# BIOACTIVE CONFORMATIONAL ENSEMBLE SERVER AND DATABASE. A PUBLIC FRAMEWORK TO SPEED UP *IN SILICO* DRUG DISCOVERY

Sanja Zivanovic[1], Genís Bayarri[1], Francesco Colizzi[1], David Moreno[1], Josep Lluís Gelpí[2,3], Robert Soliva[4], Adam Hospital[1]* and Modesto Orozco[1,3]*

Modern high-throughput structure-based drug discovery algorithms consider ligand flexibility, but typically with low accuracy, which results in a loss of performance in the derived models. Here we present the Bioactive Conformational Ensemble (BCE) server and its associated database. The server creates conformational ensembles of drug-like ligands and stores them in the BCE database, where a variety of analyses are offered to the user. The workflow implemented in the BCE server combines enhanced sampling molecular dynamics with self-consistent reaction field quantum mechanics (SCRF/QM) calculations. The server automatizes all the steps to transform 1D or 2D representation of drugs into three dimensional molecules, which are then titrated, parametrized, hydrated and optimized before being subjected to Hamiltonian replica-exchange (HREX) molecular dynamics simulations. Ensembles are collected and subjected to a clustering procedure to derive representative conformers, which are then analyzed at the SCRF/QM level of theory. All structural data is organized in a noSQL database accessible through a graphical interface and in a programmatic manner through a REST API. The server allows the user to define a private workspace and offers a deposition protocol as well as input files for "in house" calculations in those cases where confidentiality is a must. The database and the associated server are available at https://mmb.irbbarcelona.org/BCE

[1] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology (BIST). Barcelona

[2] Barcelona Supercomputing Center.

[3] Departament de Bioquímica i Biomedicina. Facultat de Biologia. Universitat de Barcelona.

[4] Nostrum Biodiscovery, Nexus II Building. Barcelona

* Correspondence to Dr. Adam Hospital (adam.hospital@irbbarcelona.org) or Prof. M.Orozco: modesto.orozco@irbbarcelona.org

# INTRODUCTION

Computer-aided drug design (CADD) speeds-up the discovery and optimization of new active compounds through a large variety of methods, usually based on predicting whether a given molecule will bind to a biological target [1,2]. A good binding is the result of a subttle balance between different terms: a small desolvation contribution, favorable ligand-receptor interaction terms and a small cost for the ligand to adopt the "bioactive" contormation. The transfer from solution to the binding sites implies a reduction in the conformational space and the sampling of conformation that may not be the most stable one in solution, which means that drugs with high affinity are usually those whose preferred conformation in solution is close to that required to bind the receptor. Thus, to be efficient, CADD approaches, either ligand- or structure-based, must be accurate at predicting the penalty paid by the ligand to achieve its protein-bound state (a process known as conformer focussing), as errors in this term ruin the accuracy of any scoring function [3]. Conformer focussing can be defined as the free energy associated to the transition of the ligand from the free state in solution to the "bound conformation", i.e. the log of the probability of finding the bioactive coformation(s) within the unbound conformational ensemble of the ligand. In other words, determining the cost of reaching the bioactive conformation requires the knowledge of the conformational ensemble of the unbound ligand in solution, something that can be complex to define for flexible ligands [4].

Currently available methods for obtaining the conformational ensembles of unbound ligands can be classified as physical or knowledge-based [5,6]. The first family of methods implement systematic or stochastic searching schemes involving the use of force-fields in expensive calculations whose quality depends on the extent of sampling and on the accuracy of the force-field. On the other hand, knowledge-based methods rely on experimental data available in public databases, and can produce a large number of conformers with relatively small computational cost [7]. A variety of computer programs implementing physical or knowledge-based algorithms for exploration of the unbound state of the ligand are available, some of them as open-source codes, other as part of commercial modelling software [8]. These packages are fast and easy to use within automatic workflows, but often the results are not accurate due to the simplicity of the knowledge rules or the energy functionals/sampling algorithms implemented.

An alternative to standard ensemble generation methods rely on the use of multi-level (ML) approaches, where the conformational space is reduced to a small set of conformations whose relative stability is evaluated by means of accurate quantum mechanical (QM) calculations [9,10]. These methods can provide high quality estimates and are reasonably free of parametrization artefacts, but are slow, and have not been implemented in automated workflows, which severely hamper their use in the context of pharmaceutical research. Furthermore, as QM calculations are not typically integrated in standard workflows and automated pipelines, the results obtained by ML calculations remain disconnected from the mainstream of rational drug design projects. Finally, as

the derived information is not maintained in an interoperable format, its re-use by external users in processes of re-parametrization, description of interacting properties, machine learning or QSAR-type is challenging.

We present here a fully automated workflow for the exploration of the ligand conformational space. The procedure starts with a simple description of the drug (e.g.: chemical drawing, SMILES, Chembl code, PDB code) and performs all the steps to sample the conformational space with Hamiltonian Replica Exchange (HREX) [11,12] molecular dynamics simulations at the classical level. The conformational space is then reduced to a small number of representative conformations by clustering, with each cluster representative being then refined through QM/SCRF calculations [13]. All the process of generating conformers is automatized and packed in a webserver to facilitate its use by non-experts. The conformational ensembles at the classical and QM levels are stored in the bioactive conformational ensemble (BCE) database, where many descriptive analyses performed on the ligands are also available. The database is free and accesible through a graphical web-based interface which allows access to a combination of interactive 3D representations and 2D plots. The server allows the user to define a private workspace to launch simulations and the REST API offers programmatic access to all the information stored in the database. A deposition protocol is available to accept the submission of new molecules so that the database can become a repository for the community. Input files are provided on demand for those cases where calculations must be carried out locally at user's computer facilities.

The BCE database and the associated input and output servers are available at: https://mmb.irbbarcelona.org/BCE/.

## BCE GENERAL STRUCTURE

The BCE framework consists in a server from which the workflow necessary to generate representative conformers can be launched from a private workspace, a database where results can be stored and a graphical interface that allows the analysis of all the data (see Figure 1). Several links allow the connection of the BCE entries with external biological databases and associated tools. Programmatic access to the database is possible to facilitate high-throughput meta-analysis of the data. BCE can incorporate data from external users, which facilitates its growth and long-term sustainability and allows the definition of subsets of data, for example, drugs targeting a given protein, or families of drugs with a given pharmacological activity. The BCE database is conceived as an open initiative working with non-confidential information, which can be shared with the community. However, the BCE server offers a private workspace, and allows downloading of input files for "in house" calculations in cases where confidentiality is a must. Downloading input files may be required to use license-protected software, or the extension of simulations beyond the defaults. The next sections describe the central parts of the framework in more detail.

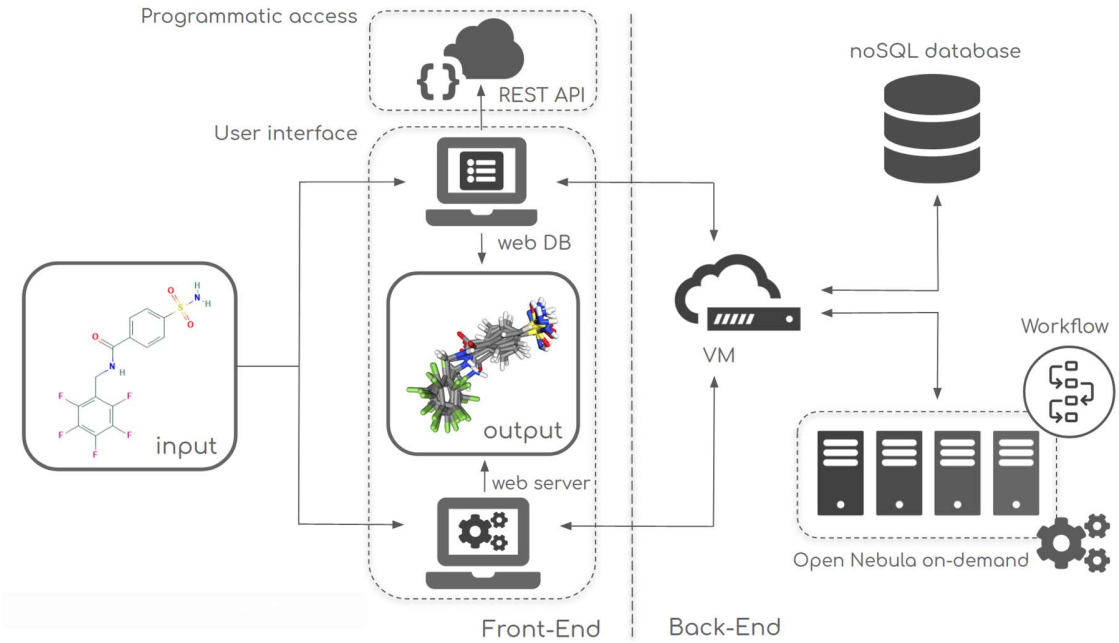



**Figure 1.** The bioactive conformational ensemble (BCE) framework. The front-end part (left-hand) shows the user interface containing the web-based graphical interfaces for the server and the database, together with the REST API programmatic access. The back-end part (right hand) contains the infrastructure behind the framework, consisting on a distributed noSQL database and a private cloud infrastructure with on-demand deployment, both connected to the front-end by means of a central Virtual Machine.

## THE BCE SERVER

The BCE server is built as a web interface, which allows the automatic launching of a workflow intended to generate reliable conformational ensembles for a particular compound. The workflow consists on a series of building blocks that generate topologies and initial structures of the molecule, define approximate force-field parameters, and derive and process classical ensembles to finally obtain representative conformations. Those conformations can be then submitted to QM calculations for further refinement (see general scheme in Figure 2 in this paper and a specific version of the workflow in a previous companion paper). Once calculations are finished, the server offers a variety of analyses on the data and the possibility to upload the results into the BCE database, where both raw data and associated analyses are accessible.

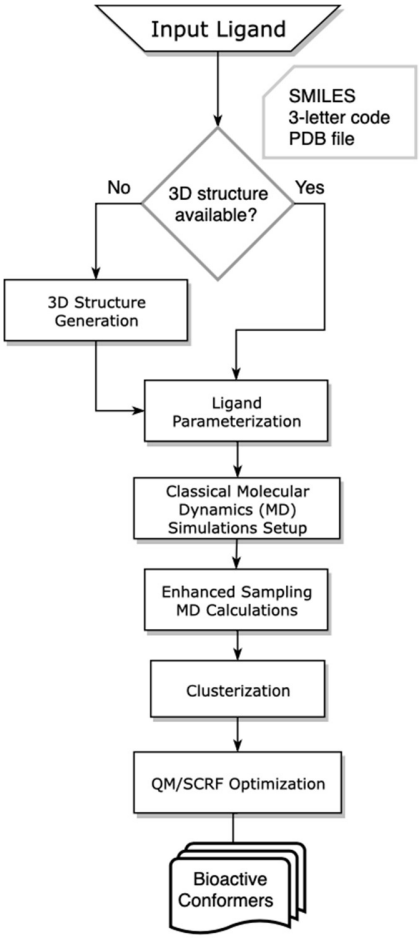



**Figure 2.** BCE pipeline schematic figure. From an input ligand, a 3D structure is generated (if needed), force-field parameters are determined, and classical ensembles computed with enhanced sampling calculations. A last step is refining the structures in a quantum level, to finally obtain accurate representative conformations

The server is defined in two main blocks: a section to create new projects and a private user workspace, where projects for a user can be explored and stored for a limited period of time.

***CREATE NEW PROJECT***

**1. Upload Data.** The pipeline outlined starts by first introducing the molecule of interest. This can be done using a variety of procedures:

- SMILES: The SMILES [14] signature can be used as input, and also as a query for searching for a similar structure (with a modifiable similarity index) or a substructure from the Chembl or PDB databases.

- Chemical Drawing: the compound can be drawn from scratch, or can be modified from a starting SMILES code, using the ChemAxon Marvin chemical editor [15] (Fig. 3A). Tanimoto´s [16] similarity search from the drawn molecule is performed to fetch similar compounds from the Chembl (for 2-D bioactive drug-like small molecules [17] ), or the PDB (for experimental 3D structures [18] ) databases.

- Browse: The server allows the user to search compounds from keywords contained in the molecule name (Chembl, PDB), or in the header/compound of a PDB structure.

- Upload PDB file: Drag and drop or choose a PDB file. Selection of the ligand to be studied (if the structure has more than one) is possible thanks to an integrated graphical viewer (Fig. 3B).

- PDB code, Ligand code: Start the project from a known PDB code and/or ligand 3-letter code. The list of ligands contained in the structure represented by the PDB code is automatically retrieved.

In all the cases, if more than one possible molecule is found (e.g. from a similarity search), a table containing the list of compounds is presented (Fig. 3D). Each of the entries from the table can be expanded to retrieve data from Chembl, PubChem [19] and PDB databases for the particular molecule, such as pharmacological information, 2D and 3D representation, and chemical descriptors.

**2. Protonation State.** Once the molecule has been chosen, all the possible (de)protonation states, are computed in the entire range of pH and graphically displayed (Fig. 3C). From all the different states generated, the major form of the molecule at pH 7.4 is highlighted. If the user uploads a known protonated input structure, this step can be skipped. After testing several tools we decided to determine protonation states using the Chemaxon tool, which also ranks all the possible tautomeric forms for a given protonation state [15]. The user can use the QM-tools coupled to the server to perform a more accurate tautomerism analysis.

**3. Settings.** Finally, the user can select a minimum set of options for the HREX simulations such as the number of replicas or the initial and final temperatures (scaling factors). Although the number of replicas and/or length of the simulations was set in our server after careful benchmark with experimental data (see companion paper ct-2020-00304y), some external user may want to extend or personalize our setup. For these cases, all input configuration files to run the simulations locally can be found and downloaded from the user workspace (see
next section). The user can also select some clustering options (defaults are provided). Due to licensing and computational restrictions we cannot offer QM calculations to external unregistered users, but we offer all the required files in Gaussian [20] format. As shown in a companion paper reasonable results can be obtained at the MST/DFT B3LYP/6-31G(d) level of theory (the default setting provided to the user), but the user is free to choose his/her own code. Once QM calculations

are done the user can integrate the data into the server to take advantage of analysis data in BCE and make calculations accessible to the community (see below).

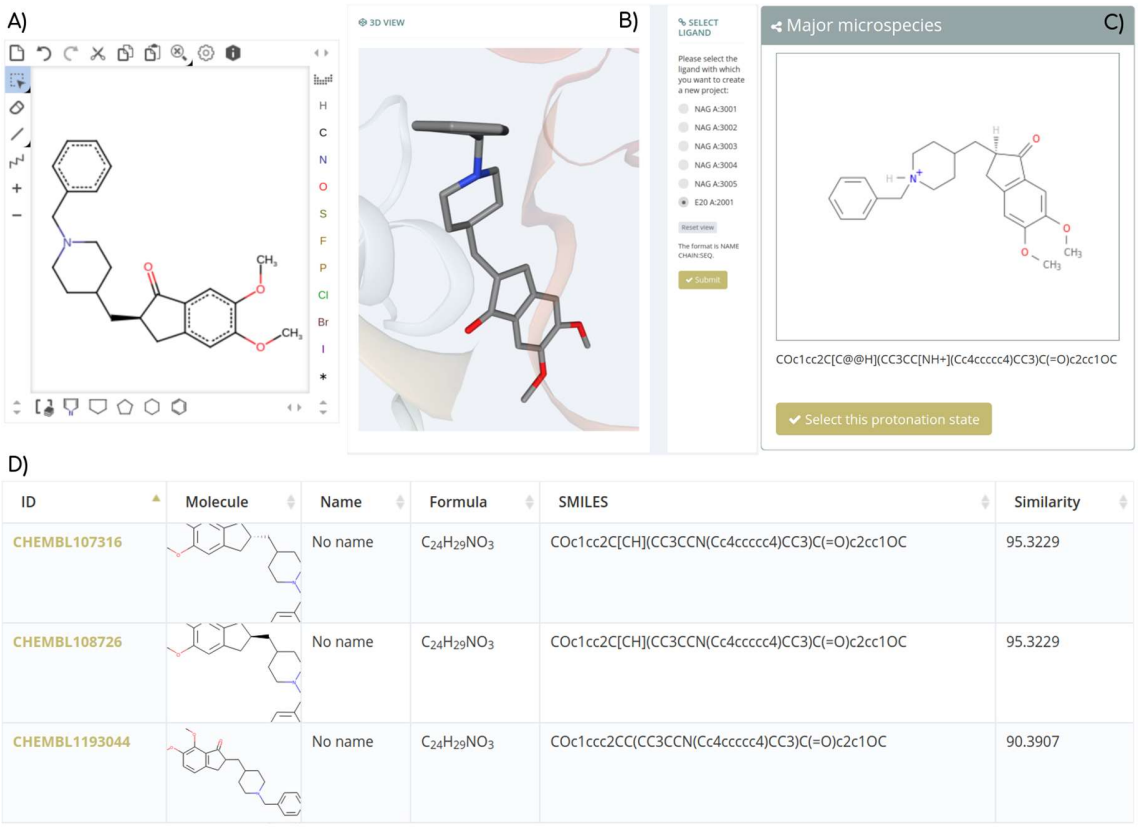



**Figure 3.** Screenshots showing examples of different web server interface sections. A) Input from a chemical drawing. ChemAxon Marvin JS is used as a chemical editor. An initial compound can be drawn from a SMILES code and modified afterwards. The compound drawn can be used as a substructure search to Chembl and PDB databases. B) Highlight of an intermediate page for a 3D-structure input. The interface automatically identifies the different ligands included in the structure, presents them to the user, and gives the option to choose one. NGL is used to interactively visualize the 3D structure. C) Possible protonation states computed by the server and offered to the user. D) Table presenting a list of results found from an initial search by substructure in the Chembl database.

***WORKSPACE***

The Bioactive Conformational Ensemble web server is built from a centralized personalized workspace (Fig. 4). Registration is not required to test the server, but results generated without registration will be kept in our disks for just 7 days. Completed projects are automatically transferred to a user workspace if a registration process is performed afterwards. Registered users have a

persistent workspace, with 2GB of disk quota, which can be extended upon request. A demonstration user is available to play with the workspace interface with pre-computed projects.

Every new project is appended to the user workspace in the files structure (Fig. 4A), mimicking a new folder from a file system environment. Files belonging to a certain project are revealed in a folder content (Fig. 4B). User jobs submitted, running or already finished are listed in the jobs section (Fig. 4C). Disk quota left is shown in the last section of the graphical workspace (Fig. 4D). Action buttons associated to every generated file offer a set of functions that varies depending on the file type. Common functionalities are “download” and “delete”. Specific functionalities include visualization of a 2D structure Kekule [21], a 3D structure NGL [22], or a plain text. Every project has an associated summary button, which opens a graphical summary of the project results, described below.

Finished projects can be directly uploaded to the BCE database to make them publicly available to the scientific community. In the case where QM-refined conformations are computed locally in the users’ facilities, the new ensemble can be uploaded and associated to a particular BCE project. In all cases, manual curation will be performed, to ensure correctness and completeness. This curation protocol and the instructions to submit new molecules can be found in the supplementary material and in the corresponding help section.

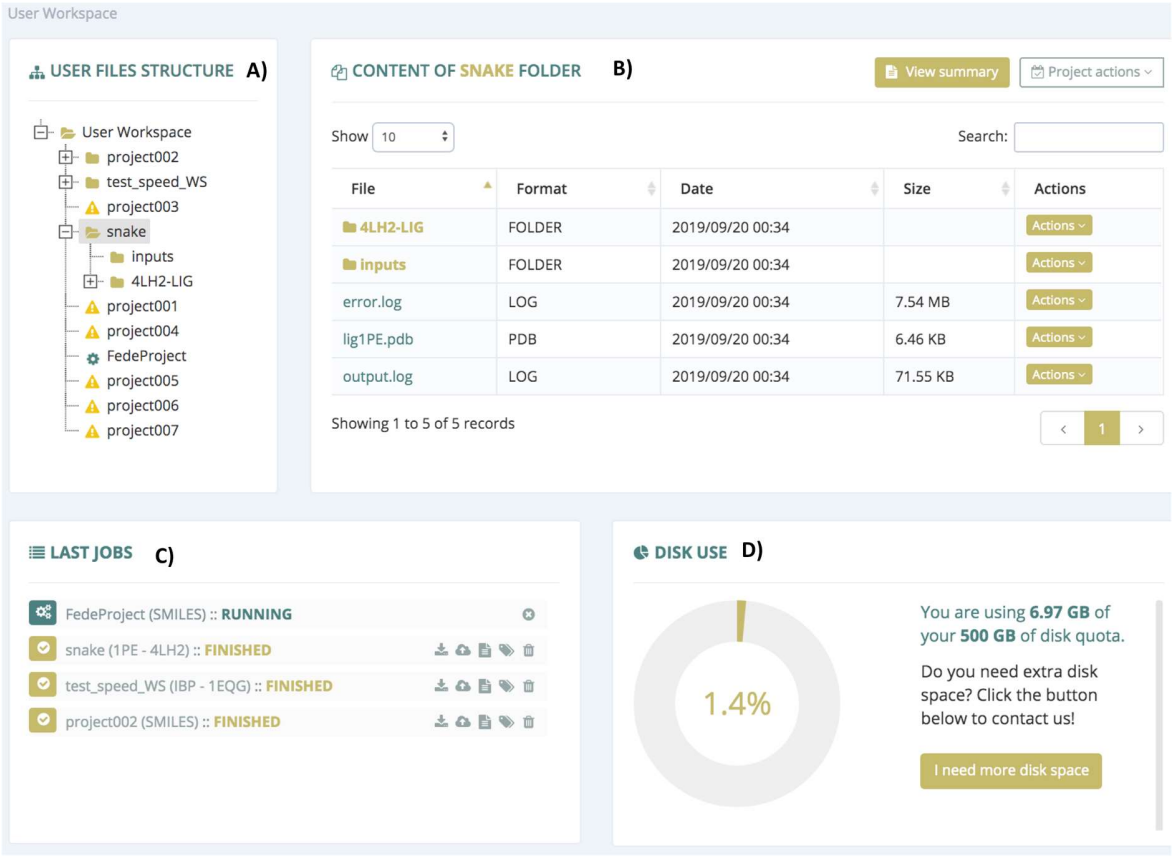


**Figure 4.** Screenshots showing examples of the web server private workspace. The files structure section (A) shows the list of projects for the particular user. All the files generated for a selected project are presented, in a file-system manner, in the folder content section (B). User jobs submitted, running or already finished are listed in the jobs section (C). Disk quota left is shown in the last section of the graphical workspace (D).

## BCE DATABASE STRUCTURE AND ASSOCIATED TOOLS

### *DATABASE CONTENT*

The BCE database contains for each molecule, a classical conformational ensemble in solution together with a set of cluster representatives derived from it, their QM-optimized geometries and energies, and several associated analyses. The "bioactive conformation" as derived from X-ray crystallography is displayed when available. For practical terms the database is divided internally in logical data sets that are saved as subsets. The entire database presented here can be then subdivided to tackle different subsets, which might have different input and output needs. For each compound representative conformations are stored and presented (data for symmetry equivalent conformers are grouped). We typically show ten diverse and highly populated conformers for which classical and QM data are displayed together with a set of analyses useful for quantitative structure–activity relationship (QSAR) and Structure-based drug design (SBDD) studies. At present time, the database contains data for around 1000 conformers of ~115 compounds and occupies around 1 Tb of disk. All trajectories were stored for further analysis using MD database recommendations [23].

### *ANALYSIS IN THE DATABASE*

Entries in the database are divided in six different sections or tabs (see Fig. 5A). General information, pharmacological properties and 2D/3D representations of the molecule are presented in the **Molecule** tab (Fig. 5B). Lipinski [24] and Veber's rules [25] are computed and shown in two independent tables. Compliance with the specific points of these rules is also integrated with the DB search engine (see search section). Links to entries from a list of pharmacologically relevant libraries for the particular compound are available, when applicable (PubChem [19], DrugBank [26], SureChembl [27], Chembl [17], ChEBI [28], Therapeutic Targets Database (TTD) [29], BindingDB [30], ZINC [31]). If the molecule studied has an associated experimental structure, a link to the RCSB PDB ligand interaction 3D view section is automatically added, to explore experimental binding mode.

The set of parameters (metadata) used in the biased MD simulation, together with a molecular representation of the total number of snapshots computed (typically 10,000) can be found in the

**Trajectory** tab (Fig. 5C). This section includes statistics of exchanges from the biased MD, and flexibility and shape adjustments through Root Mean Square deviations (RMSd), Radius of Gyration and Atomic Fluctuation analyses.

The **Principal Component Analysis** (PCA) section presents essential deformation movements [32] of the ligand as determined from the 10 first modes (those representing most of the conformational variance). Motions are represented using NGL, and projections of the snapshots to the first 3 principal components are used to easily identify the space explored by the trajectory snapshots, and the space covered by the final set of chosen conformers (Fig. 5E). Such a set is shown in the **Clusters** section. The number of different conformers is adjusted to represent 95% of the variability of the computed trajectory. Trajectory population and convergence of the chosen clusters are shown. Conformer's 3D structures are rendered in NGL in interactive plots where each individual conformer (and the experimental one if available) can be shown, hidden, or compared with the others. We have found that these plots are a fast an intuitive way to check convergence (see Suppl. Figure S1 for a side-be-side comparison of convergence as measured by looking at "reference" replica and by looking at the average population of the conformers at the different replica trajectories). Cross RMSd of cluster representative and RMSd with respect to the experimental structure (if available) are given. Relative population of cluster and their associated relative Gibbs free energies are presented in an associated table referred always to the most stable cluster. Several tests showed us that the statistical errors associated to cluster population are small and we decided then to do not include error bars in the graphs (see Suppl. Figure S2), the user can always get an estimate by fractioning the global trajectory. When the bioactive conformation is known, the strain and distortion free energies obtained from the cluster populations are shown (see companion paper ct-2020-00304y). A final subsection of the Clusters analysis represents Molecular Interaction Potential (MIPs; [33]) for the set of conformations. NGL is used again to represent the grids coming from three different MIP probes: positive, negative and neutral. The interface represents two different conformers at the same time, so that they can be easily compared.

The **Dihedrals** section allows users to explore the molecule's dihedrals flexibility (Fig. 6), which is useful to determine which torsions are contributing the most to define the molecular conformational space, something very useful for force-field refinement. By default we are using for the database generation simulations derived from the GAFF FF [34] , which has demonstrated a good performance in different tests [35–37], but in a companion paper we show how GAFF FF parameters can be improved from the QM data shown in the server. Obviously, the user is free to use other force-field in the simulations. By default, we used minimum salt condition (i.e. that required to achieve electroneutrality, which agree with the default SCRF methodology used at the QM level). However, the user is free to consider other salt conditions; in our experience for most small bioactive molecules, differences between minimum salt and physiological conditions are negligible. Using an interactive approach, users can click on a desired dihedral, or just select one from the table (categorized as flexible, flexible ring, rigid and rigid ring). A polar plot and an associated histogram are automatically generated. Dihedral values corresponding to the chosen conformers and experimental bioactive conformation (if available) are represented on top of the trajectory values.

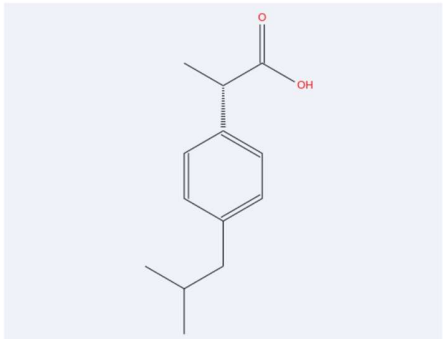
The final section, ***Quantum Mechanics*** (QM), presents the data for the different conformations optimized at the QM level. NGL is used again to show the 3D structures, and the associated table is now presenting (free) energies (internal energy and solvation free energy as determined from DFT/SCRF calculations; Fig. 5D) taking the most stable conformation as reference. As described elsewhere (see companion paper ct-2020-00304y), by default no salt effects are included in SCRF calculations that were carried out by default using our MST version of the PCM method [13]. If the bioactive compound (experimental) is available, QM tab shows an estimation of distortion and strain free energies (see companion paper ct-2020-00304y) to determine the cost required to adopt the bioactive conformation.

A)

Molecule | Trajectories | PCA | Clusters | Dihedrals | QM

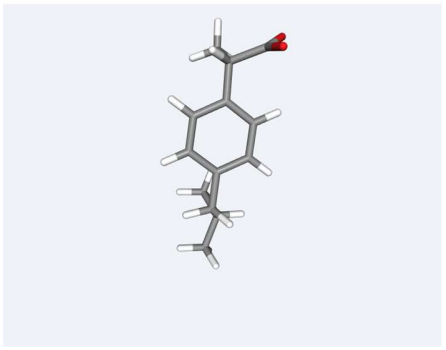
B)

2D VIEW




3D VIEW

Visualize Structure



Enable / Disable Hydrogens ON

Visualize Ligand Interaction

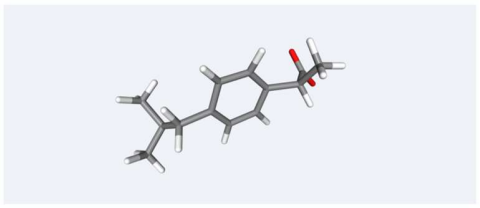
C)

METADATA

| Parameter | Value |
|---|---|
| Method | HamiltonianREMD |
| Number of MD replicas | 16 |
| Lenght of the MD simulations (ns) | 10 |
| Number of steps between exchanges | 100 |
| Force Field | General Amber Force Field (GAFF) |
| Charge scheme | Semi-Empirical (AM1) |
| REMD Progression | Geometric Progression |
| Initial (low) temperature (K) / Scaling factor | 298 / 1 |
| Final (high) temperature (K) / Scaling factor | 498 / 0.59 |
| Molecule number of atoms | 32 |

3D VIEW

Download PDB + Trajectory



666

D)

ΔG WITH RESPECT TO BIOACTIVE CONFORMATION

| B3LYP level | |
|---|---|
| $\Delta G_{dist}$ (kcal/mol) | 1.1387 |
| $\Delta G_{strain}$ (kcal/mol) | 0.0487 |

| M06 level | |
|---|---|
| $\Delta G_{dist}$ (kcal/mol) | 1.0641 |
| $\Delta G_{strain}$ (kcal/mol) | -0.052 |

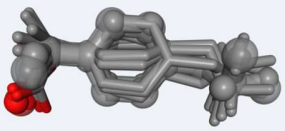
CONFORMATIONS TABLE

Show 5 Search:

| Cluster | RMSd | RMSdvsMM | $\Delta G_{B3LYP}$ | $\Delta G_{M06}$ |
|---|---|---|---|---|
| 3 | 1.4174 | 0.1287 | 0 | 0.0557 |
| 6 | 0.5174 | 0.2994 | 0.0038 | 0 |
| 5 | 0.4824 | 0.1843 | 0.0091 | 0.0341 |
| 4 | 1.2936 | 0.2095 | 0.0299 | 0.0831 |
| 2 | 1.1281 | 0.3193 | 0.0309 | 0.0574 |

3D VIEW

Actions



E)

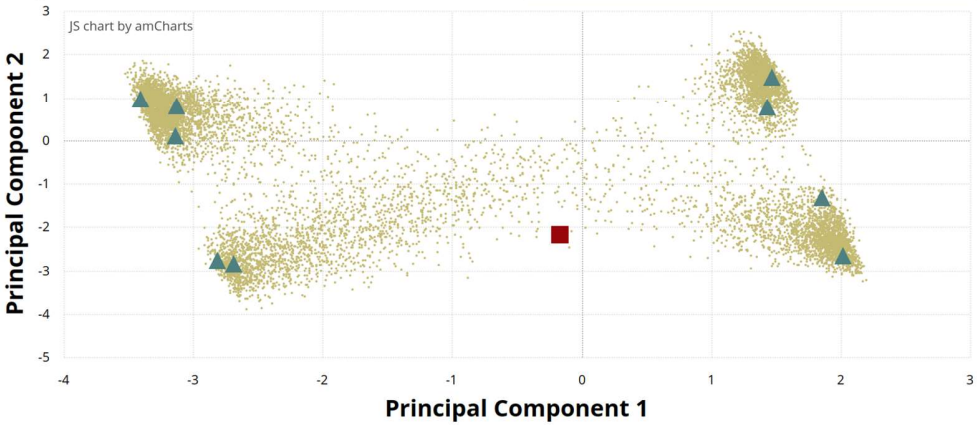


**Figure 5.** Screenshots showing examples of the different BCE database web interface sections (Compound BAC0010630, PDB id 2wd9-IBP). A) Section selection tabs. B) Molecule section: 2D and 3D representation of the molecule. C) Trajectory section: metadata information (top) and 3D trajectory structure representation (down). D) QM section: energies for all conformers chosen after QM optimization (middle table), with their structural representation (down). Distortion and strain free energies between the bioactive conformation (PDB) and the conformer with best relative energy are shown in a green colored box (top), when available. Note that the boxes change their color to red if no conformer with energy better than the experimental bioactive conformation has been found. E) PCA section: Snapshots projections to the first and second principal components. Brown dots correspond to the 10,000 trajectory snapshots. Green triangles represent the cluster conformations chosen from the trajectory, representing most of the conformational landscape. Red square belongs to the experimental bioactive conformation and is represented when available for the sake of comparison.

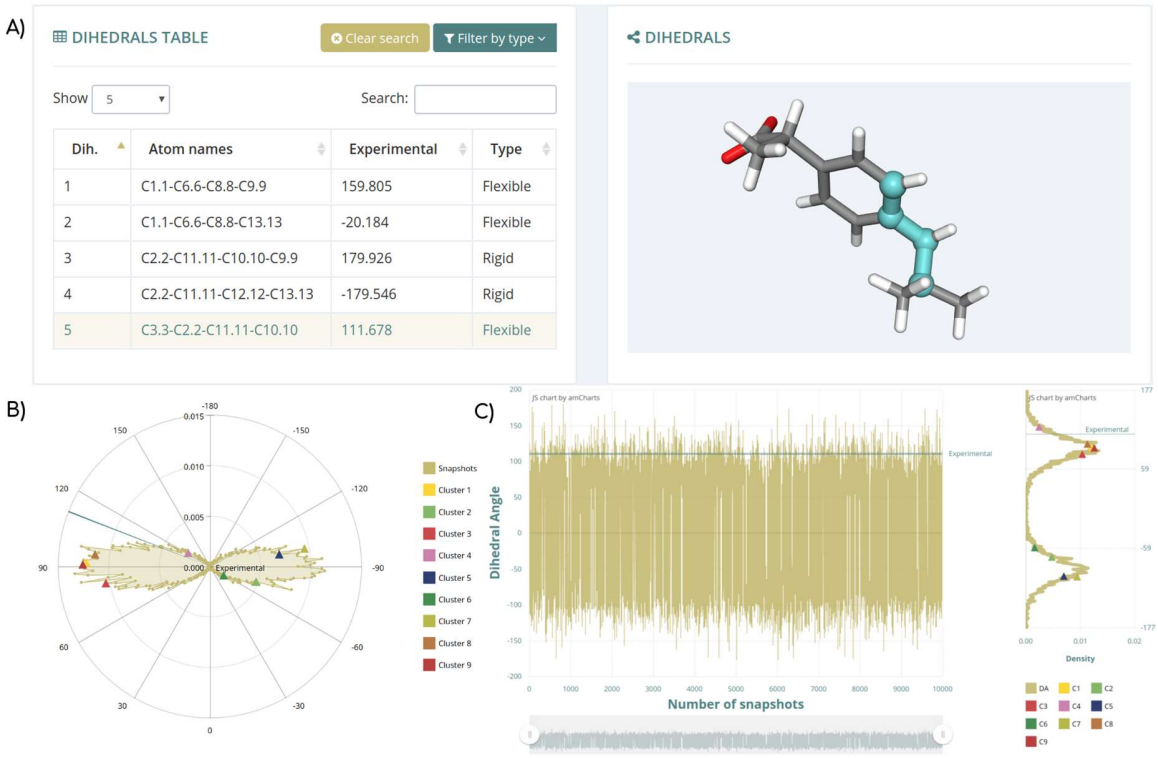



**Figure 6.** Detail of the Dihedrals analysis section. A) List of dihedrals, classified in five categories: Flexible, Rigid, Rigid-Ring and Flexible-Ring. When hovering the mouse over the table, the particular dihedral is highlighted in the structure, and the other way around. Clicking on one of the dihedrals opens two different plots: B) Polar plot representing the polar region covered by the particular torsion; and C) Line plot representing the dihedral values for the particular torsion along the whole

trajectory (10,000 snapshots in this case), with its associated histogram. Experimental (if available) and conformations torsion values are appended to the plots, identifying the regions sampled.

***SEARCHING & BROWSING***

Following the FAIR data principles [38], a unique and persistent identifier is assigned to each of the entries in the database (e.g. BAC0010630), allowing a direct link to a particular compound. The web interface allows searching the database by molecular similarity or from a set of molecular properties. In the search by similarity section, a substructure can be introduced either from a SMILES code or from a drawn 2D representation using ChemAxon Marvin chemical editor. The engine is then searching the substructure within the compound database using Tanimoto´s similarity search. In the search by parameters section, 3 different subsections allow the user to search by molecule description (e.g. name, id), molecule properties (e.g. number of heavy atoms, hydrogen bond donors/acceptors) and simulation results. The possibility to filter those molecules fulfilling Lipinski's [24] or Veber's [25] rules is also integrated, which facilitates focusing the analysis in drug-like molecules. Direct access to a browsing table listing all compounds included in each of the database's subsets is also available. Search results are presented using the same table interface, which also allows the user to search for important keywords such as BCE id, name, subset name or fulfillment of drug-like rules.

***UPLOAD OF NEW COMPOUNDS***

The incorporation of new compounds in the database is possible in two different ways: a direct file upload to BCE database web interface, or a BCE server job submission. For the direct upload, metadata should be provided in a deposition protocol. Manual curation of every entry will be done, to ensure correctness and completeness. This curation protocol and the instructions to submit new molecules to the database can be found in the supplementary material. Jobs run from the web server, when available, could be eventually made public or kept on-hold depending on project's requirements. New compounds either uploaded from the database interface or computed from the server will be added to the Users' Uploads subset, with the possibility to generate new subsets connected to a project if applicable.

***DOWNLOAD AND PROGRAMMATIC ACCESS***

The simulated trajectory, cluster structures and optimized conformations can be downloaded from the web interface, as well as be accessed programmatically via a REST API, which facilitates high throughput analysis and meta-analysis. Services are grouped following the same sections as the web server interface, with an additional division for analysis and files. The complete list of endpoints, together with examples on how to use them, can be found in Suppl. Table S1. An extensive documentation, with examples and the possibility to modify input parameters and run/test the services interactively has been prepared and is available from: https://mmb.irbbarcelona.org/BCE/api

## IMPLEMENTATION

The BCE web server has been implemented using the Slim PHP micro-framework (https://www.slimframework.com/) following a Model-view-controller (MVC) architectural pattern [39]. NGL [22] is used to visualize 3D structures and trajectories, Kekule js [21] is used to visualize 2D structures, and AmCharts javascript package (https://www.amcharts.com/), with modern data interactive visualization, was chosen to display all the analysis plots. Chemaxon Marvin is used for drawing, displaying and characterizing chemical structures, (Marvin 17.5.0, 2018, ChemAxon, http://www.chemaxon.com). The BCE API is documented using the Swagger™ (http://swagger.io) API framework.

User's workspaces and database entries information are stored using a noSQL MongoDB (https://www.mongodb.org). This ensures the flexibility in data representation needed to efficiently store and access the variety of heterogeneous data generated by the project, from the trajectory metadata and analyses to the set of generated files and their folder hierarchy (MongoDB's GridFS). Complete replicas of the Protein Data Bank [18], Uniprot [40] and Chembl [17] are stored in the same MongoDB infrastructure for efficiency purposes. The server is linked to these DB mirrors, with additional queries to the original DBs through REST API calls within error control flow. Values shown in the molecule section of the analysis division were automatically fetched from PDB [18] and PubChem (PUG-REST) REST [41] APIs at the time of upload, and stored in the internal database for the sake of completeness and efficiency. Persistence of the workspace is guaranteed with user registration. Single Sign On (SSO) access control is implemented, allowing the possibility to login using a Google, LinkedIn or ELIXIR account.

Input molecules are converted to 3D structures (when needed) using OpenBabel chemistry toolbox [42]. Protonation states are determined using ChemAxon Calculator module [43]. Details on the pipeline implemented in the BCE server to generate the conformational ensemble and the software used are presented in a companion paper. Briefly, all MD simulations are run using GROMACS [44] patched with PLUMED [45,46] and the HREX implementation [47]. Principal Component Analysis is computed using pcasuite package [48]. Molecular Interaction Potential grids are computed using CMIP program [33]. RMSd, radius of gyration, atomic fluctuation, conformational clusters and dihedral angles are computed using the analysis tools integrated in the Gromacs MD package [44]. Symmetry-corrected RMSds are computed using cpptraj tool [49] from the Ambertool MD analysis package [50].

Jobs are queued using Open Grid Scheduler (OGS) manager (http://gridscheduler.sourceforge.net), and served in an on-demand processing model performed by Virtual Machines automatically deployed in an Open Nebula [51] OneFlow cloud environment for multi-tiered applications.

## SOME EXAMPLE APPLICATIONS

In this section we provide a few examples of the power of the database and associated server. The low throughput methodology used to generate the BCE database yields ensembles capturing details that are out of reach for the traditional, quick conformational generation methods. Exhaustive sampling coupled to quantum mechanical calculations allows spotting errors in the assignment of conformations in X-ray crystal structures as well as reproducing important phenomena considered by medicinal chemists while optimizing their chemical series such as intramolecular repulsion, sigma hole effects [52], unconventional geometries for saturated rings or the effect of a "magic methyl" [53]. Here are a few examples of the insight derivable from the BCE database.

**Example 1. Quality check.**

A search for compounds with a $\Delta\Delta G > 5$ kcal/mol between the lowest energy conformer as predicted by the HREX/QM calculations and the experimental receptor-bound conformation as found in the PDB yields a total of 8 instances. As expected, most of them are very flexible ligands, but BAC0010568 (Donepezil, 1eve-E20), with only 6 rotatable bonds (excluding the piperidine ring) is also in the list, as simulations suggest that the most stable conformation (carbon atoms in pink in Figure 7) is at about 5 kcal/mol below the bioactive conformation (carbon atoms in cyan Figure 7). Analysis of the X-ray structure of the acetylcholinesterase (AchE)-E20 complex shows that Donepezil is completely embedded into the gorge of AchE [54], which might justify the large conformational penalty [55]. When rigidly docked [56] the global minimum conformation nicely fits into the AchE gorge also involving a direct cation-pi interaction with AchE Phe330 via its piperidine nitrogen (Figure 7A, green dotted line). More interestingly, the BCE-generated most stable conformation is also compatible with the experimental electron density of Donepezil (Figure 7B). Thus, the insight from using the BCE ensemble is twofold: i) quantify the energetic penalty paid by Donepezil to achieve its protein-bound state; ii) suggest that the AchE-E20 complex could include an E20 conformation different from the one found in the PDB.

**Example 2. Sigma hole effect**

The role of non-covalent sulfur interactions in protein-ligand recognition is often explained through the "σ-hole" interaction [52], which is generated by the interaction of positive electrostatic potential around sulfur atoms in heterocycles such as thiophene or thiazole with the lone pair of hydrogen bond acceptors (e.g. nitrogen and oxygen atoms). It is expected that "σ-hole" interactions might have also a role in determining ligand conformation, but in the absence of specific corrections to the force-field, this is not captured in classical simulations. We found a nice example of "σ-hole" mediated conformation in the system 3tki-S25 (BAC0010644). The S25 ligand binds Chk1 kinase in a bioactive conformation where the thiazole sulfur atom is engaged in an intra "σ-hole" interaction with the pyrimidine nitrogen atom (Fig. 7C). The application of our pipeline to this ligand yields an ensemble where the lowest energy conformer (Fig. 7D) indeed has an intramolecular sigma hole 1,5 N···S interaction. Interestingly, the conformers without this sigma hole interaction are less favored energetically, with at least a penalty of $\Delta G_{strain}$=3.22 kcal/mol (Fig. 7E). Thus, an accurate potency estimation for this congeneric series is beyond reach when ignoring this intramolecular interaction.

**Example 3. Magic methyl**

Medicinal chemists name "magic methyl" effect to the dramatic increase in potency caused by the introduction of a methyl substituent at specific positions. It has been shown [53] that such an increase is not only due to hydrophobic contacts, but also to conformational biases towards the bioactive conformations. One example we found in our database was pdb complex 3d7z-GK5 (BAC0010631). The GK5 inhibitor binds p38MAP kinase in a bioactive conformation where the adjacent phenyl rings must be placed at ca. 90 degrees for optimized potency (Fig. 7F). GK5 molecule and its close de-methylated counterpart (BAC0010632) were run through the pipeline and their ensembles studied in detail. Our results show (Fig. 7G) that the second most stable QM conformer for the pdb ligand is equivalent to the bioactive conformation, while reaching the bioactive conformation implies a large penalty ($\Delta G_{strain}$= 2.75 kcal/mol) for the demethylated compound where the aromatic rings tend to be coplanar (Fig. 7H).

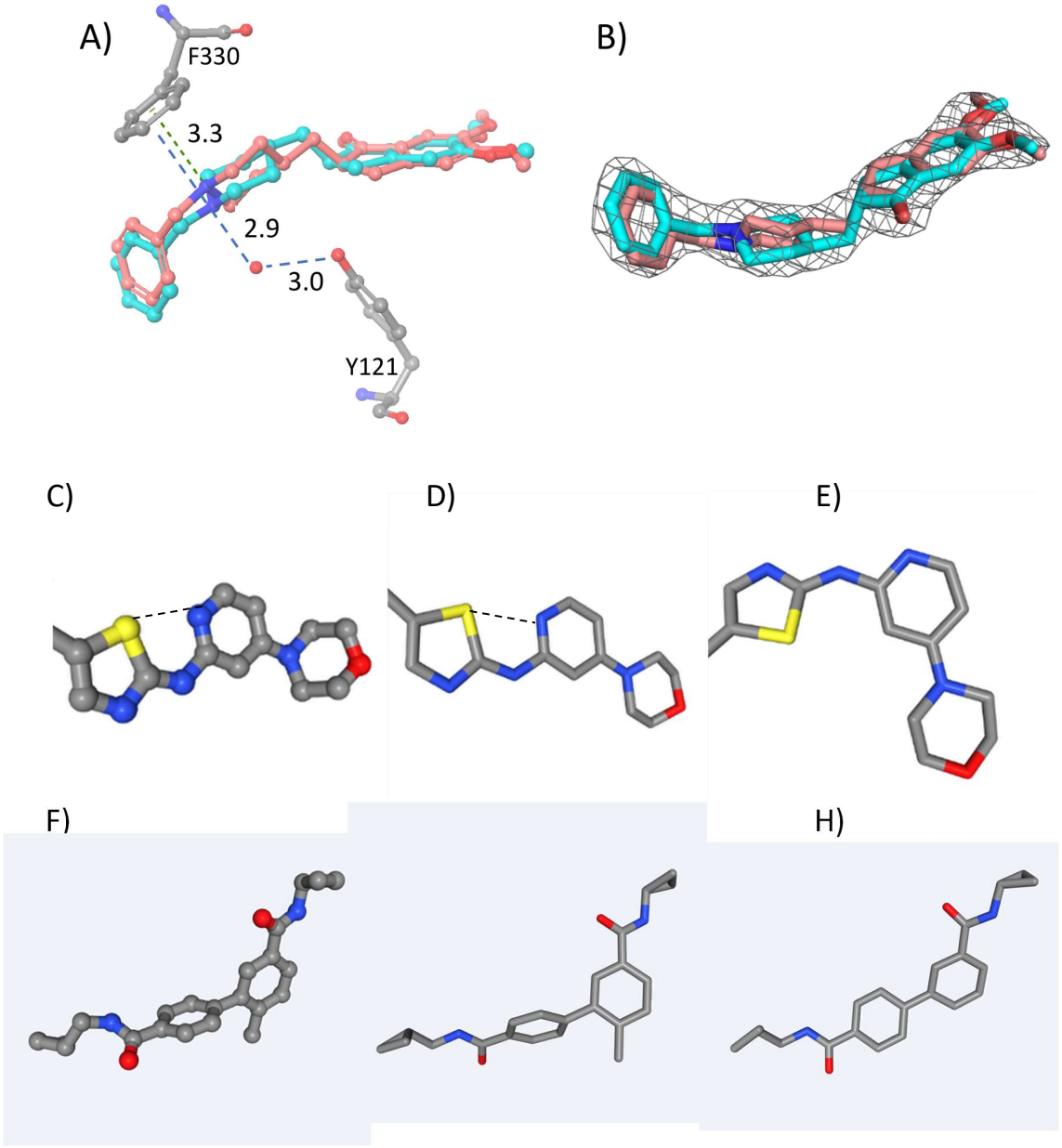



**Figure 7. (A, B)** Application of BCE server to study the complex between Acetylcholinesterase and Donepezil (AchE-E20). A) Lowest QM energy conformer (pink color; $\Delta G_{strain}$=0) of Donepezil rigidly docked into AchE and the X-ray bioactive structure (in cyan; $\Delta G_{strain}$=5.21 kcal/mol) with proximal AchE residues engaging with piperidine Nitrogen; dotted green and blue lines highlight interactions between the piperidine Nitrogen in the docked most-stable conformer and X-ray binding mode, respectively; distances are shown in Angstroms. B) Superposition of the BCE-generated most stable

conformer (pink color) and the crystallographic conformation of Donepezil (in cyan); the experimental $2f_o$-$f_c$ electron density map contoured at 2.0 sigma is shown as grey mesh. **(C, D, E)** Example of sigma hole effect in the 3tki-S25 system (BAC0010644). C) Close up on the crystallographic protein-bound conformation of ligand S25; "σ-hole" interaction highlighted by dotted line. D) Lowest energy conformer as found in the BCE ensemble, with the "σ-hole" interaction also present (highlighted by dotted line); E) A conformer where no "σ-hole" interaction takes place, leading to a penalty of 3.22 kcal/mol. **(F, G, H)** Example of magic methyl present in 3d7z-GK5 (BAC0010631) and absent in 3d7z-GKN (BAC0010632). F) Experimental bioactive conformation for BAC0010631; G) Second lowest energy conformer for the same molecule. The phenyl rings are almost perpendicular to each other, similarly to the orientation observed experimentally for the protein-bound conformer; H) Second lowest energy conformer when methyl is absent showing the aromatic rings tending to coplanarity.

## CONCLUSIONS AND FUTURE DIRECTIONS

We present a new structure-based drug discovery framework to generate accurate bioactive conformational ensembles of small molecules obtained by a theoretical multilevel strategy combining MM and QM. A web-based private workspace allows running the pipeline on a desired ligand, whereas the associated database and API allows the generated conformers to be publicly available in a FAIR-compliant way. The number of conformers offered by the BCE database will grow with new subsets currently being computed in our group (e.g. Central Nervous System-related compounds analyzed in the context of the Human Brain Project). Besides, upload of new ensembles is possible for the community to extend the database. The BCE framework is designed to become a repository for the computer-aided drug design community to find and deposit accurate structural conformers, speeding up *in silico* drug discovery.

**ACKNOWLEDGMENTS**

We are indebted to the support of the Spanish Ministry of Science [RTI2018-096704-B-100], the Catalan SGR, the Instituto Nacional de Bioinformática; European Union's Horizon 2020 research and innovation program [BioExcel-2 project], Biomolecular and Bioinformatics Resources Platform (ISCIII PT 13/0001/0030) co-funded by the Fondo Europeo de Desarrollo Regional (FEDER) (all awarded to M.O.) as well as the MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona) and the CDTI (Neotec grant−EXP 00094141/SNEO-20161127) (awarded to Nostrum Biodiscovery). NDB is sponsored by the Fundación Botín (Mind the Gap Program). We are indebted to the support of BSC. MO is an ICREA Research Professor. FC has received funding from the Horizon 2020 research and innovation programme of the European Union under the Marie Skłodowska-Curie grant agreement No. 752415.

## REFERENCES


(1) Ooms, F. Molecular Modeling and Computer Aided Drug Design. Examples of Their Applications in Medicinal Chemistry. *Curr. Med. Chem.* **2000**, *7* (2), 141–158.

(2) Jorge Moura Barbosa, A.; Del Rio, A. Freely Accessible Databases of Commercial Compounds for High-Throughput Virtual Screenings. *Curr. Top. Med. Chem.* **2012**, *12* (8), 866–877.

(3) Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energies for Protein− Ligand Binding. *J. Med. Chem.* **2006**, *49* (20), 5880–5884.

(4) Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57* (11), 2719–2728.

(5) Taylor, R.; Cole, J.; Korb, O.; McCabe, P. Knowledge-Based Libraries for Predicting the Geometric Preferences of Druglike Molecules. *J. Chem. Inf. Model.* **2014**, *54* (9), 2500–2514.

(6) Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57* (3), 529–539.

(7) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50* (4), 534–546.

(8) Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57* (8), 1747–1756.

(9) Juárez-Jiménez, J.; Barril, X.; Orozco, M.; Pouplana, R.; Luque, F. J. Assessing the Suitability of the Multilevel Strategy for the Conformational Analysis of Small Ligands. *J. Phys. Chem. B* **2014**, *119* (3), 1164–1172.

(10) Forti, F.; Cavasotto, C. N.; Orozco, M.; Barril, X.; Luque, F. J. A Multilevel Strategy for the Exploration of the Conformational Flexibility of Small Molecules. *J. Chem. Theory Comput.* **2012**, *8* (5), 1808–1819.

(11) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian Replica Exchange Method for Efficient Sampling of Biomolecular Systems: Application to Protein Structure Prediction. *J. Chem. Phys.* **2002**, *116* (20), 9058–9067.

(12) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151.

(13) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Luque, F. J. Extension of the MST Model to the IEF Formalism: HF and B3LYP Parametrizations. *J. Mol. Struct. THEOCHEM* **2005**, *727* (1–3), 29–40.

(14) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.

(15) Marvin Was Used for Drawing, Displaying and Characterizing Chemical Structures, Substructures and Reactions, Marvin 17.21.0, ChemAxon (Https://Www.Chemaxon.Com).

(16) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.* **2015**, *7* (1), 20.

(17) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B. ChEMBL: A Large-Scale Bioactivity Database

for Drug Discovery. *Nucleic Acids Res.* **2011**, *40* (D1), D1100–D1107.
(18) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S. The Protein Data Bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2002**, *58* (6), 899–907.
(19) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual reports in computational chemistry*; Elsevier, 2008; Vol. 4, pp 217–241.
(20) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H. Gaussian 16. Gaussian, Inc. Wallingford, CT 2016.
(21) Jiang, C.; Jin, X.; Dong, Y.; Chen, M. Kekule. Js: An Open Source Javascript Chemoinformatics Toolkit. *J. Chem. Inf. Model.* **2016**, *56* (6), 1132–1138.
(22) Rose, A. S.; Bradley, A. R.; Valasatava, Y.; Duarte, J. M.; Prlić, A.; Rose, P. W. NGL Viewer: Web-Based Molecular Graphics for Large Complexes. *Bioinformatics* **2018**, *34* (21), 3755–3758.
(23) Hospital, A.; Battistini, F.; Soliva, R.; Gelpí, J. L.; Orozco, M. Surviving the Deluge of Biosimulation Data. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* e1449.
(24) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23* (1–3), 3–25.
(25) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623.
(26) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Res.* **2013**, *42* (D1), D1091–D1097.
(27) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N. SureChEMBL: A Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res.* **2015**, *44* (D1), D1220–D1228.
(28) Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites. *Nucleic Acids Res.* **2015**, *44* (D1), D1214–D1219.
(29) Li, Y. H.; Yu, C. Y.; Li, X. X.; Zhang, P.; Tang, J.; Yang, Q.; Fu, T.; Zhang, X.; Cui, X.; Tu, G. Therapeutic Target Database Update 2018: Enriched Resource for Facilitating Bench-to-Clinic Research of Targeted Therapeutics. *Nucleic Acids Res.* **2017**, *46* (D1), D1121–D1127.
(30) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* **2015**, *44* (D1), D1045–D1053.
(31) Sterling, T.; Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337.
(32) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins Struct. Funct. Bioinforma.* **1993**, *17* (4), 412–425.
(33) Gelpí, J. L.; Kalko, S. G.; Barril, X.; Cirera, J.; de la Cruz, X.; Luque, F. J.; Orozco, M. Classical Molecular Interaction Potentials: Improved Setup Procedure in Molecular Dynamics Simulations of Proteins. *Proteins Struct. Funct. Bioinforma.* **2001**, *45* (4), 428–437.
(34) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174.
(35) Sprenger, K. G.; Jaeger, V. W.; Pfaendtner, J. The General AMBER Force Field (GAFF) Can

Accurately Predict Thermodynamic and Transport Properties of Many Ionic Liquids. *J. Phys. Chem. B* **2015**, *119* (18), 5882–5895.

(36) Caleman, C.; van Maaren, P. J.; Hong, M.; Hub, J. S.; Costa, L. T.; van der Spoel, D. Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant. *J. Chem. Theory Comput.* **2012**, *8* (1), 61–74.

(37) Zhu, S. Validation of the Generalized Force Fields GAFF, CGenFF, OPLS-AA, and PRODRGFF by Testing Against Experimental Osmotic Coefficient Data for Small Drug-Like Molecules. *J. Chem. Inf. Model.* **2019**, *59* (10), 4239–4247.

(38) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. data* **2016**, *3*.

(39) Burbeck, S. Applications Programming in Smalltalk-80 (Tm): How to Use Model-View-Controller (Mvc). *Smalltalk-80 v2* **1992**, *5*, 1–11.

(40) UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2016**, *45* (D1), D158–D169.

(41) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Bryant, S. H. PUG-SOAP and PUG-REST: Web Services for Programmatic Access to Chemical Information in PubChem. *Nucleic Acids Res.* **2015**, *43* (W1), W605–W611.

(42) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3* (1), 33.

(43) ChemAxon, M. ChemAxon. Ltd 2015.

(44) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447.

(45) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604–613.

(46) Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G.; Bonas, P.; Barducci, A.; Bernetti, M.; Bolhuis, P. G.; Bottaro, S.; Branduardi, D. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **2019**, *16* (8), 670–673.

(47) Bussi, G. Hamiltonian Replica Exchange in GROMACS: A Flexible Implementation. *Mol. Phys.* **2014**, *112* (3–4), 379–384.

(48) Meyer, T.; Ferrer-Costa, C.; Pérez, A.; Rueda, M.; Bidon-Chanal, A.; Luque, F. J.; Laughton, C. A.; Orozco, M. Essential Dynamics: A Tool for Efficient Trajectory Compression and Management. *J. Chem. Theory Comput.* **2006**, *2* (2), 251–258.

(49) Roe, D. R.; Cheatham III, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.

(50) Case, D. A.; Babin, V.; Berryman, J.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham III, T. E.; Darden, T. A.; Duke, R. E.; Gohlke, H. AMBER 14, 2014. *Univ. California, San Fr.* **2014**.

(51) Milojičić, D.; Llorente, I. M.; Montero, R. S. Opennebula: A Cloud Management Tool. *IEEE Internet Comput.* **2011**, *15* (2), 11–14.

(52) Beno, B. R.; Yeung, K.-S.; Bartberger, M. D.; Pennington, L. D.; Meanwell, N. A. A Survey of the Role of Noncovalent Sulfur Interactions in Drug Design. *J. Med. Chem.* **2015**, *58* (11), 4383–4438.

(53) Angell, R.; Aston, N. M.; Bamborough, P.; Buckton, J. B.; Cockerill, S.; deboeck, S. J.; Edwards, C. D.; Holmes, D. S.; Jones, K. L.; Laine, D. I. Biphenyl Amide P38 Kinase Inhibitors 3: Improvement of Cellular and in Vivo Activity. *Bioorg. Med. Chem. Lett.* **2008**, *18* (15), 4428–4432.

(54) Kryger, G.; Silman, I.; Sussman, J. L. Structure of Acetylcholinesterase Complexed with E2020 (Aricept®): Implications for the Design of New Anti-Alzheimer Drugs. *Structure* **1999**, *7* (3), 297–307.
(55) Fu, Z.; Li, X.; Miao, Y.; Merz Jr, K. M. Conformational Analysis and Parallel QM/MM X-Ray Refinement of Protein Bound Anti-Alzheimer Drug Donepezil. *J. Chem. Theory Comput.* **2013**, *9* (3), 1686–1693.
(56) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749.