ChEMBL & RDKit

Marrying Open Data with Open Tools

Rodrigo Ochoa
Mark Davies
Francis Atkinson
George Papadatos
John Overington

EMBL-EBI

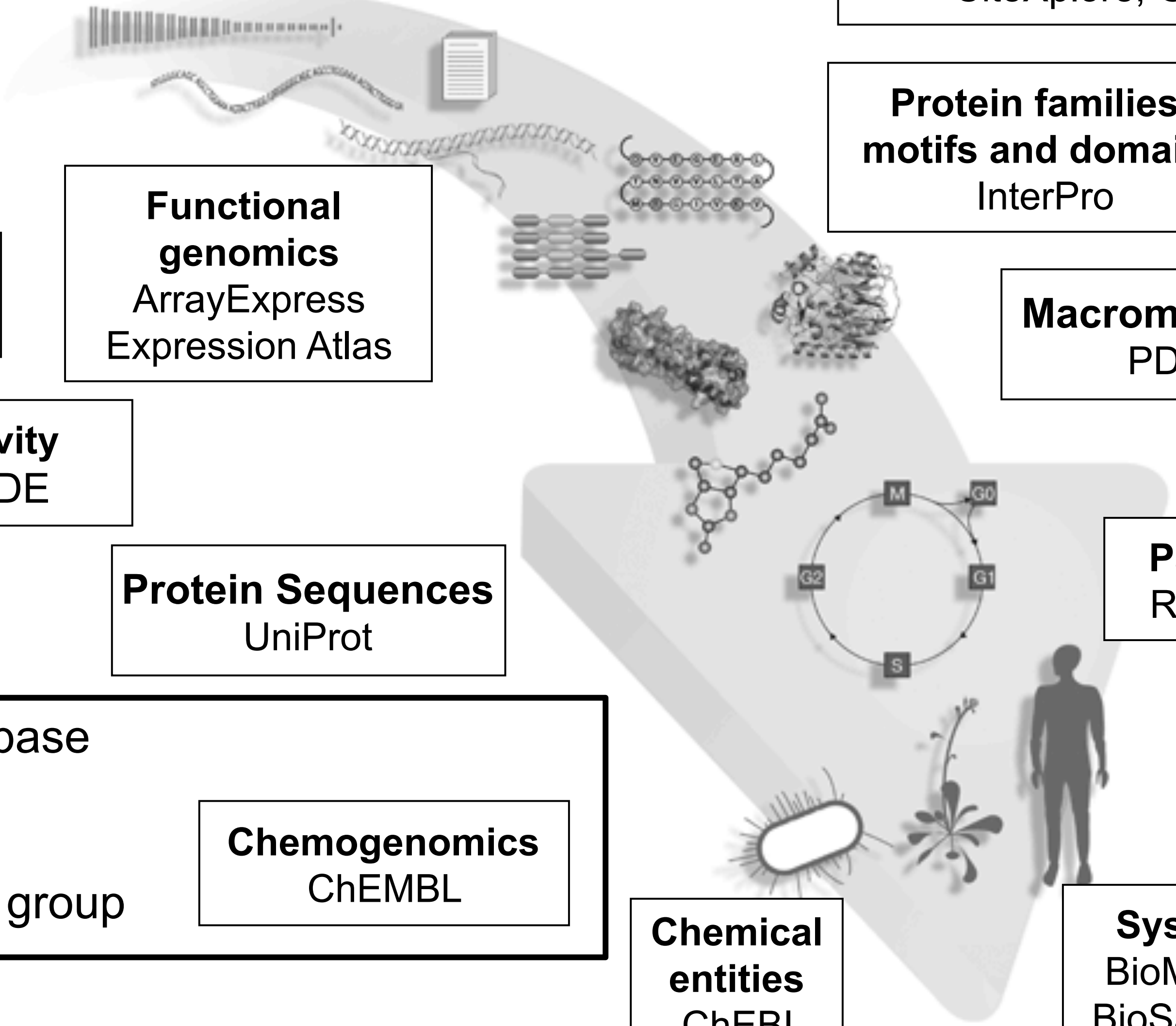
Outline

- Introduction to EBI and ChEMBL
- Loading ChEMBL (PostgreSQL + RDKit)
- Comparative studies between RDKit and Symyx
- Web interface

ChEMBL

INTRODUCTION

EBI structure



Genomes
Ensembl
Ensembl Genomes
EGA

Literature and ontologies
CiteXplore, GO

Nucleotide sequence
ENA

Functional genomics
ArrayExpress
Expression Atlas

Protein families, motifs and domains
InterPro

Macromolecular
PDBe

Protein activity
IntAct , PRIDE

Protein Sequences
UniProt

Pathways
Reactome

- ChEMBL database
  - Curation
  - Interface
  - Research group

Chemogenomics
ChEMBL

Chemical entities
ChEBI

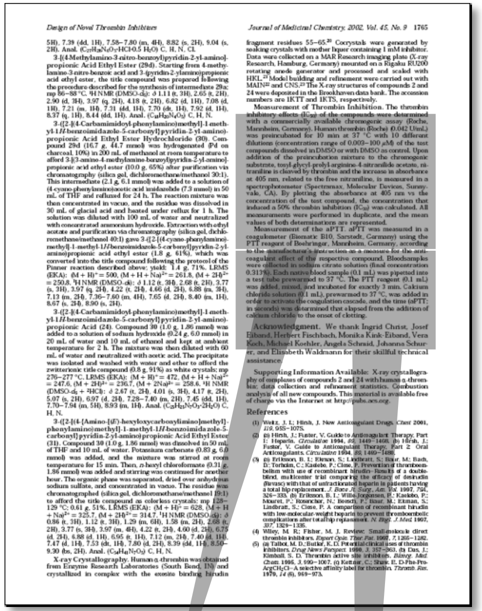
Systems
BioModels
BioSamples

What is the ChEMBL database?

- A freely-available, curated source of small molecules, targets, assays and bioactivity data
- Core data is from the primary Med Chem literature
  - J. Med. Chem., Bioorg. Med. Chem. Lett., J. Nat. Prod.

- Information extracted
  - Compounds tested
  - Assays performed
  - Biological targets of assays
  - Activities of compounds in assays

- Structures and data curated in-house to ensure quality



5H), 7.39 (dd, 1H), 7.58–7.80 (m, 4H), 8.82 (s, 2H), 9.04 (s, 2H). Anal. ($C_{27}H_{28}N_6O_3$·HCl·0.5 H_2O) C, H, N, Cl.

3-[(4-Methylamino-3-nitro-benzoyl)pyridin-2-yl-amino]-propionic Acid Ethyl Ester (29d). Starting from 4-methylamino-3-nitro-benzoic acid and 3-(pyridin-2-ylamino)propionic acid ethyl ester, the title compound was prepared following the procedure described for the synthesis of intermediate 29a: mp 86–88 °C. ^{1}H NMR (DMSO-d_6): δ 1.11 (t, 3H), 2.65 (t, 2H), 2.90 (d, 3H), 3.97 (q, 2H), 4.18 (t, 2H), 6.82 (d, 1H), 7.08 (d, 1H), 7.21 (m, 1H), 7.31 (dd, 1H), 7.70 (dt, 1H), 7.92 (d, 1H), 8.37 (q, 1H), 8.44 (dd, 1H). Anal. ($C_{18}H_{20}N_4O_5$) C, H, N.

3-({2-[(4-Carbamimidoyl-phenylamino)methyl]-1-methyl-1*H*-benzoimidazole-5-carbonyl}pyridin-2-yl-amino)-propionic Acid Ethyl Ester Hydrochloride (30). Compound 29d (16.7 g, 44.7 mmol) was hydrogenated (Pd on charcoal, 10%) in 200 mL of methanol at room temperature to afford 3-[(3-amino-4-methylamino-benzoyl)pyridin-2-yl-amino]-propionic acid ethyl ester (10.0 g, 65%) after purification via chromatography (silica gel, dichloromethane/methanol 30:1). This intermediate (2.1 g, 6.1 mmol) was added to a solution of (4-cyano-phenylamino)acetic acid imidazolide (7.3 mmol) in 50 mL of THF and refluxed for 24 h. The reaction mixture was then concentrated in vacuo, and the residue was dissolved in 30 mL of glacial acid and heated under reflux for 1 h. The solution was diluted with 100 mL of water and neutralized with concentrated ammonium hydroxide. Extraction with ethyl acetate and purification via chromatography (silica gel, dichloromethane/methanol 40:1) gave 3-({2-[(4-cyano-phenylamino)-methyl]-1-methyl-1*H*-benzoimidazole-5-carbonyl}pyridin-2-yl-amino)propionic acid ethyl ester (1.8 g, 61%), which was converted into the title compound following the protocol of the Pinner reaction described above; yield: 1.4 g, 71%. LRMS (EKA): $(M + H)^+ = 500$, $(M + H + Na)^{2+} = 261.8$, $(M + 2H)^{2+} = 250.8$. ^{1}H NMR (DMSO-d_6): δ 1.12 (t, 3H), 2.68 (t, 2H), 3.77 (s, 3H), 3.97 (q, 2H), 4.22 (t, 2H), 4.66 (d, 2H), 6.88 (m, 3H), 7.13 (m, 2H), 7.36–7.60 (m, 4H), 7.65 (d, 2H), 8.40 (m, 1H), 8.67 (s, 2H), 8.90 (s, 2H).

3-({2-[(4-Carbamimidoyl-phenylamino)methyl]-1-methyl-1*H*-benzoimidazole-5-carbonyl}pyridin-2-yl-amino)-propionic Acid (24). Compound 30 (1.0 g, 1.86 mmol) was added to a solution of sodium hydroxide (0.24 g, 6.0 mmol) in 20 mL of water and 10 mL of ethanol and kept at ambient temperature for 2 h. The mixture was then diluted with 60 mL of water and neutralized with acetic acid. The precipitate was isolated and washed with water and ether to afford the zwitterionic title compound (0.8 g, 91%) as white crystals; mp 276–277 °C. LRMS (EKA): $(M + H)^+ = 472$, $(M + H + Na)^{2+} = 247.6$, $(M + 2H)^{2+} = 236.7$, $(M + 2Na)^{2+} = 258.6$. ^{1}H NMR (DMSO-d_6 + ^{2}HCl): δ 2.67 (t, 2H), 4.01 (s, 3H), 4.17 (t, 2H), 5.07 (s, 2H), 6.97 (d, 2H), 7.28–7.40 (m, 2H), 7.45 (dd, 1H), 7.70–7.94 (m, 5H), 8.93 (m, 1H). Anal. ($C_{25}H_{25}N_7O_3$·2H_2O) C, H, N.

3-({2-[(4-{Amino-[(*E*)-hexyloxycarbonylimino]methyl}-phenylamino)methyl]-1-methyl-1*H*-benzoimidazole-5-carbonyl}pyridin-2-yl-amino)propionic Acid Ethyl Ester (31). Compound 30 (1.0 g, 1.86 mmol) was dissolved in 50 mL of THF and 10 mL of water. Potassium carbonate (0.83 g, 6.0 mmol) was added, and the mixture was stirred at room temperature for 15 min. Then, *n*-hexyl chloroformate (0.31 g, 1.86 mmol) was added and stirring was continued for another hour. The organic phase was separated, dried over anhydrous sodium sulfate, and concentrated in vacuo. The residue was chromatographed (silica gel, dichloromethane/methanol 19:1) to afford the title compound as colorless crystals: mp 128–129 °C; 0.61 g, 51%. LRMS (EKA): $(M + H)^+ = 628$, $(M + H + Na)^{2+} = 325.7$, $(M + 2H)^{2+} = 314.7$. ^{1}H NMR (DMSO-d_6): δ 0.86 (t, 3H), 1.12 (t, 3H), 1.29 (m, 6H), 1.58 (m, 2H), 2.68 (t, 2H), 3.77 (s, 3H), 3.97 (m, 4H), 4.22 (t, 2H), 4.60 (d, 2H), 6.75 (d, 2H), 6.88 (d, 1H), 6.95 (t, 1H), 7.12 (m, 2H), 7.40 (d, 1H), 7.47 (d, 1H), 7.53 (dt, 1H), 7.80 (d, 2H), 8.39 (dd, 1H), 8.50–9.30 (bs, 2H). Anal. ($C_{34}H_{41}N_7O_5$) C, H, N.

X-ray Crystallography. Human α-thrombin was obtained from Enzyme Research Laboratories (South Bend, IN) and crystallized in complex with the exosite binding hirudin fragment residues 55–65.[20] Cocrystals were generated by soaking crystals with mother liquor containing 1 mM inhibitor. Data were collected on a MAR Research imaging plate (X-ray Research, Hamburg, Germany) mounted on a Rigaku RU200 rotating anode generator and processed and scaled with HKL.[21] Model building and refinement were carried out with MAIN[22] and CNS.[23] The X-ray structures of compounds 2 and 24 were deposited in the Brookhaven data bank. The accession numbers are 1KTT and 1KTS, respectively.

Measurement of Thrombin Inhibition. The thrombin inhibitory effects (IC_{50}) of the compounds were determined with a commercially available chromogenic assay (Roche, Mannheim, Germany). Human thrombin (Roche) (0.042 U/mL) was preincubated for 10 min at 37 °C with 10 different dilutions (concentration range of 0.003–100 μM) of the test compounds dissolved in DMSO or with DMSO as control. Upon addition of the preincubation mixture to the chromogenic substrate, tosyl-glycyl-prolyl-arginine-4-nitranilide acetate, nitraniline is cleaved by thrombin and the increase in absorbance at 405 nm, related to the free nitraniline, is measured in a spectrophotometer (Spectramax, Molecular Devices, Sunnyvale, CA). By plotting the absorbance at 405 nm vs the concentration of the test compound, the concentration that induced a 50% thrombin inhibition (IC_{50}) was calculated. All measurements were performed in duplicate, and the mean values of both determinations are represented.

Measurement of the aPTT. aPTT was measured in a coagulometer (Biomatic B10, Sarstedt, Germany) using the PTT reagent of Boehringer, Mannheim, Germany, according to the manufacturer's instruction as a measure for the anticoagulant effect of the respective compound. Bloodsamples were collected in sodium citrate solution (final concentration 0.313%). Each native blood sample (0.1 mL) was pipetted into a test tube prewarmed to 37 °C. The PTT reagent (0.1 mL) was added, mixed, and incubated for exactly 3 min. Calcium chloride solution (0.1 mL), prewarmed to 37 °C, was added in order to activate the coagulation cascade, and the time (aPTT; in seconds) was determined that elapsed from the addition of calcium chloride to the onset of clotting.

Acknowledgment. We thank Ingrid Christ, Josef Eiband, Herbert Fischbach, Monika Kink-Eiband, Vera Koch, Michael Koehler, Angela Schmid, Johanna Schurer, and Elisabeth Waldmann for their skillful technical assistance.

Supporting Information Available: X-ray crystallography of complexes of compounds 2 and 24 with human α-thrombin; data collection and refinement statistics. Combustion analysis of all new compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Weitz, J. I.; Hirsh, J. New Anticoagulant Drugs. *Chest* 2001, *119*, 95S–107S.

(2) (a) Hirsh, J.; Fuster, V. Guide to Anticoagulant Therapy. Part 1: Heparin. *Circulation* 1994, *89*, 1449–1468. (b) Hirsh, J.; Fuster, V. Guide to Anticoagulant Therapy. Part 2: Oral Anticoagulants. *Circulation* 1994, *89*, 1469–1480.

(3) (a) Eriksson, B. I.; Ekman, S.; Lindbratt, S.; Baur, M.; Bach, D.; Torholm, C.; Kaelebo, P.; Close, P. Prevention of thromboembolism with use of recombinant hirudin–Results of a double-blind, multicenter trial comparing the efficacy of desirudin (Revasc) with that of unfractionated heparin in patients having a total hip replacement. *J. Bone Jt. Surg., Am. Vol.* 1997, *79A*, 326–333. (b) Eriksson, B. I.; Wille-Jorgensen, P.; Kaelebo, P.; Mouret, P.; Rosencher, N.; Boesch, P.; Baur, M.; Ekman, S.; Lindbratt, S.; Close, P. A comparison of recombinant hirudin with low-molecular-weight heparin to prevent thromboembolic complications after total hip replacement. *N. Engl. J. Med.* 1997, *337*, 1329–1335.

(4) Wiley, M. R.; Fisher, M. J. Review: Small-molecule direct thrombin inhibitors. *Expert Opin. Ther. Pat.* 1997, *7*, 1265–1282.

(5) (a) Talbot, M. D.; Butler, K. D. Potential clinical uses of thrombin inhibitors. *Drug News Perspect.* 1990, *3*, 357–363. (b) Das, J.; Kimball, S. D. Thrombin active site inhibitors. *Bioorg. Med. Chem.* 1995, *3*, 999–1007. (c) Kettner, C.; Shaw, E. D-Phe-Pro-ArgCH$_2$Cl–A selective affinity label for thrombin. *Thromb. Res.* 1979, *14* (6), 969–973.

Design of Novel Thrombin Inhibitors

Journal of Medicinal Chemistry, 2002, Vol. 45, No. 9 1765

5H), 7.39 (dd, 1H), 7.58–7.80 (m, 4H), 8.82 (s, 2H), 9.04 (s, 2H). Anal. ($C_{27}H_{28}N_6O_3$·HCl·0.5 H_2O) C, H, N, Cl.

3-[(4-Methylamino-3-nitro-benzoyl)pyridin-2-yl-amino]-propionic Acid Ethyl Ester (29d). Starting from 4-methylamino-3-nitro-benzoic acid and 3-(pyridin-2-ylamino)propionic acid ethyl ester, the title compound was prepared following the procedure described for the synthesis of intermediate 29a: mp 86–88 °C. ^{1}H NMR (DMSO-d_6): δ 1.11 (t, 3H), 2.65 (t, 2H), 2.90 (d, 3H), 3.97 (q, 2H), 4.18 (t, 2H), 6.82 (d, 1H), 7.08 (d, 1H), 7.21 (m, 1H), 7.31 (dd, 1H), 7.70 (dt, 1H), 7.92 (d, 1H), 8.37 (q, 1H), 8.44 (dd, 1H). Anal. ($C_{18}H_{20}N_4O_5$) C, H, N.

3-({2-[(4-Carbamimidoyl-phenylamino)methyl]-1-methyl-1*H*-benzoimidazole-5-carbonyl}pyridin-2-yl-amino)-propionic Acid Ethyl Ester Hydrochloride (30). Compound 29d (16.7 g, 44.7 mmol) was hydrogenated (Pd on charcoal, 10%) in 200 mL of methanol at room temperature to afford 3-[(3-amino-4-methylamino-benzoyl)pyridin-2-yl-amino]-propionic acid ethyl ester (10.0 g, 65%) after purification via chromatography (silica gel, dichloromethane/methanol 30:1). This intermediate (2.1 g, 6.1 mmol) was added to a solution of (4-cyano-phenylamino)acetic acid imidazolide (7.3 mmol) in 50 mL of THF and refluxed for 24 h. The reaction mixture was then concentrated in vacuo, and the residue was dissolved in 30 mL of glacial acid and heated under reflux for 1 h. The solution was diluted with 100 mL of water and neutralized with concentrated ammonium hydroxide. Extraction with ethyl acetate and purification via chromatography (silica gel, dichloromethane/methanol 40:1) gave 3-({2-[(4-cyano-phenylamino)-methyl]-1-methyl-1*H*-benzoimidazole-5-carbonyl}pyridin-2-yl-amino)propionic acid ethyl ester (1.8 g, 61%), which was converted into the title compound following the protocol of the Pinner reaction described above; yield: 1.4 g, 71%. LRMS (EKA): $(M + H)^+ = 500$, $(M + H + Na)^{2+} = 261.8$, $(M + 2H)^{2+} = 250.8$. ^{1}H NMR (DMSO-d_6): δ 1.12 (t, 3H), 2.68 (t, 2H), 3.77 (s, 3H), 3.97 (q, 2H), 4.22 (t, 2H), 4.66 (d, 2H), 6.88 (m, 3H), 7.13 (m, 2H), 7.36–7.60 (m, 4H), 7.65 (d, 2H), 8.40 (m, 1H), 8.67 (s, 2H), 8.90 (s, 2H).

3-({2-[(4-Carbamimidoyl-phenylamino)methyl]-1-methyl-1*H*-benzoimidazole-5-carbonyl}pyridin-2-yl-amino)-propionic Acid (24). Compound 30 (1.0 g, 1.86 mmol) was added to a solution of sodium hydroxide (0.24 g, 6.0 mmol) in 20 mL of water and 10 mL of ethanol and kept at ambient temperature for 2 h. The mixture was then diluted with 60 mL of water and neutralized with acetic acid. The precipitate was isolated and washed with water and ether to afford the zwitterionic title compound (0.8 g, 91%) as white crystals; mp 276–277 °C. LRMS (EKA): $(M + H)^+ = 472$, $(M + H + Na)^{2+} = 247.6$, $(M + 2H)^{2+} = 236.7$, $(M + 2Na)^{2+} = 258.6$. ^{1}H NMR (DMSO-d_6 + ^{2}HCl): δ 2.67 (t, 2H), 4.01 (s, 3H), 4.17 (t, 2H), 5.07 (s, 2H), 6.97 (d, 2H), 7.28–7.40 (m, 2H), 7.45 (dd, 1H), 7.70–7.94 (m, 5H), 8.93 (m, 1H). Anal. ($C_{25}H_{25}N_7O_3$·2H_2O) C, H, N.

3-({2-[(4-{Amino-[(*E*)-hexyloxycarbonylimino]methyl}-phenylamino)methyl]-1-methyl-1*H*-benzoimidazole-5-carbonyl}pyridin-2-yl-amino)propionic Acid Ethyl Ester (31). Compound 30 (1.0 g, 1.86 mmol) was dissolved in 50 mL of THF and 10 mL of water. Potassium carbonate (0.83 g, 6.0 mmol) was added, and the mixture was stirred at room temperature for 15 min. Then, *n*-hexyl chloroformate (0.31 g, 1.86 mmol) was added and stirring was continued for another hour. The organic phase was separated, dried over anhydrous sodium sulfate, and concentrated in vacuo. The residue was chromatographed (silica gel, dichloromethane/methanol 19:1) to afford the title compound as colorless crystals: mp 128–129 °C; 0.61 g, 51%. LRMS (EKA): $(M + H)^+ = 628$, $(M + H + Na)^{2+} = 325.7$, $(M + 2H)^{2+} = 314.7$. ^{1}H NMR (DMSO-d_6): δ 0.86 (t, 3H), 1.12 (t, 3H), 1.29 (m, 6H), 1.58 (m, 2H), 2.68 (t, 2H), 3.77 (s, 3H), 3.97 (m, 4H), 4.22 (t, 2H), 4.60 (d, 2H), 6.75 (d, 2H), 6.88 (d, 1H), 6.95 (t, 1H), 7.12 (m, 2H), 7.40 (d, 1H), 7.47 (d, 1H), 7.53 (dt, 1H), 7.80 (d, 2H), 8.39 (dd, 1H), 8.50–9.30 (bs, 2H). Anal. ($C_{34}H_{41}N_7O_5$) C, H, N.

X-ray Crystallography. Human α-thrombin was obtained from Enzyme Research Laboratories (South Bend, IN) and crystallized in complex with the exosite binding hirudin fragment residues 55–65.[20] Cocrystals were generated by soaking crystals with mother liquor containing 1 mM inhibitor. Data were collected on a MAR Research imaging plate (X-ray Research, Hamburg, Germany) mounted on a Rigaku RU200 rotating anode generator and processed and scaled with HKL.[21] Model building and refinement were carried out with MAIN[22] and CNS.[23] The X-ray structures of compounds 2 and 24 were deposited in the Brookhaven data bank. The accession numbers are 1KTT and 1KTS, respectively.

Measurement of Thrombin Inhibition. The thrombin inhibitory effects (IC_{50}) of the compounds were determined with a commercially available chromogenic assay (Roche, Mannheim, Germany). Human thrombin (Roche) (0.042 U/mL) was preincubated for 10 min at 37 °C with 10 different dilutions (concentration range of 0.003–100 μM) of the test compounds dissolved in DMSO or with DMSO as control. Upon addition of the preincubation mixture to the chromogenic substrate, tosyl-glycyl-prolyl-arginine-4-nitranilide acetate, nitraniline is cleaved by thrombin and the increase in absorbance at 405 nm, related to the free nitraniline, is measured in a spectrophotometer (Spectramax, Molecular Devices, Sunnyvale, CA). By plotting the absorbance at 405 nm vs the concentration of the test compound, the concentration that induced a 50% thrombin inhibition (IC_{50}) was calculated. All measurements were performed in duplicate, and the mean values of both determinations are represented.

Measurement of the aPTT. aPTT was measured in a coagulometer (Biomatic B10, Sarstedt, Germany) using the PTT reagent of Boehringer, Mannheim, Germany, according to the manufacturer's instruction as a measure for the anticoagulant effect of the respective compound. Bloodsamples were collected in sodium citrate solution (final concentration 0.313%). Each native blood sample (0.1 mL) was pipetted into a test tube prewarmed to 37 °C. The PTT reagent (0.1 mL) was added, mixed, and incubated for exactly 3 min. Calcium chloride solution (0.1 mL), prewarmed to 37 °C, was added in order to activate the coagulation cascade, and the time (aPTT; in seconds) was determined that elapsed from the addition of calcium chloride to the onset of clotting.

Acknowledgment. We thank Ingrid Christ, Josef Eiband, Herbert Fischbach, Monika Kink-Eiband, Vera Koch, Michael Koehler, Angela Schmid, Johanna Schurer, and Elisabeth Waldmann for their skillful technical assistance.

Supporting Information Available: X-ray crystallography of complexes of compounds 2 and 24 with human α-thrombin; data collection and refinement statistics. Combustion analysis of all new compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Weitz, J. I.; Hirsh, J. New Anticoagulant Drugs. *Chest* 2001, *119*, 95S–107S.

(2) (a) Hirsh, J.; Fuster, V. Guide to Anticoagulant Therapy. Part 1: Heparin. *Circulation* 1994, *89*, 1449–1468. (b) Hirsh, J.; Fuster, V. Guide to Anticoagulant Therapy. Part 2: Oral Anticoagulants. *Circulation* 1994, *89*, 1469–1480.

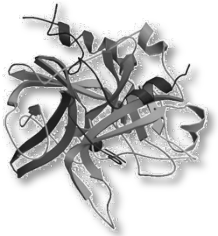
(3) (a) Eriksson, B. I.; Ekman, S.; Lindbratt, S.; Baur, M.; Bach, D.; Torholm, C.; Kaelebo, P.; Close, P. Prevention of thromboembolism with use of recombinant hirudin–Results of a double-blind, multicenter trial comparing the efficacy of desirudin (Revasc) with that of unfractionated heparin in patients having a total hip replacement. *J. Bone Jt. Surg., Am. Vol.* 1997, *79A*, 326–333. (b) Eriksson, B. I.; Wille-Jorgensen, P.; Kaelebo, P.; Mouret, P.; Rosencher, N.; Boesch, P.; Baur, M.; Ekman, S.; Lindbratt, S.; Close, P. A comparison of recombinant hirudin with low-molecular-weight heparin to prevent thromboembolic complications after total hip replacement. *N. Engl. J. Med.* 1997, *337*, 1329–1335.

(4) Wiley, M. R.; Fisher, M. J. Review: Small-molecule direct thrombin inhibitors. *Expert Opin. Ther. Pat.* 1997, *7*, 1265–1282.

(5) (a) Talbot, M. D.; Butler, K. D. Potential clinical uses of thrombin inhibitors. *Drug News Perspect.* 1990, *3*, 357–363. (b) Das, J.; Kimball, S. D. Thrombin active site inhibitors. *Bioorg. Med. Chem.* 1995, *3*, 999–1007. (c) Kettner, C.; Shaw, E. D-Phe-Pro-ArgCH$_2$Cl–A selective affinity label for thrombin. *Thromb. Res.* 1979, *14* (6), 969–973.

Compounds

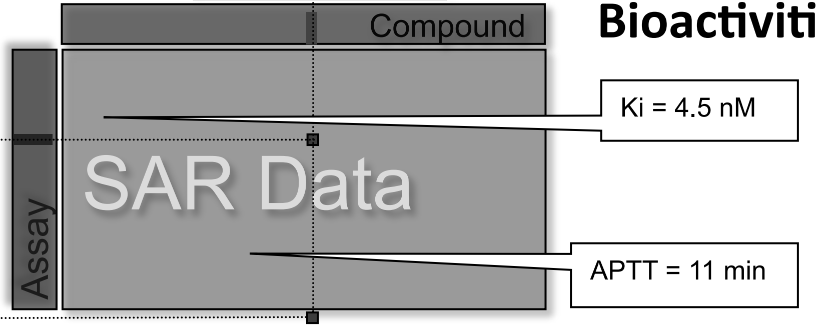
The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.



```
>Thrombin
MAHVRGLQLPGCLALAALCSLVHSQHVFLAPQQARSLLQRVRRANTFLEEVRKGNLERECVEETCS
YEEAFEALESSTATDVFWAKYTACETARTPRDKLAACLEGNCAEGLGTNYRGHVNITRSGIECQLW
RSRYPHKPEINSTTHPGADLQENFCRNPDSSTTGPWCYTTDPTVRRQECSIPVCGQDQVTVAMTPR
SEGSSVNLSPPLEQCVPDRGQQYQGRLAVTTHGLPCLAWASAQAKALSKHQDFNSAVQLVENFCRN
PDGDEEGVWCYVAGKPGDFGYCDLNYCEEAVEEETGDGLDEDSDRAIEGRTATSEYQTFFNPRTFG
SGEADCGLRPLFEKKSLEDKTERELLESYIDGRIVEGSDAEIGMSPWQVMLFRKSPQELLCGASLI
SDRWVLTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLDRD
IALMKLKKPVAFSDYIHPVCLPDRETAASLLQAGYKGRVTGWGNLKETWTANVGKGQPSVLQVVNL
PIVERPVCKDSTRIRITDNMFCAGYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGEGC
DRDGKYGFYTHVFRLKKWIQKVIDQFGE
```

Targets

Bioactivities



ChEMBL 14 (latest version)

ChEMBL14

Compounds: **1,213,239**

Assays: **644,734**

Targets: **9,003**

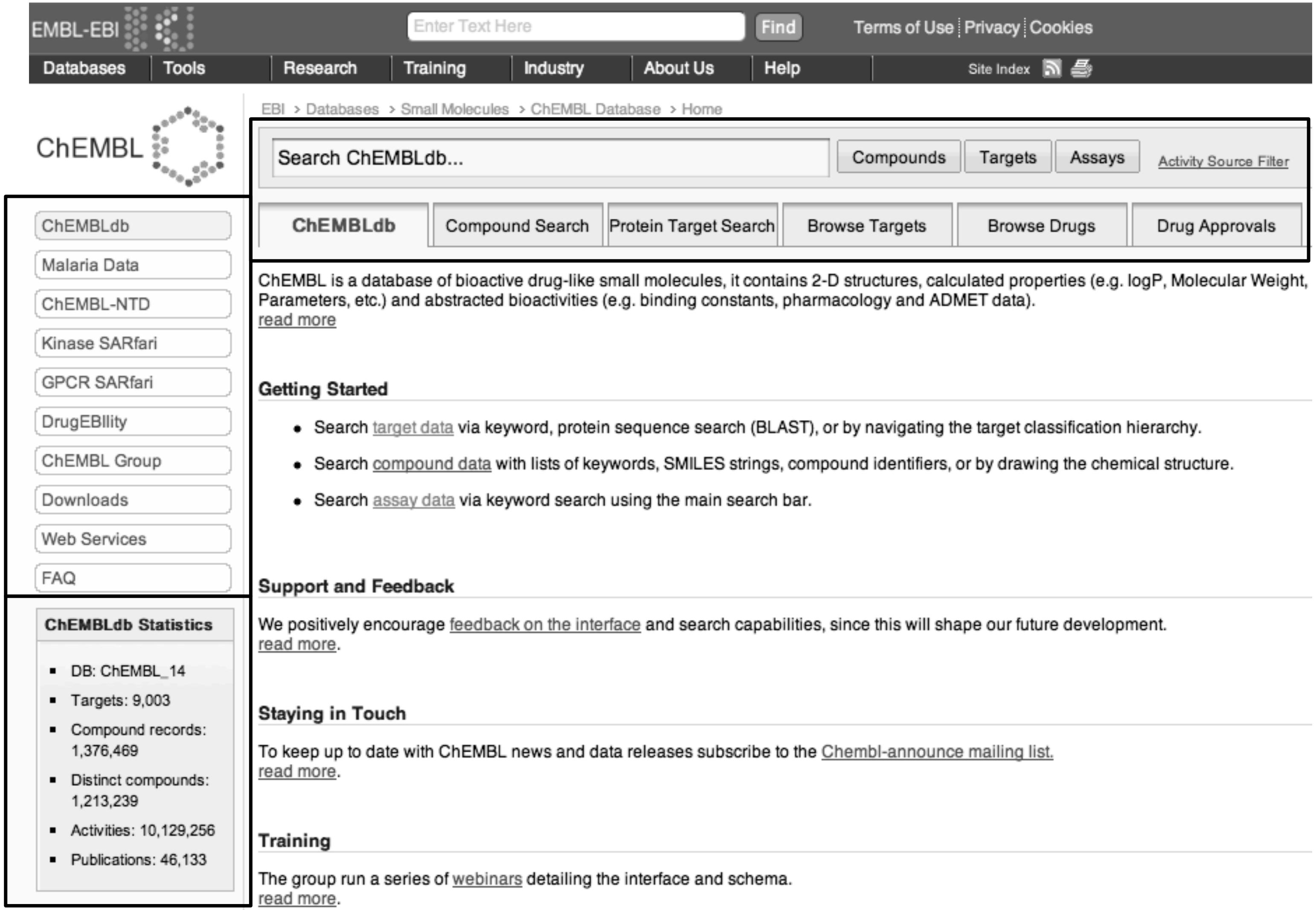
Publications: **46,133**

Activities: **10,129,256**

Data sources: **10**

Increase of >200,000 compounds
from literature since ChEMBL01

ChEMBL interface



ChEMBLdb Statistics

- DB: ChEMBL_14
- Targets: 9,003
- Compound records: 1,376,469
- Distinct compounds: 1,213,239
- Activities: 10,129,256
- Publications: 46,133

ChEMBL is a database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties (e.g. logP, Molecular Weight, Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data).
read more

Getting Started

- Search target data via keyword, protein sequence search (BLAST), or by navigating the target classification hierarchy.
- Search compound data with lists of keywords, SMILES strings, compound identifiers, or by drawing the chemical structure.
- Search assay data via keyword search using the main search bar.

Support and Feedback

We positively encourage feedback on the interface and search capabilities, since this will shape our future development.
read more.

Staying in Touch

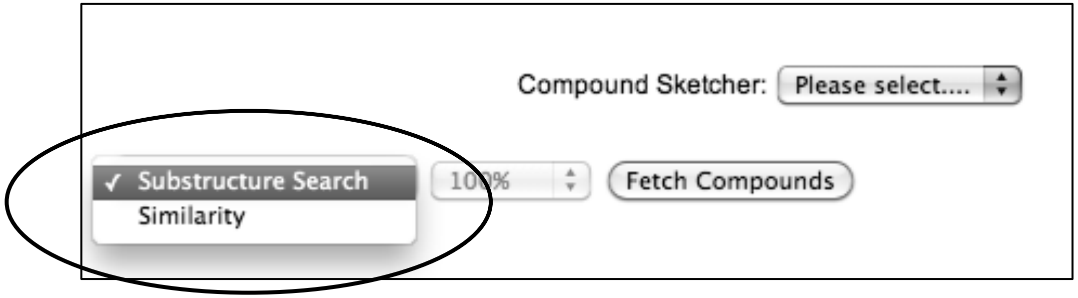
To keep up to date with ChEMBL news and data releases subscribe to the Chembl-announce mailing list.
read more.

Training

The group run a series of webinars detailing the interface and schema.
read more.

Chemical search in ChEMBL

- A variety of sketchers (Marvin, JDraw, JME)

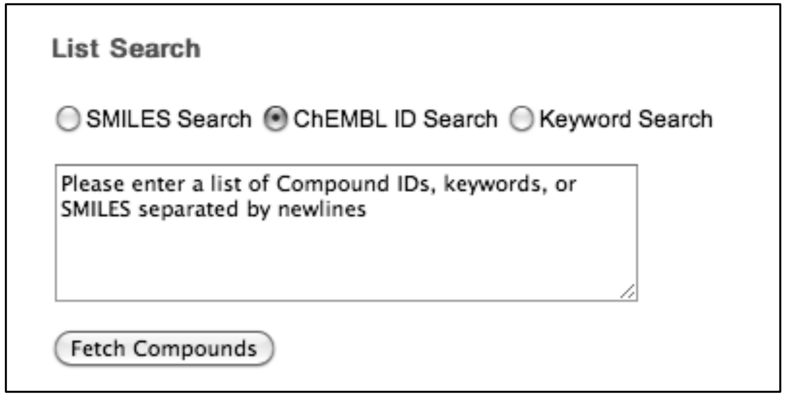


Compound Sketcher: Please select....

✓ Substructure Search

Similarity

100%



List Search

○ SMILES Search ● ChEMBL ID Search ○ Keyword Search

Please enter a list of Compound IDs, keywords, or SMILES separated by newlines

- ChEMBL chemical cartridge:

accelrys®

Symyx

PostgreSQL

LOADING CHEMBL

RDKit overview

- An open source collection of chemoinformatics and machine-learning software written in C++ and Python

- There is available a PostgreSQL cartridge

(http://code.google.com/p/rdkit/wiki/DatabaseCartridgeReferenceDocumentation)

RDKit

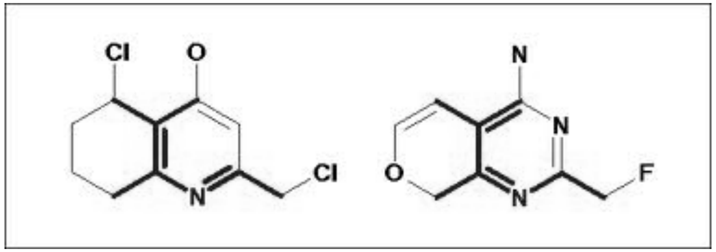
Open-Source Cheminformatics
and Machine Learning

RDKit PostgreSQL cartridge

- Molecular conversion (SMILES, SMARTS, CTAB)
- Substructure search
- Similarity Search

  Fingerprints: Morgan (ECFP-like), Atom-Pair, Torsion …

  Similarity Coefficients: Tanimoto, Dice
- Molecular properties calculation



Bingo chemical cartridge

- Open Source chemical cartridge provided by Indigo
- Support PostgreSQL and Oracle schemas

Unfortunately it could not be compared with the others chemical cartridges (bug reported in the Google group forum)

GGA Software Services

PostgreSQL ChEMBL version

- ChEMBL originally is available in Oracle (Based on this version a MySQL dump file is available)

Task: Migrate the Oracle Schema to a PostgreSQL schema

Tool:

Ora2Pg

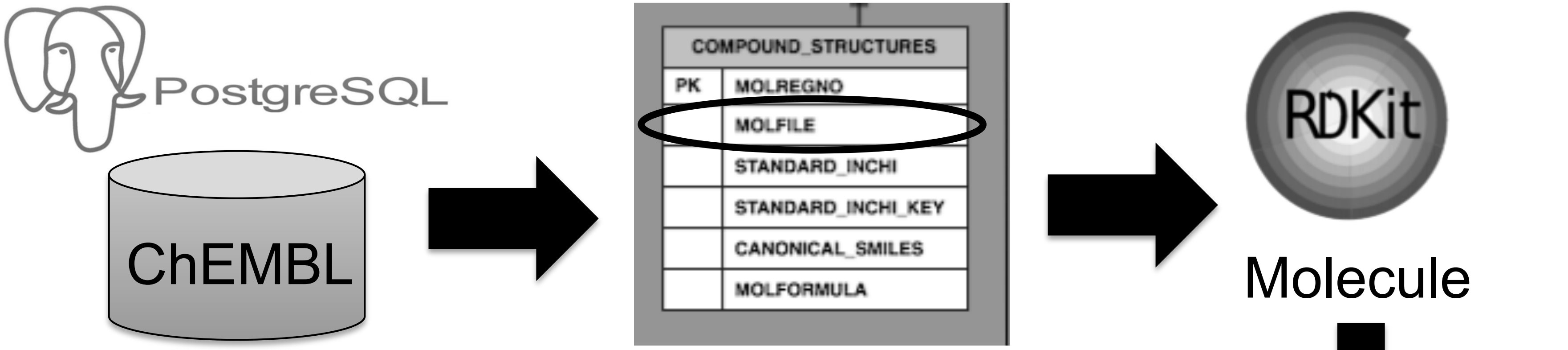
Moves Oracle databases to PostgreSQL

Open Source Kit of Perl libraries

Highly configurable

http://chembl.blogspot.co.uk/2012/08/chembl-postgresql.html

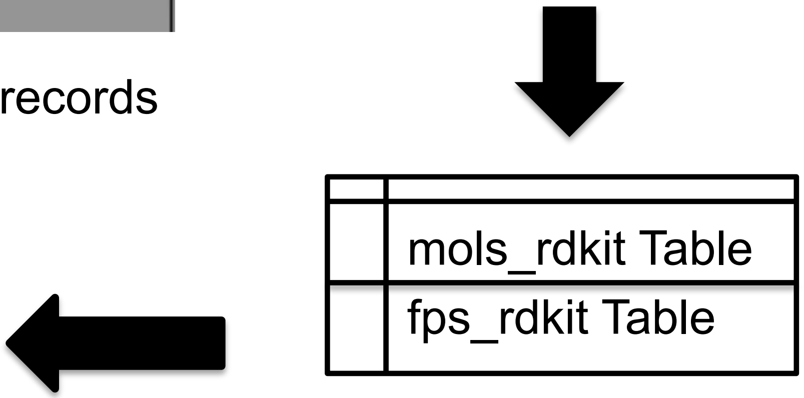
ChEMBL + cartridges



COMPOUND_STRUCTURES	
PK	MOLREGNO
	MOLFILE
	STANDARD_INCHI
	STANDARD_INCHI_KEY
	CANONICAL_SMILES
	MOLFORMULA

ChEMBL version 14

1.213.239 records



Chemical Cartridge	# Molecules correctly built
RDKit	1.146.045
Symyx	1.147.790
Bingo	1.138.682

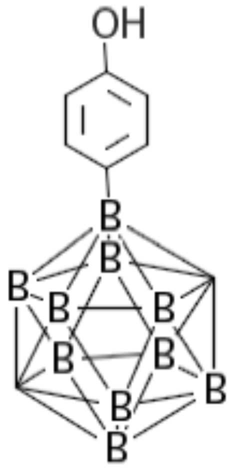
is_valid_ctab() RDKit function

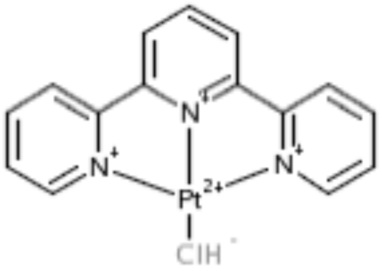
Building differences

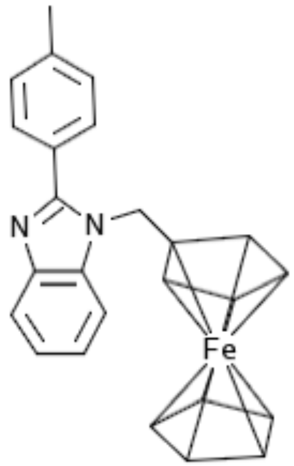
Not built by RDKit (1565)

Non pure organics (1506)

- Mainly boron clusters & organometallics

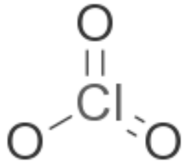


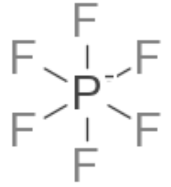




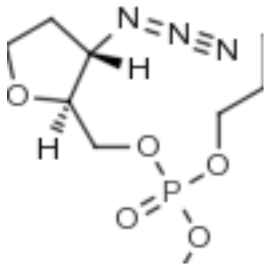
High valency (164; overlap with inorganics)

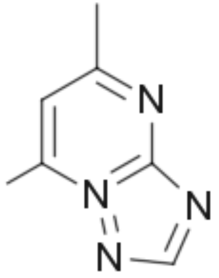
Some are legitimate structures and probably should be handled:



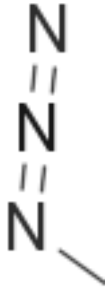


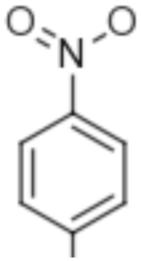
Some need standardizing in ChEMBL to *e.g.* charge-separated form:

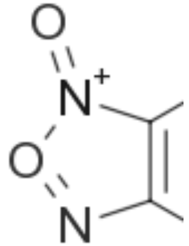




A few need fixing in ChEMBL:

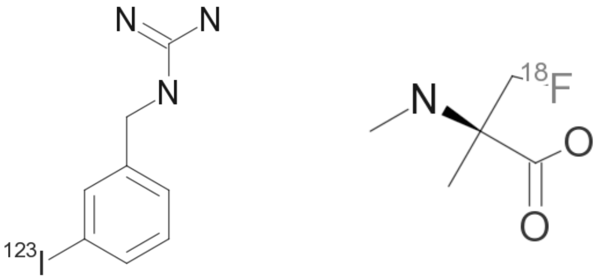




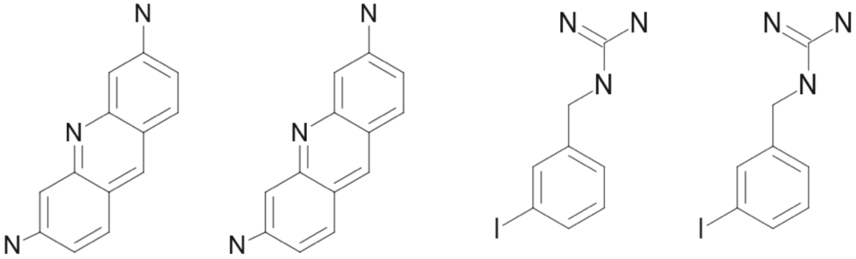


Not built by Symyx (50)

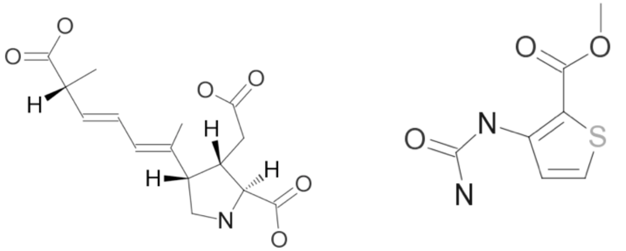
Unusual isotopes (10); used in PET, as tracers *etc*:



Duplicated structures (35):



No obvious reason for failure (5):



1,145,994 parent structures built by both systems
i.e. problems with only a very small minority of structures!

KNIME

COMPARISON

Cartridge comparison

- Chemical cartridges: RDKit and Symyx
- Type of queries: SMILES
- Objective: Detect discrepancies and the reasons behind them.
- RDKit configuration:

Version: 2012_06

OS: Linux Ubuntu 12.04 64 bits (Virtual Machine)

PostgreSQL version: 9.1.4

- Symyx configuration:

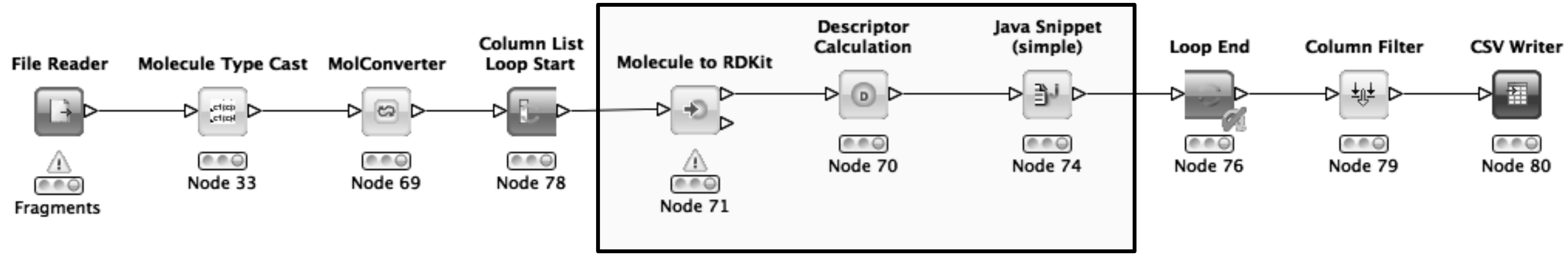
Version: 6.2

OS: Red Hat 5.8 64 bits (Server)

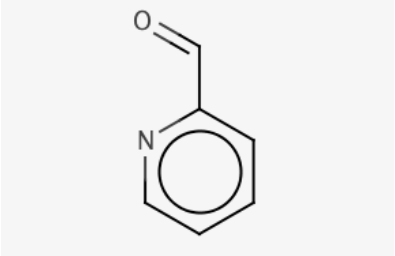
Oracle version: 11.1.0.7.0

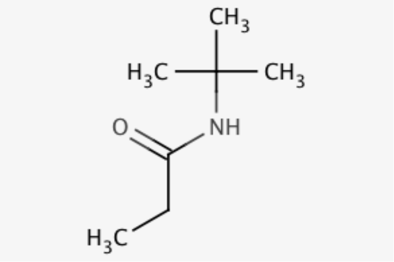
Fragment library

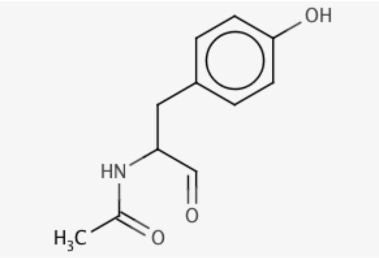
- From a set of ChEMBL molecules, RECAP* algorithm was applied.
- A filter based on the number of heavy atoms and molecular weight was done using the RDKit nodes in KNIME

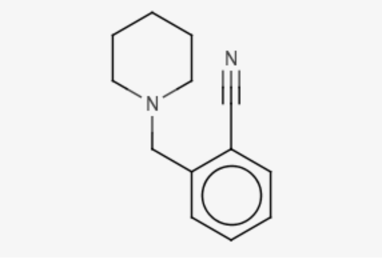


- 40 fragments were chosen



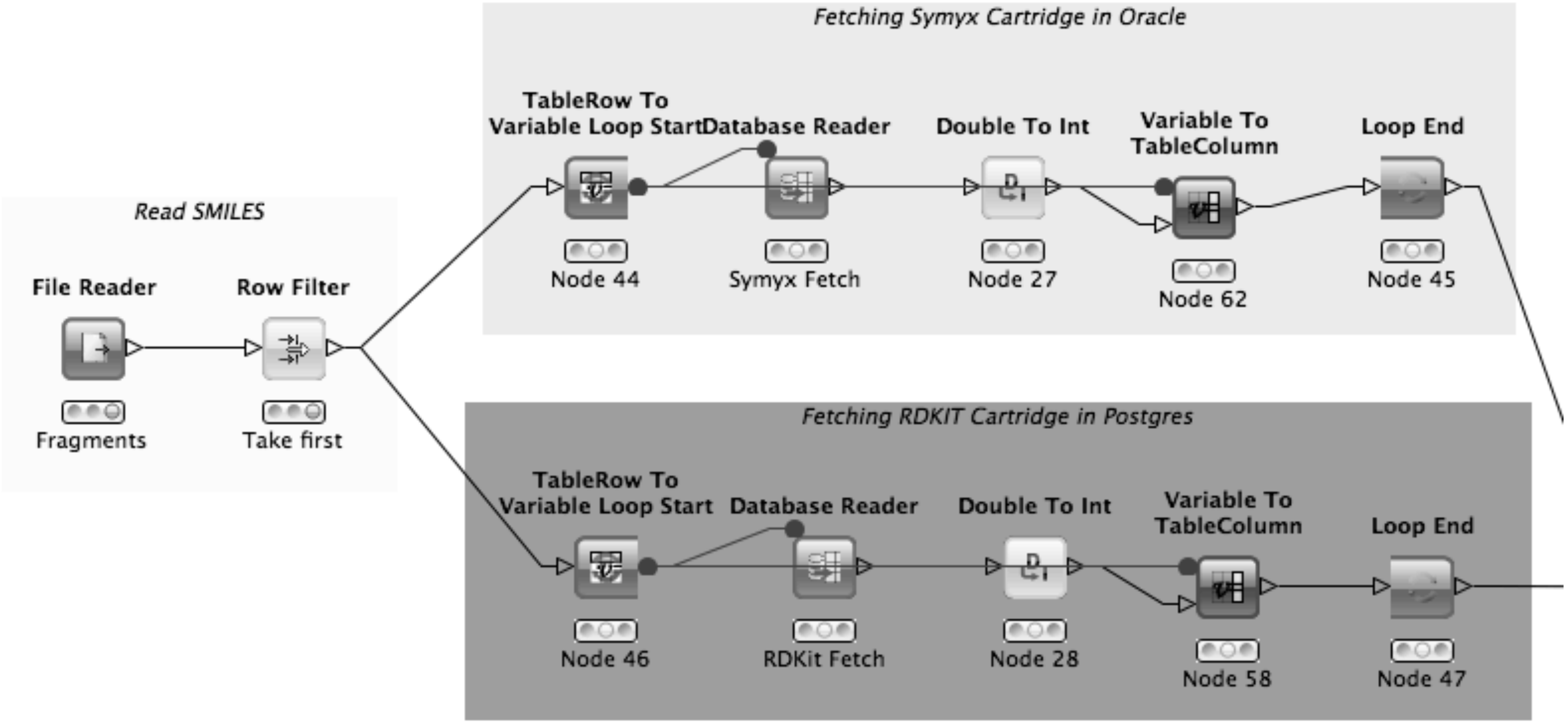




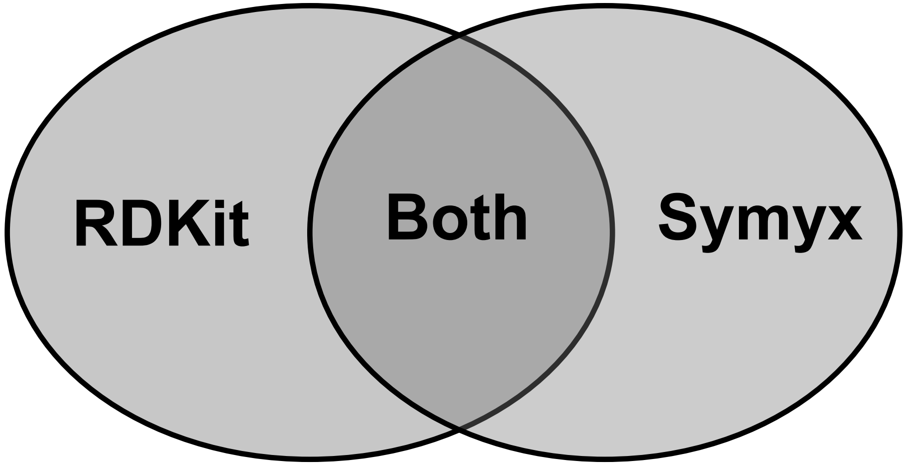


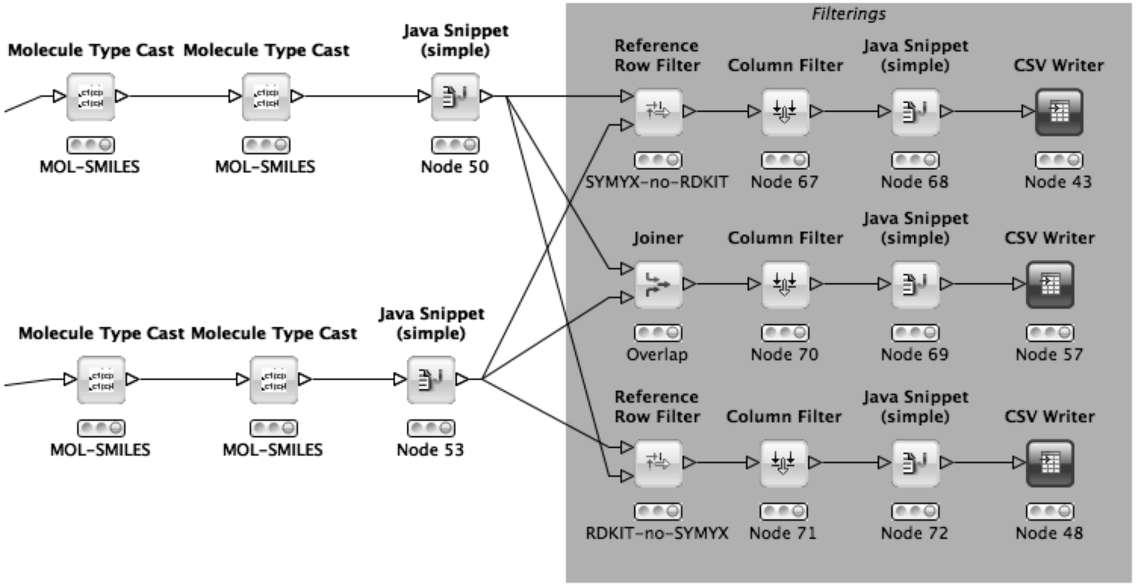
* Lewell XQ, Judd DB, Watson SP, Hann MM. RECAP - Retrosynthetic Combinatorial Analysis Procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry*. Journal of Chemical Information and Computer Sciences* 1998, 38:511-522.

KNIME protocol

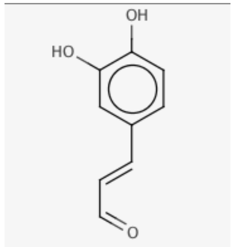
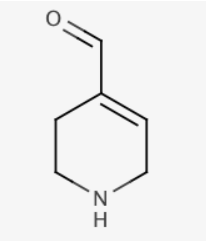
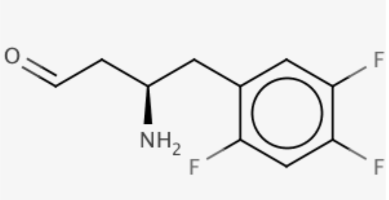
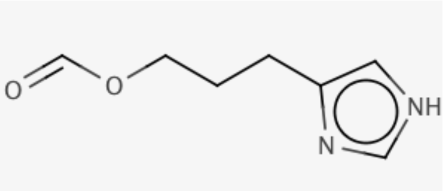


KNIME results

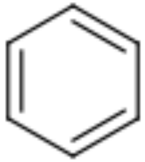




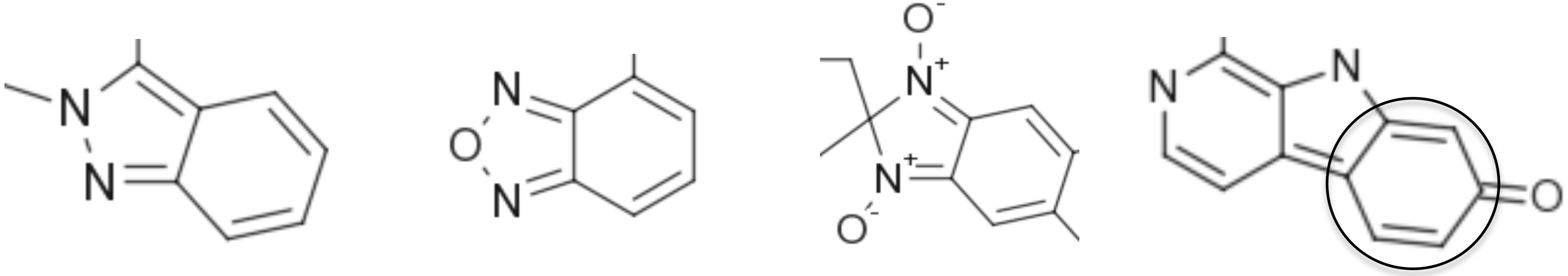
Some results …

Fragment	Molecule	RDKit no Symyx	Symyx no RDKit	Join Symyx RDKit	Total RDKit	Total Symyx
1		1595	5565	25404	26999	30969
5		52	268	18247	18299	18515
9		263	6	199	462	205
14		62546	2455	73510	136056	75965

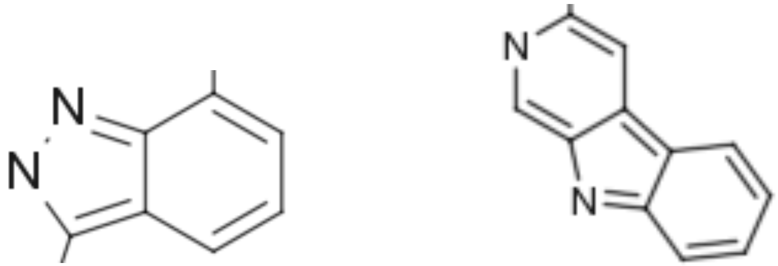
Aromaticity perception: benzene

RDKit SMARTS: `c1ccccc1` Symyx molfile: 

- 1445 structures found by RDKit but not by Symyx
- Due to differing aromaticity models

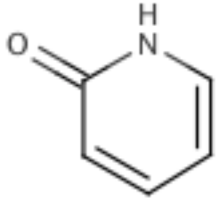


- Sometimes highlights unusual choice of tautomer

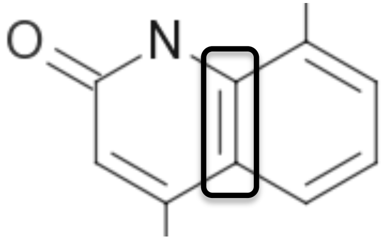


Aromaticity perception: pyridone

RDKit SMARTS: `c1cccnc1=O` Symyx molfile

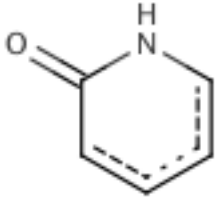


- 6760 structures hit by RDKit but not by Symyx

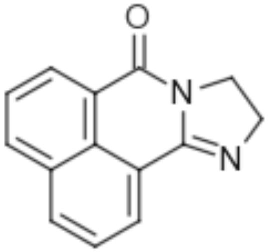


Bond is now aromatic and doesn't match double bond in query

- Can adjust Symyx query to catch these:



- Reduces discrepancy to 179…



Aromaticity models

- Differing aromaticity models mean it can be difficult to code queries that give same hits in both systems

  Not a new problem!

  e.g. Symyx/MDL *vs*. Daylight

- MarvinSketch can be used to generate SMARTS queries appropriate for RDKit

  User would need to be aware of issues if using advanced query features

- Unlikely to be a real problem once transitioning users become accustomed to new conventions

Stereochemistry

Chiral matching not currently handled by RDKit

Symyx *can* handle chiral queries

However, there are issues with chiral searching in ChEMBL

- Some inconsistences in setting of chiral flag
- Issue is being investigated

Best avoided for now !!!

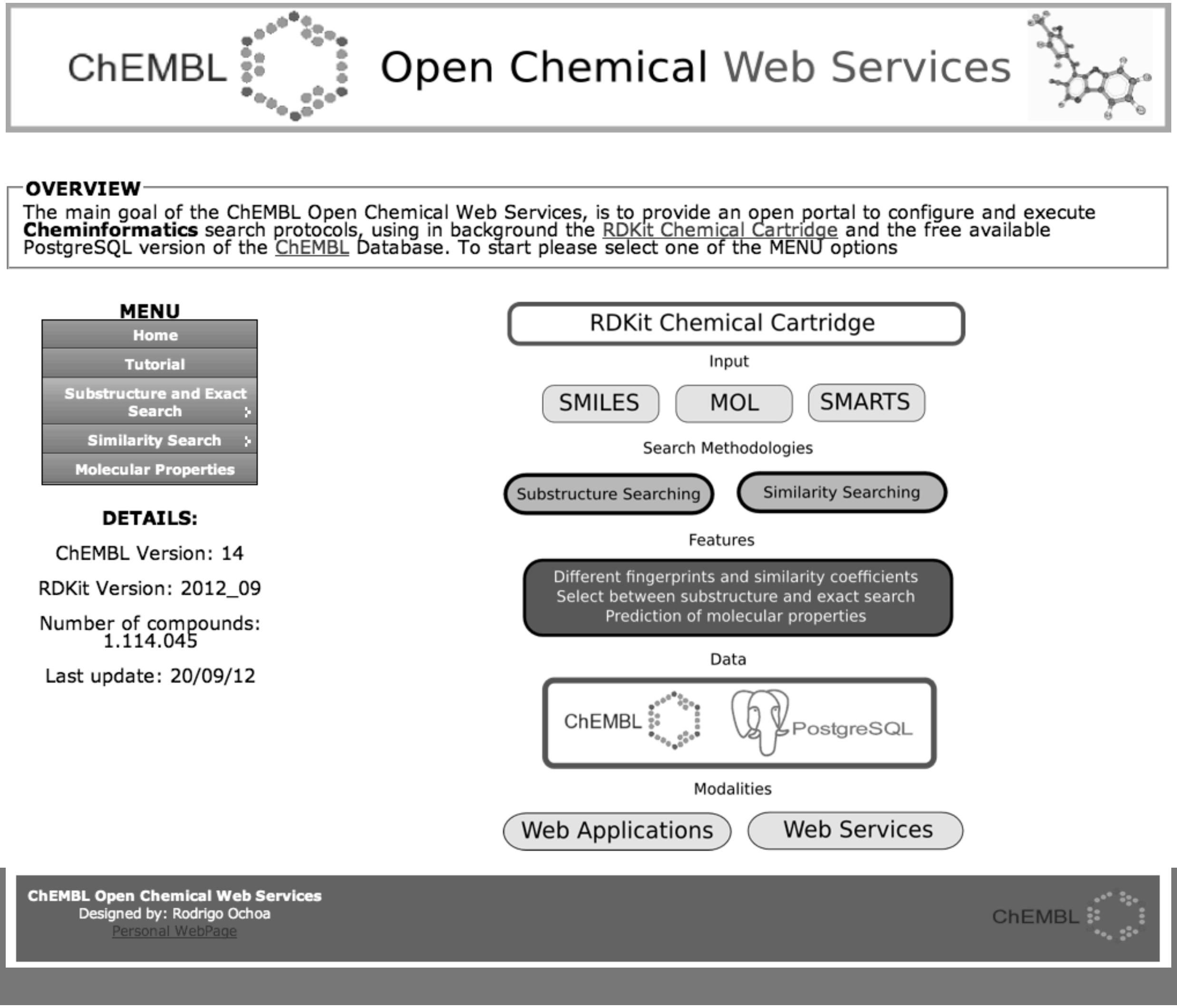
ChEMBL + PostgreSQL + RDKit

Open-Source Cheminformatics and Machine Learning

INTERFACE

Interface

- Substructure, Similarity search against ChEMBL database
- Additionally, calculation of molecular properties



ChEMBL Open Chemical Web Services

OVERVIEW

The main goal of the ChEMBL Open Chemical Web Services, is to provide an open portal to configure and execute **Cheminformatics** search protocols, using in background the RDKit Chemical Cartridge and the free available PostgreSQL version of the ChEMBL Database. To start please select one of the MENU options

DETAILS:

ChEMBL Version: 14

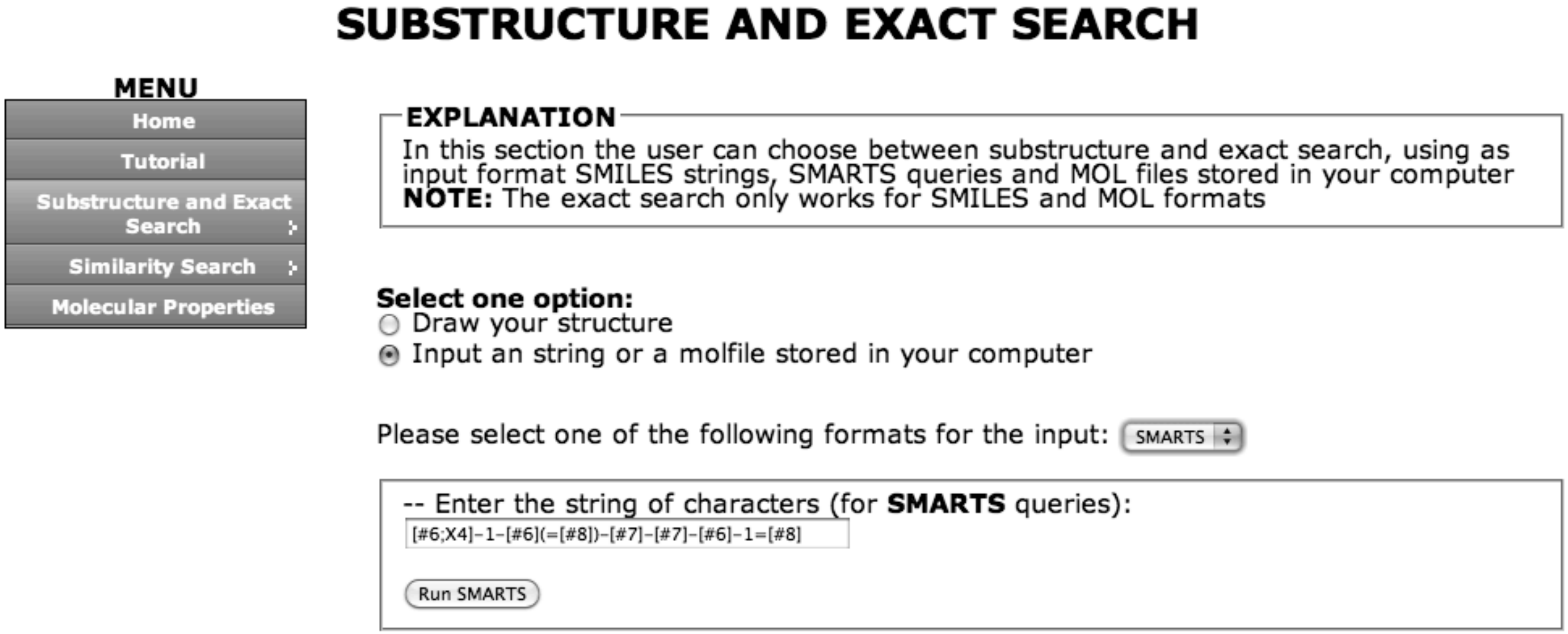
RDKit Version: 2012_09

Number of compounds: 1.114.045

Last update: 20/09/12

ChEMBL Open Chemical Web Services
Designed by: Rodrigo Ochoa
Personal WebPage

- Substructure search using SMARTS, SMILES or MolFiles



SUBSTRUCTURE AND EXACT SEARCH

EXPLANATION

In this section the user can choose between substructure and exact search, using as input format SMILES strings, SMARTS queries and MOL files stored in your computer
NOTE: The exact search only works for SMILES and MOL formats

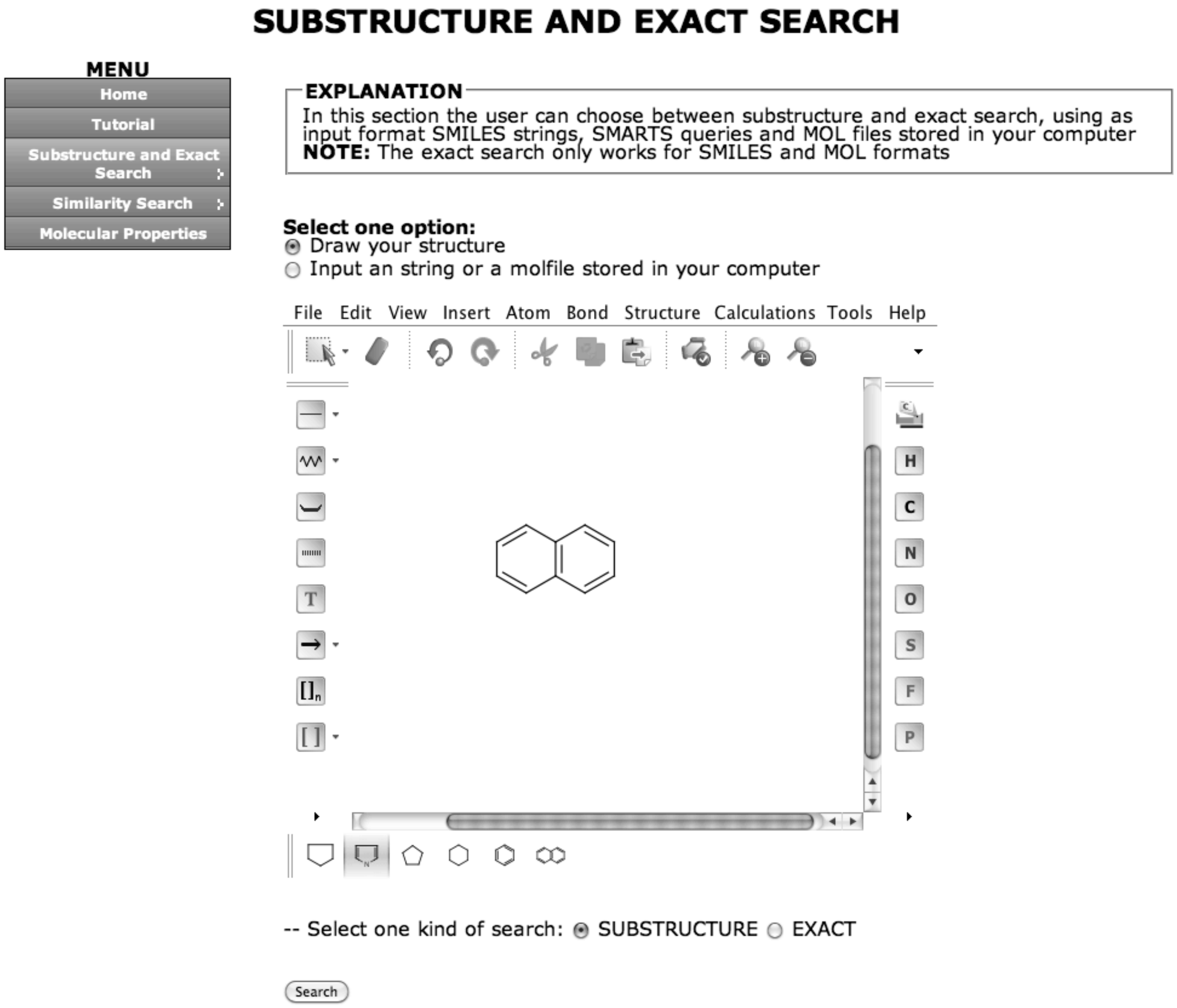
Select one option:

○ Draw your structure
● Input an string or a molfile stored in your computer

Please select one of the following formats for the input: SMARTS

-- Enter the string of characters (for **SMARTS** queries):

[#6;X4]-1-[#6](=[#8])-[#7]-[#7]-[#6]-1=[#8]



SUBSTRUCTURE AND EXACT SEARCH

MENU
Home
Tutorial
Substructure and Exact Search
Similarity Search
Molecular Properties

EXPLANATION

In this section the user can choose between substructure and exact search, using as input format SMILES strings, SMARTS queries and MOL files stored in your computer
NOTE: The exact search only works for SMILES and MOL formats

Select one option:

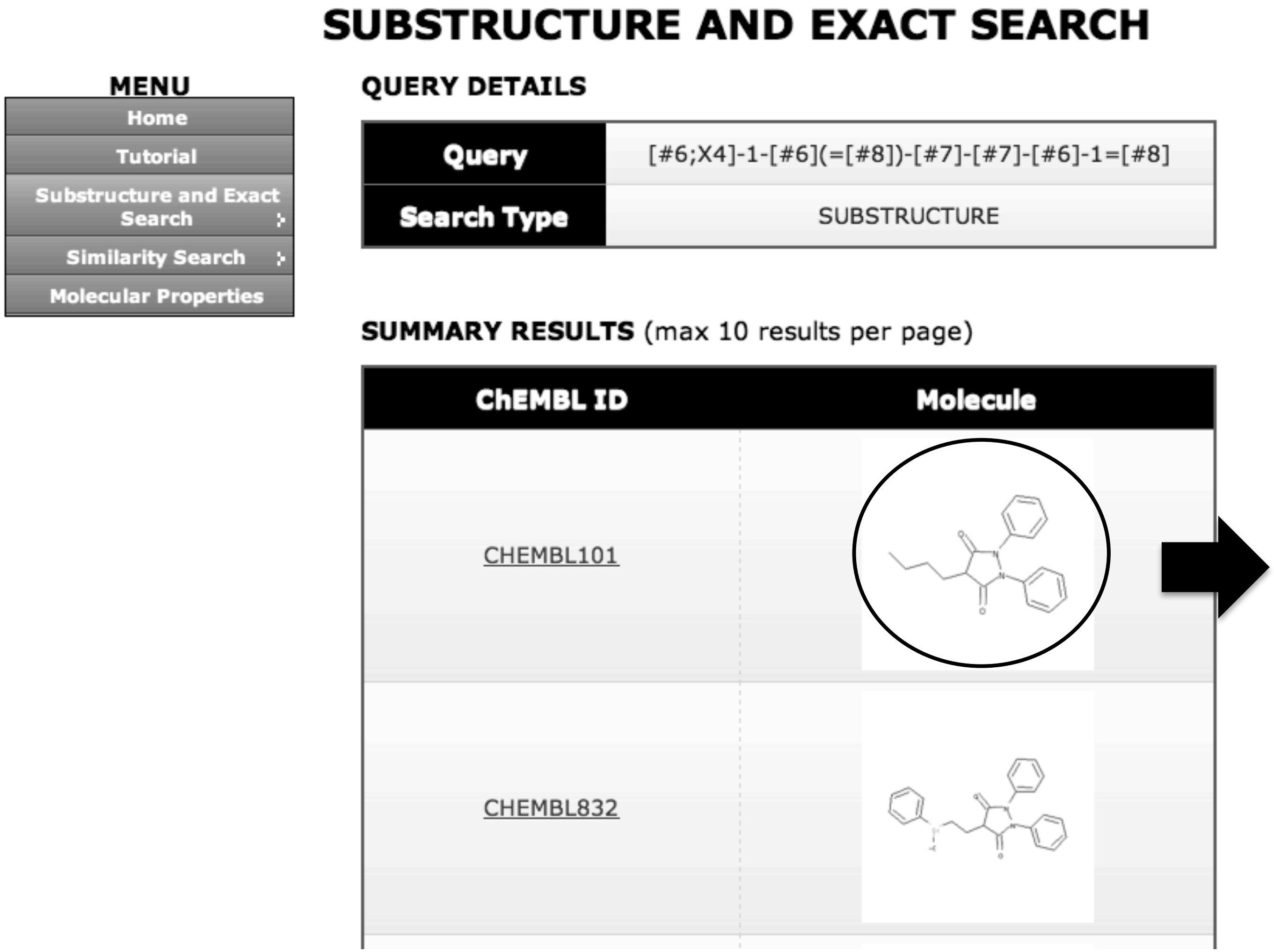
● Draw your structure

○ Input an string or a molfile stored in your computer

File Edit View Insert Atom Bond Structure Calculations Tools Help
T
[]n
[ ]
H
C
N
O
S
F
P

-- Select one kind of search: ● SUBSTRUCTURE ○ EXACT

- Query results using substructure search



SUBSTRUCTURE AND EXACT SEARCH

QUERY DETAILS

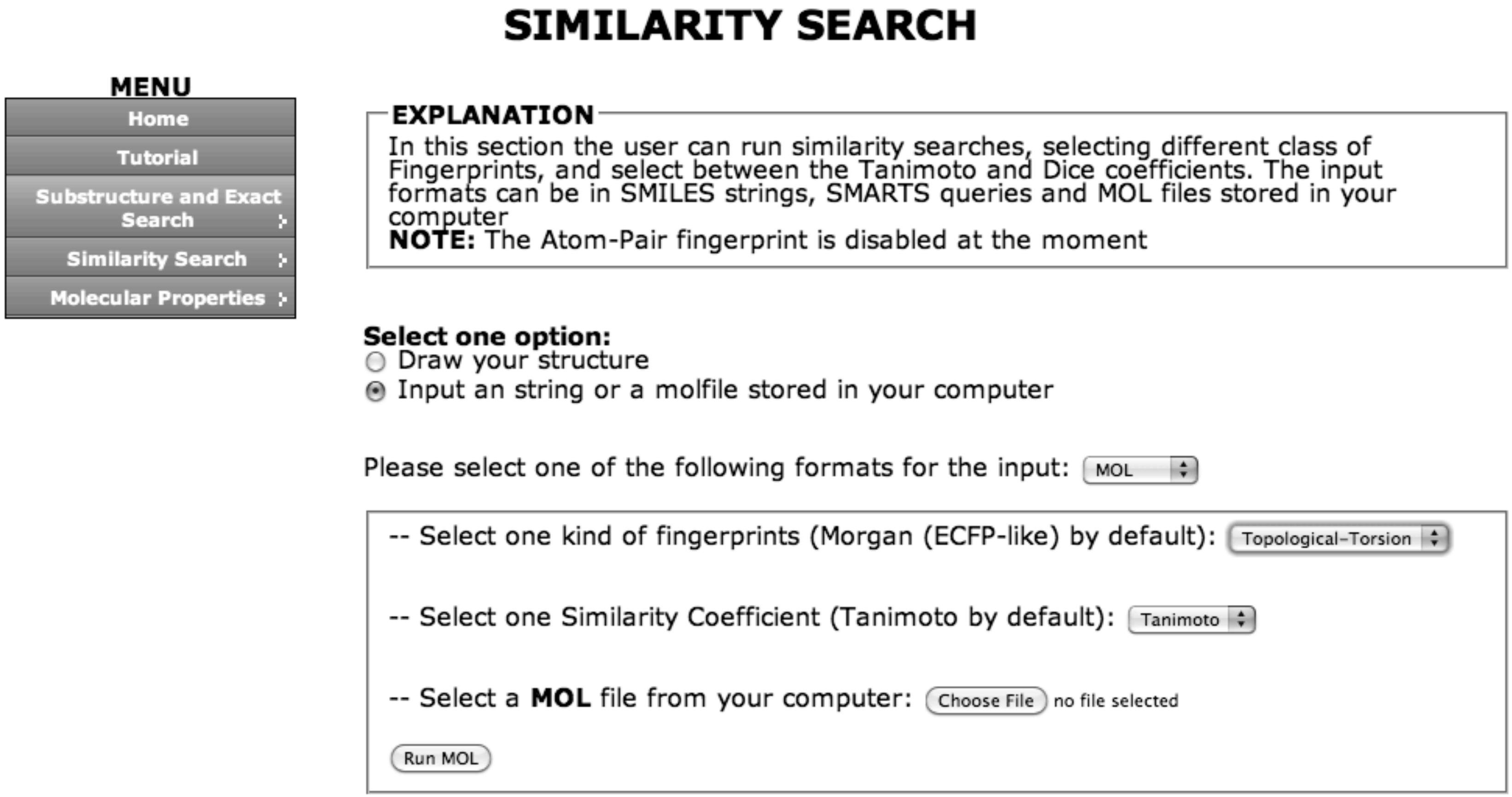
Query	[#6;X4]-1-[#6](=[#8])-[#7]-[#7]-[#6]-1=[#8]
Search Type	SUBSTRUCTURE

SUMMARY RESULTS (max 10 results per page)

ChEMBL ID	Molecule
CHEMBL101	
CHEMBL832	

Generated with RDKit

- Similarity search using different Fingerprints and Similarity Coefficients



SIMILARITY SEARCH

EXPLANATION

In this section the user can run similarity searches, selecting different class of Fingerprints, and select between the Tanimoto and Dice coefficients. The input formats can be in SMILES strings, SMARTS queries and MOL files stored in your computer

NOTE: The Atom-Pair fingerprint is disabled at the moment

Select one option:

○ Draw your structure

● Input an string or a molfile stored in your computer

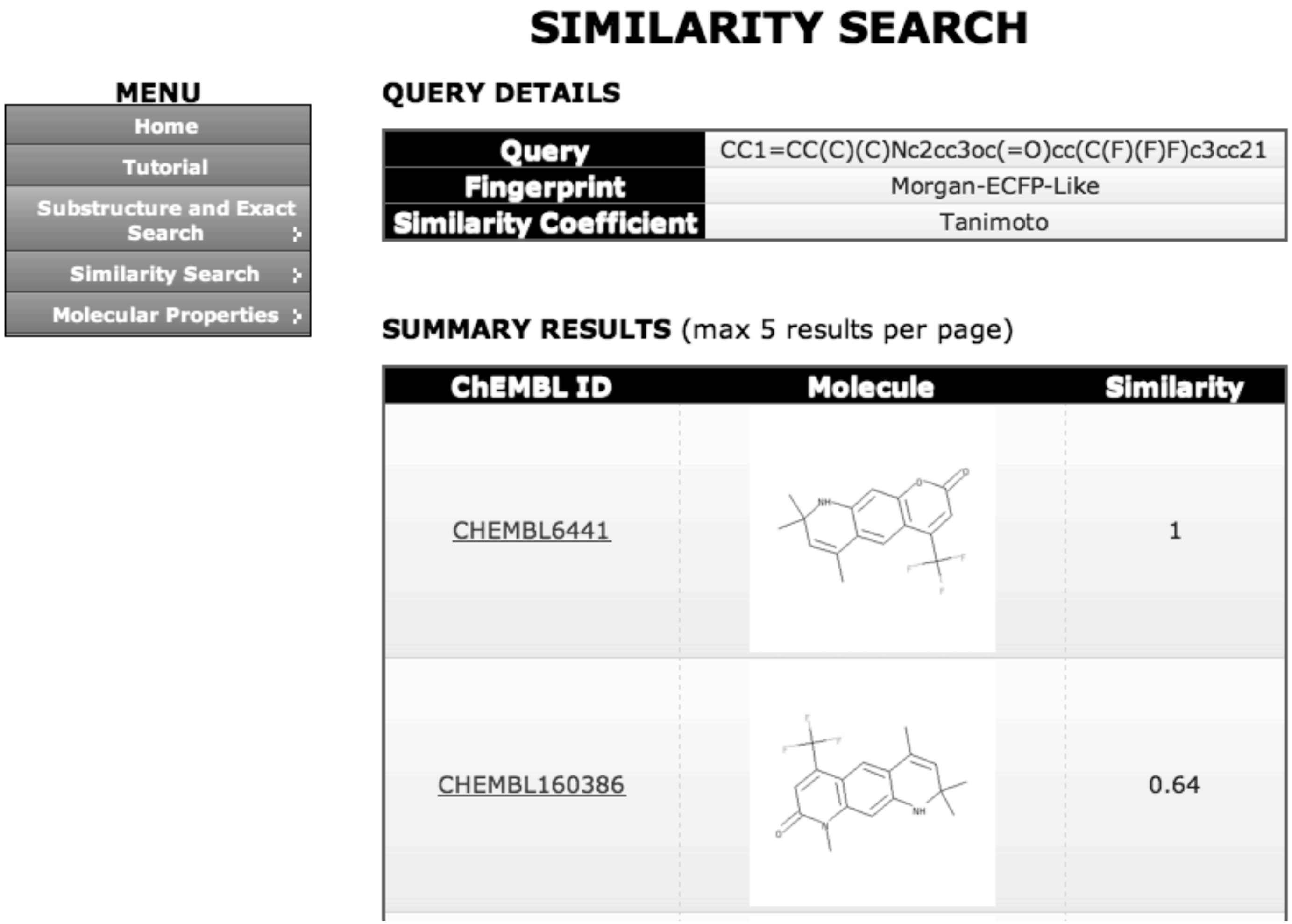
Please select one of the following formats for the input: MOL

-- Select one kind of fingerprints (Morgan (ECFP-like) by default): Topological-Torsion

-- Select one Similarity Coefficient (Tanimoto by default): Tanimoto

-- Select a **MOL** file from your computer: Choose File no file selected

- Query results using Similarity search



SIMILARITY SEARCH

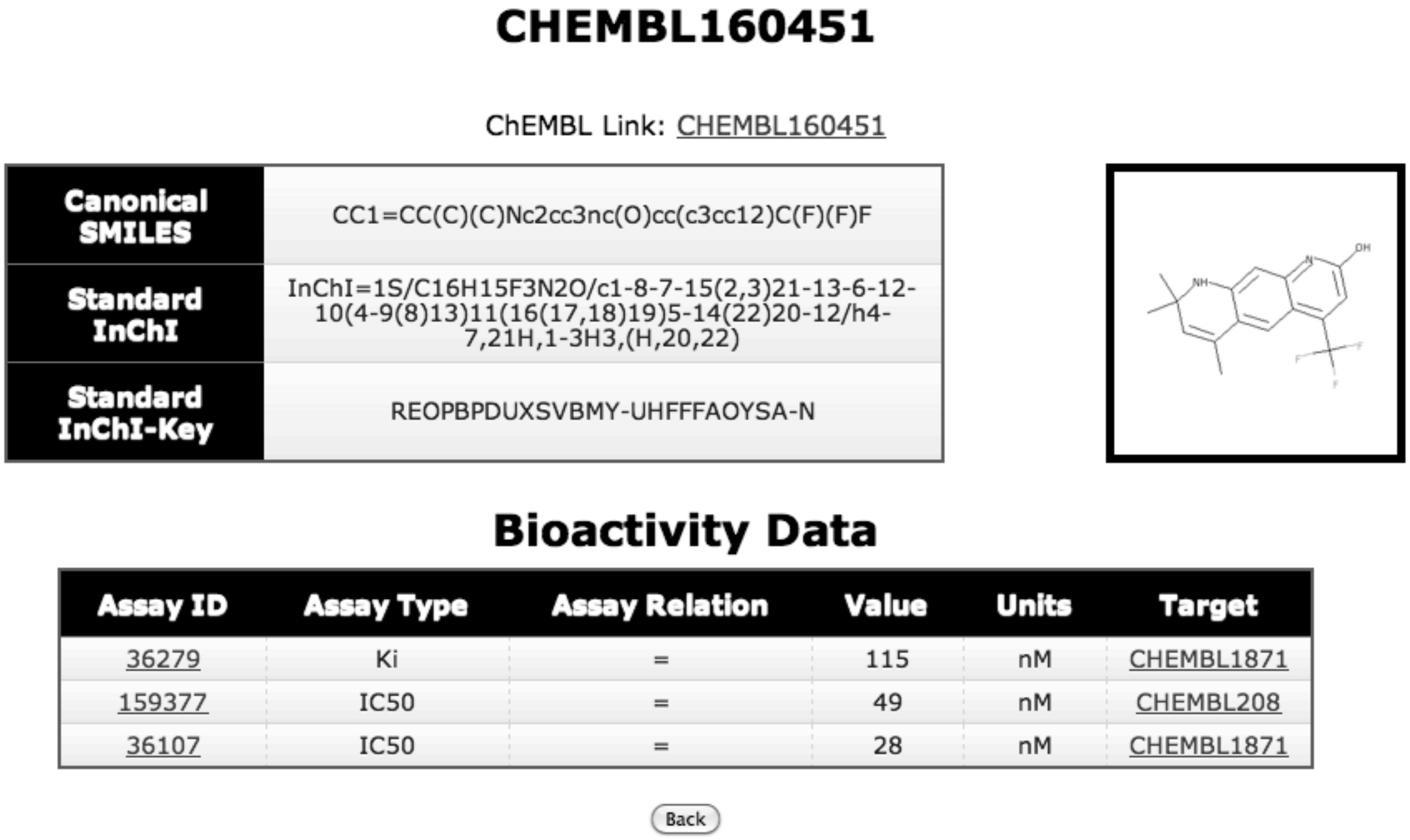
QUERY DETAILS

Query	CC1=CC(C)(C)Nc2cc3oc(=O)cc(C(F)(F)F)c3cc21
Fingerprint	Morgan-ECFP-Like
Similarity Coefficient	Tanimoto

SUMMARY RESULTS (max 5 results per page)

ChEMBL ID	Molecule	Similarity
CHEMBL6441		1
CHEMBL160386		0.64

- Compound report (with links to the ChEMBL website)



CHEMBL160451

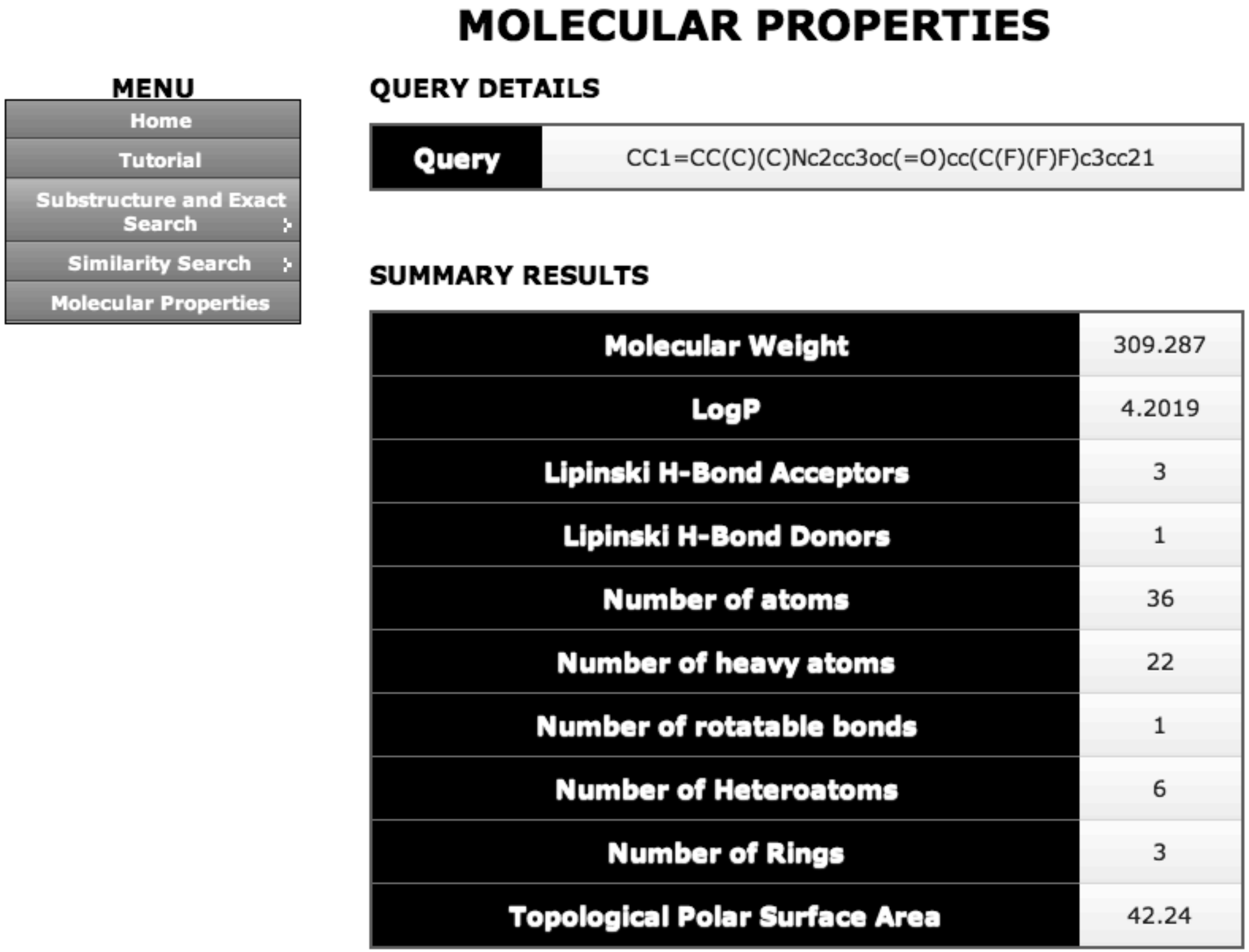
ChEMBL Link: CHEMBL160451

Canonical SMILES	CC1=CC(C)(C)Nc2cc3nc(O)cc(c3cc12)C(F)(F)F
Standard InChI	InChI=1S/C16H15F3N2O/c1-8-7-15(2,3)21-13-6-12-10(4-9(8)13)11(16(17,18)19)5-14(22)20-12/h4-7,21H,1-3H3,(H,20,22)
Standard InChI-Key	REOPBPDUXSVBMY-UHFFFAOYSA-N

Bioactivity Data

Assay ID	Assay Type	Assay Relation	Value	Units	Target
36279	Ki	=	115	nM	CHEMBL1871
159377	IC50	=	49	nM	CHEMBL208
36107	IC50	=	28	nM	CHEMBL1871

- Molecular properties calculation



MOLECULAR PROPERTIES

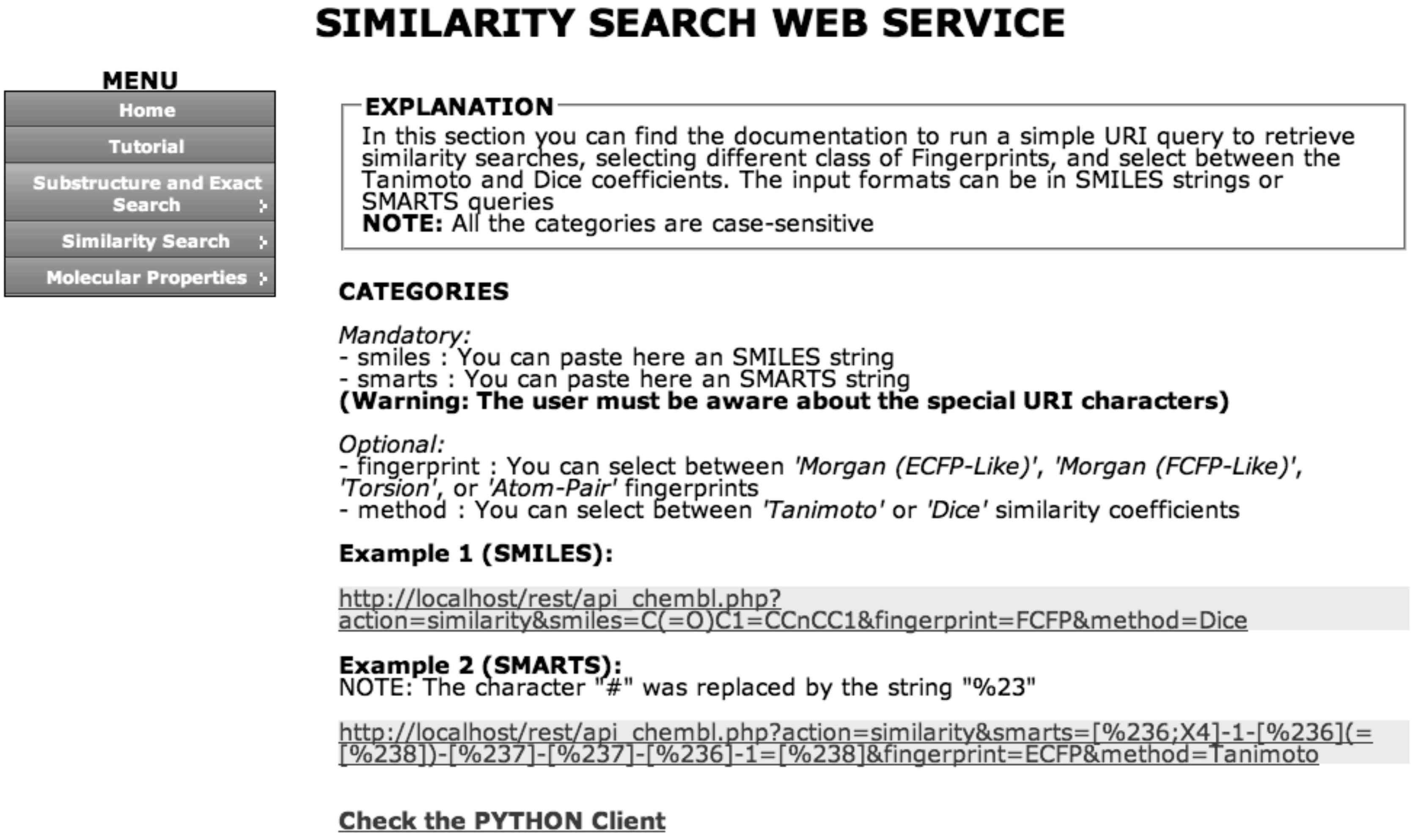
QUERY DETAILS

Query	CC1=CC(C)(C)Nc2cc3oc(=O)cc(C(F)(F)F)c3cc21

SUMMARY RESULTS

Molecular Weight	309.287
LogP	4.2019
Lipinski H-Bond Acceptors	3
Lipinski H-Bond Donors	1
Number of atoms	36
Number of heavy atoms	22
Number of rotatable bonds	1
Number of Heteroatoms	6
Number of Rings	3
Topological Polar Surface Area	42.24

- Similarity search RESTful web service (also for substructure)



SIMILARITY SEARCH WEB SERVICE

EXPLANATION

In this section you can find the documentation to run a simple URI query to retrieve similarity searches, selecting different class of Fingerprints, and select between the Tanimoto and Dice coefficients. The input formats can be in SMILES strings or SMARTS queries
NOTE: All the categories are case-sensitive

CATEGORIES

Mandatory:
- smiles : You can paste here an SMILES string
- smarts : You can paste here an SMARTS string
(Warning: The user must be aware about the special URI characters)

Optional:
- fingerprint : You can select between *'Morgan (ECFP-Like)'*, *'Morgan (FCFP-Like)'*, *'Torsion'*, or *'Atom-Pair'* fingerprints
- method : You can select between *'Tanimoto'* or *'Dice'* similarity coefficients

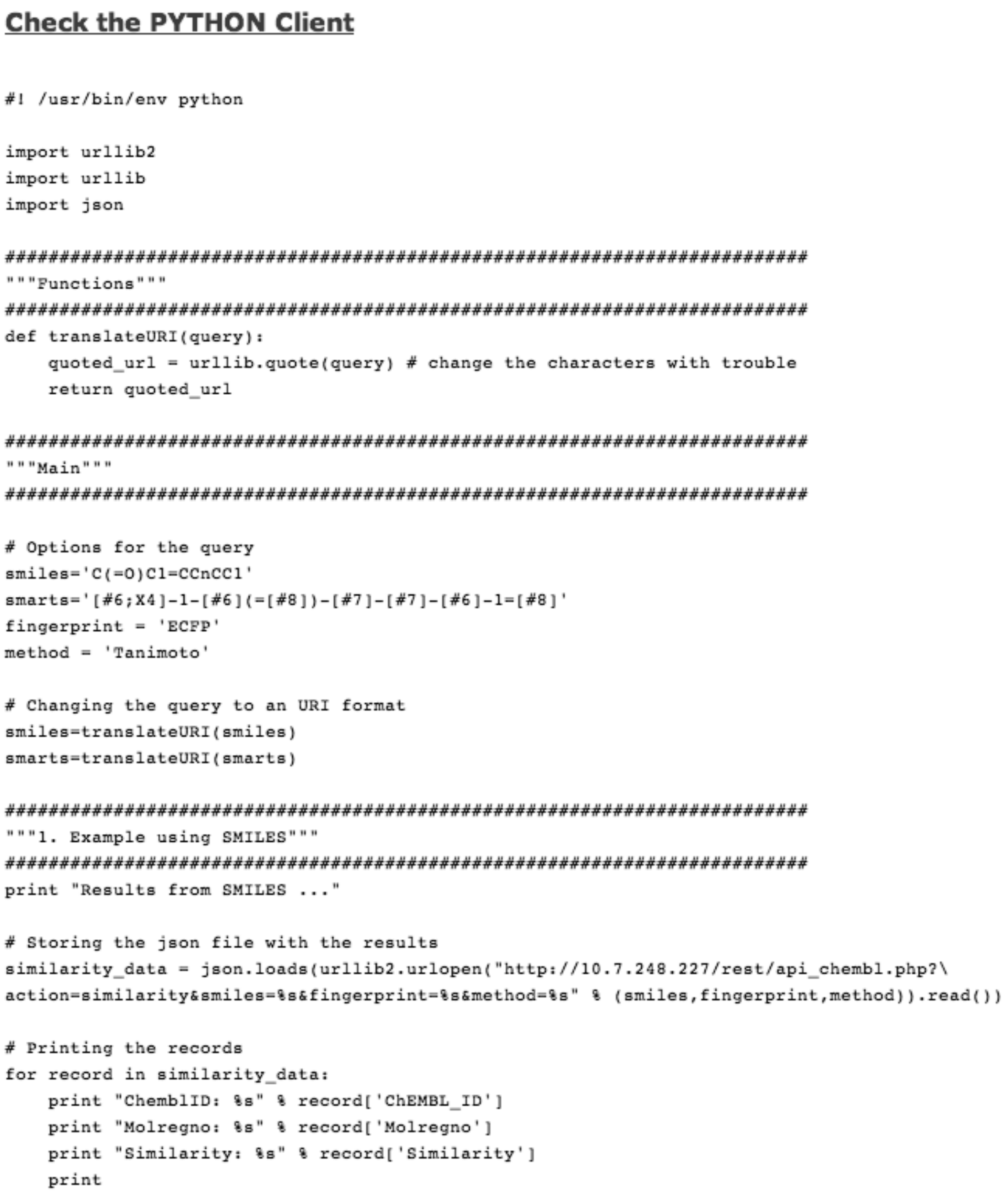
Example 1 (SMILES):

http://localhost/rest/api_chembl.php?action=similarity&smiles=C(=O)C1=CCnCC1&fingerprint=FCFP&method=Dice

Example 2 (SMARTS):
NOTE: The character "#" was replaced by the string "%23"

http://localhost/rest/api_chembl.php?action=similarity&smarts=[%236;X4]-1-[%236](=[%238])-[%237]-[%237]-[%236]-1=[%238]&fingerprint=ECFP&method=Tanimoto

Check the PYTHON Client



Check the PYTHON Client

```
#! /usr/bin/env python

import urllib2
import urllib
import json

###########################################################################
"""Functions"""
###########################################################################
def translateURI(query):
    quoted_url = urllib.quote(query) # change the characters with trouble
    return quoted_url

###########################################################################
"""Main"""
###########################################################################

# Options for the query
smiles='C(=O)C1=CCnCC1'
smarts='[#6;X4]-1-[#6](=[#8])-[#7]-[#7]-[#6]-1=[#8]'
fingerprint = 'ECFP'
method = 'Tanimoto'

# Changing the query to an URI format
smiles=translateURI(smiles)
smarts=translateURI(smarts)

###########################################################################
"""1. Example using SMILES"""
###########################################################################
print "Results from SMILES ..."

# Storing the json file with the results
similarity_data = json.loads(urllib2.urlopen("http://10.7.248.227/rest/api_chembl.php?\
action=similarity&smiles=%s&fingerprint=%s&method=%s" % (smiles,fingerprint,method)).read())

# Printing the records
for record in similarity_data:
    print "ChemblID: %s" % record['ChEMBL_ID']
    print "Molregno: %s" % record['Molregno']
    print "Similarity: %s" % record['Similarity']
    print
```

Summary and future plans

- Provide a complete PostgreSQL version of ChEMBL per new release, available in the FTP site

  https://www.ebi.ac.uk/chembldb/index.php/downloads

- Integrate the Open Chemical Cartridge (RDKit) with open chemogenomics data (ChEMBL) in a Virtual Machine available for users, who can personalize the configuration and update the data.

- Provide a personal interface, with web applications and RESTful web services.

Acknowledgements

ChEMBL

Training Opportunities

Mark Davies
Francis Atkinson
George Papadatos
John Overington

GGA

GGA Software Services

accelrys®

Symyx

RDKit

Open-Source Cheminformatics
and Machine Learning

KNIME