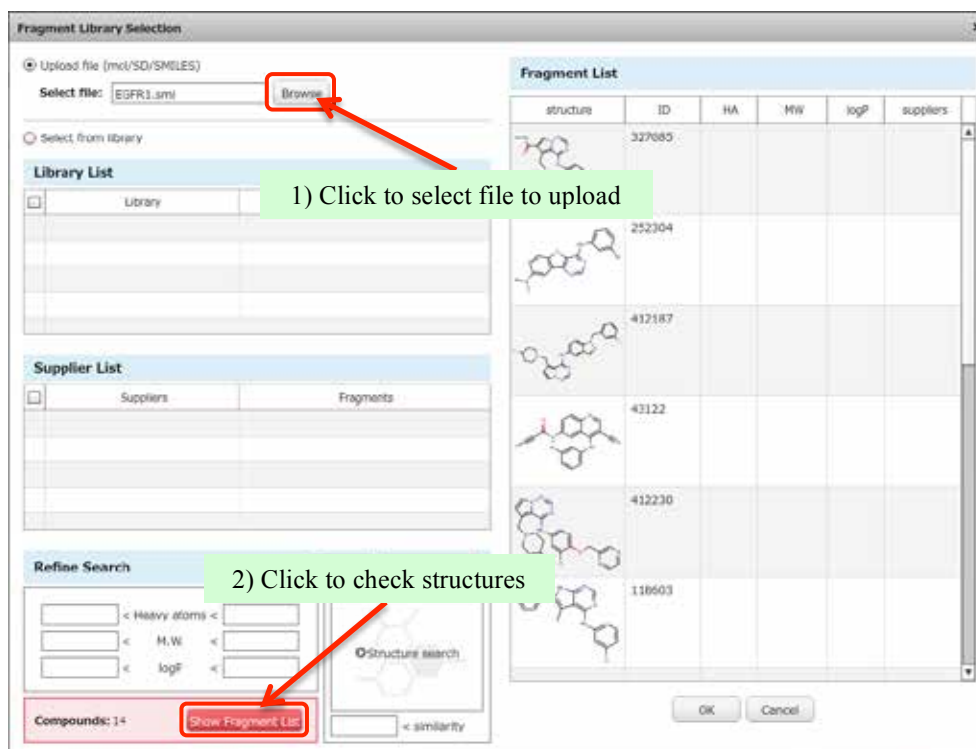


Figure 36. Uploading structure via the fragment library selection window

Results for the above calculation is shown in Figure 37 below.

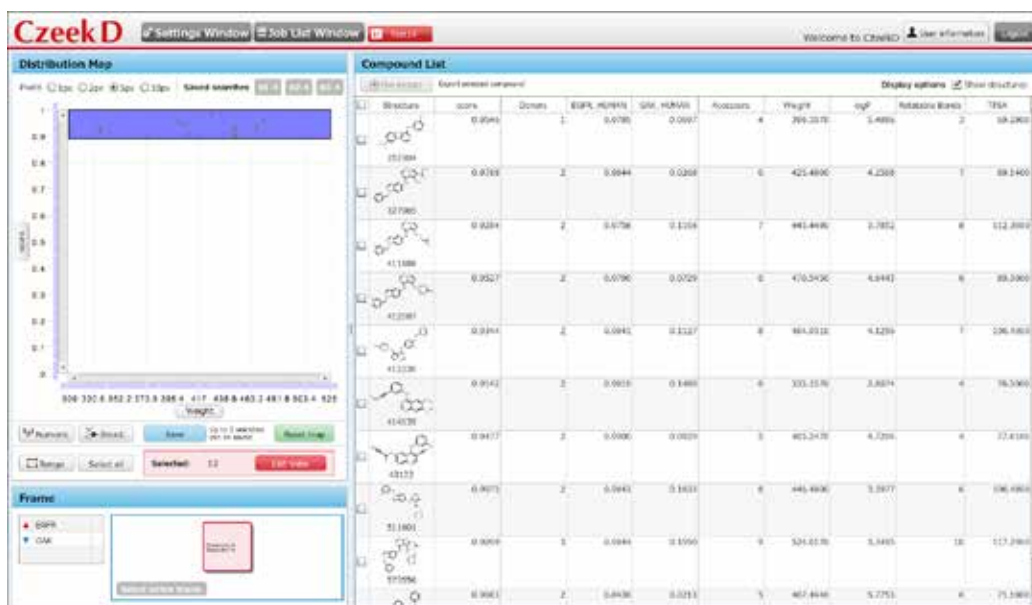


Figure 37. Calculation results for compound structures uploaded from file

### 5.3 Assigning several core structures

Several core structures can be assigned using the Marvin Sketch drawing tool. As shown in Figure 38, it can be done by drawing several structures within Marvin Sketch [Figure 38 (a)]. Structures not connected together by

bonds are recognized as independent structures [Figure 38(b)]. There are two important points to remember when using this method:

- ① number of R groups should be the same for all the structures
- ② R group symbols should be the same for all the structures

For the second condition above, if 2 R groups are present, the R1 and R2 symbols must be used for all the structures and not R1/R2 for one, then R3/R4 or R5/R6 and so on for the succeeding structures. Failure to follow the conditions above will lead to unexpected results. Please note the system does not give a warning if you commit such errors and will accept the invalid settings and continue on to calculation.

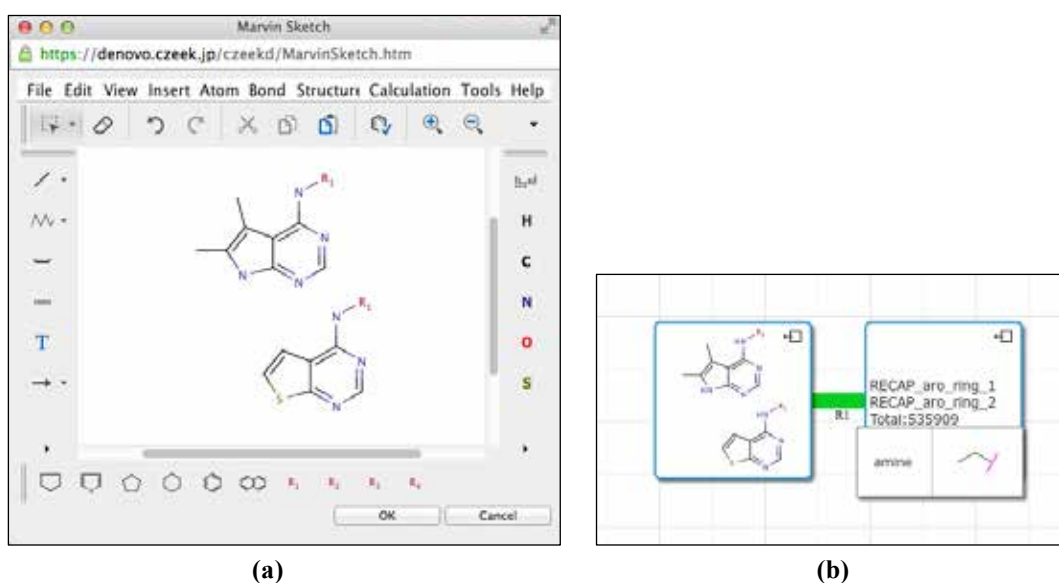


Figure 38. Drawing more than one structure within the Marvin Sketch plugin

#### 5.4 Sharing calculation results

Users belonging to the same group (defined during creation of accounts) are allowed to share their calculation results with each other. This can be easily done by clicking on the purple icon containing white callout symbols (Figure 39). This will make the gray button located at the leftmost column to turn to red color indicating that the results for that particularly calculation is currently being shared with everybody within the same group. Clicking on the same icon will un-share the data. Only the owner of the calculation results can share and un-share the data.

Unchecking the “Show shared data” checkbox will reset the job list to display only the currently logged-in user’s data. Checking the box again will show currently shared data by other group members.

Share	User	Job name	Status	Start time	End time	Running time	Expected completion time	Operation
	会井 千里	test kinase egfr	正常終了	2014/8/21 11:30	2014/8/21 19:15	07:44:57	2014/8/21 18:53	[Icons]
	会井 千里	test kinase egfr	正常終了	2014/8/22 10:42	2014/8/22 14:54	04:12:33	2014/8/22 14:27	[Icons]
●	Taylor Edmund	EGFR_GAK_Test1	Normal-end	2014/12/16 10:35	2014/12/16 18:13	07:38:11	2014/12/16 18:09	[Icons]
●	Taylor Edmund	Geftinib_EGFR-GAK_Test	Normal-end	2014/12/18 15:01	2014/12/18 15:02	00:01:01		[Icons]
●	Taylor Edmund	Test100	Normal-end	2014/12/22 16:58	2014/12/22 16:59	00:01:07		[Icons]
●	Taylor Edmund	Test114	Normal-end	2014/12/22 17:14	2014/12/22 17:15	00:00:57		[Icons]

2) Changes to red color indicating shared status

1) Click here to share data for this job

Figure 39. Sharing calculation data

### 5.5 Expanding the range of compounds that can be generated

The synthesis frame can be designed in such way that the number of virtual compounds (the chemical space) that can be generated will increase. As shown in Figure 40, adding a unit to the synthesis frame will expand the calculation at the R1 side. In this case, the chemical space increased from  $1.83 \times 10^8$  to  $1.96 \times 10^{14}$  and so is the variation of the chemical structures. However, increasing the number of units will also increase the time required to perform the calculation.

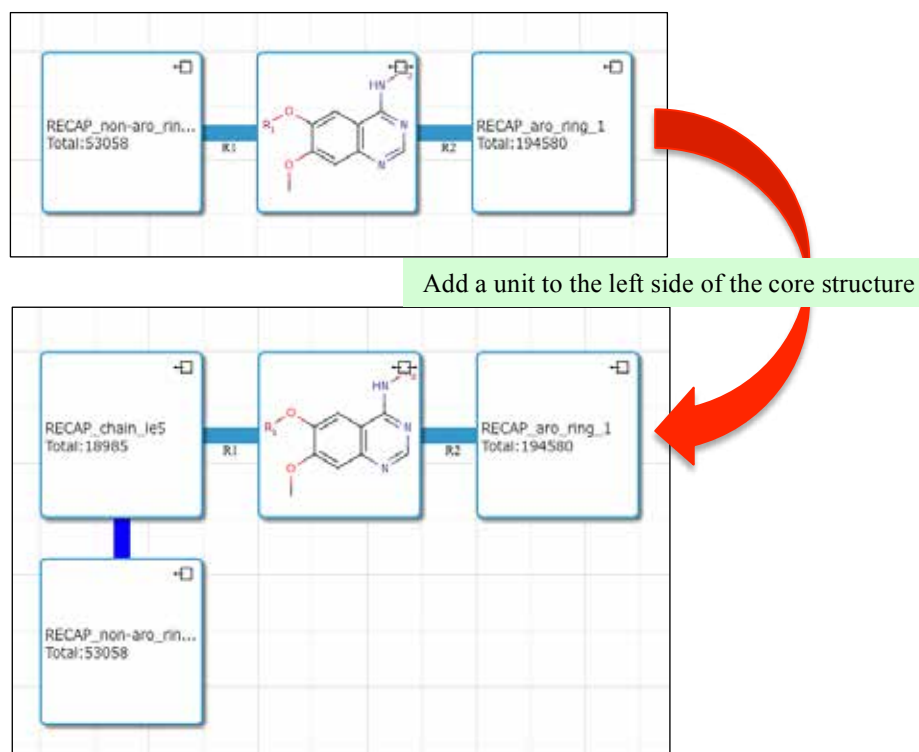


Figure 40. Extending the synthesis frame

### 5.6 Using custom-made fragments

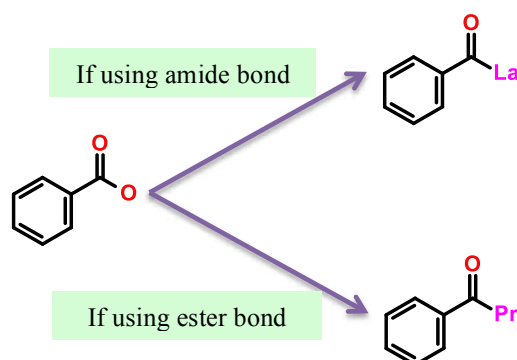
As described in section 5-2, structures can be uploaded by using the SD/MILES/mol file upload feature. The same method can be used to upload user designed custom fragments. However, this type of usage requires understanding of the RECAP rules.

As discussed in section 2, fragments provided in CzeekD are labeled with heavy metal element symbols that serve as markers for proper binding based on RECAP rules. These markers are generally removed during the binding process. The custom fragments must, therefore, contain these same markers before being uploaded into the CzeekD system. Additionally, the first line of information for every compound present in the SD file should contain the fragment name written with a maximum of 10 letters. This is required to prevent unexpected results.

As an example, benzoic acid is set as the fragment as shown in Figure 41. Since it is a carboxylic acid, based on RECAP rules, the bond used should be the amide bond. However, in CzeekD both the amide bond and the ester bond can be used so they are presented separately. In Figure 41, in case of amide bond the OH group is replaced with LA. On the other hand, if the ester bond is to be used the OH group is replaced with Pr. If both bonds are to be used during the calculation, both fragments with the appropriate labels should be included in the upload file.

Bond Type	Element 1	Element 2
amide	La	Ce
ester	Pr	Nd
amine	Am	Pm
urea	Cm	Sm
ether	Eu	Bk
olefin	Gd	Gd
quaternary nitrogen	Np	Pu
aromatic N carbon	Dy	Ac
lactam N carbon	U	Pa
aromatic carbon	Yb	Yb
sulphonamide	Lu	Th

**Table 3. Heavy metal marker pairs that allow binding**



**Figure 41. Selecting the heavy metal markers**

## 5.7 Advanced calculation parameters

Clicking on the “Advanced” button in the settings windows will open a popup window similar to that shown in

Figure 42. The upper left hand area of the window allows the selection of the optimization algorithm (PSO or GA) to be used in the calculation. CzeekD uses PSO by default, so in this case just keep the PSO radio button checked. The default number of iterations is set at 2000 but can be set higher if a more refined chemical space search is required. However, increasing the number of iterations will also increase the amount of time required for calculation. It is recommended not to drastically increase the number of iterations. If a much broader search range is required, performing the same calculation several times is a better option. You can however, make changes to the settings when the need arises.

Parameter	Value
Algorithm	PSO
Iterations	2000
Particles	128
PSO type	PCA
n_dimensions	5
n_sampling	30
n_interval	200
damping_coef	1
Vcoef	2.0
Vmax	-1
w	0.9
c1	1.5
c2	1.5
k1	1.0
k2	1.0
u	1.0
t	1.0
GA Iterations	500
GA Particles	1000
GA n_block	10
GA scale_factor	5
GA P_roulette	0.3
GA P_mutation	0.1
GA crossing_method	2point

Figure 42. Optimization algorithm settings